Sources of error in uncovering of the low level mitochondrial heteroplasmy landscape



Daina Šadurska

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2024

Abstract

As organelles descendant from an ancient endosymbiont, mitochondria retain their own short circular genomes, with multiple copies of the mitochondrial DNA (mtDNA) distributed through the mitochondrial network within a cell. The multi-copy nature of the mitochondrial genome enables variant mtDNA copies to coexist within the same cell or organism. This well documented phenomenon, known as heteroplasmy, is a key feature of mitochondrial genetics that distinguishes mtDNA from its nuclear counterpart, and requires careful application of population genetics principles to the study of mitochondrial genetics.

Presence of multiple mitochondrial genome copies in a cell may allow deleterious variants to emerge at sub-phenotypic frequencies, with their dynamics governed by the principles of by mutation, selection and genetic drift. Unlike the nuclear genome, mitochondrial DNA is also subject to bottlenecks during cell division capable of altering heteroplasmic allele balance in the descendant cell populations.

A sub-threshold heteroplasmy landscape could exist within an organism, reflecting the balance between the opposing forces of new variant introduction via DNA damage or replication errors on one side, and the maintenance of the genome integrity through repair and quality control mechanics on the other side, with the patterns of such a landscape being be a direct reflection of the equilibrium dynamics. This thesis aims to assess whether such a low level mitochondrial heteroplasmy landscape can be observed and described in the human population using large scale, publicly available whole genome sequence datasets.

Sequencing errors present a challenge for studying very low frequency heteroplasmic variants, with false positives resulting from systematic context-dependent sequencing errors (CSEs) being especially problematic for any investigation into heteroplasmy dependence on sequence context. We address this by systematically measuring CSE levels in our dataset, reporting CSE rates associated with up to 8bp long sequence motifs. The patterns of sequencing error rate variability between samples and between motifs are assessed in a first-of-its-kind systematic analysis of population-level CSE dynamics. Applying our insights in CSEs to mitochondrial heteroplasmy analyses, we propose an additional sequence data clean-up step intended to exclude sequence reads most likely to result from CSEs. Application of the additional data clean-up step results in a demonstrable decrease in the measured minor allele frequency estimates of the detected heteroplasmies. This suggests a possible benefit for applications such as low-level heteroplasmy quantification.

This thesis aims to provides a framework for systematic CSE variability analysis on a population scale and describes a possible strategy for mitigation of CSE-induced false positive risks in low frequency variant discovery, that can be applicable in mitochondrial or other genetic system. The proposed strategy in itself does not however mitigate against other artefacts besides CSEs, and is intended to supplement other data quality control steps.

Table of Contents

Abbreviations 5

List of figures 6

1 Introduction and background 8

1.1 Mitochondrial genetics background 8

1.2 Mitochondrial heteroplasmy 12

1.3 Research questions addressed in this thesis 19

2 Materials and methods 20

2.1 Additional methods, supplementary information, codebase and data availability 20

2.2 1000 Genomes Project phase 3 core dataset mapped to GRCh37 20

2.3 Base and motif enrichment analysis using cumulative frequency curves 21

2.4 Sequence context dependent error rate estimation 23

2.5 1000 Genomes Project extended cohort high coverage dataset mapped to GRCh38 27

2.6 Heteroplasmy identification in high coverage NYGC dataset 28

2.7 Heteroplasmy literature review 30

3 Results 1 – The non-reference read frequency landscape 32

3.1 Introduction and research question 32

3.2 %NR levels and landscapes 33

3.2.1 Overall %NR level of an individual 33

3.2.2 The shape of the MT genome %NR landscape 36

3.2.3 Site-specific %NR behaviour across the entire analysed population 40

3.2.4 Section conclusions 44

3.3 Link between position-specific %NR behaviour and the mtDNA sequence 45

3.3.1 Base composition of mtDNA position subsets with different %NR behaviours 45

3.3.2 Quantification of the association between %NR behaviour and the DNA sequence features 48

3.4 Sequence features associated with %NR behaviour differences between sequencing directions 53

3.4.1 Link between sequencing direction %NR behaviour differences and the position base type 54

3.4.2 Link between the discrepancy in sequencing direction specific %NR behaviour and the adjacent DNA sequence 56

3.5 Chapter conclusion 61

4 Results 2 – Context dependent DNA sequencing errors 62

4.1 Introduction 62

4.2 Motifs associated with high level of context specific sequencing error 64

4.2.1 Motif spaces investigated 64

4.2.2 Motif sets with strongest association to sequencing errors 66

4.2.3 Generalising sequence patterns associated with high error rates 71

4.2.4 Sequencing error rate differences between closely related motifs 73

4.2.5 Predictive power of motif patterns for identification of error sites 78

4.2.6 Comparison of identified error prone contexts to previous literature 81

4.3 Motif specific sequencing error rate variation 86

4.3.1 Relationship between motif population ΣMER level and the MER variation between samples 87

4.3.2 Variation in the sequencing error rate profiles observed between different samples 88

4.4 Chapter discussion 92

5 Results 3 – Heteroplasmy analysis in a high coverage dataset 96

5.1 Introduction 96

5.2 Heteroplasmy detection using consensus read frequencies 97

5.2.1 Putative heteroplasmy counts at different allele filtration stringency levels 98

5.2.2 Putative heteroplasmy MAF at different filtration stringencies 102

5.2.3 Putative heteroplasmy set major allele base composition 105

5.3 Comparison with literature 110

5.3.1 Heteroplasmy frequency reported in literature 110

5.3.2 Most common heteroplasmic positions 113

5.3.3 Further comparison between heteroplasmies reported in this study with those reported in literature 118

5.4 The landscape of low level heteroplasmy prevalence in the population 127

5.4.1 The landscape of low level heteroplasmy frequency in the population 128

5.4.2 Positions associated with highest heteroplasmy incidence 131

5.5 Heteroplasmy levels in the population 142

5.5.1 Average measured MAF in the population 143

5.5.2 Average MAF across heteroplasmic individual subset 144

5.5.3 MAF and the base type 150

5.5.4 Positions associated with highest average MAF 152

5.5.5 Positions associated with largest number of high MAF heteroplasmies 157

5.6 Relationship between the observed mitochondrial heteroplasmy and its genomic context 160

5.6.1 Heteroplasmy differences between mitochondrial genome functional elements 161

5.6.2 Heteroplasmy patterns in protein coding genes 166

5.6.3 Relationships between heteroplasmy and the base type in different genomic contexts 169

5.7 Prevalence of mtDNA variants linked mitochondrial disease 174

5.8 Chapter discussion 181

6 Discussion and conclusions 184

7 References 196

Acknowledgements 208

Supplementary Information 209

Appendix 1 Supplementary information for Chapter 3 210

Appendix 2 Supplementary information for Chapter 4 230

Appendix 3 Supplementary information for Chapter 5 259

Abbreviations

General

A,C,G,T	Adenine, Thymine, Cytosine, Guanine
CR	Control region
CSE	Context-dependent sequencing error
G1K	1000 genomes project
G4	G-quadruplex
Indel	Insertion - deletion mutation
IQR	Interquartile range
MAF	Minor allele frequency
MT	Mitochondrial
mtDNA	Mitochondrial DNA
nDNA	Nuclear DNA
NC	Non-coding nucleotides
NGS	Next generation sequencing
NUMTs	Nuclear sequences of mitochondrial origin
OXPHOS	Oxidative phosphorylation
P_{50}, P_{95}	Median, 95th percentile
PCR	Polymerase chain reaction
rCRS	Revised Cambridge reference sequence
ROS	Reactive oxygen species
SD	Standard deviation
SNP	Single nucleotide polymorphism
WES	Whole exome sequencing
WGS	Whole genome sequencing

Results 1

%NR	Non-reference read frequency
E, E_N	Enrichment score

Results 2

MER	Motif sequencing error rate
ΣMER	Population level motif sequencing error rate
KS2, KS-2	Kolmogorov-Smirnov 2 sample test

Results 3

MAF	Minor allele frequency
HDR	Heteroplasmy detection rate
dN/dS	Nonsynonymous-to-synonymous ratio
Ti/Tv	Transition to transversion ratio
LCL	Lymphoblastoid cell line

Mitochondrial diseases

CPEO/CEO	Chronic progressive external ophthalmoplegia syndrome / progressive external ophthalmoplegia
LHON	Leber hereditary optic neuropathy
LS	Leigh syndrome
MELAS	Mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes
MERRF	Myoclonic epilepsy associated with ragged red fibres
MIDD	Maternally inherited diabetes and deafness
NIDDM	Noninsulin-dependent diabetes mellitus
NARP	Neuropathy, ataxia and pigmentary retinopathy

Mitochondrial genome functional elements

Protein coding

ATP6	ATP synthase F0 subunit 6
ATP8	ATP synthase F0 subunit 8
CO1	Cytochrome c oxidase subunit I
CO2	Cytochrome c oxidase subunit II
CO3	Cytochrome c oxidase subunit III
CYB	Cytochrome b
ND1	NADH dehydrogenase subunit 1
ND2	NADH dehydrogenase subunit 2
ND3	NADH dehydrogenase subunit 3
ND4	NADH dehydrogenase subunit 4
ND4L	NADH dehydrogenase subunit 4L
ND5	NADH dehydrogenase subunit 5
ND6	NADH dehydrogenase subunit 6

rRNA

RNR2	16S ribosomal RNA
RNR1	12S ribosomal RNA

tRNA

TL1, TL2	Leucine
TS1, TS2	Serine
TN	Asparagine
TQ	Glutamine
TF	Phenylalanine
TK	Lysine
TV	Valine
TI	Isoleucine
TA	Alanine
TH	Histidine
TE	Glutamic acid
TM	Methionine
TW	Tryptophan
TD	Aspartic acid
TG	Glycine
TP	Proline
TC	Cysteine
TY	Tyrosine
TT	Threonine
TR	Arginine

Control region

HV1, HV2, HV3	Hypervariable segments
CSB1, CSB2, CSB3	Conserved sequence blocks
TFH, TFL, TFX, TFY	mtTF1 binding sites
OHR, OHR57	H-strand origins
TAS	Termination associated sequence
HPR	Replication primer
LSP	L-strand promoter
HSP1	Major H-strand promoter
MT3H, MT3L, MT4H, MT5	Control elements

Other

HSP2	Minor H-strand promoter
OLR	L-strand origin
TER	Transcription terminator
ATT	Membrane attachment site
Hum	Humanin peptide
RNR3	5S-like sequence

List of figures

Introduction

Figure 1.1 F1: Mitochondrial network and the architecture of the human mitochondrial genome..... 8
Figure 1.1 F2: Types of DNA damage and the mitochondrial DNA repair mechanisms 10
Figure 1.1 F3: Non B-DNA secondary structures 11
Figure 1.1 F4: G-quadruplex structure and possible location in the mitochondrial genome 11
Figure 1.2 F1: Pathogenic mtDNA point mutations associated with mitochondrial diseases 13

Materials and methods

Figure 2.3 F1: Simple adjacent sequence patterns consisting of a single base type around the position of interest....22
Figure 2.4 F1: Motif context-dependent sequencing error rate measurement approach 25
Figure 2.4 F2: Summary of the analysed motif spaces 26
Figure 2.6 F1: Heteroplasmic variant filtration criteria.....29

Chapter 3

Figure 3.2.1 F1: Mitochondrial genome non-reference read frequency distributions in 6 individuals.....34
Figure 3.2.1 F2: Mitochondrial genome %NR distribution summary statistics across all analysed individual 35
Figure 3.2.2 F1: Visualisation of the mitochondrial genome %NR landscape fragment.....37
Figure 3.2.2 F2: Mitochondrial genome %NR landscape peak distribution on the population scale.....39
Figure 3.2.3 F1: Position-specific population %NR distributions at 20 consecutive mtDNA positions.....41
Figure 3.2.3 F2: Population %NR distribution summary statistics across all analysed mtDNA positions.....43
Figure 3.3.1 F1: Base composition of positions displaying different %NR population distribution behaviours 46
Figure 3.3.2 F1: Sequence motifs linked to different site-specific %NR population distribution median levels 49
Figure 3.3.2 F2: Sequence motifs linked to different site-specific %NR population distribution 95th percentile level ..50
Figure 3.3.2 F3: Sequence motifs linked to different site-specific %NR population distribution shape atypicality.....51
Figure 3.4.1 F1: Base composition differences between positions showing different extent of deviation between strand-specific %NR distribution behaviours.....55
Figure 3.4.2 F1: Sequence motifs linked to high strand-specific %NR distribution differences.....57
Figure 3.4.2 F2: Sequence motifs linked to low strand-specific %NR distribution differences.....58
Figure 3.4.2 F3: Sequence patterns linked to high %NR level or atypicality but not to high %NR strand bias 60

Chapter 4

Figure 4.2.1 F1: Motif spaces investigated for association with context dependent DNA sequencing errors 65
Figure 4.2.2 F1: 2bp to 5bp mtDNA sequence motifs linked to most significant sequencing error rates 66
Figure 4.2.2 F2: 6bp mtDNA sequence motifs linked to most significant sequencing error rates 67
Figure 4.2.2 F3: 7bp mtDNA sequence motifs linked to most significant sequencing error rates 68
Figure 4.2.2 F4: 8bp mtDNA sequence motifs linked to most significant sequencing error rates 69
Figure 4.2.3 F1: The common features of the most error prone sequence motif sets of different lengths.....72
Figure 4.2.4 F1: Sequencing error rate differences between motifs closely related to the 10 highest CSE 8bp motifs 74
Figure 4.2.4 F2: Sequencing error rate differences between motifs closely related to the 10 highest CSE 7bp motifs 75
Figure 4.2.4 F3: Sequencing error rate differences between motifs closely related to the 10 highest CSE 6bp motifs 76
Figure 4.2.5 F1: GGT-like patterns as predictors for high motif associated sequencing error rate 79
Figure 4.2.6 F1: Overlap between our identified error-prone 4bp contexts with those reported in an earlier study...82
Figure 4.2.6 F2: Overlap between our identified error-prone 8bp contexts with those reported in an earlier study...83
Figure 4.3.2 F1: Most error prone 6bp motifs: sequencing error rate variability between samples 89
Figure 4.3.2 F2: Most error prone 6bp motifs: variation of sample relative ranking by MER at different motifs 90

Chapter 5

Figure 5.2.1 F1: Criteria used for putative heteroplasmy filtration....98
Figure 5.2.1 F2: Heteroplasmies identified at different allele filtration stringency levels....99
Figure 5.2.2 F1: Median minor allele frequency of the identified heteroplasmies....102
Figure 5.2.2 F2: Median MAF of heteroplasmies identified by both methods versus standard method only....103
Figure 5.2.3 F1: Major allele base composition of heteroplasmies passing different filtration stringency levels....106
Figure 5.3.1 F1: Heteroplasmy incidence reported in literature....111
Figure 5.3.1 F2: Mitochondrial genome positions with previously reported heteroplasmies found in literature....112
Figure 5.3.2 F1: Mitochondrial genome positions most frequently reported in literature as heteroplasmic....114
Figure 5.3.2 F2: Base composition of known heteroplasmic mtDNA positions reported in previous literature....116
Figure 5.3.3 F1: Heteroplasmic sequence fraction and discovery rate comparison between the present study and comparable previous studies....118
Figure 5.3.3 F2: Overlap between previously reported heteroplasmic mtDNA positions and the heteroplasmies identified in the present study....121
Figure 5.3.3 F3: Correlation between heteroplasmic sample fraction found in the present study and the frequency of previous heteroplasmy reports in literature....122
Figure 5.3.3 F4: Relationship between heteroplasmies found in the present study at different MAF and the frequency of previous heteroplasmy reports in literature....125
Figure 5.4.1 F1: Detected heteroplasmy prevalence at each mitochondrial genome position....129
Figure 5.4.2 F1: 107 positions with 100 or more identified heteroplasmies with MAF ≥0.1%....132
Figure 5.4.2 F2: Base composition at and around the sites with ≥100 identified heteroplasmies with MAF ≥0.1%....133
Figure 5.4.2 F3: Adjacent sequence contexts around 107 sites with ≥100 heteroplasmies with MAF ≥0.1%....134
Figure 5.4.2 F4: 88 positions with 10 or more identified heteroplasmies with MAF ≥1%....136
Figure 5.4.2 F5: Base composition of and around the 88 sites with ≥10 identified heteroplasmies with MAF>1%....137
Figure 5.4.2 F6: Adjacent sequence contexts around 88 sites with ≥10 identified heteroplasmies with MAF ≥1%....138
Figure 5.4.2 F7: Overlap between the most frequently found heteroplasmic positions in the present study with the positions most often reported as heteroplasmic in the literature....140
Figure 5.5.1 F1: The average MAF observed across the mitochondrial genome....143
Figure 5.5.2 F1: Mean MAF of heteroplasmies passing different filtration stringency levels at each mtDNA site....145
Figure 5.5.2 F2: Overview of observed heteroplasmy mean MAF....146
Figure 5.5.2 F3: Observed heteroplasmy MAF population distribution shape variability between positions....147
Figure 5.5.3 F1: Heteroplasmic position base composition and distribution....150
Figure 5.5.3 F2: Relationship between position base type and the average heteroplasmy MAF....151
Figure 5.5.4 F1: Relationship between the number of heteroplasmies and the average heteroplasmy MAF level....152
Figure 5.5.4 F2: Features of 1,477 mtDNA positions with high observed heteroplasmy median MAF ≥1%....154
Figure 5.5.4 F3: Features of 191 mtDNA positions with very high observed heteroplasmy median MAF ≥10%....156
Figure 5.5.5 F1: Base composition mtDNA position sets where high to very high heteroplasmy levels (MAF≥1%, 5% or 10%) are found in ≥5 samples....157
Figure 5.6.1 F1: Mitochondrial heteroplasmy incidence observed in different genomic contexts....162
Figure 5.6.1 F2: Previously reported mitochondrial heteroplasmy incidence in different genomic contexts....163
Figure 5.6.2 F1: Observed heteroplasmy distribution between protein coding gene codon positions....166
Figure 5.6.2 F2: Nonsynonymous to synonymous heteroplasmy ratio....168
Figure 5.6.3 F1: Heteroplasmies per site found in different genomic contexts at positions with different reference bases....169
Figure 5.6.3 F2: Observed heteroplasmy major allele base composition differences between gnomic contexts....170
Figure 5.6.3 F3: Nonsynonymous-to-synonymous heteroplasmy ratios depending on position reference base type173
Figure 5.7 F1: Prevalence of mtDNA variants associated with mitochondrial disease....175
Figure 5.7 F2: Correlation between observed and reported disease associated-variant prevalence patterns....178

Discussion and conclusions

Figure 6 F1: Major allele base type relationship to observed heteroplasmic variant MAF....186
Figure 6 F2: Heteroplasmy incidence at suspected DNA damage hotspot sites....188

1 Introduction and background

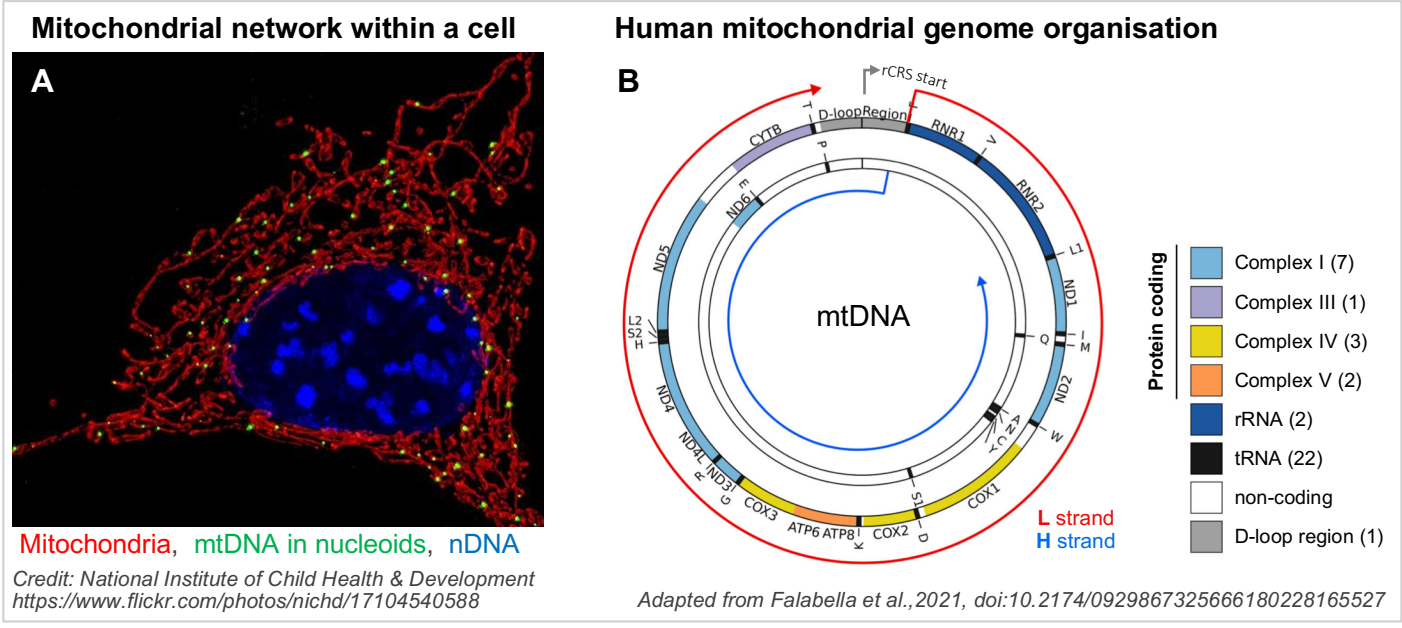
1.1 Mitochondrial genetics background

Mitochondria are highly metabolically active organelles capable of forming complex interconnected dynamic networks within eukaryotic cells (Ježek *et al.*, 2009), with a role in multiple cell functions including oxidative phosphorylation (OXPHOS), apoptosis, calcium signalling, phospholipid synthesis and others (Hopper *et al.*, 2006; Wang *et al.*, 2009; Osellame *et al.*, 2012). As organelles of presumed endosymbiotic origin (Archibald, 2015) mitochondria have retained their own small circular double stranded genome. The human mitochondrial genome is a 16,568 nucleotides long circular molecule encoding 22 tRNAs, 2 rRNAs and 13 proteins essential for OXPHOS function (Anderson *et al.*, 1981).

Unlike the nuclear DNA, the mammalian mitochondrial genome exists in multiple copies per cell, localised in discrete nucleoids that are distributed throughout the mitochondrial network. Each nucleoid typically contain 1-2 mtDNA copies (Kukat *et al.*, 2011), with the total mtDNA copy number per cell ranging between ~100 to over 250,000 copies per cell, depending on the tissue type (Santos *et al.*, 2006; Osellame *et al.*, 2012; Falabella *et al.*, 2021). Higher mitochondrial genome copy number is often observed in more metabolically active tissues (Osellame *et al.*, 2012). A notable exception to this is the mammalian oocyte, a cell type that becomes arrested during meiosis and remains quiescent for prolonged period of time (Mehlmann, 2005), yet may contain over 250 000 copies of mtDNA (Santos *et al.*, 2006). High mtDNA copy number is directly linked to oocyte viability (Santos *et al.*, 2006), enabling the maturing oocyte to meet the high energy needs required for resuming meiosis prior to ovulation (Cotterill *et al.*, 2013) and constituting a ready pool of mtDNA copies to be distributed between the dividing cells of the early pre-implantation embryo (Spikings *et al.*, 2007).

Figure 1.1 F1 panel A shows an example of the cellular mitochondrial network in a fibroblast with visible nucleoids where the mitochondrial DNA is localised. The cellular mitochondrial network is a dynamic structure undergoing fission and fusion (Westermann, 2010). Its morphology is closely linked to the bioenergetic state of the cell, with more interconnected mitochondrial networks characteristic to metabolically active cells, while mitochondria in less active cells tend to exist as multiple smaller, more functionally distinct compartments (Ježek *et al.*, 2009; Westermann, 2010).

Figure 1.1 F1: Mitochondrial network[1] and the architecture of the human mitochondrial genome



[1] Image credit: U. Manor, NICHD. CC BY 2.0

mtDNA replication and propagation in cell division

Mitochondrial genome undergoes asynchronous replication by DNA polymerase gamma. Replication begins at the H strand primary origin within the D-loop region with the nascent DNA strand displacing the original H strand. Once the replication fork passes the primary L strand origin, the light strand replication begins in the opposite direction (Clayton, 2000). As a result of this asynchronicity, during the replication process parts of the H strand temporarily exist as single stranded DNA and thus being potentially exposed to higher risk of DNA damage (Holt *et al*., 2012). Furthermore, only around 5% of mtDNA replication initiation events are thought to proceed all the way around the circle synthesising the full mtDNA molecule. The remaining 95% terminate prematurely producing approximately 650nt long single-stranded 7S DNA fragment that remains bound to the parental template strand resulting in a triple stranded D-loop structure (Falkenberg, 2018). The exact function of the 7S DNA remains debated, as are the details of the mtDNA replication model. For review of mtDNA replication models and the outstanding questions see Falkenberg, 2018.

Since the mtDNA replication is autonomous and takes place within the mitochondria, some debate exists in the literature to what extent it is coupled to the cell cycle. Some studies point to coordination between the mtDNA replication and the cell cycle, showing that the bulk of mtDNA synthesis takes place during the S phase (Sasaki *et al*., 2017) with the initiation of mtDNA replication slightly preceding nDNA synthesis (Chatre *et al*., 2013). Other research however argues against such synchronisation, attributing the spike in mitochondrial DNA synthesis in the S phase to partial replication producing excess of short-lived single stranded 7S DNA fragments while full replication of the mitochondrial genome remains independent of the cell cycle (Antes *et al*., 2010).

During cell division the pool of existing mitochondrial genomes needs to be split between the daughter cells as the mitochondrial network undergoes fission (Adebayo *et al*., 2021). Whether this is random or somehow regulated remains an active area of study. Recent research has found links between mitochondrial fission initiation sites and the locations of nucleoids containing recently replicated mtDNA, suggesting mitochondrial network dynamics as a potential mechanism for promoting segregation of mtDNA copies between daughter cells (Lewis *et al*., 2016; Nicholls *et al*., 2018).

Mitochondrial DNA damage and repair

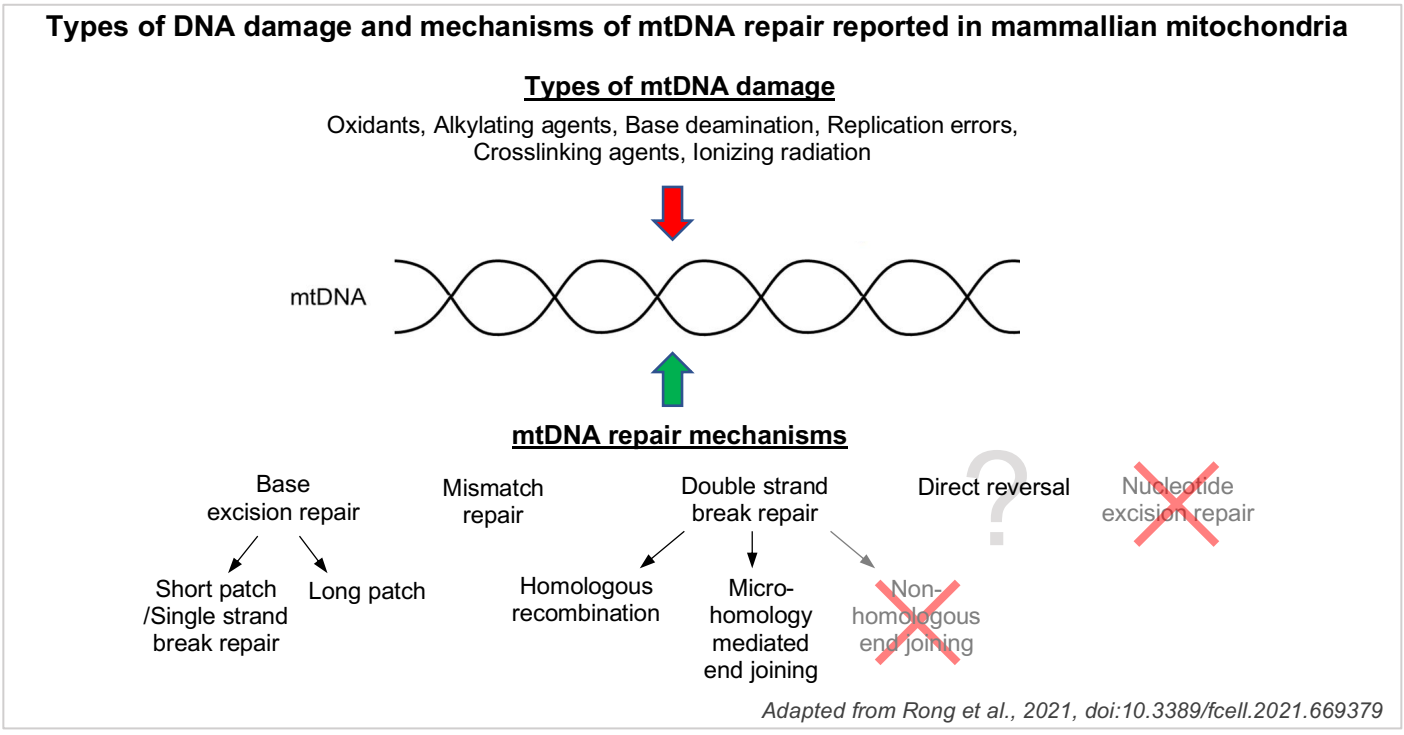
The mutation rate of the human mitochondrial genome exceeds that of the nuclear DNA (Rand, 2001; Lynch, 2010) with accumulation of somatic mitochondrial mutations linked to aging, neurodegenerative diseases and cancers (Chinnery *et al*., 2002; Schon *et al*, 2003; de Souza-Pinto *et al*., 2008; Jeppesen *et al*, 2011; Sondheimer *et al*., 2011). Historically, this used to be attributed to the highly reactive environment of the mitochondrial matrix, where, due to its close proximity to OXPHOS and lack of protective histones, mtDNA may be exposed to high levels of damaging agents capable of introducing DNA lesions (Shigenaga *et al*, 1994; Alexeyev *et al*., 2013). The types of DNA damage reported in mitochondria include alkylation and hydrolytic damage especially affecting cytosines, oxidative damage, formation of adducts, DNA strand breaks, and mismatched base introduction during DNA replication (LeDoux *et al*., 1999; Jeong *et al*., 2005; Zheng *et al*., 2006; Jiao *et al*., 2008; Ricci *et al*., 2008; Liu *et al*., 2010; Cline, 2012; Alexeyev *et al*., 2013; Wauchope *et al*., 2018).

Oxidative mtDNA damage induced by reactive oxygen species (ROS), a bi-product of OXPHOS, in particular was originally posited as the main driver of the mitochondrial mutation accumulation, forming the basis of the free radical theory of aging (Harman, 1956; Shigenaga *et al*., 1994). This theory proposed that mtDNA mutations caused by ROS-induced DNA damage result in OXPHOS dysfunction, which in turn increases ROS further leading to more mutations and further cellular dysfunction, ultimately resulting in cell senescence and death.

More recent research however points to the G>T/C>A transversion mutations, which are characteristic to ROS damage, constituting only a minority of mtDNA mutations found in aging tissues (Kennedy *et al*., 2013; Itsara *et al*., 2014; Kauppila *et al*., 2015), prompting reassessment of the role ROS-induced mtDNA damage plays in aging-related mitochondrial dysfunction. The abundance of G > A/C > T transitions and the distinct strand asymmetry of the mutations instead point to replication errors as the more likely reason for the aging-associated somatic mtDNA accumulation (Zheng *et al*., 2006; Kennedy *et al*., 2013).

Historically, despite the presumed high levels of oxidative damage, mitochondria were also thought to possess only a very limited repertoire of DNA damage repair mechanisms (Druzhyna *et al*., 2008). In recent decades however, evidence of numerous mtDNA repair and maintenance pathways has been found. mtDNA repair pathways well documented in human mitochondria include short- and long-patch base excision repair (Rong *et al*., 2021), single strand break repair (Das *et al*., 2010; Sykora *et al*., 2011) and mismatch repair (de Souza-Pinto *et al*., 2009). Evidence has also been emerging of DNA double strand break repair in human mitochondria via homologous recombination (Sage *et al*., 2010; Dahal *et al*., 2018) and microhomology-mediated end joining (Tadi *et al*., 2016). Non-homologous end joining repair of double strand breaks is thought to be largely absent from mitochondria (Tadi *et al*., 2016), although presence of some proteins involved in this pathway have been reported (Rong *et al*., 2021). Nucleotide excision repair, a key mechanism for chemically damaged nucleotide removal from DNA in the nucleus (Reardon *et al*., 2005), is also absent from mitochondria (Clayton *et al*., 1974; Cline *et al*., 2010), and whether direct mtDNA damage reversal takes place in humans remains unclear (Cai *et al*., 2005). **Figure 1.1 F2** summarises the types of DNA damage and mtDNA repair pathways documented in mammals.

Figure 1.1 F2: Types of DNA damage and the mitochondrial DNA repair mechanisms

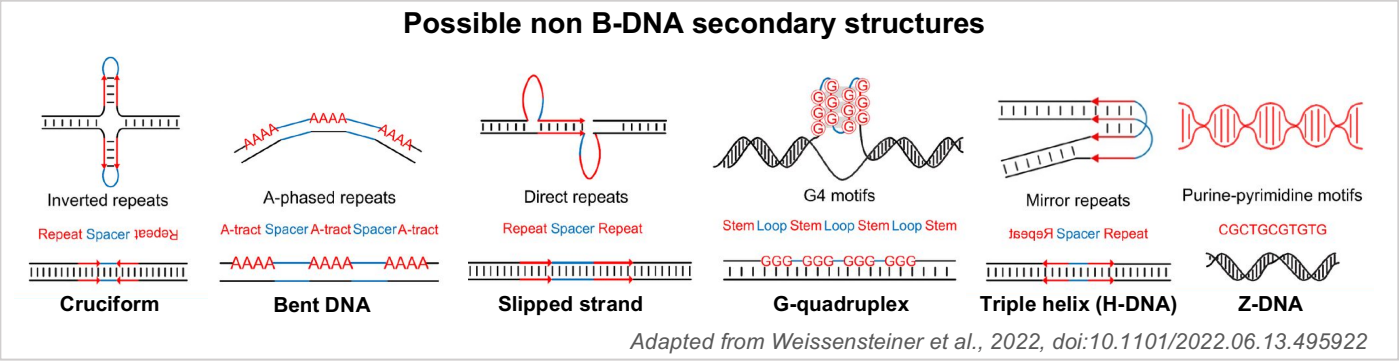


In addition to the damage repair pathways, several other routes for maintaining mtDNA integrity exist. Elimination of chemically damaged, pre-mutagenic free nucleotides from the mitochondria contributes to mtDNA damage prevention (Liu *et al*., 2010). The dynamic nature of the mitochondrial network also provides a route for recycling of entire damaged mtDNA copies (Westermann, 2010), with defective mitochondria isolated through fission and recycled via mitophagy pathways (Twig *et al*., 2008; Narendra *et al*., 2010). Evidence from several model systems points to mitophagy being able to selectively target and remove deleterious mtDNA variants during, for example, oogenesis (Cox *et al*., 2021; Palozzi *et al*., 2022).

DNA secondary structure role in mtDNA mutation

While most of the DNA exists in the canonical right handed double helix structure with 10bp/turn, certain DNA sequences may form non-canonical secondary structures, summarised in **figure 1.1 F3** below.

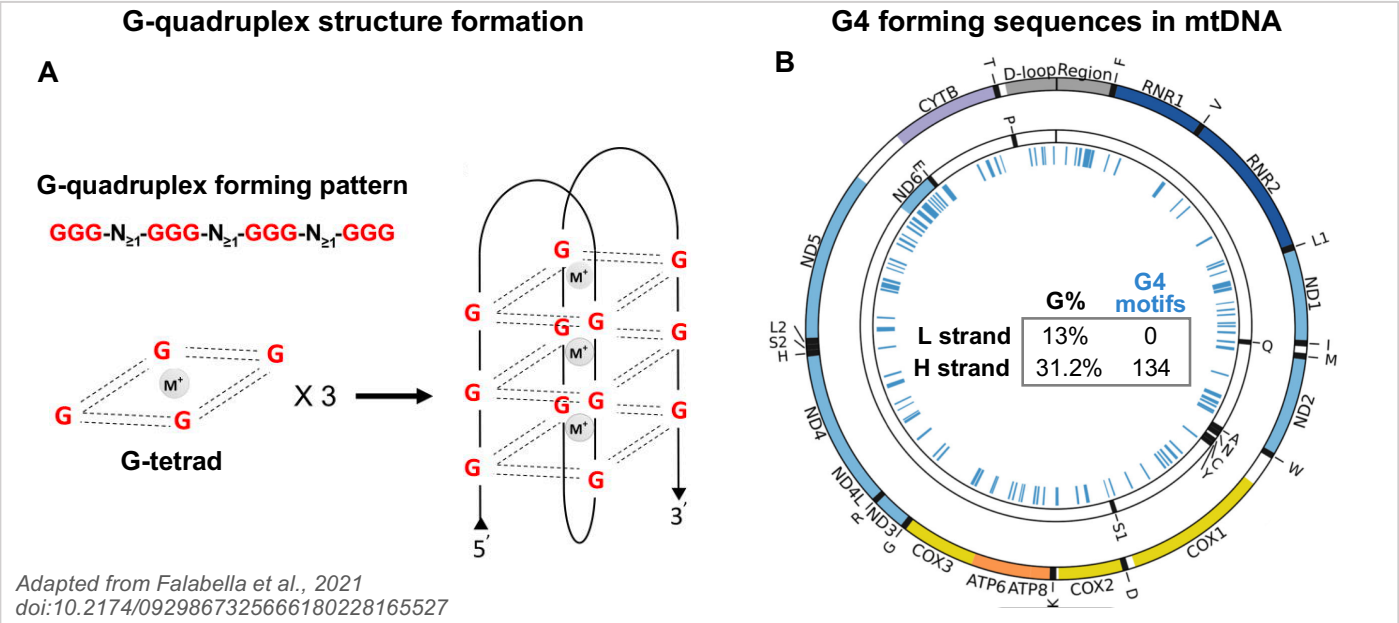
Figure 1.1 F3: Non B-DNA secondary structures



Adapted from Weissensteiner et al., 2022, doi:10.1101/2022.06.13.495922

Several studies have reported an association between mtDNA mutations and sequence elements capable of forming some of these non-canonical secondary structures. In particular, left-handed zig-zag DNA structures known as Z-DNA formed at alternating pyrimidine-purine tracts have been linked to mitochondrial deletions (Damas *et al*., 2012), while G-quadruplex (G4) structures which form in G rich regions have been implicated in single base mismatch formation via stalling of the polymerase (Butler *et al*., 2020). G4 secondary structure of the DNA and RNA are particularly fascinating in the mitochondrial context, where these structures are not only associated with mutations, but also thought to be heavily involved in the regulation of the mtDNA replication and transcription (Wanrooij *et al*., 2010; Falabella *et al*., 2021), and formation of DNA-RNA hybrid G4 structures might be one of the factors modulating heteroplasmic variant segregation bias over time (Røyrvik *et al*., 2020). **Figure 1.1 F4 panel A** illustrates how G4 structures form between four G-rich stems interspersed with non-G loops, and the validated mtDNA locations where G4 formation has been confirmed are shown in **Panel B** (Bedrat *et al*., 2016). Interestingly, the large G content differences between mtDNA L and H strands mean that G4 formation is largely restricted to the H strand.

Figure 1.1 F4: G-quadruplex structure and possible location in the mitochondrial genome



Adapted from Falabella et al., 2021
doi:10.2174/0929867325666180228165527

1.2 Mitochondrial heteroplasmy

Mitochondrial heteroplasmy is a phenomenon where mtDNA molecules coexist within the same cell. It is a direct consequence of the multi-copy nature of the mitochondrial genome (Rand, 2001), with important implications for mitochondrial genome evolution, mitochondrial diseases, aging and cancer.

Threshold effect of mtDNA mutations

The ability for mitochondria for form interconnected networks through fusion enables mitochondria with wild type mtDNA to compensate for defects from mutant mtDNA copies, creating an environments more permissive to the presence of low levels of deleterious variants compared to the nucleus (Rong *et al*., 2021). Because of this, mtDNA mutations tend to show a threshold effects with sufficiently high variant allele frequency necessary for a phenotype to emerge. The majority of pathogenic mtDNA variants represent loss-of-function, with the reported disease thresholds ranging between 40% to over 90% depending on the mutation and the tissue type (Rossignol, *et al*., 2003), although at least one pathogenic variant in a tRNA gene, m.5545C>T, has been reported where clinical disease phenotype emerges at minor allele frequency (MAF) <25% and potentially as low as 4% (Sacconi *et al*., 2008), suggesting a much rarer gain-of-function mode of action.

Due to the threshold effects typical to most mtDNA mutations, on an organism level even potentially deleterious heteroplasmy may exist without a detrimental effect provided that deleterious variants remain below a phenotypic threshold and a sufficient proportion of wild type mtDNA copies are passed onto daughter cells during any cell division (Rossignol, *et al*., 2003). However, if the distribution of mitochondrial genomes between daughter cells is biased, or if the number of mitochondrial genomes passed onto the daughter cells is small, deviations from the original heteroplasmy levels can be extremely rapid. A classic example of this is the mitochondrial bottleneck effect observed in mammalian oogenesis, where mtDNA copy number may drop to as low as 10-100 copies before rapidly expanding to over 100,000 copies in mature oocytes (Shoubridge, 2000). Such dramatic reduction in the effective population size of mitochondrial genomes present in the cell leads to often dramatic differences in heteroplasmy levels between mother and offspring, with significant implications for mitochondrial disease inheritance (Monnot *et al*., 2011; Rebolledo-Jaramillo *et al*., 2014).

Mitochondrial diseases

The term 'mitochondrial diseases' is used to encompass a group of clinically heterogeneous syndromes with a range of clinical symptoms, often affecting high energy demand organs, such as the brain, heart and skeletal muscle (Gorman *et al*., 2016). **Figure 1.2 F1** below, adapted from Ryzhkova *et al*., 2018, lists some of the known disease causing mtDNA point mutations. The same variants may be associated with different syndromes depending on the heteroplasmy levels in different tissues (Rossignol, *et al*., 2003). Examples of this include variants m.3243 A>G (MELAS, CPEO/PEO, cardiomyopathy/encephalomyopathy (Chinnery *et al*., 1997; Vydt *et al*., 2007)), m.8363 G>A (MERRF, cardiomyopathy/encephalopathy (Lorenzoni *et al*., 2014)), m.8993 T>C/G (LS, NARP (Trounce *et al*., 1994; Nunnari *et al*., 2012)), m.10197 G>A (MELAS, LS (Leng *et al*., 2015)), m13042 A>T (MELAS, MERRF (Lorenzoni *et al*., 2014; Schinwelski *et al*., 2015)) and m.14484 A>G (LHON, cardiomyopathy/encephalopathy (Johns *et al*., 1993; Cao *et al*., 2010)).

Deleterious mutations at sub-threshold MAF have been found to be relatively common in the population, with around 1 in 200 healthy individuals found harbouring at least one pathogenic variant at a low MAF (Elliott *et al*., 2008). The population frequencies observed by Elliott *et al*., 2008, the MELAS-associated variant 3243A>G and 4 other common pathogenic variants are listed in **figure 1.2 F1**. Heteroplasmic 3243A>G mutation is the most common

pathogenic variant observed in the human population, with the reported incidence in population, depending on the source population and tissue, ranging between 0.01% to over 1.6% (Li *et al*., 2022), although heteroplasmies above phenotypic threshold are much rarer (Schaefer *et al*., 2008), indicating strong selection against deleterious variants at higher frequencies.

Figure 1.2 F1: Pathogenic mtDNA point mutations associated with mitochondrial diseases

Mitochondrial genome point mutations associated with mitochondrial diseases *(adapted from Ryzhkova et al., 2018)*

Disease	Gene	Position	Mutation	%	Disease	Gene	Position	Mutation	%
MELAS	MT-TL1	3243	A>G	*0.14%*	LS	MT-ATP6, complex V	8993	T>C/G	
		3256	C>T			MT-ND3, complex I	10197	G>A	
		3271	T>C			MT-ND5, complex I	13513	G>A	
	MT-ND3, complex I	10197	G>A		Aminoglycoside induced hearing disorders	MT-RNR1	1095	T>C	
	MT-ND5, complex I	13042	A>T				1494	C>T	
	MT-CYTB, complex III	15533	A>G				1555	A>G	*0.07%*
MERRF	MT-TK	8344	A>G		CPEO/PEO	MT-TV	1658	T>C	
		8356	T>C			MT-TL2	12315	G>A	
		8363	G>A			MT-TL1	3243	A>G	*0.14%*
	MT-ND5, complex I	13042	A>T			MT-ND1, complex I	3421	A>G	
NIDDM	MT-RNR2	3200	T>C		NARP	MT-ATP6, complex V	8993	T>G/C	
	MT-TL1	3242	G>A		Cardiomyopathy and encephalo-myopathy	MT-TL1	3243	A>G	*0.14%*
		3252	A>G				3260	A>G	
		3264	T>C			MT-TI	4269	A>G	
	MT-ND1, complex I	3316	G>A			MT-CO2, complex IV	7587	T>C	
		3394	T>C			MT-TK	8296	A>G	
	MT-ND6, complex I	14577	T>C				8348	A>G	
	MT-ND2, complex I	14833	A>G				8363	G>A	
LHON	MT-ND1, complex I	3460	G>A	*0.11%*		MT-CO3, complex IV	9957	T>C	
	MT-CO3, complex IV	9804	G>A			MT-TG	9997	T>C	
	MT-ND4, complex I	11778	G>A	*0.11%*		MT-TH	12192	G>A	
	MT-ND6, complex I	14459	G>A			MT-TL2	12297	C>T	
		14484	A>G	*0.11%*		MT-ND6, complex I	14484	A>G	*0.11%*
	MT-CYTB, complex III	15257	G>A			MT-CYTB, complex III	15059	G>A	

Abbreviations: *MELAS – mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes; CPEO/PEO – chronic progressive external ophthalmoplegia syndrome/progressive external ophthalmoplegia; MERRF – myoclonic epilepsy associated with ragged red fibres; NIDDM – noninsulin-dependent diabetes mellitus; LS – Leigh syndrome; LHON – Leber hereditary optic neuropathy; MIDD – maternally inherited diabetes and deafness; NARP – neuropathy, ataxia and pigmentary retinopathy; % - reported prevalence in population by Elliot at al., 2008*

Mitochondria in cancers

The role of somatic mtDNA mutations and mitochondrial dysfunction in cancers is a complex and extensively studied area. Increased reliance on mitochondrial respiration has been found in various cancers, including pancreatic cancers, Hodgkin's lymphoma, and myeloid leukaemia (Lagadinou *et al*., 2013; Viale *et al*., 2014; Birkenmeier *et al*., 2016; Kuntz *et al*., 2017), highlighting OXPHOS a potential therapeutic target (Ashton *et al*., 2018). Yet in many others, the opposite is true, and reduced mitochondrial function is observed in for example acute myeloid leukaemia, and prostate, kidney and colorectal cancers (Larman *et al*., 2012; Sriskanthadevan *et al*., 2015; Smith *et al*., 2020; Yuan *et al*., 2020; Bezwada *et al*., 2024). In fact, cancers affecting the same tissue or organ system may exhibit varied relationships to mitochondrial function. For example, some ovarian and breast cancers are linked to impaired mitochondrial respiration (Larman *et al*., 2012; Putignani *et al*., 2012) while others exhibit increased reliance on OXPHOS (Whitaker-Menezes *et al*., 2011; Wu *et al*., 2022). Similarly, diffuse large B-cell lymphomas are also classed into high- and low-OXPHOS subgroups (Caro *et al*., 2012).

It is evident that the exact involvement of mitochondrial OXPHOS (dis)function is highly dependent on the organ system affected, the specific type of cancer, and its stage of progression, with the characteristics of the mtDNA mutation signatures found in cancers being equally varied.

Given the high proliferation of cancer cells, strongly deleterious mtDNA mutations, such as those resulting in OXPHOS protein truncation are thought to be strongly selected against in most cancers, at least at phenotypic

heteroplasmy thresholds (Yuan *et al*., 2020). This is not universal however, with enrichment of nonsense mtDNA mutations at higher-than-expected MAF having been found in kidney, colorectal and thyroid cancers (Yuan *et al*., 2020). Inactivation of the mitochondrial function may be important for these cancer types. This is supported by 13C-labelling experiments showing reduced mitochondrial metabolism and OXPHOS dysfunction in primary renal cell carcinomas with OXPHOS Complex I mutations (Bezwada *et al*., 2024). Interestingly though, the same study also points to metabolic profile divergence between the primary RCC tumour and its metastasis. In RCC mouse models OXPHOS function is not important for primary tumour growth yet is critical for metastasis forming ability, and in human RCCs higher tumour cell OXPHOS activity correlates with worse clinical outcomes (Bezwada *et al*., 2024).

Somatic mtDNA mutations that disrupt OXPHOS function are also common in prostate and colorectal cancers where they can persist at high heteroplasmy levels (Petros *et al*., 2005; Schöpf *et al*., 2020; Smith *et al*., 2020). A downstream consequence of OXPHOS disruptions is compensatory remodelling of the cellular metabolism to promote cell survival, which enables cancer cells to overcome OXPHOS deficiency (Schöpf *et al*., 2020; Smith *et al*., 2020). Smith *et al*., who focus on colorectal cancers, posit that reorganisation of metabolism may in fact confer a selective advantage to tumour cells. This is backed by experiments in mouse models showing that OXPHOS defects can induce metabolic remodelling in colon epithelium and lead to increased cancer cell proliferation (Smith *et al*., 2020). Similarly, increased growth after introduction of OXPHOS-inhibiting mutations has also been demonstrated for prostate cancer cells in vitro (Petros *et al*., 2005), with clinical data linking higher non-synonymous mtDNA mutation loads and metabolic remodelling to worse patient outcomes (Kalsbeek *et al*., 2018; Schöpf *et al*., 2020).

Increased ROS signalling has been posited as a possible mechanism that could confer selective advantage during tumorigenesis to cells carrying OXPHOS-impairing mtDNA mutations (Petros *et al*., 2005; Park *et al*., 2009). Inhibition of OXPHOS pathways increases cellular ROS production, which, despite being toxic at higher concentrations, can promote cell proliferation at lower levels (Sabharwal *et al*., 2014). There are however costs to OXPHOS impairments, as functional respiration is required for cell proliferation. The resulting cost-benefit balance may explain observations made by Park *et al*., that link tumorigenesis specifically to heteroplasmic and not homoplasmic mutations (Park *et al*., 2009). Focusing on frameshift mutations in the ND5 gene found in colorectal cancers, study found that these mutations promote cell proliferation in vitro when found at a heteroplasmic level, yet result in reduced cell viability inhibiting tumorigenesis when present at a homoplasmic level (Park *et al*., 2009). In vivo, such ND5 frameshift mutations tend to be observed at homoplasmic levels in late stage tumours, which Park *et al*. argue is likely the result of changing selective pressures during cancer progression. Early on, mutations may accumulate due to the selective advantage conferred by increased ROS but are prevented from reaching heteroplasmy by the necessity to maintain functional respiration needed for rapid proliferation. Only later, as tumours adapt to hypoxic environment, do OXPHOS-inhibiting mutations become free to reach homoplasmic levels, which we see in late stage tumours (Park *et al*., 2009). This view is also consistent with findings in kidney cancers, where functional OXPHOS was shown to be required for metastasis formation (Bezwada *et al*., 2024). Similarly to initial rapid proliferation, metastatic invasion is likely to favour cells that have functional OXPHOS and are robust enough to survive and proliferate in the new environment of the secondary site. Heteroplasmy dynamics and varied mtDNA mutation patterns are therefore likely are a reflection of the changing selective pressures during cancer progression.

Whether OXPHOS-impairing mutations a actively selected for during tumorigenesis is however disputed. Park *et al*. highlight that, despite being an attractive explanation, the accumulation of deleterious mtDNA mutations observed in cancers in itself does not prove positive selection for OXPHOS impairments (Park *et al*., 2009). The levels of OXPHOS-impairing mtDNA mutations observed could also be a consequence of an overall genome instability characteristic to cancers (Park *et al*., 2009).

A 2014 analysis of the spectrum and heteroplasmy of mtDNA point mutations across 31 different cancer types also found no evidence of positive selection for OXPHOS-impairing mutations (Ju *et al*., 2014). Missense mutations were found to behave largely neutrally drifting to fixation, whereas the more severe mutation introducing stop codon in protein coding genes or disrupting tRNA anticodons, tended to remain heteroplasmic, which authors argue is evidence of negative selection against such mutations (Ju *et al*., 2014). The fact that these observations holds true regardless of cancer type, including for the cancers with some of the highest overall mtDNA mutation burdens observed (e.g. prostate, colorectal), is especially important, and casts doubts on the OXPHOS-impairment positive selection hypothesis.

Mutation signatures found by Ju *et al*., in mtDNA show extreme asymmetry between the L and H stands, with by far the most common substitutions being C>T on H strand and T>C on L strand (Ju *et al*., 2014). The striking similarity to the signatures of the human mtDNA evolution over time responsible for the GC content differences between strands (Nikolaou *et al*., 2006) points to similar intrinsic factors linked to mtDNA replication as the primary explanation behind both the somatic mtDNA variation in cancers as well as the evolution of the human mitochondrial genome over time (Ju *et al*., 2014).

Population genetics within a cell

Whether one is studying cancers, aging, congenital diseases or other aspects of mitochondrial biology, the mtDNA mutation dynamics must be considered in the light of the multicopy nature of the mitochondrial genome. Mitochondrial genetics are characterised by multiple hierarchical levels where selection can act on the mtDNA: at the level of multiple mtDNA molecules within the mitochondrion, at the level of multiple mitochondria per cell, at the level of different cells within the tissue of an organism, to finally the level of different organisms within the population. At each of these levels, the composition of the smaller units affect the fitness of the larger unit, and thus are subject to the population genetics principles of mutation, selection and drift (Rand, 2001; Shtolz *et al*., 2019).

The high mutation rates and a lack of routine recombination experienced by the mammalian mtDNA, which would otherwise lead to proliferation of deleterious variants via Muller's ratchet (Muller, 1964), is counteracted by genetic drift and selection (both within and between cells) either fixing or eliminating new variants. mtDNA evolution is thus dependent on transient heteroplasmy levels that are a direct result of the mutation-selection balance (Rand, 2001; Shtolz *et al*., 2019). The evolution of the mtDNA is also impacted by genetic bottlenecks which reduce the effective population size of mtDNA genome copies within the cell (Radzvilavicius *et al*., 2022), such as the already discussed bottleneck in oogenesis where mtDNA copy number drops by ∼3 orders of magnitude (Ježek *et al*., 2009; Zhang *et al*., 2018). Other mtDNA bottlenecks have been found during differentiation of specific tissue lineages (Tang *et al*., 2022), and even a cell division that simply splits the existing mtDNA population evenly between two daughter cells can be looked at as a more minor bottleneck.

One of the consequences of genetic drift and random mtDNA segregation during the bottlenecks of oogenesis and cell division is clonal expansion of mitochondrial variants, which may allow mutated mtDNA molecules to become the dominant variants in some cells, leading to mosaic tissue heteroplasmy levels (Lawless *et al*., 2020). For example, this is demonstrated in a 2013 study showing extreme tissue specific differences for the heteroplasmy m.16093T>C, with the observed heteroplasmy levels ranging from 5% to 62% between different tissues (Krjutškov *et al*., 2014). Another example of a genetic bottleneck, albeit at the cell rather than individual mtDNA copy level is metastasis formation in cancer, which can lead to large shifts in both the nDNA and mtDNA mutational profiles between the primary and secondary tumours (Chambers *et al*., 2002)

There are benefits to high mtDNA mutation rates, such as an enhanced ability to explore mutational space, which in combination with biased mtDNA segregation during cell division can enable mtDNA to faster adapt to changing environmental pressures (Castellana *et al*. 2011; Radzvilavicius *et al*., 2022), with evidence of adaptive selection observed in mitochondrial tRNA and rRNA evolution (Ruiz-Pesini *et al*., 2006).

The dynamics of heteroplasmy associated with aging are also interesting in the context of adaptation. While in actively dividing cells point mutation heteroplasmy tends to either get eliminated or become fixed, the opposite actually appears to happen in aging non-mitotic cells, where point mutation heteroplasmies accumulate over time, with long standing evidence of heteroplasmic mutation accumulation in ageing (Sondheimer *et al*., 2011). Research tracing somatic mtDNA mutation accumulation suggests that there even exists some positive selection for non-synonymous mutation, with a significant increase in such heteroplasmic mutation prevalence in individuals older than 60 (Hübner *et al*., 2019). A possible explanation of this apparent paradox may lie in the hypothesis of the 'survival of the slowest', which suggests that somatic mutations reducing OXPHOS activity actually provide aging cells a selective advantage, as the reduced ROS levels mean slower degradation compared to cells with higher OXPHOS activity and ROS levels (De Grey, 1997).

While we generally expect mtDNA variants conferring a selective fitness advantage to propagate better compared to other variants, there are cases where deleterious heteroplasmic variants are observed to outcompete their wild type counterparts. For example, we see this with mtDNA molecules harbouring large deletions, where the smaller molecule size is thought to confer a relative selective advantage through faster replication despite the variant being deleterious to the cell fitness overall (Diaz *et al*., 2002; Dubie *et al*., 2019). Another interesting case of such selfish mtDNA is associated with human mtDNA control region mutation in the CSB2 element that modifies the ability of this region to form a G-quadruplex structure with the nascent transcript, which in turn modulates the balance between replication and transcription (Røyrvik *et al*., 2020). In this case, 'selfish' variants that shift the balance towards replication outcompete the wild type alternatives.

Heteroplasmy landscape of the human population

Numerous studies have focused on identifying and studying mitochondrial heteroplasmies found in the human population in a wide variety of contexts from the aforementioned bottleneck effects to forensics to mitochondrial diseases (Stoneking, 2000; He *et al*., 2010; Li *et al*., 2010; Goto *et al*., 2011; Sondheimer *et al*., 2011; Sosa *et al*., 2012; Calabrese *et al*., 2014; Diroma *et al*., 2014; Ye *et al*., 2014; Rebolledo-Jaramillo *et al*., 2014; Ding *et al*., 2015; Skonieczna *et al*., 2015; Liu *et al*., 2018; Duan *et al*., 2019; Wei *et al*., 2019; Bolze *et al*., 2020; Calabrese *et al*., 2022; Laricchia *et al*., 2022). For an in depth review of heteroplasmy landscapes reported across these studies see **chapter 5.3**, but briefly the earliest studies were limited in scale and largely focused on heteroplasmies found in the hypervariable mtDNA segments within the D-Loop region, but with advances of sequencing technologies many large scale studies of MT genome-wide heteroplasmy have been carried out, reporting a landscape of heteroplasmic variants found across the mitochondrial genome, showing that heteroplasmy is a relatively common phenomenon. However due to the intrinsic difficulties of resolving very low heteroplasmy levels the vast majority of population scale studies have focused on heteroplasmies at frequencies above at least 1% and often above 5% to even 10% levels only, with only one study analysing heteroplasmies of as low as 0.2% (Duan *et al*., 2019) albeit in only 2 individuals. Thus little is known about the behaviour of very low level heteroplasmies below ~1% MAF.

The principal goal of our study thus is the identification and analysis of such a low-level heteroplasmy landscape of the human population.

Low level heteroplasmy landscape as a readout for changes to selection-mutation equilibrium

The existing transient sub-phenotypic threshold heteroplasmy landscape (Rand, 2001) provides a snapshot of the mutation-selection balance of the mitochondrial genome : the equilibrium between processes introducing mutations (e.g. DNA damage or replication errors) and those eliminating mutations (e.g. mitophagy or direct DNA damage repair). One may speculate that changes to any of the processes impacting mutation rate or selective pressure could leave a 'signature' in the mitochondrial low level heteroplasmy landscape in the form of heteroplasmy pattern or frequency changes. One example of this could be mutations tweaking the efficacy of some DNA damage repair machinery component. Considering the multi-copy nature of the mitochondrial genome, it is not unreasonable that changes to genome maintenance could more easily manifest as low level mitochondrial heteroplasmy compared to emergence of nuclear DNA variants. If this is the case, the mitochondrial genetic system could be a useful readout for various aspects of genome maintenance processes. The present study is motivated by this.

Another factor that makes mitochondrial heteroplasmy attractive as a potential readout for studying for example DNA repair machinery is the abundance of mitochondrial sequences generated as an off-target by-product from completely unrelated sequencing experiments. In terms of the copy number, the mitochondrial genome is much more abundant compared to the nuclear genome, therefore unless steps are taken to remove mitochondria during sample preparation, high number of reads from mitochondrial genome will be present in the sequencing output. Such by-product sequence data could prove to be a treasure trove of readily available data representing a wide range of relevant mutations one might be interested in studying in the context of DNA repair.

Limitations to low level heteroplasmy detection

There are however technological limitations that complicate low level heteroplasmy detection. Any experiment is bound to be associated with some level of noise, and when attempting to detect low level heteroplasmy, the signal to noise ratios can become problematic, as we are trying to detect very rare events. Different sequencing platforms vary, but for Illumina platforms, which we focus on in this study, per base error rates between 0.001 to 0.01 are cited (Meacham *et al*., 2011; Loman *et al*., 2012) , with per base error rates of ∼0.001 reported in a recent study for HiSeq and NovaSeq instruments (Stoler *et al*., 2021). Sequencing methods differ in their associated error profiles, with the propensity for specific types of errors and the characteristics of the problematic sequence patterns depending on the specific system used (Becker *et al*., 2012; Quail *et al*., 2012; Stoler *et al*., 2021). In case of Illumina platforms, the mismatch rates are found to be strongly linked to preceding sequence contexts, especially sequence patterns such as 'GGC' and 'GGT' (Nakamura *et al*., 2011; Stoler *et al*., 2021).

Another potential complication to low level heteroplasmy detection is presented by nuclear sequences of mitochondrial origin (NUMTs) – nuclear DNA regions related to the mitochondrial genome generated through multiple events of organelle-to-nuclear DNA transfer in our evolutionary past (Hazkani-Covo *et al*., 2010). NUMTs vary in size, with some of the largest known NUMTs covering around 90% of the mitochondrial genome (Hazkani-Covo *et al*., 2010). The similarity between NUMTs and the actual mitochondrial genome presents several potential problems for accurate heteroplasmy identification. Firstly, there is evidence that if mtDNA is PCR amplified prior to sequencing, primer placement in NUMT-like regions can inadvertently result in NUMT rather than mtDNA amplification (Parr *et al*., 2006). Another aspect in which NUMTs may interfere with mitochondrial heteroplasmy analyses is during sequence alignment, with reported tendency for reads containing heteroplasmic variants to be preferentially mapped to nuclear rather than mitochondrial DNA, reducing the available sequence coverage and potentially rendering some heteroplasmies invisible, although this is improved by using paired end and longer reads (Maude *et al*., 2019). At the same time, other studies indicate that NUMT presence actually introduce erroneous low-level apparent 'heteroplasmies' in the data, but again, this is mitigated through use of longer sequence reads

(Albayrak *et al*., 2016). All together this highlights the importance of the sequence alignment strategy choice, and necessity for a careful analysis and consideration of possible alternative erroneous sources for any putative heteroplasmy we may observe.

1.3 Research questions addressed in this thesis

In this study we use publicly available whole genome sequencing datasets to attempt to describe and analyse the low level human mitochondrial heteroplasmy landscapes on the human population scale, and to discover what the common patterns linked to low level heteroplasmy landscape behaviours might be.

The first results chapter of this focuses on determining whether such low level heteroplasmy landscape can be detected, and if so, whether we can identify sequence patterns associated with elevated heteroplasmy levels on the population scale. We also assess how reliably we are detecting genuine heteroplasmy signals rather than false positives arising from sequencing errors.

The second results chapter addresses the sequencing error component in heteroplasmy detection via systematic analysis of context dependent sequencing error rates associated with the Illumina HiSeq platform. A major focus of this chapter is analysis of systematic error variability patterns, with the aim to develop such a data clean-up strategy for removing the false positive signal component that we could apply to analyse low level heteroplasmy patterns in existing sequencing datasets.

The third and final results chapter shows application of our devised strategy for false positive signal reduction allowing us to detect and describe low level heteroplasmy landscape of minor allele frequencies (MAF) >0.1% across 3202 individuals. We present a comparison between putative heteroplasmy levels detected with and without the false positive signal reduction strategy, and demonstrate an observable decrease in MAF and inter-sample variability as a result of our developed approach.

2 Materials and methods

2.1 Additional methods, supplementary information, codebase and data availability

Supplementary data including appendices, additional methods, codebase and source datasets may be accessed online at: https://github.com/DainaSadurska/MT_sources_of_error_public

2.2 1000 Genomes Project phase 3 core dataset mapped to GRCh37

Sequence alignments of the 1000 genomes project phase 3 core dataset were used for analyses described in chapters 3 and 4. This dataset represents 2535 whole genome sequences from across 26 geographically diverse populations, all sequenced using Illumina HiSeq 2000 and 2500 platforms (1000 Genomes Project Consortium, 2015). 2 samples were eventually discarded due to insufficient mtDNA coverage, leading to the total analysed sample size of 2533 individuals.

Sequence alignment data generated as part of 1000 genomes phase 3 release was retrieved from IGSR ftp site ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/

We retrieved data in the form of BAM files representing alignments of 2,535 samples generated from 59,824 quality control passing sequencing runs aligned with BWA v0.5.915 to the GRCh37 (hg19) human genome reference assembly (GenBank accession: GCA_000001405.14) (1000 Genomes Project Consortium, 2015).

Samtools (Li *et al*., 2009) software was used to extract aligned reads mapped to the mitochondrial genome and generate files in pileup format. This format combines all available base calls mapped to a specific genome position are combined into a single character string, convenient for easy parsing of the data for variant calling and analyses. Two samples with alignment data for less than 75% of mitochondrial genome were discarder at this stage. Sample-level pileup files were parsed using a custom python scripts (see online for codebase) and sample position-specific read counts were combined into population level base count matrices with the dimensions of 2535 samples x 16569 positions. One population level read count matrix was generated for each base type and each sequencing direction. The population level base count matrices where then used to calculate reference and non-reference read counts and frequencies used in our analyses, using MATLAB software (see online for codebase).

As sample reference mtDNA sequences in our analyses we used the mitochondrial consensus sequences generated during phase 3 of the 1000 genomes project (see additional methods online), available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/MT/chrMT_sequences_2534.20160505.fasta.gz

2.3 Base and motif enrichment analysis using cumulative frequency curves

Enrichment score

Enrichment score or E score, developed in-house for this project, is a measure we devised for quantifying the strength of the association between any variable used to rank mtDNA positions (e.g. median non-reference base frequency (%NR) level) and any position property of interest (e.g. base type at position).

To calculate the enrichment score, first we rank all positions using whichever variable we are interested in. n is the total number of positions analysed and r represents rank of each position, with $r = 1$ indicating highest ranked position and $r = n$ indicating the lowest ranked position. Next, for each ranked position we determine whether or not it is a hit (i.e. a match) for our property of interest and construct a vector $H = \{h_1, h_2, h_3, \dots h_n\}$ of 1s and 0s representing ranked position hit status such that $h_r = 1$ if position ranked r matches the property of interest, otherwise $h_r = 0$. At each rank value r we calculate the observed fraction of total hits represented by positions ranked r or above as

$$CF_{obs} = \frac{\sum_{i=1}^{r} H_i}{\sum H}$$

where $\sum_{i=1}^{r} H_i$ is the cumulative hit count at rank r and $\sum H$ is the total number of hits found. The relationship between ranking variable and the property of interest can be visualised as an enrichment curve by plotting CF_{obs} plotted against r.

To quantify the observed enrichment of hits for our property of interest, we need to compare our observed enrichment curve to the neutral baseline expected under the null hypothesis of no association between our property of interest and ranking variable. If no association exists, hit probability is independent of position rank and therefore under null hypothesis rank r we expect the cumulative hit fraction $CF_{null} = \frac{r}{n}$. At each rank r we can calculate the difference between our observed cumulative hit fraction from this neutral baseline.

Enrichment score E describing how different on average our observed hit enrichment pattern is from the neutral baseline can be calculated as

$$E = \frac{\sum_{r=1}^{n} \left(\frac{\sum_{i=1}^{r} H_i}{\sum H} - \frac{r}{n} \right)}{n}$$

where $\frac{\sum_{i=1}^{r} H_i}{\sum H} - \frac{r}{n}$ is the difference between the observed cumulative hit fraction CF_{obs} and the neutral baseline cumulative hit fraction CF_{null} expected at each rank r under the null hypothesis.

Statistical significance of the enrichment score

Statistical significance of the observed enrichment score is estimated using a permutation test. E score null distribution is generated by randomly permuting hit vector H 50,000 times. One tailed P value is calculated from the null distribution as

$$p = \frac{N_{(E_{perm} > E_{obs})} + 1}{N_{perm} + 1} \text{ if } E_{obs} > 0 \text{ or } p = \frac{N_{(E_{perm} < E_{obs})} + 1}{N_{perm} + 1} \text{ if } E_{obs} < 0$$

where E_{perm} is E score calculated based on permuted H vector, N_{perm} is the total number of permutations and $N_{(E_{perm} > E_{obs})}$ or $N_{(E_{perm} < E_{obs})}$ is the number of permutations resulting in more extreme E score than the observed E score (Phipson *et al*., 2010).

Simple sequence motifs investigated for enrichment among differentially sorted mitochondrial genome positions

We investigated the association between %NR behaviour and the adjacent DNA sequence patterns. The patterns investigated encompass sets of related simple sequence motifs where the position of interest is either preceded or followed by up to 4 consecutive positions of a specific base type. The construction of the investigated motif sets is summarised in **figure 2.3 F1.** Two types of wildcards were used in motif definitions – either a fully unrestricted wildcard 'N' allowing any base type, or a wildcard 'not-Y' which restricts the variable positions to any base other than the base in the defined motif positions. In total, 416 such motifs (192 with wildcard 'N', 192 with wildcard 'not-Y' and 32 without wildcards) were analysed, with enrichment statistical significance determined via permutation test, with Bonferroni correction used to account for multiple comparisons.

Figure 2.3 F1: Simple adjacent sequence patterns consisting of a single base type around the position of interest

Families of simple single base type adjacent sequence patterns investigated for sequence context association with different position specific %NR population distribution behaviours

(a) Motif family: 7 related up to 4bp long motifs of single base type adjacent to position of interest

Motif	**Position of interest**	**Upstream / downstream positions adjacent to position of interest**			
	Position 0	Position +/-1	Position +/-2	Position +/-3	Position +/-4
1	defined base	defined base	variable base	variable base	variable base
2	defined base	defined base	defined base	variable base	variable base
3	defined base	defined base	defined base	defined base	variable base
4	defined base	defined base	defined base	defined base	defined base
5	defined base	defined base	defined base	variable base	defined base
6	defined base	defined base	variable base	defined base	defined base
7	defined base	variable base	defined base	defined base	defined base

defined base – *defined base type at position of interest*
defined base – *defined adjacent sequence motif base type (same base type in all defined positions)*
variable base – *wildcard: variable positions with either fully unrestricted or semi-restricted base types*

(b) Related family set formed by 4 motif families with the same defined position bases

UPSTREAM MOTIFS

	Motif family A	*Motif positions* -4	-3	-2	-1	*Position of interest*
ANY BASE at variable positions	*motif 1a*	N	N	N	Y	X
	motif 2a	N	N	Y	Y	X
	motif 3a	N	Y	Y	Y	X
	*motif 4a/b**	Y	Y	Y	Y	X
	motif 5a	Y	N	Y	Y	X
	motif 6a	Y	Y	N	Y	X
	motif 7a	Y	Y	Y	N	X
RESTRICTED BASES at variable positions	**Motif family B**	*Motif positions* -4	-3	-2	-1	*Position of interest*
	motif 1b	notY	notY	notY	Y	X
	motif 2b	notY	notY	Y	Y	X
	motif 3b	notY	Y	Y	Y	X
	*motif 4a/b**	Y	Y	Y	Y	X
	motif 5b	Y	notY	Y	Y	X
	motif 6b	Y	Y	notY	Y	X
	motif 7b	Y	Y	Y	notY	X

DOWNSTREAM MOTIFS

Position of interest	*Motif positions* +1	+2	+3	+4	**Motif family C**
X	Y	N	N	N	*motif 1c*
X	Y	Y	N	N	*motif 2c*
X	Y	Y	Y	N	*motif 3c*
X	Y	Y	Y	Y	*motif 4c/d**
X	Y	Y	N	Y	*motif 5c*
X	Y	N	Y	Y	*motif 6c*
X	N	Y	Y	Y	*motif 7c*
Position of interest	*Motif positions* +1	+2	+3	+4	**Motif family D**
X	Y	notY	notY	notY	*motif 1d*
X	Y	Y	notY	notY	*motif 2d*
X	Y	Y	Y	notY	*motif 3d*
X	Y	Y	Y	Y	*motif 4c/d**
X	Y	Y	notY	Y	*motif 5d*
X	Y	notY	Y	Y	*motif 6d*
X	notY	Y	Y	Y	*motif 7d*

X – *defined base type at position of interest, either A,C,G, or T*
Y – *defined adjacent sequence motif base type, either A,C,G, or T*
N – *wildcard for variable position of any base type without restriction, denoted as '·' or 'n' in specific motif sequences*
notY – *wildcard for semi-restricted variable position of any of three base types other than the defined motif base, shown in motif sequences as 'b' when not A, 'd' when not C, 'h' when not G, 'v' when not T*
* – *motifs 4a/b and motifs 4c/d are shared between two families each*

2.4 Sequence context dependent error rate estimation

The context dependent sequencing error (CSE) rate measurement approach we use relies on the assumption that during the sequencing by synthesis process the context dependent error rate should primarily be determined by the sequence context upstream of the base being added in the specific extension cycle. If a sequence motif upstream of a particular position is associated with higher probability of an incorrect base call at that position, a higher proportion of variant base calls should be found coming from reads in the forward sequencing direction relative to the motif orientation compared to reads in the opposite direction. Our approach for CSE rate estimation employs this directionality and is similar to a method used by Allhoff *et al.*, 2013. The approach for motif error rate measurement we use is as follows:

Step 1: Identification of motif locations and minimum read depth filtration

First, both sample mtDNA sequence strands are searched for all instances of the motif of interest. At each instance where the motif is found we denote the last motif position (at motif 3' end) as position 0 and extract all reads corresponding to this site . Any motif locations where less than 500 reads in each sequencing direction are found at position 0 are excluded at this stage.

Step 2: Local contingency table generation

For each of the remaining motifs of interest we separate the extracted reads into forward and reverse directions relative to motif orientation such that forward reads correspond to the same DNA strand on which the motif is found, while reverse reads correspond to the opposite strand. Reads in each direction are then split into matches (i.e. same base as individuals reference sequence) and mismatches (different base from individuals reference sequence) and local contingency tables constructed with match/mismatch defined on the horizontal dimension and read relative direction defined as the vertical dimension.

Step 3: Reverse mismatch fraction filtration

At this point reverse mismatch fraction filtration is applied to exclude sites where the motif of interest on one strand coincides with a different high-error motif on the opposite direction. To do this, for each motif instance identified we calculate the mismatch fraction of all reverse reads, and discard from analysis any motif instances where mismatches form more than 1% of reverse reads. Exclusion of motif sites where the reverse mismatch rate is above 1% means that we are limiting our error rates calculation to only those instances of the motif of interest where it is found in isolation, i.e. not obscured by a different error prone sequence on the opposite strand. A threshold of 1% was chosen based on similar thresholds used in previous literature and the reported error rates for Illumina platform . (Meacham *et al*., 2011; Loman *et al*., 2012)

Step 4: Sample-level contingency table construction and motif error rate calculation

After exclusion of any motif instances that fail the reverse mismatch fraction filter, we construct a sample-level contingency tables by summing the respective forward match (F_M), forward mismatch (F_{MM}), reverse match (R_M) and reverse mismatch (R_{MM}) read counts across the remaining motif instances within the sample. These sample-level contingency table values are then used to calculate motif forward and reverse mismatch rates or fractions, denoted as F_{MM}^{Rate} and R_{MM}^{Rate}, as follows:

$$F_{MM}^{Rate} = F_{MM}/(F_M + F_{MM})$$
$$R_{MM}^{Rate} = R_{MM}/(R_M + R_{MM})$$

Sample-level motif error rate (MER) is calculated as the difference between the forward and reverse read mismatch fractions

$$MER = F_{MM}^{Rate} - R_{MM}^{Rate} = \frac{F_{MM}}{F_M + F_{MM}} - \frac{R_{MM}}{R_M + R_{MM}}$$

Step 5: Population-level motif error rate calculation

In parallel, we also calculate population-level motif error rate. This is done by adding up the respective sample-level motif contingency table F_M , F_{MM} , R_M and R_{MM} values across all samples to obtain population level match and mismatch counts, denoted as $\sum F_M$, $\sum F_{MM}$, $\sum R_M$ and $\sum R_{MM}$. These global counts are then used to calculate population global motif error rate ($\sum$MER) as follows

$$\sum \text{MER} = \sum F_{MM}{}^{\text{Rate}} - \sum R_{MM}{}^{\text{Rate}} = \frac{\sum F_{MM}}{\sum F_M + \sum F_{MM}} - \frac{\sum R_{MM}}{\sum R_M + \sum R_{MM}}$$

Step 6: Statistical significance testing and correction for multiple comparisons

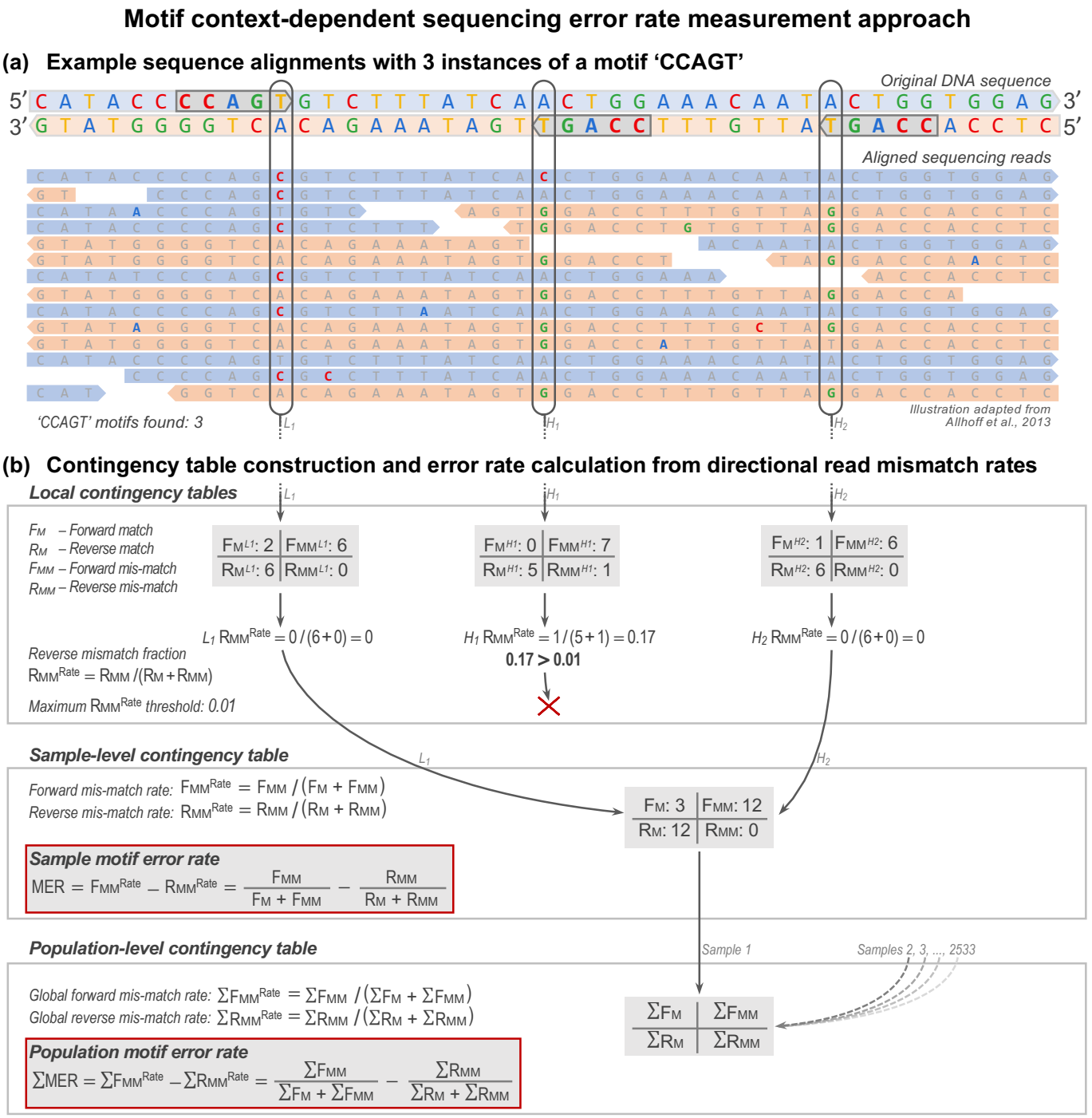
Statistical significance testing was carried out preferentially using either Fisher's exact test, except in cases with contingency table values exceeding 107, where Chi squared test was used instead. At values this high Fisher's exact test becomes computationally expensive, and Chi squared test is a generally accepted alternative for large sample sizes). To determine whether a motif is associated with significant strand bias at a sample-level we calculate p-values from sample-level contingency tables of relative direction specific reference and non-reference read counts, whereas error rate significance at a population level is assessed by computing p-values from the population level contingency tables. Null hypothesis assumes independence between read mismatch fraction and sequencing direction variables, i.e. no evidence of direction-specific error rate association with that motif. P-value computed from the contingency table represents the probability of at least as extreme contingency table being observed under the null hypothesis. We reject the null hypothesis when p-value is below a significance threshold $\alpha = 0.05$. Correction for multiple comparison is performed using Bonferroni method with an adjusted significance threshold $\alpha_{\text{MER}} = 0.05/(\textit{number of motifs} \times \textit{number of samples})$ used for sample-level motif error rates and adjusted threshold $\alpha_{\sum \text{MER}} = 0.05/\textit{number of motifs}$ used for population level error rates.

Figure 2.4 F1 below provides an overview of our motif error rate measurement approach using a hypothetical example of sequence reads aligned to a mtDNA sequence fragment containing three instances of a motif 'CCAGT'. **Panel A** shows an example sequence read alignment, with **panel B** illustrating read count contingency table construction and describing the sample-level and population-level error rate (denoted as MER and $\sum$MER respectively) calculation principles. **Panel C** walks through an example calculation of sample-level error rate associated with such a 'CCAGT' motif given the read alignment example used.

Panel C walks through an example calculation of the sample-level error rate that would be associated with such a 'CCAGT' motif based on the given read alignment example from **panel A**. In this hypothetical example three motif instances are found. However at one of these motif instances (labelled H_1) the reverse mismatch fraction (1 out of 6 reads) exceeds the maximum threshold of 0.01. Because of this, we exclude this motif instance and calculate the sample-level motif error rate from the other two motif instances only (labelled L_1 and H_2), yielding a result of 0.8. This very high error rate is a result of extreme strand bias where a much higher proportion of non-reference base 'G' is found at the motif position 0 in forward reads containing the motif compared to reverse reads without the motif.

It should be noted that the motif 'CCAGT' is used here as a completely hypothetical example, and in reality this motif is associated with a different error rate.

Figure 2.4 F1: Motif context-dependent sequencing error rate measurement approach



(c) Example error rate calculation for the motif 'CCAGT'

Contingency table values		Forward and reverse mis-match fractions	Motif error rate
FM = 2+1 = 3	FMM = 6+6 = 12	FMM Rate = FMM / (FM + FMM) = 12 / (3+12) = 0.8	MER = FMM Rate - RMM Rate = 0.8 - 0 = **0.8**
RM = 6+6 = 12	RMM = 0+0 = 0	RMM Rate = RMM / (RM + RMM) = 0 / (12+0) = 0	

Reads corresponding to L strand shown in blue and reads corresponding to H strand shown in orange;
L_1 – motif instance on L strand; H_1,H_2 – two motif instances on H strand; Ovals mark motif end position of interest;
'Match' / 'mis-match' – number of reads with the same / different base compared to sample reference sequence;
'Forward' / 'reverse' – sequencing read direction relative to motif orientation on L or H strands

Seven motif spaces investigated

We systematically measured motif CSE association for all possible permutations of 2bp to 8bp sequences, which we refer to as motif spaces. **Figure 2.4 F2** below provides a summary of the 7 investigated motif spaces, and the number of motifs of each length found and analysed in our dataset. The difference between 'possible' and 'analysed' motif counts for the larger motif spaces arises from the small size of the human mitochondrial genome. Not all possible sequence combinations of 6bp and longer are represented in the mtDNA.

Figure 2.4 F2: Summary of the analysed motif spaces

Summary of analysed motif counts across samples from each motif sapce

(a) Motif spaces analysed

Motif space	N2	N3	N4	N5	N6	N7	N8
Motif length	*2bp*	*3bp*	*4bp*	*5bp*	*6bp*	*7bp*	*8bp*
Possible motifs	*16*	*64*	*256*	*1,024*	*4,096*	*16,384*	*65,536*
Analysed motifs	*16*	*64*	*256*	*1,024*	*4,088*	*15,454*	*42,375*

(b) Total number of motif instances analysed across the population

Total instances	N2	N3	N4	N5	N6	N7	N8
Mean	3,620,820	905,205	226,301	56,575	14,172	3,748.7	1,367.2
Median	3,717,369	941,008	228,344	53,389	11,662	2,505.5	1,420
SD	795,326	282,776	95,902	30,679	10,047	3,623.1	1,648.5
IQR	821,183	276,253	149,249	45,672	13,501	3,963.8	1,955
Range	3,099,095	1,085,080	466,280	153,941	59,401	31,261	15,325

(c) Number of samples with the motif

Samples	N2	N3	N4	N5	N6	N7	N8
Mean	2,033.3	2,032.4	2,030.7	2,026.4	1,950.8	1,516.7	982.6
Median	2,033	2,033	2,031	2,028	2,018	1954	1,402
SD	1.0	0.9	2.4	7.5	316.9	792.8	930.8
IQR	1	1	2	5	24	392	1,933
Range	3	5	29	120	2,032	2,032	2,030

(d) Average number of motif instances found per sample

Avg. per sample	N2	N3	N4	N5	N6	N7	N8
Mean	1,780.8	445.4	111.4	27.9	7.1	2.1	1.2
Median	1,829.0	462.8	112.4	26.3	5.8	1.4	1
SD	391.2	139.1	47.2	15.1	4.9	1.6	0.5
IQR	403.1	135.8	73.4	22.4	6.6	1.8	0.004
Range	1,525.5	534.3	229.3	75.7	28.2	14.5	6.6

2.5 1000 Genomes Project extended cohort high coverage dataset mapped to GRCh38

For the analysis presented in **Chapter 5** we used sequence alignments generated as part of a new whole genome sequencing study by New York Genome Center (NYGC) (Byrska-Bishop *et al.*, 2022). The dataset of 3202 high-depth whole genome sequences made available by Byrska-Bishop *et al.* represents re-sequencing of the original 1000 genomes project phase 3 release panel of 2504 samples together with 698 additional, previously un-sequenced individuals related to the core panel. Sequences were generated using Illumina NovaSeq 6000 platform with 2x150bp reads to the target nuclear depth of 30x and aligned to the GRCh38 human genome reference assembly (GenBank accession: GCA_000001405.15) with BWA-MEM v0.7.15. The initial alignments at lane level were generated using BWA-MEM algorithm (Li, 2013) (https://github.com/lh3/bwa, v0.7.15), with Piccard tools (https://broadinstitute.github.io/picard, v2.4.1) then used to fix mate-pair information, merge lane-level BAM files into a single sample-level file, mark duplicate reads and coordinate sort the sample-level BAM file. Base quality scores were recalibrated using known SNPs with GATK (https://github.com/broadinstitute/gatk/, v3.5), and the final alignment files in CRAM format with their associated CRAI index files generated using Samtools (https://www.htslib.org/, v1.3.1).
The description of the full pipeline used by Byrska-Bishop *et al.* available online at:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/20190405_NYGC_b38_pipeline_description.pdf

This NYGS high coverage dataset was made publicly available in 2019. Sequence alignment data for the 3202 NYGC samples was retrieved from European Nucleotide Archive ftp site ftp://ftp.sra.ebi.ac.uk. Index of the NYGC high coverage dataset files is available on the IGSR ftp site:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/

Mitochondrial alignments were extracted from remote CRAM files and local BAM files were generated using samtools (Li *et al*., 2009). Properly mapped paired reads (SAM flags 99,147,83,163) with minimum base q 30 and mapping Q 25 were used to generate sample-level pileups. The read and mapping quality score choice was based on upper-range (i.e. most stringent) of quality score choices typically found in the previously published heteroplasmy literature.

Pileups (an alignment data format where all basecalls associated with a specific position are combined into a single string of characters) from each sample were parsed to obtain the number of calls of each base type at every position in F & R directions. Indel heteroplasmy analysis is beyond the scope of this study, therefore deletion and insertion counts were retained for overview purposes only.

Base count matrices (16569 positions x 3202 samples) were generated from parsed individual sample count sets for each base type for F and R directions separately, as well as for the sum counts across both directions, and used in subsequent steps for read frequency calculation and any subsequent heteroplasmy analysis.

Quality filtered read pileups were also used for consensus base calling using Samtools to determine individuals consensus sequences. These consensus sequences were later used to flag and discard likely indel positions from subsequent analysis, as only SNP heteroplasmy is analysed in this study.

2.6 Heteroplasmy identification in high coverage NYGC dataset

Heteroplasmy identification was performed using two slightly different approaches that we compare – a standard method and a consensus frequency based method. Both approaches share the initial alignment filtration steps such as exclusion of duplicate reads, removal of reads with insufficient mapping quality and exclusion of low quality base calls. In addition, we also exclude any positions where sample consensus sequences indicate indels as indel heteroplasmy analysis is beyond the scope of this study.

After alignment quality filtration, we apply a minimum read depth filter to our data, but this is where the standard versus consensus frequency based heteroplasmy identification methods diverge. With consensus frequency based method, prior to read depth filtration we incorporate an additional step where the read fraction not supported in both sequencing directions (i.e. "excess reads") is excluded leaving only the reads with bidirectional support (i.e. "consensus reads"). (The principle of excess read filtration is explained in detail later). Minimum read depth threshold is then applied to the consensus read count. In contrast, with standard method no excess read exclusion step takes place, and minimum read depth filter is applied to the total base/mapping QC passing read count.

From this point onward, the rest of allele filtration steps applied are the same between methods, except that wherever 'number of reads' is considered with standard method, the 'number of consensus reads' instead is the relevant variable with consensus frequency based method.

The minimum read depth thresholds we applied were 500 reads minimum in each sequencing direction with standard method and 500 consensus read minimum with consensus frequency based method.

The initial full high coverage dataset consisted of 3202 samples with 16568 analysed positions each, which is a total of 53,050,736 sites. After deletion site exclusion and application of minimum read depth thresholds 97.8% of the original 53,050,736 sites were retained, with only insignificant differences between methods. With standard method 51,883,365 sites pass initial filtration and with consensus method 51,882,772 sites pass initial filtration, which in only 593 sites fewer than standard method.

Allele filtration steps applied with standard method

For each site the total read counts across both sequencing directions were used to calculate frequencies of A,C,G and T base calls. In order to reduce the potential impact of sequencing artefacts, the dataset was filtered using similar strategies to those employed in recent literature (reviewed in **chapter 5**).

For each base type we determined whether that base type reads at any given position passed filter criteria of 7 different stringency levels. The minimum filter criteria at each level are summarised in **figure 2.6 F1** below.

Seven filtered datasets containing reads of only those base types at each position which passed the set criteria were generated. Each of the 7 datasets was checked for sufficient read depth, and any sites found to have less than 500 filter-passing reads in each sequencing direction were excluded from the analysis.

We then identified major and minor alleles for each filtered dataset. For each sample, we defined major allele as the most common base type at each position across both sequencing directions. Any other lower frequency base types at that position were defined as minor alleles 1, 2 and 3 in decreasing frequency order.

Within each sample, frequencies of major and any minor alleles for each position were then calculated. The major allele was used to define that individuals 'reference' sequence, with any minor alleles representing putative heteroplasmy. The putative heteroplasmy level is defined as the sum minor allele frequency (MAF) at each position.

MAF values can range between 0 and 50%, and can differ between sequencing directions. Therefore MAF was calculated at each site for the forward and reverse sequencing direction separately as well as overall across both directions. Matrices of the overall MAF at each position and each sample were then used for subsequent heteroplasmy analysis.

Figure 2.6 F1: Heteroplasmic variant filtration criteria

Filter level	Number of reads of the specific base type	Frequency of the specific base type
Level 1	1 read total	-
Level 2	1 read in each sequencing direction	-
Level 3	5 reads in each sequencing direction	-
Level 4	5 reads in each sequencing direction	0.1% total across all reads
Level 5	5 reads in each sequencing direction	0.1% of reads in each sequencing direction
Level 6	5 reads in each sequencing direction	1% total across all reads
Level 7	5 reads in each sequencing direction	1% of reads in each sequencing direction

The additional excess read filtration step applied with consensus frequency based method

Consensus frequency based heteroplasmy identification method introduces an additional step of data filtration, in order to remove as many artefact reads from the dataset as possible. We refer to this step as excess read filtration. For each site the separate read counts for each sequencing direction were used to calculate sequencing direction specific frequencies of A,C,G and T base calls. For each base, frequencies between sequencing directions were compared. Consensus frequency was defined as the lower of the two direction specific base frequencies, and all reads of this base from the lower frequency sequencing direction were included into the consensus read count.

For the opposite sequencing direction, an excess frequency of the base was calculated as the total base frequency in that direction minus the lower, consensus frequency of the same base in the opposite direction. Any excess reads contributing to such excess base frequency in one sequencing direction only were excluded from further analysis, only retaining the consensus fraction of reads for each base type for heteroplasmy identification and analysis.

After the excess read removal, we determined whether the remaining consensus reads of each base type at any given position passed the same 7 variable stringency filter criteria as used in standard method approach, and 7 datasets of only the filter-passing consensus reads were generated.

Each of the 7 datasets was checked for sufficient read depth, and any sites found to have less than 500 filter-passing consensus reads in each sequencing direction were excluded from the analysis.

We calculated allele frequencies and identified major and minor alleles at each position for each filtered consensus read dataset in the same way as for standard method. The putative consensus heteroplasmy for any given position in an individual was defined as the sum of the minor allele consensus frequencies (MAF) at that site and matrices of the consensus MAF at each position and each sample were used for subsequent heteroplasmy analysis.

Unlike with the standard method, consensus frequency based method results in allele consensus frequencies being the same between sequencing directions. Allele consensus read counts however can differ between sequencing directions, and may not necessarily be integers in the directions from which excess read fraction has been removed.

2.7 Heteroplasmy literature review

Aims for reviewing heteroplasmy literature

1. To determine what a 'typical range' of mitochondrial sequence data clean-up and filtration steps is that is typically used in heteroplasmy research literature by other authors. Based on this we can design our heteroplasmy identification 'standard' approach to use filtration steps and thresholds similar to those used in previous literature. Our eventual aim is to determine how useful our proposed heteroplasmy identification method that used consensus frequencies rather than raw read counts is, for which a heteroplasmy landscape identified using a more 'standard' method functions as an internal control.
2. To compile as comprehensive as possible a list of mitochondrial genome positions where heteroplasmies have already been reported previously in literature.
3. To assess how much results from different heteroplasmy studies typically tend to overlap, and determine whether a set of very commonly reported heteroplasmies exists that we should also be expecting to identify in our study.
4. To determine what the typical range is for heteroplasmy detection rates reported in literature, and to assess how our detected heteroplasmy levels compare against it.

The available literature was searched for research articles reporting heteroplasmy in the human mitochondrial genome. Combinations of search keywords used included "human", "mitochondrial genome", "mitochondrial DNA", "mtDNA", "heteroplasmy", "heteroplasmic", "site", "position", "variant". Reference lists of any relevant articles (review or original research) found were also examined for further original research papers with these and similar keywords.

Criteria used for including studies for review

- Study investigates heteroplasmy in the human mitochondrial genome; studies investigating mitochondrial heteroplasmy in other organisms are excluded.
- Study provides sufficiently detailed information about data processing and clean-up steps used, and clearly states what thresholds were set for classifying variants as heteroplasmic.
- Study provides a complete list of heteroplasmic positions or variants identified. Studies describing overall heteroplasmy levels but only including detailed information for some of the heteroplasmic variants identified are excluded.
- Study states how many samples the reported heteroplasmic variants or positions are found in.

In total we included 20 papers published between 2000 and 2022. Of these 20 studies, one study, by Bolze *et al*., 2020, is a pre-print published on BioRXiv server still awaiting peer review.

We have also included a review paper by Stoneking, 2000, combining results from several earlier studies focused on the D-Loop region only due to pre-NGS technology scale limitations. We checked each of the papers reviewed by Stoneking separately, and excluded any results from studies with insufficient information or otherwise not matching our inclusion criteria. But rather than treating these papers separately for our review purposes, we have chosen to combine this earlier research within the Stoneking, 2000, review umbrella. This is so as not to bias our review outcome towards the relatively larger number of smaller early studies done on a smaller sequence fragment.

Another study, by Sondheimer *et al*., 2011, that we chose to include uses a genotyping microarray rather than a sequencing approach, and therefore somewhat similarly to Stoneking, 2000, does not reports heteroplasmies from the entire mitochondrial genome. Unlike Stoneking, 2000, where heteroplasmies from a contiguous 955 base pair

region are reported, Sondheimer *et al.,* 2011, report heteroplasmies from 138 mitochondrial SNP locations throughout mitochondrial genome covered by HumanHap550 and HumanHap610 genotyping arrays.

Both for Stoneking, 2000, and Sondheimer *et al.,* 2011, we are accounting for the smaller sequence length analysed when calculating their reported heteroplasmy discovery rates.

3 Results 1 – The non-reference read frequency landscape

3.1 Introduction and research question

In the introduction we touched upon the idea that the heteroplasmy landscape we may observe across an individual's mitochondrial genome is a reflection of the mutation-selection balance representing the equilibrium of the ongoing introduction and removal of new mutations. Some of the heteroplasmic mitochondrial mutations must be arising by chance, while others may be a consequence of some intrinsic properties of the mitochondrial genome sequence or the cellular environment itself, making certain sequence contexts systematically more vulnerable to, for example, replication errors. While heteroplasmies with relatively high frequencies of the variant allele are generally quite rare and seemingly arise randomly, a dataset of several thousands of mitochondrial genome sequences gives us the opportunity to look for lower level mitochondrial heteroplasmy patterns potentially shared among many individuals. If such common low-level heteroplasmy landscape can be detected, it could potentially inform us about the underlying mechanisms by which the integrity of the mitochondrial genome is protected, whether this is by detection and correction of DNA damage, or purifying mitophagy removing entire genomes based on loss of function.

Thus the overarching aim of this thesis is to further our understanding of the low-level mitochondrial heteroplasmy levels, patterns and variability found across the human mitochondrial genome among a large number of individuals. The results presented in this thesis are split into three chapters. The present chapter presents an initial exploration of the observable landscapes of variant (i.e. non-reference) read frequencies across the mitochondrial genomes of 2533 individuals sequenced as part of the 1000 genomes project phase 3. Then, in chapter 4 we focus on identification and quantification of the sequencing error components found in our datasets. Finally, using these insights, in chapter 5 we present our best attempt of artefact exclusion and uncovering of as much of the true low-level mitochondrial heteroplasmy landscapes as possible.

For the purposes of this chapter, we shall be using the term 'non-reference read frequency' or '%NR' when referring to the proportion of variant reads observed at a genome position that indicate a different base from the reference sequence for that individual. The reason for using this term rather than 'minor allele frequency' or 'heteroplasmy' is primarily to avoid implying that our measured non-reference read frequencies entirely represent genuine heteroplasmy rather than some combination of real biological signals together with experimental artefacts.

Assuming that the %NR levels we detect arise from genuine biological variants compounded by systematic artefacts and noise, it is not unreasonable that the impact on the overall %NR level from each of these components would vary between mitochondrial genome positions, leading to potentially different %NR behaviours on the population scale (e.g. variability between individuals, peak behaviour relative to the overall %NR level of the individual, etc.). Such %NR population-level behaviour differences may give us a starting point for deciphering what proportion of %NR is the genuine heteroplasmy which we are ultimately interested in studying versus other contributing factors such as noise and systematic errors.

Thus in this chapter we aim to (i) describe the overall shapes of the mitochondrial non-reference read frequency landscapes observed across a population, (ii) address the %NR landscape variability between individuals and (iii) investigate the mitochondrial genome sequence patterns associated with different %NR behaviours on the population scale. In addition, we set out to highlight sequence patterns most linked to high levels of sequencing error versus true heteroplasmy.

3.2 %NR levels and landscapes

In this section we investigate the overall mitochondrial non-reference read frequency levels detected in the population and address the variability between samples and between positions. In this chapter we define 'landscape' as the value of a parameter (often %NR) graphed by mitochondrial genome position (often with respect to the entire mitochondrial genome).

The mitochondrial non-reference read frequencies analysed in this chapter represent a complex dataset of 2533 individual landscapes of each individuals' %NR level measured at each of the 16568 mitochondrial genome positions. There are multiple ways in which the data can be sliced asking different questions. We can look at the population-level %NR behaviour differences between MT genome positions, and search for outlier genome sites. Or we can look at the MT genome-level %NR behaviour differences between samples assessing %NR behaviour variability in the population. We can ask questions about the overall %NR levels through distributions or histograms, or we can look at the data as landscapes where the %NR levels are considered in the context of the MT genome position information.

Using these approaches, the aim of this section is to assess whether some common population-level %NR landscape exists, and if it does, to describe its features. Questions we address are:

- How variable individuals are in terms of their overall %NR levels averages and distribution shapes
- How much variation exists between MT genome positions in terms of the population average %NR levels or the distribution shapes
- Are individuals' %NR landscapes obviously structured in some way or are they more random.

3.2.1 Overall %NR level of an individual

To assess the overall MT genome %NR levels across the analysed samples we looked at the population distributions of sample MT genome mean and median %NR levels. Across our analysed 2533 samples the average MT genome mean %NR level was 0.6% and the average MT genome median %NR level was 0.2%. Both %NR level central tendency measures follow asymmetric distributions, with very low average MT genome %NR levels observed in most samples. MT genome mean %NR level below 0.5% is found in 37.9% of analysed samples and below 1% in 94.1% of samples. In comparison the MT genome median %NR level is below 0.5% in 94.3% of samples and below 1% in 99.3% of samples. This indicates that we are likely looking at quite skewed %NR distributions within most individuals with %NR across the majority of positions found below 0.5% level. Visual inspection of individuals' %NR distributions confirms that this is the case. This is illustrated in **figure 3.2.1 F1** below which shows histograms representing MT genome %NR level distributions in 6 representative samples.

Figure 3.2.1 F1: Mitochondrial genome non-reference read frequency distributions in 6 individuals

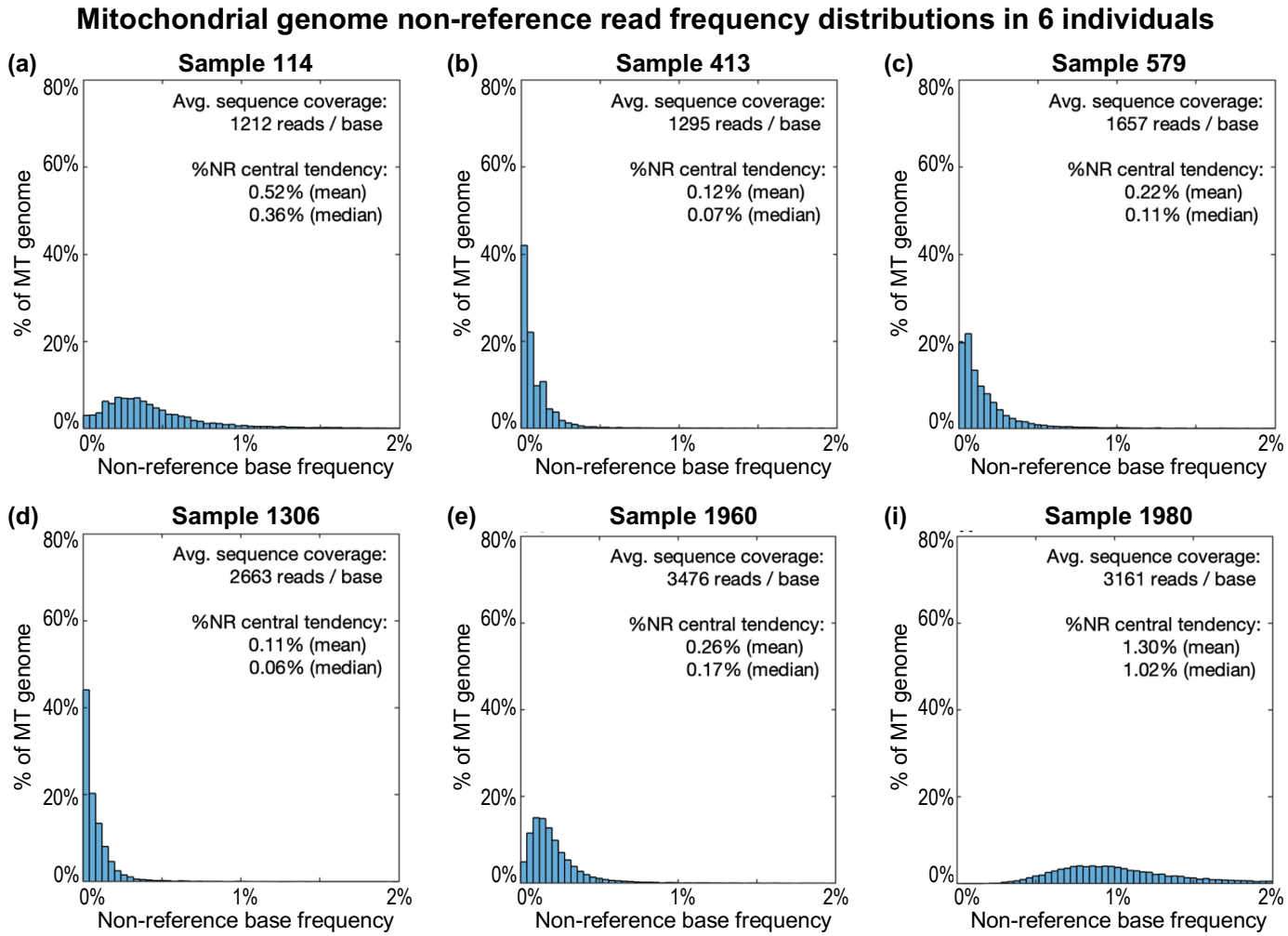
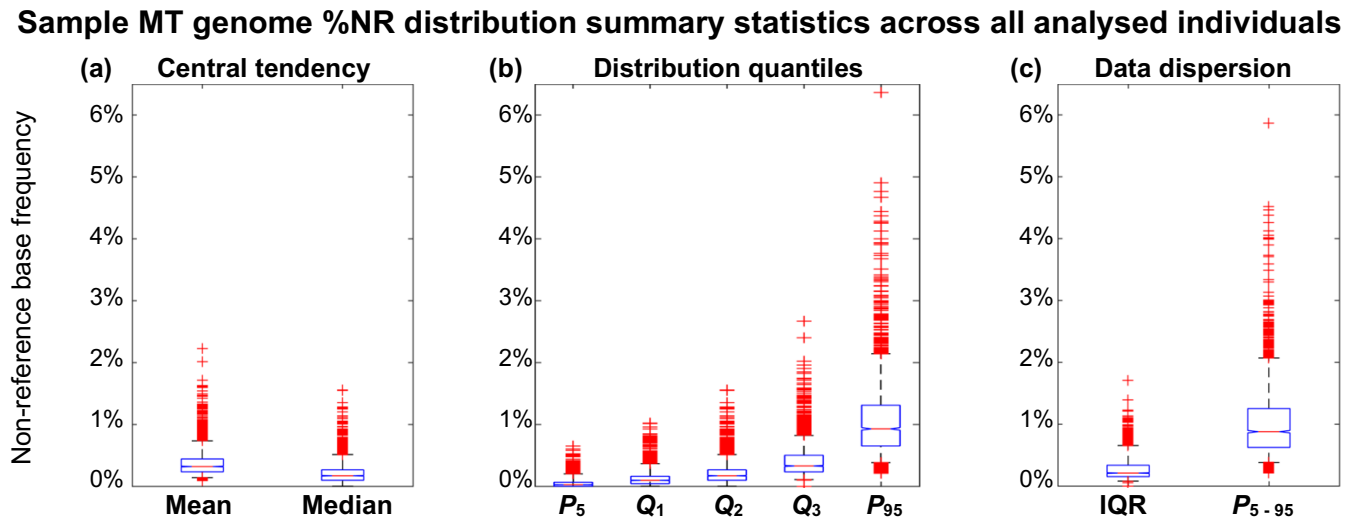


Figure 3.2.1 F1 illustrates that on individual sample level, MT genome %NR levels follow a skewed distribution where only a smaller number of sites where high %NR values are found while the bulk of MT genome is characterised by lower %NR levels well below 1%. The mean MT genome %NR is lower than the median in all 6 examples, as well as in every other analysed sample, as expected for skewed distributions. The set of 6 visualised samples shows a considerable variability both in terms of where the respective %NR distributions are centred around and how spread out they are. This indicates both a variability in sample baseline %NR levels as well as variation in the extent of within-sample %NR level variability.

While informative, **figure 3.2.1 F1** does not allow for generalised conclusions of what the overall mitochondrial non-reference base frequency levels are for the entire population and how these vary between samples. To address this on a population scale, we determined a range of summary statistics describing MT genome %NR distributions in each analysed sample and visualised the population distributions of each of these summary statistic values via box-and-whiskers plots in **figure 3.2.1 F2**.

Figure 3.2.1 F2: Mitochondrial genome %NR distribution summary statistics across all analysed individual



	Mean	Median	P_5	Q_1	Q_2	Q_3	P_{95}	IQR	P_{5-95}
Max	2.23%	1.56%	0.65%	1.02%	1.56%	2.67%	6.37%	1.71%	5.87%
P_{95}	0.73%	0.51%	0.20%	0.36%	0.51%	0.82%	2.14%	0.65%	2.07%
Box Hi - Q_3	0.44%	0.27%	0.06%	0.16%	0.27%	0.50%	1.31%	0.33%	1.25%
Median	0.32%	0.17%	0.03%	0.09%	0.17%	0.33%	0.93%	0.21%	0.88%
Box Lo - Q_1	0.23%	0.10%	0%	0.04%	0.10%	0.23%	0.65%	0.15%	0.62%
P_5	0.14%	0%	0%	0%	0%	0.11%	0.38%	0.08%	0.38%
Min	0.09%	0%	0%	0%	0%	0%	0.21%	0	0.21%

Figure 3.2.1 F2 confirms that on the population scale the bulk of MT genome is characterised by low %NR levels below 1%, with %NR ≤0.51% across half of mtDNA positions and ≤0.82% across three quarters of mtDNA positions observed in 95% of analysed samples. In half of our analysed samples 95% of mtDNA positions were found to have %NR levels ≤0.93%. **Figure 3.2.1 F2** also demonstrates that the analysed samples exhibit more variability in terms of %NR distribution spread and extremes than they do in terms of their MT genome baseline %NR levels, suggesting that it is the minority of high %NR peak positions that differ most from sample to sample.

It is evident that for the majority of the population most of the mitochondrial genome is characterised by relatively low non-reference base frequencies with higher %NR levels present in only some smaller proportion of MT genome sites. This however raises questions about the %NR peak site locations and patterns in the mitochondrial genome. Are the %NR peak locations consistent between individuals, and are there any obvious patterns we may be able to identify. We address these questions in the next section.

3.2.2 The shape of the MT genome %NR landscape

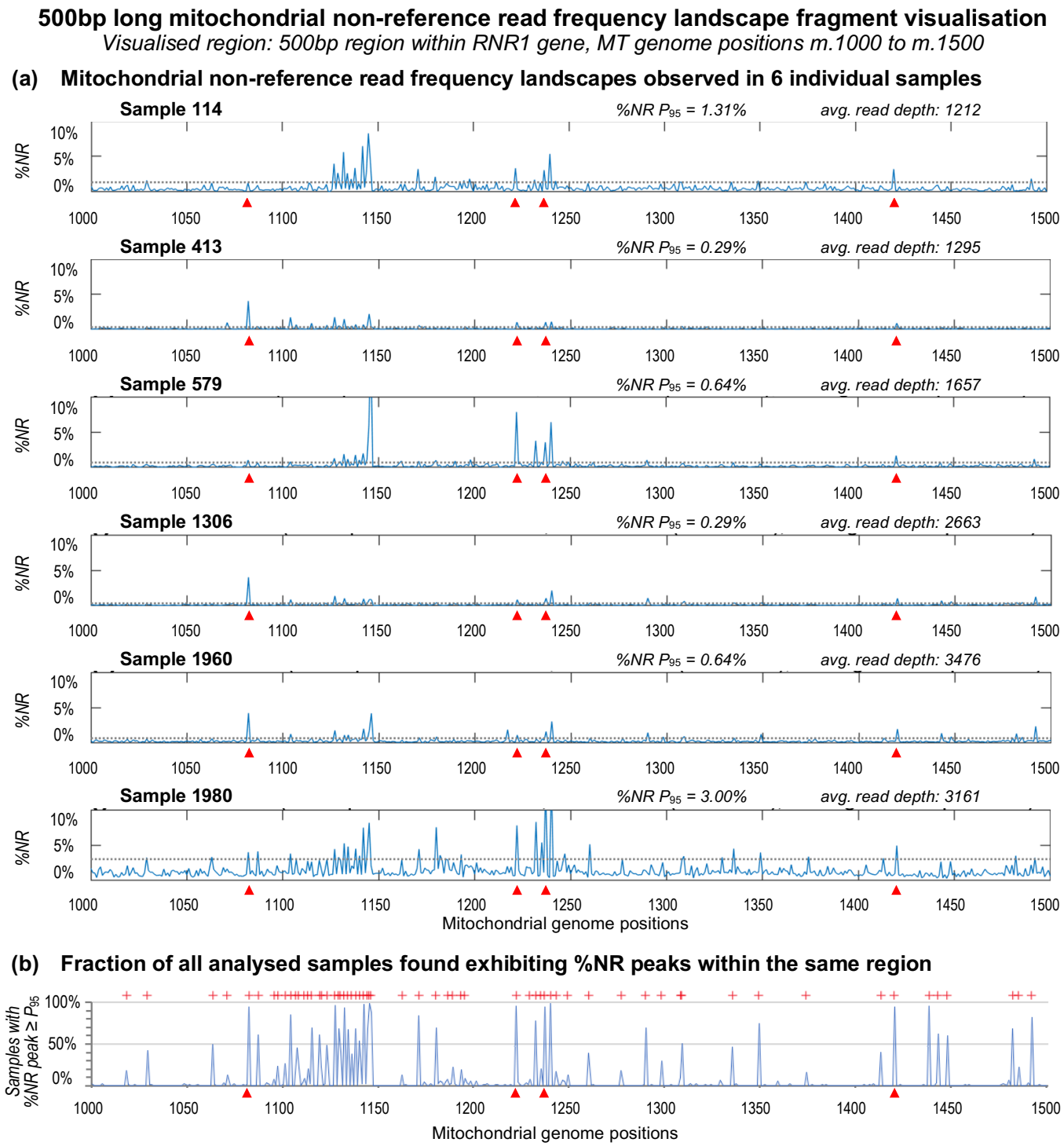
In the previous section we showed that in the majority of analysed samples the mitochondrial genome %NR levels follow skewed distributions. In this section we also consider MT genome position information and look at the mitochondrial %NR level landscape shapes. The term 'landscape' is used here to mean the pattern on %NR levels found at specific mtDNA positions plotted against the position coordinates. The question we address is whether or not similar %NR landscape patterns are shared among the individuals, and whether the landscapes are non-random.

Assuming that genuine heteroplasmy makes up some proportion of our observed %NR levels, we would expect positions with genuinely elevated heteroplasmy to also show elevated %NR level. Should heteroplasmy level patterns be consistent across the population, we would expect the same common peaks to be present within %NR landscapes of different individuals.

First, we perform visual inspection of a short %NR landscape segment across a small number of select individuals. A 500bp long mtDNA section between positions m.1000 and m.1500 within the RNR1 gene was chosen arbitrarily, and we use the same 6 individuals as used in the previous section. The %NR landscape fragment of each of these 6 individuals is shown in **Figure 3.2.2 F1 panel A** below. Each person's MT genome %NR level 95th percentile is marked by a dotted line, and any positions where that sample %NR landscape extends above this line are considered %NR peak positions (i.e. within the top 5% of highest %NR level sites in that person). **Figure 3.2.2 F1 panel B** provides a population-level look at the same MT genome fragment, in this case on the Y axis we show the fraction of all analysed samples in which %NR peaks are observed at any given position.

Supplementary figure 3.2.2 S1 listing the highest %NR peak incidence positions and observed %NR peak heights within the visualised 500bp region is provided in the **appendix 1.1**.

Figure 3.2.2 F1: Visualisation of the mitochondrial genome %NR landscape fragment



%NR – Non reference read frequency; %NR peaks defined as %NR values above sample %NR 95th percentile; Sample %NR 95th percentile marked with a dotted line; '▲' – Positions specifically addressed in the main text: m.1082, m.1222, m.1237, m.1420; '+' – Positions where population level %NR peak count is significantly higher than expected if %NR peak patterns were random.

While not by any means an exhaustive analysis, **figure 3.2.2 F1** provides several important insights into the behaviour and variability of %NR level landscapes among individual samples. In general, the visualisations of individual sample %NR landscapes in **panel A** indicate that %NR levels at adjacent positions can be very different and the majority of %NR peaks are single base wide. The distribution of peaks is not visually uniform across the region with several peak clusters visible, with **Panel B** indicating that many of the %NR peaks within these clusters are indeed present in a large proportion of the analysed samples. Specific positions discussed in the following paragraph correspond to visible peaks.

Not all %NR peaks are present in every sample however, and some high %NR peaks in some individuals can be completely absent from others (e.g. m.1082, marked in the figure). In contrast, some other peaks that only slightly exceed the surrounding %NR landscape are present in all 6 samples (e.g. m.1420, marked in the figure), showing that peak height differences among samples are not consistent between positions. We see that within an individual %NR peak heights can be highly variable and high %NR value at one position does not necessarily imply a comparably high %NR at another different peak location elsewhere and the samples vary considerably in terms of which peaks are higher and which are lower (e.g. m.1222 compared to m.1237, both marked in the figure).

The data visualised in **figure 3.2.2 F1** suggests that the landscape of %NR peak locations is non-random. To test this statistically, we performed a permutation test with the null hypothesis of sample MT genome %NR peak pattern being random and independent of MT genome locations. The dataset of the observed %NR locations in each of the 2533 analysed samples was randomised and the probability of observing as extreme %NR peak count at each position as seen in our dataset was calculated across 500,000 dataset permutations.

Positions within the MT genome region m.1000-m.1500 visualised in **Figure 3.2.2 F1 panel B** where our observed %NR peak count is found to be significant higher than expected under the null hypothesis ($p<0.05$ after correcting for multiple comparisons) are marked with '+'. We found total of 62 such positions within the visualised 500bp MT genome segment. See **supplementary figures 3.2.2 S2** and **3.2.2 S3** for visualisation of the population level %NR peak incidence landscape across the entire mitochondrial genome, and for examples of the entire MT genome %NR landscapes in two specific individuals.

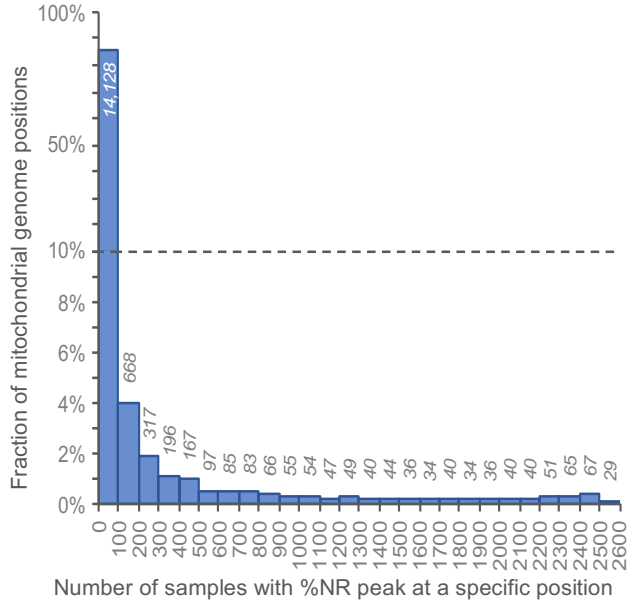
While **figure 3.2.2 F1** only focuses on a short MT genome section, the whole mitochondrial genome wide population-level %NR peak landscape behaviour is addressed in **figure 3.2.2 F2**.

Figure 3.2.2 F2 panel A shows that across most of the MT genome %NR peaks are observed in a relatively small number of the analysed samples (median 11, IQR 5 - 34) and only 15% of MT genome positions are found to be %NR peaks in more than 100 samples. There is however a subset of MT genome positions, summarised in **panel B**, where %NR peak prevalence in the population is very high exceeding 50% of the analysed samples. The reference sequence base composition of these highest %NR peak incidence position subsets (**Panel C**) is strongly skewed towards A and to a lesser extent T, suggesting a link between position base type and %NR peak behaviour at that site. Lastly, in **panel D** we address the statistical significance of the %NR peak pattern observations. In total we found that 11.2% of MT genome positions were associated with significantly higher %NR peak incidence than expected under the null hypothesis assuming independence between %NR level and MT genome location, while 83.4% of MT genome show significantly lower %NR peak incidence than expected under the null hypothesis (Permutation test, n=500,000, α = 0.05 after correction for multiple comparisons). This data supports our observations that the %NR landscapes are unlikely to be a result entirely of random chance and that there is a link between MT genome location and the %NR level behaviour. The links between %NR behaviour and MT genome position base types will be addressed in later sections.

Figure 3.2.2 F2: Mitochondrial genome %NR landscape peak distribution on the population scale

MT genome %NR landscape peak distribution on the population scale

(a) Distribution of %NR peaks* per MT genome position

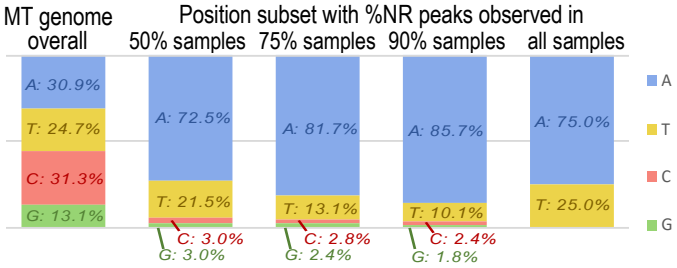


***%NR peaks observed per position**: mean 126.4, SD 390.1*
median 11, IQR 5 to 34

(b) MT genome fraction with high incidence of %NR peaks*

%NR peak present in	Number of MT positions	MT genome fraction
50% samples	573	3.46%
75% samples	327	1.97%
90% samples	168	1.01%
100% samples	4	0.02%

(c) Base composition of high %NR peak incidence sites



(d) %NR peak prevalence deviation from expected assuming no link between %NR behaviour and MT genome position**

Observed %NR peak count	Positions	% of MT genome
Significantly above expected	1,852 positions	11.2%
Significantly below expected	13,811 positions	83.4%
No significant difference	905 positions	5.5%

** %NR peak – mtDNA position where the observed %NR value falls within the 95th percentile for that sample.*
*** Null hypothesis assumes independence between mtDNA positions and the presence or absence of %NR peaks in each sample. Under the null hypothesis the distribution of %NR peak locations within each sample are expected to be random throughout the MT genome. Statistical significance estimated via permutation test (500,000 permutations, α = 0.05 after correction for multiple comparisons).*

3.2.3 Site-specific %NR behaviour across the entire analysed population

In this section we shall address the population level %NR behaviour at specific mtDNA sites, and address the level of site-specific population %NR distribution variability on the whole mitochondrial genome scale.

First, **Figure 3.2.3 F1** presents 20 position-specific %NR level histograms representing the distributions of the %NR values observed at each of these positions across all our analysed samples. For this we selected 20 consecutive positions from m.1101 to m.1120 that fall within the 500bp MT genome fragment investigated in the earlier **figure 3.2.2 F1.** This region contains a mixture of both high and low %NR peak incidence sites, allowing us to compare the population-wide %NR distribution behaviours between sites with variable %NR landscape features. For position specific population %NR distributions at a different set of 20 randomly drawn mtDNA positions see **Supplementary figure 3.2.3 S1.**

Figure 3.2.3 F1 illustrates considerable variability in the population %NR distribution shapes among the subset of 20 visualised positions, with %NR distributions ranging from very tight peaks near zero (m.1116A, m.1117A) to much more spread-out distributions over a wider range of %NR values (m.1104A). As expected, the three positions with highest population %NR peak incidence within this 20bp region (m.1104A, m.1115T and m.1119T with %NR peaks above sample P_{95} level in 84.8%, 69.5% and 61.0% of samples respectively) also exhibit the most extreme population %NR level distribution shift away from 0, as well as the highest %NR distribution median and 95th percentile levels of all 20 sites. There are, however noticeable visual differences in the %NR distribution shapes between these three peak positions. The population %NR values at position m.1104A appear more uniformly spread-out with a higher population median %NR level compared to positions m.1115T and m.1119T which are associated with lower population median but higher 95th percentile %NR levels.

Figure 3.2.3 F1: Position-specific population %NR distributions at 20 consecutive mtDNA positions

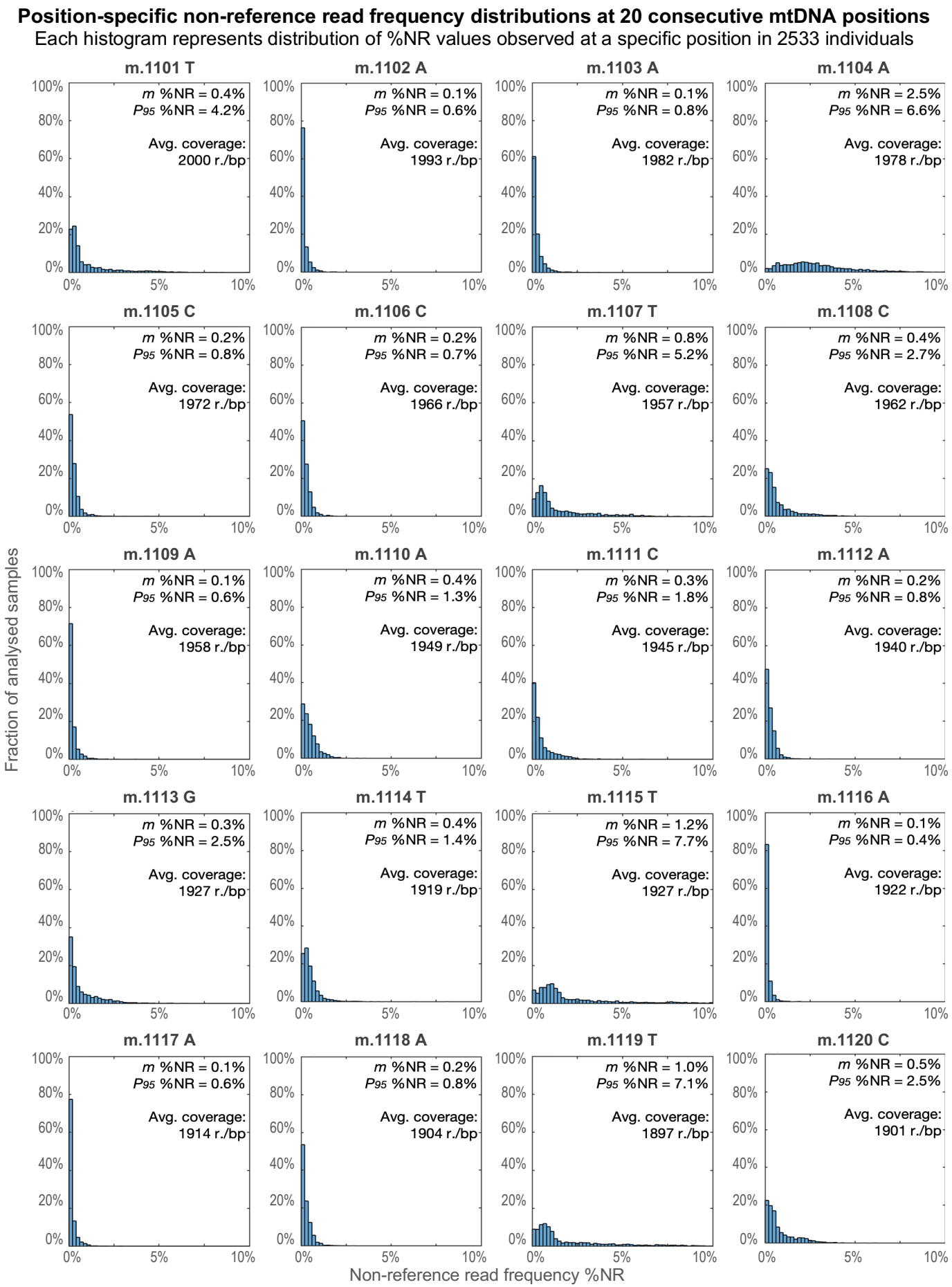


Figure 3.2.3 F1 is illustrative of substantial variation between MT genome position in terms of the %NR behaviour on the population scale, where some sites may be characterised by commonly observed %NR peaks of a more intermediate magnitude while others appear as less common but more prominent %NR peaks. This figure however

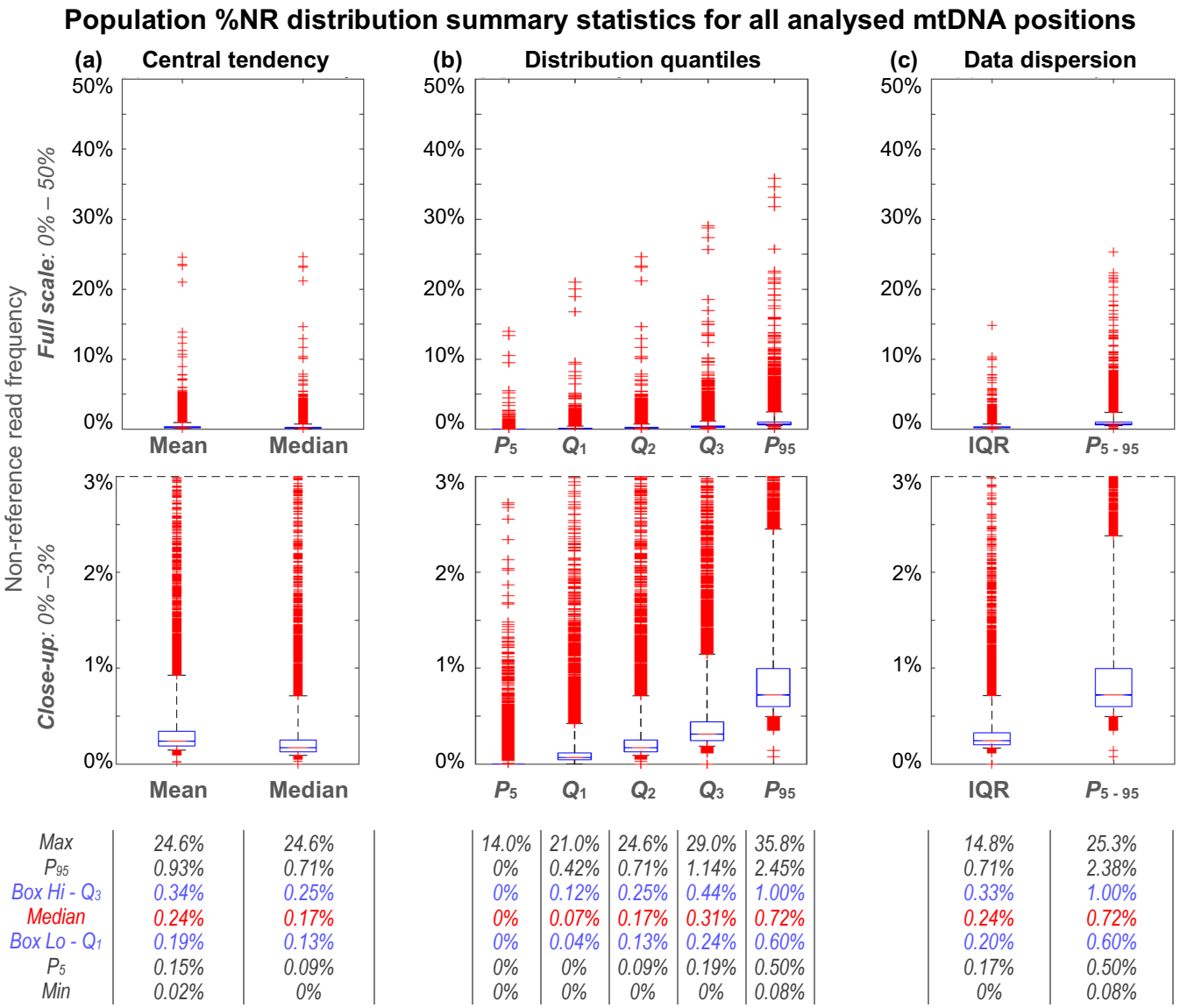
only focuses on a small fragment of the MT genome which cannot be generalised to the rest of the MT genome. The position-specific %NR level population distribution behaviours and variability on the whole MT genome scale is addressed in the next **figure 3.2.3 F2**.

Figure 3.2.3 F2 uses a similar approach to that employed earlier in **Figure 3.2.1 F2** where we described the MT genome %NR distribution variability between samples, except in this case we are looking to describe the %NR distribution variability between positions rather than between samples. For this, we determined a range of descriptive statistics describing our observed population %NR distribution shapes at each MT genome position, and then we visualise each descriptive statistic value distribution across all analysed MT genome sites as a separate box and whiskers plot.

Figure 3.2.3 F2 shows that on the MT genome scale site-specific population %NR levels tend to follow skewed distributions, and that more variability is found between positions in terms of %NR distribution spread and upper quantiles than between distribution central tendencies. While 95% of MT genome is associated with population median %NR ≤0.71% and the population 95th percentile %NR level ≤2.45%, we also see outlier positions associated with much higher %NR levels. Such outliers also include the three peak %NR positions m.1104A, m.1115T and m.1119T we focused on in the previous **figure 3.2.3 F1**, which based on data shown in **Figure 3.2.3 F2** fall within the top 5% of all analysed MT genome sites both in terms of population %NR median and 95th percentile levels.

The data shown in **figure 3.2.3 F2** is intended to illustrate the overall %NR distribution shapes and behaviours, and as such it confirms that the variation of %NR distribution shapes observed among the 20 position subset looked at in **figure 3.2.3 F1** is reasonably illustrative of the %NR distribution shape variability across the wider MT genome.

Figure 3.2.3 F2: Population %NR distribution summary statistics across all analysed mtDNA positions



	Mean	Median	P5	Q1	Q2	Q3	P95	IQR	P5 - 95
Max	24.6%	24.6%	14.0%	21.0%	24.6%	29.0%	35.8%	14.8%	25.3%
P95	0.93%	0.71%	0%	0.42%	0.71%	1.14%	2.45%	0.71%	2.38%
Box Hi - Q3	0.34%	0.25%	0%	0.12%	0.25%	0.44%	1.00%	0.33%	1.00%
Median	0.24%	0.17%	0%	0.07%	0.17%	0.31%	0.72%	0.24%	0.72%
Box Lo - Q1	0.19%	0.13%	0%	0.04%	0.13%	0.24%	0.60%	0.20%	0.60%
P5	0.15%	0.09%	0%	0%	0.09%	0.19%	0.50%	0.17%	0.50%
Min	0.02%	0%	0%	0%	0%	0%	0.08%	0%	0.08%

3.2.4 Section conclusions

Overall, data presented through this section demonstrates the existence of a distinct landscape of variable non-reference read frequency levels across the human mitochondrial genome on a population scale. Despite considerable variation between individuals, common clearly non-random %NR overall landscape patterns are evident throughout the population. On the level of specific mitochondrial genome positions, variable non-reference frequency population distribution shapes emerge. Genome sites with consistently low %NR levels are interspersed with some positions where population as the whole exhibits elevated non-reference frequencies as well as sites where very high %NR levels are present in only some individuals but not others.

The variability between MT genome positions in terms of the population %NR distribution shapes means that simple population averages may not be sufficient for quantifying and studying the exact %NR landscape behaviour on a genome scale. Instead, population-wide position-specific %NR distribution shape quantification and comparison are likely to be more informative.

At the start of our work, we hypothesised that some low-frequency sequence variation (i.e. heteroplasmy) might exist throughout the human mitochondrial genome. We speculated that such a heteroplasmy landscape, if present, should be detectable upon sufficiently deep sequencing and should manifest as a pattern of low frequency non-reference base read presence throughout the sequencing output.

The results presented in **section 3.2** demonstrate that the sequence data from 2533 individuals contains a complex low-level non-reference base frequency landscape throughout the mitochondrial genome, which could originate in various ways, including genuine heteroplasmy, sequencing error and background noise. We show that the overall incidence of non-reference reads is low: from the population of 2533, in 95% of individuals the mitochondrial genome median %NR levels are between 0 and 0.93%. If we consider genome non-reference read frequency 95th percentile as the upper boundary of the %NR 'typical range' in an individual, on the population scale in only 5% of individuals this genome 'typical range' upper boundary exceeds 2.14% non-reference read frequency. We also show that positions where non-reference read frequency levels exceed an individual's 'typical range' %NR levels, can have non-reference frequencies several times above the genome baseline level for that particular individual. We demonstrate that such outlier non-reference read frequency positions are not distributed randomly throughout individuals' genomes. Instead, sets of such high non-reference read frequency positions in particular individuals overlap considerably between individuals, leading to presence of %NR peak positions where much higher proportion of samples show %NR peaks than should be expected under a null hypothesis of no association between the MT genome location and the %NR level. On a population scale, certain genome positions have demonstrably elevated non-reference frequency levels throughout the population compared to other non-outlier mitochondrial genome sites. Yet we also show considerable non-reference frequency landscape variation between individuals, and not all individuals present with the same %NR landscape peak patterns over any given genome section.

Together, these findings paint a picture of a persistent non-random non-uniform low-level mitochondrial genome non-reference frequency landscape that likely is produced through a combination of several underlying mechanisms including genuine mitochondrial heteroplasmy, systematic sequencing errors and background noise.

In the next section we investigate whether any discernible links can be identified between the observed non-reference base frequency landscape patterns and mitochondrial genome sequence.

3.3 Link between position-specific %NR behaviour and the mtDNA sequence

In the previous section we demonstrated a high level of variability between positions in terms of non-reference read frequency population distribution behaviour leading to what appears to be a non-random population level %NR landscape across the mitochondrial genome.

Our starting hypothesis is that low level heteroplasmy exists, and accounts for some part of the observed non-reference frequency landscape. Our first aim in attempting to disentangle a true heteroplasmy landscape from other causes of non-reference frequency is to investigate the differences between position subsets associated with different population %NR distribution behaviours.

The most basic difference between positions is the identity of the base found at each position, and our data presented in the previous section has already hinted at a link between %NR landscape peak incidence and the MT genome sequence base identity at the position(Figure 3.2.2 F2 panel C). Therefore this is our starting point, and in this section we aim to quantify the link between the physical makeup of the MT genome sequence and the differences in the observable population %NR landscape behaviours.

So far we have mainly focused on two measures for describing %NR distribution behaviour at a specific MT genome position – the population %NR distribution median and 95th percentile levels. However as demonstrated these two measures may not always be congruent. Therefore in this section we will shall also include a third measure of %NR behaviour, namely the %NR distribution shape overall atypicality at the site as quantified by Kolmogorov-Smirnov two sample (KS2) test statistic comparing site-specific %NR distribution shape to the %NR distribution averaged across all sites in the MT genome. KS2 test statistic has been chosen for this due to the test being designed to be sensitive to a range of differences between distributions being compared. Thus %NR behaviour 'atypicality' expressed by KS2 test statistic could potentially be more sensitive to a wider range of %NR behaviour differences than measures such as %NR median or 95th percentile levels.

In **section 3.3.1** we shall focus on the link between the various measures of %NR behaviour and the base type at the site of interest while **section 3.3.2** will address the connection between %NR behaviour and the wider sequence patterns around the site of interest.

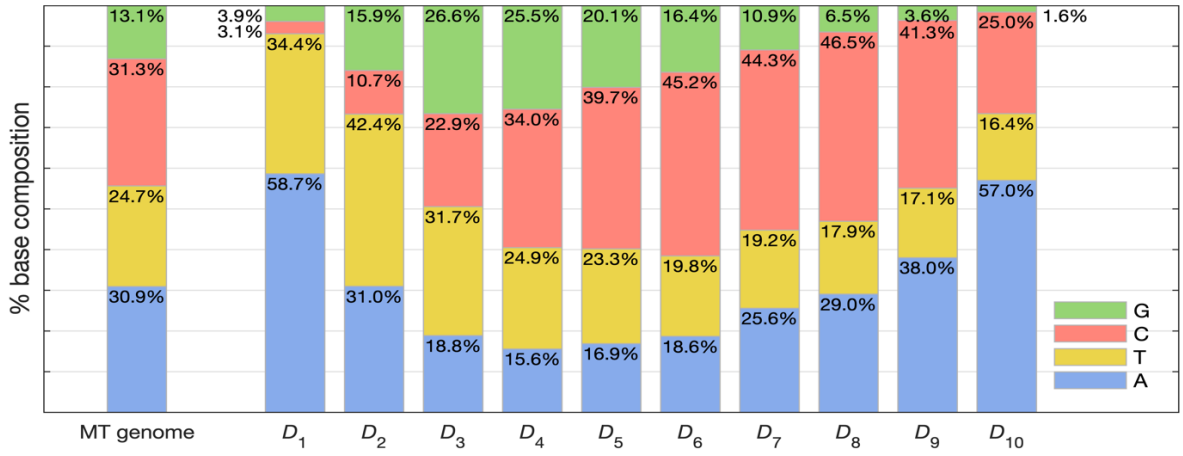
3.3.1 Base composition of mtDNA position subsets with different %NR behaviours

The data presented earlier in **figure 3.2.2 F2 panel C** has already hinted at a link between certain %NR behaviours and DNA base type found at the site. **Figure 3.3.1 F1** below explores this further in terms of three different site-specific %NR distribution behaviour measures: %NR 'average' level as expressed by the median %NR at that site(**panel A**), %NR upper 'typical range' as expressed by %NR 95th percentile at that site (**panel B**) and %NR distribution shape overall 'atypicality' as expressed by KS2 test statistic (**panel C**). In each panel all 16568 MT genome positions have been ranked by one of these %NR distribution shape measures (median, 95th percentile or shape 'atypicality') from highest to lowest. The resulting ranked position list has then been split into deciles (9 sets of 1657 positions + 1 set of 1656 positions), and the base composition of each set (i.e. decile) visualised in the figure below.

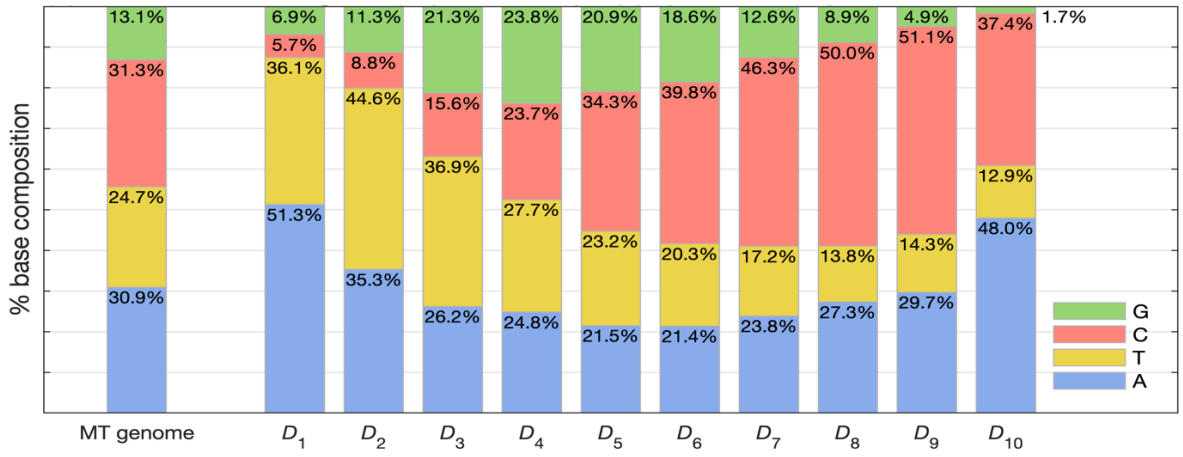
Figure 3.3.1 F1: Base composition of positions displaying different %NR population distribution behaviours

Base composition variation across deciles of mitochondrial genome positions ranked by different measures of position-specific %NR behaviour across 2533 samples

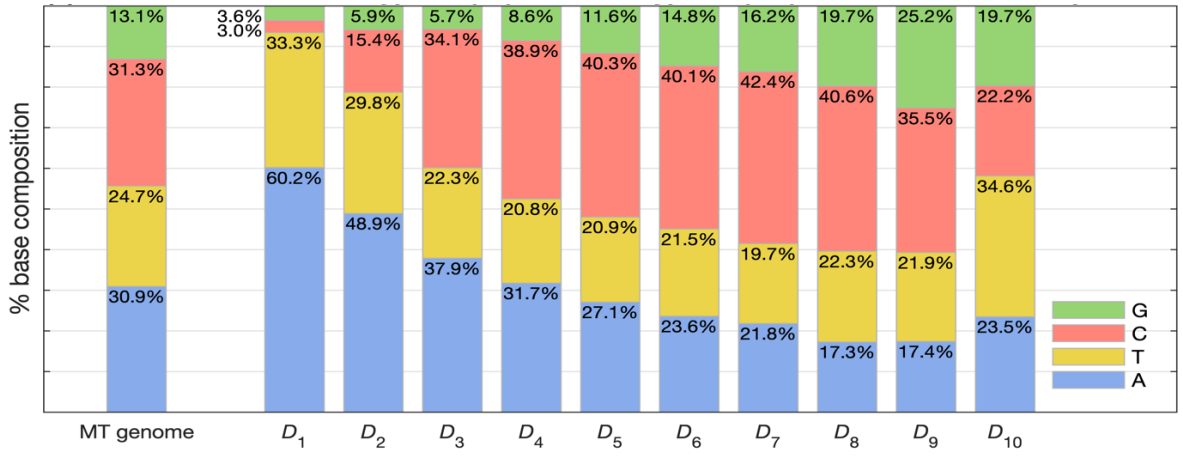
(a) Rank order: Highest (D_1) to lowest (D_{10}) position-specific %NR distribution median



(b) Rank order: Highest (D_1) to lowest (D_{10}) position-specific %NR distribution 95th percentile



(c) Rank order: Highest (D_1) to lowest (D_{10}) position-specific %NR distribution atypicality by K-S2 t.st.



D_1, $D_{2,...}D_{10}$ – deciles of ranked mtDNA positions; Sample size 2533 individuals; %NR atypicality measured via KS2 test statistic

Figure 3.3.1 F1 confirms that when mtDNA positions are ranked by any of the three proposed %NR behaviour measures, there is a clear base composition variation across the ranked deciles, supporting a link between mtDNA position reference base type and the %NR behaviours at that position. Regardless of the position ranking method used, the top decile of ranked positions shows increased frequency of As and Ts compared to Cs and Gs similarly to what we saw when we looked at positions with highest %NR peak incidence in the earlier sections.

When we compare the outcomes between the different ranking methods, the resulting ranked mtDNA position decile base composition when positions are ranked by the %NR distribution median versus by %NR 95th percentile is generally quite similar. In both cases we see a very high proportion of As and Ts compared to Cs and Gs in decile 1, which then slowly decreases over the next few deciles. Interestingly, while the frequency change of Cs and Gs is monotonic as the %NR median / 95th percentile decreases, the same is not true for Gs and As, with A being highly overrepresented and G distinctly underrepresented both within the highest and the lowest ranked deciles.

The position ranking by %NR distribution shape atypicality results in a slightly different ranked decile base composition compared to the other two position ranking methods. However this is expected due to how we defined the %NR distribution shape 'atypicality'. The KS2 test statistic that we use as the atypicality measure tells us how dissimilar the %NR distribution is at a certain position from an 'average' %NR distribution. Therefore sites with both an unusually high and an unusually low %NR levels will be classed as highly atypical and ranked high by our %NR shape atypicality measure. In contrast, when ranked directly by %NR median or 95th percentile level, high %NR sites and low %NR sites will be found at opposite ends of the ranked position list, similar to 'one-sided' versus 'two-sided' tests.

The differences in MT genome position ranking by each of the three %NR behaviour measures is addressed in more depth in **appendix 1.2** (see **supplementary figures 3.3.2 S1** - **S2**). It is worth noting that any divergence between position ranking by %NR quantiles versus distribution shape atypicality mostly appears to affect deciles 2 to 10. The top decile of the highest ranked positions appears to have roughly the same base composition irrespective of the ranking method used. This is confirmed in the supplementary data in the **appendix 1.2** showing that the top 10% of the positions with the most atypical %NR distribution shapes consist almost entirely of sites with higher than average %NR median and 95th percentile levels.

Overall the data presented in **Figure 3.3.1 F1** points to a link between DNA base type and the %NR distribution behaviour, suggesting that high %NR levels may be more strongly linked to positions where the reference sequence base is an A or a T rather than a C or a G. The results are not clear cut however, and bases A and C also appear to be associated with positions where %NR levels are especially low, suggesting that the base identity at the site of interest is unlikely to be the only factor affecting the %NR behaviour, and there are likely other mechanisms modifying the %NR behaviour at any given site. An obvious candidate for one such %NR behaviour modifier is the adjacent DNA sequence context around the site of interest, which is explored in the next section.

3.3.2 Quantification of the association between %NR behaviour and the DNA sequence features

An approach for quantifying the link between DNA sequence features and the %NR behaviour can be devised based on principles similar to those used by Gene Set Enrichment Analysis (Subramanian *et al.*, 2005), whereby association is measured by quantifying the enrichment of data points matching a particular feature of interest within a ranked list of all data points. A similar principle can be applied to our dataset to measure the enrichment of specific DNA sequence characteristics within a list of MT genome positions ranked by various %NR distribution behaviour measures. The strength of association can be estimated via an enrichment score E with theoretical range from -0.5 to +0.5, with E>0 indicating positive association and E<0 indicating negative association. This enables identification of sequence features most strongly linked with specific %NR behaviours of interest. For analyses details of the enrichment score calculation and statistical significance assessment see **Materials and Methods 2.3.**

In this section we present the results of the enrichment of simple adjacent sequence patterns within mitochondrial genome positions ranked by their %NR behaviours. The %NR behaviour measured used for position ranking were the same as discussed in the previous section: site-specific %NR population median, 95th percentile and the overall distribution shape atypicality as measured by KS2 test statistic. The sequence motifs investigated were up to 4bp long simple sequences consisting of a single base type with a variable number of wildcards upstream or downstream of the position of interest. For detailed description of the set of 416 sequence motifs investigated see **Materials and Methods figure 2.3 F1**.

The following figures list motifs found to show statistically significant positive or negative enrichment among positions ranked by different %NR behaviour measures. Motifs associated with different %NR median level behaviours are listed in **figure 3.3.2 F1** while motifs linked to different %NR distribution 95th percentile level behaviours are listed in **figure 3.3.2 F2**. Motif sets found to be most and least associated with %NR distribution shape atypicality as expressed by KS2 test statistic are provided in **figure 3.3.2 F3**. Only motifs with statistically significant positive or negative enrichment with p<0.05 after correction for multiple comparisons are included.

Figure 3.3.2 F1: Sequence motifs linked to different site-specific %NR population distribution median levels

Sequence motifs associated with different site-specific %NR distribution median levels

Motifs linked to increased %NR median

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	E(P50)	Sequence -4 bp to +4 bp	*n*	E(P50)
G.GG**T**....	*15*	0.475	....**A**CCCC	*79*	0.491
.GGG**T**....	*18*	0.470	....**A**CCC.	*199*	0.478
GG.G**T**....	*17*	0.454	....**A**CC.C	*190*	0.469
..GG**T**....	*80*	0.444	....**A**CC..	*515*	0.457
...G**T**....	*419*	0.383	....**A**C.CC	*184*	0.454
CCCC**T**....	*63*	0.326	....**T**CCCC	*50*	0.436
AAAA**T**....	*41*	0.323	....**T**CCC.	*108*	0.418
G.GG**A**....	*15*	0.321	....**A**C...	*1495*	0.398
GG.G**A**....	*20*	0.298	....**T**CC.C	*125*	0.391
.GGG**A**....	*17*	0.296	....**T**C.CC	*158*	0.368
.CCC**T**....	*190*	0.279	....**G**CCCC	*26*	0.367
C.CC**T**....	*176*	0.256	....**T**CC..	*361*	0.363
CC.C**T**....	*154*	0.245	....**T**GGG.	*18*	0.340
CCCC**G**....	*24*	0.244	....**G**CCC.	*93*	0.304
AAA.**T**....	*125*	0.243	....**G**CC.C	*79*	0.276
.AAA**T**....	*110*	0.243	....**G**C.CC	*53*	0.268
CCC.**T**....	*160*	0.239	....**G**CC..	*271*	0.261
T.TT**G**....	*29*	0.238	....**T**C...	*1204*	0.261
..GG**G**....	*72*	0.234	....**G**.CCC	*68*	0.256
..CC**T**....	*542*	0.220	....**T**AAAA	*46*	0.253
.CCC**G**....	*47*	0.210	....**T**GG.G	*22*	0.248
..GG**A**....	*123*	0.203	....**G**AAAA	*27*	0.234
C.CC**G**....	*57*	0.202	....**T**.CCC	*135*	0.230
AAA.**G**....	*59*	0.199	....**G**AA.A	*66*	0.221
CCCC**A**....	*68*	0.198	....**T**GG..	*100*	0.200
A.AA**T**....	*125*	0.186	....**G**C...	*711*	0.19
..CC**G**....	*141*	0.182	....**T**A.AA	*110*	0.182
CC.C**G**....	*53*	0.177	....**G**A.AA	*69*	0.178
..TT**G**....	*116*	0.174	....**G**AA..	*201*	0.174
...C**G**....	*435*	0.162	....**G**AAA.	*68*	0.161
..AA**T**....	*376*	0.154	....**T**.AAA	*133*	0.158
...C**T**....	*1440*	0.146	....**T**AAA.	*129*	0.155
...G**G**....	*426*	0.141	....**A**.CCC	*220*	0.149
AA.A**T**....	*116*	0.139	....**T**AA.A	*149*	0.146
C.CC**A**....	*162*	0.138	....**G**A...	*613*	0.120
.CCC**A**....	*163*	0.131	....**T**G...	*513*	0.115
...T**G**....	*513*	0.128	....**G**....	*2169*	0.098
CC.C**A**....	*191*	0.103	....**T**....	*4094*	0.087
....**G**....	*2169*	0.098	....**T**AA..	*414*	0.084
CCC.**A**....	*198*	0.094	....**C**AA..	*465*	0.068
...G**A**....	*613*	0.094	....**T**A...	*1373*	0.048
....**T**....	*4094*	0.087			
..CC**A**....	*464*	0.081			
...A**T**....	*1230*	0.071			

Motifs linked to decreased %NR median

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	E(P50)	Sequence -4 bp to +4 bp	*n*	E(P50)
CCC.**C**....	*205*	-0.254	....**A**AAA.	*182*	-0.389
CCCC**C**....	*69*	-0.216	....**A**AAAA	*70*	-0.377
.TTT**T**....	*63*	-0.214	....**A**AA..	*524*	-0.356
C.CC**C**....	*201*	-0.197	....**A**AA.A	*172*	-0.347
.CCC**C**....	*224*	-0.193	....**A**A.AA	*187*	-0.343
..TT**T**....	*251*	-0.186	....**A**A...	*1604*	-0.297
CC.C**C**....	*201*	-0.172	....**A**.AAA	*175*	-0.181
...A**C**....	*1495*	-0.167	....**T**TTT.	*63*	-0.166
AAAA**A**....	*70*	-0.141	....**T**TT.T	*65*	-0.144
T.TT**T**....	*69*	-0.135	....**T**T.TT	*78*	-0.132
..CC**C**....	*624*	-0.134	....**T**TT..	*251*	-0.126
AAA.**A**....	*172*	-0.121	....**C**C...	*1771*	-0.109
A.AA**A**....	*175*	-0.116	....**A**T...	*1230*	-0.109
.AAA**A**....	*182*	-0.106	....**A**G...	*795*	-0.100
AA.A**C**....	*137*	-0.105	....**C**....	*5181*	-0.091
AA.A**A**....	*187*	-0.102	....**C**.AAA	*147*	-0.089
...T**T**....	*1004*	-0.102	....**C**TT..	*318*	-0.086
..AA**C**....	*495*	-0.093	....**T**T...	*1004*	-0.081
....**C**....	*5181*	-0.091	....**C**CC..	*624*	-0.064
..AA**A**....	*524*	-0.082	....**A**....	*5124*	-0.019
...C**C**....	*1771*	-0.075			
...A**A**....	*1604*	-0.067			
...T**C**....	*1204*	-0.057			
...T**A**....	*1373*	-0.056			
....**A**....	*5124*	-0.019			

E(P50) – Motif enrichment index among positions ranked by site-specific population %NR distribution median; n – instances of motif found in the mitochondrial genome; '•' – wildcard indicating motif positions where base type is not restricted; Only motifs with statistically significant positive or negative enrichment with $p < 0.05$ after correction for multiple comparisons are shown

Figure 3.3.2 F2: Sequence motifs linked to different site-specific %NR population distribution 95th percentile level

Sequence motifs associated with different site-specific %NR distribution 95th percentile levels

Motifs linked to increased %NR 95th percentile

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	$E_{(P50)}$	Sequence -4 bp to +4 bp	*n*	$E_{(P50)}$
G.GG**T**....	*15*	0.463	....**A**CCCC	*79*	0.481
.GGG**T**....	*18*	0.459	....**A**CCC.	*199*	0.469
..GG**T**....	*80*	0.430	....**A**CC.C	*190*	0.463
GG.G**T**....	*17*	0.425	....**A**CC..	*515*	0.446
CCCC**T**....	*63*	0.355	....**A**C.CC	*184*	0.434
...G**T**....	*419*	0.346	....**T**CCCC	*50*	0.428
AAAA**T**....	*41*	0.317	....**T**CCC.	*108*	0.407
.CCC**T**....	*190*	0.303	....**T**GGG.	*18*	0.387
CCC.**T**....	*160*	0.281	....**T**CC.C	*125*	0.383
C.CC**T**....	*176*	0.280	....**A**C...	*1495*	0.369
CC.C**T**....	*154*	0.269	....**T**C.CC	*158*	0.363
AAA.**T**....	*125*	0.268	....**T**CC..	*361*	0.355
.AAA**T**....	*110*	0.251	....**G**CCCC	*26*	0.344
CCCC**G**....	*24*	0.242	....**T**GG.G	*22*	0.313
..CC**T**....	*542*	0.239	....**T**.GGG	*18*	0.311
CCCC**A**....	*68*	0.236	....**T**AAAA	*46*	0.296
T.TT**G**....	*29*	0.228	....**G**CCC.	*93*	0.289
..GG**G**....	*72*	0.203	....**G**CC.C	*79*	0.265
.CCC**G**....	*47*	0.192	....**T**C...	*1204*	0.259
C.CC**G**....	*57*	0.188	....**G**C.CC	*53*	0.256
A.AA**T**....	*125*	0.183	....**T**.CCC	*135*	0.252
..GG**A**....	*123*	0.179	....**G**CC..	*271*	0.243
CC.C**G**....	*53*	0.171	....**T**GG..	*100*	0.233
C.CC**A**....	*162*	0.164	....**G**.CCC	*68*	0.227
.CCC**A**....	*163*	0.161	....**T**AAA.	*129*	0.218
...C**T**....	*1440*	0.161	....**T**A.AA	*110*	0.211
..AA**T**....	*376*	0.161	....**A**.CCC	*220*	0.194
AAA.**G**....	*59*	0.156	....**T**AA.A	*149*	0.193
AA.A**T**....	*116*	0.151	....**T**.AAA	*133*	0.189
CCC.**A**....	*198*	0.151	....**G**AA.A	*66*	0.169
..CC**G**....	*141*	0.145	....**G**C...	*711*	0.164
..TT**G**....	*116*	0.138	....**T**AA..	*414*	0.130
CC.C**A**....	*191*	0.126	....**T**G...	*513*	0.127
...G**G**....	*426*	0.121	....**G**AA..	*201*	0.125
....**T**....	*4094*	0.117	....**T**....	*4094*	0.117
...C**G**....	*435*	0.111	....**G**....	*2169*	0.074
..CC**A**....	*464*	0.093	....**T**A...	*1373*	0.072
...G**A**....	*613*	0.090	....**G**A...	*613*	0.057
...A**T**....	*1230*	0.086			
....**G**....	*2169*	0.074			
...T**G**....	*513*	0.069			

Motifs linked to decreased %NR 95th percentile

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	$E_{(P50)}$	Sequence -4 bp to +4 bp	*n*	$E_{(P50)}$
CCC.**C**....	*205*	-0.258	....**A**AAA.	*182*	-0.305
CCCC**C**....	*69*	-0.223	....**A**AAAA	*70*	-0.283
.CCC**C**....	*224*	-0.212	....**A**AA..	*524*	-0.268
C.CC**C**....	*201*	-0.208	....**A**AA.A	*172*	-0.255
...A**C**....	*1495*	-0.183	....**A**A.AA	*187*	-0.240
CC.C**C**....	*201*	-0.178	....**C**T...	*1440*	-0.211
..CC**C**....	*624*	-0.165	....**A**A...	*1604*	-0.198
...T**C**....	*1204*	-0.149	....**A**.AAA	*175*	-0.155
....**C**....	*5181*	-0.133	....**C**....	*5181*	-0.133
TT.T**C**....	*75*	-0.126	....**C**C...	*1771*	-0.123
..AA**C**....	*495*	-0.110	....**C**TT..	*318*	-0.122
AAA.**A**....	*172*	-0.105	....**C**.CCC	*201*	-0.118
...C**C**....	*1771*	-0.100	....**C**C.CC	*201*	-0.098
AA.A**C**....	*137*	-0.092	....**C**CC..	*624*	-0.094
..TT**T**....	*251*	-0.091	....**A**T...	*1230*	-0.091
...G**C**....	*711*	-0.080	....**C**A...	*1534*	-0.087
...T**A**....	*1373*	-0.036	....**A**G...	*795*	-0.084
			....**C**G...	*435*	-0.073

$E_{(P95)}$ – Motif enrichment index among positions ranked by site-specific population %NR distribution 95th percentile; n – instances of motif found in the mitochondrial genome; '•' – wildcard indicating motif positions where base type is not restricted; Only motifs with statistically significant positive or negative enrichment with $p < 0.05$ after correction for multiple comparisons are shown

Figure 3.3.2 F3: Sequence motifs linked to different site-specific %NR population distribution shape atypicality

Sequence motifs associated with different site-specific %NR distribution shape atypicality

Motifs linked to highly atypical %NR behaviour

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	$E_{(KS2)}$	Sequence -4 bp to +4 bp	*n*	$E_{(KS2)}$
..GG**T**....	*80*	0.434	....**A**CCCC	*79*	0.491
G.GG**T**....	*15*	0.433	....**A**CCC.	*199*	0.481
.GGG**T**....	*18*	0.422	....**A**CC.C	*190*	0.473
GG.G**T**....	*17*	0.407	....**A**CC..	*515*	0.452
...G**T**....	*419*	0.321	....**A**C.CC	*184*	0.437
.AAA**A**....	*182*	0.190	....**T**CCCC	*50*	0.399
AAAA**A**....	*70*	0.171	....**T**CCC.	*108*	0.372
A.AA**A**....	*175*	0.170	....**A**C...	*1495*	0.342
..AA**A**....	*524*	0.166	....**T**CC.C	*125*	0.323
CCCC**T**....	*63*	0.165	....**T**CC..	*361*	0.269
T.TT**A**....	*88*	0.161	....**T**C.CC	*158*	0.268
AAA.**A**....	*172*	0.155	....**G**CCCC	*26*	0.244
..TT**A**....	*329*	0.152	....**A**AAAA	*70*	0.238
...T**A**....	*1373*	0.152	....**A**AAA.	*182*	0.235
TT.T**A**....	*96*	0.144	....**A**AA.A	*172*	0.186
AA.A**A**....	*187*	0.134	....**A**A.AA	*187*	0.184
.CCC**A**....	*163*	0.133	....**A**AA..	*524*	0.180
...C**A**....	*1534*	0.126	....**A**.AAA	*175*	0.169
..CC**A**....	*464*	0.126	....**A**.CCC	*220*	0.133
CCC.**C**....	*205*	0.115	....**T**C...	*1204*	0.131
....**A**....	*5124*	0.114	....**T**.CCC	*135*	0.129
C.CC**A**....	*162*	0.105	....**A**....	*5124*	0.114
...A**A**....	*1604*	0.093	....**A**A...	*1604*	0.112
CC.C**A**....	*191*	0.081			
...G**A**....	*613*	0.054			

Motifs linked to most typical %NR behaviour

Upstream motifs			Downstream motifs		
Sequence -4 bp to +4 bp	*n*	$E_{(KS2)}$	Sequence -4 bp to +4 bp	*n*	$E_{(KS2)}$
...A**G**....	*795*	-0.172	....**G**TTT.	*29*	-0.222
.AAA**G**....	*75*	-0.169	....**G**TT..	*104*	-0.185
A.AA**G**....	*61*	-0.168	....**G**GG..	*72*	-0.183
.TTT**C**....	*70*	-0.164	....**G**G...	*426*	-0.180
..AA**G**....	*209*	-0.164	....**G**T...	*419*	-0.167
AA.A**G**....	*97*	-0.160	....**C**AAAA	*39*	-0.166
T.TT**C**....	*91*	-0.149	....**G**.TTT	*41*	-0.163
AAAA**C**....	*52*	-0.146	....**C**AAA.	*145*	-0.156
....**G**....	*2169*	-0.145	....**G**A...	*613*	-0.156
AAA.**C**....	*168*	-0.144	....**G**....	*2169*	-0.145
..GG**C**....	*151*	-0.144	....**C**AA.A	*150*	-0.140
...G**G**....	*426*	-0.138	....**C**AA..	*465*	-0.138
...T**G**....	*513*	-0.130	....**G**A.AA	*69*	-0.136
...G**C**....	*711*	-0.120	....**G**AA..	*201*	-0.131
...C**G**....	*435*	-0.113	....**C**.TTT	*67*	-0.125
..TT**C**....	*308*	-0.101	....**C**GG..	*80*	-0.119
..CC**G**....	*141*	-0.099	....**C**A.AA	*145*	-0.108
...C**C**....	*1771*	-0.089	....**C**G...	*435*	-0.108
.AAA**C**....	*157*	-0.087	....**C**A...	*1534*	-0.106
..AA**C**....	*495*	-0.066	....**G**C...	*711*	-0.098
....**C**....	*5181*	-0.057	....**A**TT..	*330*	-0.094
...T**C**....	*1204*	-0.054	....**A**GG..	*174*	-0.090
			....**C**CC..	*624*	-0.066
			....**C**C...	*1771*	-0.066
			....**C**....	*5181*	-0.057
			....**A**G...	*795*	-0.056
			....**A**T...	*1230*	-0.052

$E_{(KS2)}$ – Motif enrichment index among positions ranked by site-specific population %NR distribution shape atypicality; n – instances of motif found in the mitochondrial genome; '•' – wildcard indicating motif positions where base type is not restricted; Only motifs with statistically significant positive or negative enrichment with $p < 0.05$ after correction for multiple comparisons are shown

Figures 3.3.2 F1-F3 show that regardless of whether MT genome positions are ranked by %NR median, 95th percentile or overall %NR distribution shape atypicality, highest positive enrichment is shown by similar sets of motifs: Ts at position 0 preceded by runs of Gs, and As or Ts at position 0 followed by runs of Cs. Gs flanked by Cs on either side also appear among high enrichment motif sets.

A notable difference between MT genome position ranking methods is the motif pattern where A at position 0 is flanked by other As. This pattern is highly associated with %NR distribution shape atypicality while being also linked to low ow %NR median and 95th percentile levels. Generally, a common feature among motifs linked to low %NR median and 95th percentile levels is the base at position 0 being flanked that same base runs on one or the other side. However, unlike runs of As, continuous runs of the other three bases do not appear among motifs strongly linked to atypical %NR distribution shapes.

Overall, groups of related motifs tend to show similar association with %NR behaviours, although there are a few exceptions. For example, if any one of the C within the downstream motif TCCCC is replaced by a wildcard N which means it can be any base, changes to the position -1 have by far the strongest effect on the motif enrichment score (TCCCN E_{p50} = 0.41, TCCNC E_{p50} = 0.39, TCNCC E_{p50} = 0.37, TNCCC E_{p50}=2.3). The downstream motif ACCCC is an even more extreme example, with motif variant ACNCC being associated with E_{p50} = 0.45 while the E_{p50} for the motif variant ANCCC is just 0.15. These are not the only examples, and it appears that the positions -1 and +1 generally tend to have the biggest impact on how strongly the motif correlates with different %NR behaviours.

It should be noted that in the figures above we are looking at the overall enrichment of different motifs against a neutral baseline. However, in several instances even just a single base at position 0 without any consideration for the adjacent sequence context can show significantly enrichment across the ranked position sets. Therefore in order to separate the effect of the adjacent sequence context from the effect of the base at position 0, we can measure motif enrichment relative to the enrichment of the type found at position 0. **Supplementary figures 3.3.2 S3-S5** listing the most significant motifs identified based on relative enrichment accounting for the base type at position 0 are provided in **appendix 1.2** with the data generally following similar patterns to those described in figures above.

The data presented here highlights an important observation: the relationships between different motifs and their association to specific %NR behaviours demonstrates directionality where the same upstream versus downstream sequence patterns results in different motif enrichment whereas upstream and downstream patterns that are reverse complements of each other result in similar enrichment levels. The most obvious examples of this are the GGGGT/ TGGGG and CCCCA / ACCCC motif families.

The directionality we see suggests that whatever process is resulting in high or atypical %NR levels at some sites, its link to the adjacent DNA sequence must be asymmetric, with bases to one side from the position of interest clearly being more important than base on the other side of it. This in itself is not unexpected, as many processes *in vivo* interact with the DNA in a directional manner (transcription, replication). However we must also consider another potential modulator of %NR levels that is highly directional: the sequencing process itself, which can plausibly produce artificially inflated %NR levels following particularly hard-to-sequence motifs. This in combination with sequence contexts such as 'GGT' having been previously reported in literature as more prone to sequencing errors (Abnizova *et al*., 2012; Schirmer *et al*., 2016; Stoler *et al*., 2021), requires addressing to what extent the %NR behaviour patterns we have observed so far may be a result of artefacts introduced during sequencing process. This is the focus of the next section.

3.4 Sequence features associated with %NR behaviour differences between sequencing directions

In this section we address the differences in the observed %NR behaviour when comparing the %NR distributions associated with sequence reads arising from mtDNA H and L strands separately.

Presence of genuine non-reference bases in the sequenced samples should result in similar measured %NR regardless of the sequencing read direction, whereas different strand-specific %NR distribution behaviours at a particular site can be an indicator of one sequencing direction being more prone to sequencing artefacts. Finding certain sequence motifs enriched among MT genome positions with high levels of %NR strand bias would allow isolation of sequence contexts where high observed %NR is more likely resulting from sequencing artefacts rather than genuine heteroplasmy.

The overall aim of this section is to identify sequence features that are common to MT genome positions associated with highly different strand-specific %NR behaviours, as well as to compare this to sequence features found in the previous section to be linked to high or atypical %NR overall levels. Comparing such sequence patterns to those reported in the previous section as being associated with high or atypical overall %NR levels provides some indication of which groups of identified motifs discussed in the previous section may be more likely associated with sequencing errors versus genuine heteroplasmy

To achieve this, we use a similar approach to that employed in the previous section where the MT genome positions were ranked by some measure of %NR distribution behaviour and the ranked position list investigated for enrichment of specific sequence features among positions ranked highest by the %NR distribution behaviour of interest. Except unlike the previous section which addressed the behaviour of the overall %NR distribution at each position, here we are now separating the %NR data obtained from reads arising from the H and L stands, and looking at the differences between the strand specific %NR distributions behaviours at each site. Three different measures are used for ranking MT genome positions by their strand-specific %NR behaviour differences:

- The difference between H and L strand %NR distribution medians (ΔP_{50})
- The difference between H and L strand %NR distribution 95th percentiles (ΔP_{95})
- The differences between H and L strand %NR distribution overall shapes as expressed by Kolmogorov-Smirnov 2 sample test statistic (KS2).

The ranked position list is then analysed to identify common sequence features associated with positions being ranked particularly high or low based on the strand-specific %NR differences.

Section 3.4.1 specifically focuses on the link between strand-specific %NR differences and the base type found at the position of interest, while **section 3.4.2** presents simple adjacent sequence contexts found to be enriched among the positions with highest and lowest levels of %NR behaviour differences.

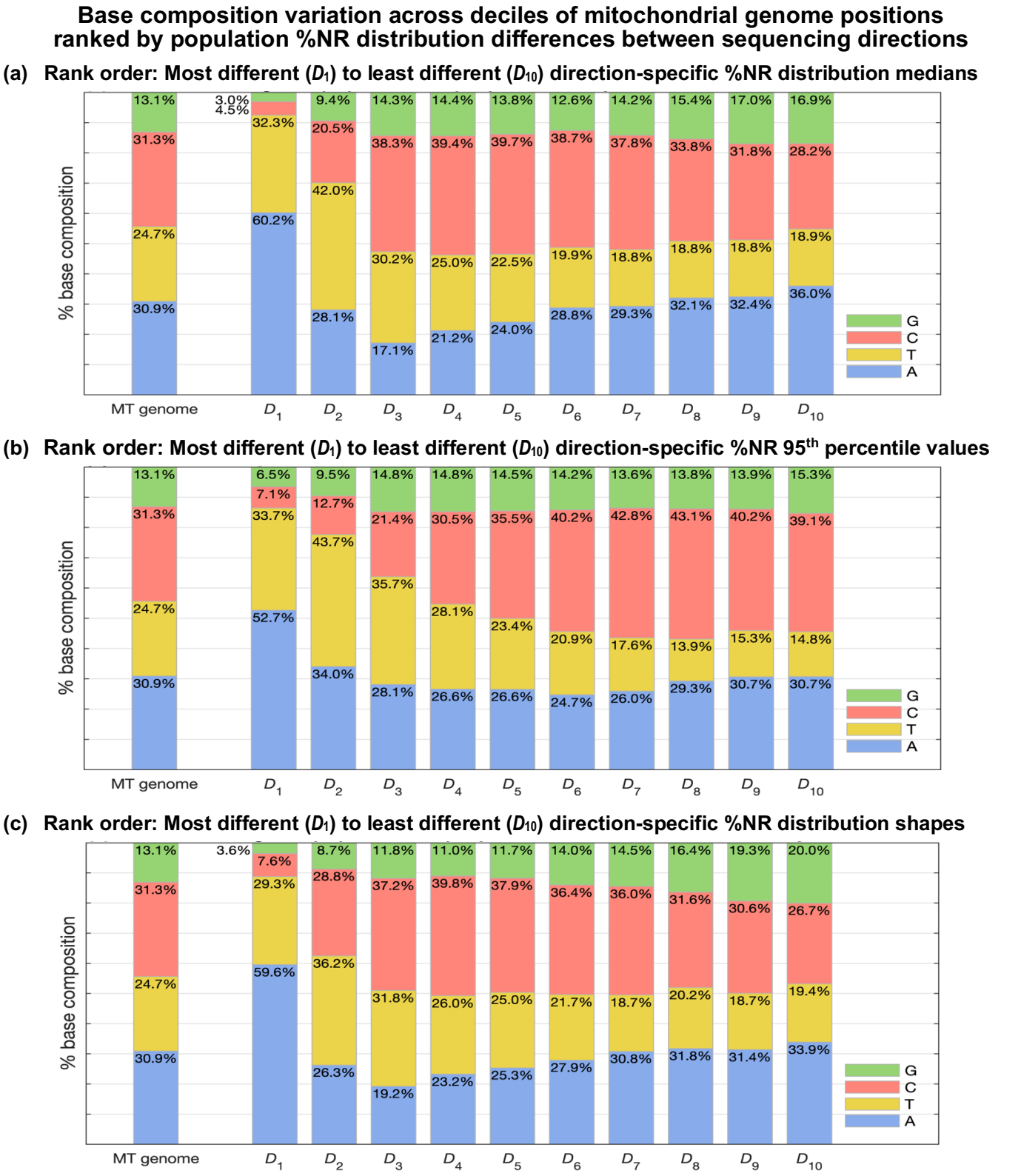
For a comparison between the resulting MT genome position rank order by all three measures see **Supplementary figure 3.4 S1.**

3.4.1 Link between sequencing direction %NR behaviour differences and the position base type

Figure 3.4.1 F1 below addresses how base type at a position relates to the directional %NR behaviour differences employing a similar data visualisation to the earlier **figure 3.3.1 F1.** Each **figure 3.4.1 F1** panel shows the base composition of ranked deciles of MT genome positions ranked by one of three measures for the strand-specific %NR distribution behaviour differences: the difference between H and L strand %NR distribution medians (**panel A**) , 95th percentiles (**panel B**) and the overall distribution shapes as expressed by KS2 test statistic (**panel C**).

Figure 3.4.1 F1 shows a clear link between the base type and %NR behaviour differences between sequencing directions regardless of whether positions are ranked by differences between strand-specific %NR distribution medians, 95th percentiles or the distribution overall shapes. In all cases overrepresentation of A in the top ranked decile and T across the deciles 1 to 4 indicates that bases A and T are linked to higher %NR behaviour strand bias than bases C and G. This is consistent with findings from previous literature. When quantified using enrichment score (see **Materials & Methods 2.3**), both A and T show significant positive enrichment among positions ranked highest by directional %NR differences. For supplementary figures addressing base type enrichment quantification as well as discussion of the enrichment patterns, see **Supplementary figures 3.4.1 S1-S2.**

Figure 3.4.1 F1: Base composition differences between positions showing different extent of deviation between strand-specific %NR distribution behaviours



D_1, $D_{2,..}D_{10}$ – deciles of ranked mtDNA positions; Position ranked based on behaviour of %NR distribution across all 2533 samples

3.4.2 Link between the discrepancy in sequencing direction specific %NR behaviour and the adjacent DNA sequence

To explore the link between strand-specific %NR behaviour differences and the mtDNA sequence features around the position of interest we studied the enrichment of the same set of 416 simple sequence motifs already discussed in **section 3.3.2** among positions ranked based on observed differences in sequencing direction specific %NR behaviours. For details of the motif set investigated see **Materials & Methods 2.3**. The three measures for directional %NR behaviour discrepancy were the difference between strand-specific %NR medians, the difference between strand-specific %NR 95th percentiles and the difference between the strand-specific %NR distribution overall shapes expressed via KS2 test statistic.

In total we identified 40 upstream sequences and 34 downstream sequences showing significant positive association to differences in strand-specific %NR behaviours by at least one of the three different measures used. These sets of positively enriched motifs are presented in **figure 3.4.2 F1**.

We also found 24 upstream and 33 downstream sequence motifs showing a significant negative relationship to strand-specific %NR behaviour differences. These negatively enriched motifs are listed in **figure 3.4.2 F2**.

Figure 3.4.2 F1: Sequence motifs linked to high strand-specific %NR distribution differences

Sequence motifs associated with large differences between %NR strand-specific distributions

Upstream motifs Sequence -4 bp to +4 bp	*n*	Enrichment score $E_{(\Delta P50)}$	$E_{(\Delta P95)}$	$E_{(KS2)}$	Also linked to high / atypical overall %NR
GGGG T	*4*		0.599		
G.GG T	*15*	0.472	0.483	0.465	x
.GGG T	*18*	0.463	0.465	0.457	x
..GG T	*80*	0.435	0.385	0.413	x
GG.G T	*17*	0.398	0.322	0.343	x
...G T	*419*	0.332	0.240	0.272	x
AAAA T	*41*	0.227	0.241		x
AAA. T	*125*	0.201	0.233	0.169	x
.AAA T	*110*	0.164	0.194	0.134	x
A.AA T	*125*	0.162	0.182	0.136	x
..AA T	*376*	0.125	0.147	0.105	x
AA.A T	*116*		0.104		x
...A T	*1230*	0.054	0.084	0.047	x
CCCC T	*63*	0.175	0.179		x
.CCC T	*190*	0.171	0.165	0.092	x
CC.C T	*154*	0.158	0.172	0.092	x
C.CC T	*176*	0.144	0.156		x
CCC. T	*160*	0.130	0.153		x
..CC T	*542*	0.125	0.143	0.074	x
...C T	*1440*	0.104	0.132	0.080	x
...T T	*1004*		0.035		
.... T	*4094*	0.073	0.104	0.059	x
GG.G A	*20*	0.286		0.266	x
..GG A	*123*	0.135	0.103		x
...G A	*613*	0.069	0.049	0.047	x
CCCC A	*68*	0.197	0.210	0.180	x
C.CC A	*162*	0.134	0.162	0.130	x
.CCC A	*163*	0.134	0.134	0.122	x
CCC. A	*198*	0.103	0.137	0.094	x
CC.C A	*191*	0.086	0.106		x
..CC A	*464*	0.071	0.097	0.071	x
...C A	*1534*		0.043		x*
...A A	*1604*		0.055	0.027	x*
.... A	*5124*	0.015	0.037	0.019	x*
AAAA C	*52*	0.192		0.217	
AAA. C	*168*	0.171	0.105	0.192	
.AAA C	*157*	0.106		0.130	
A.AA C	*150*	0.101		0.132	
..AA C	*495*			0.068	
..GG G	*72*	0.135			x

Downstream motifs Sequence -4 bp to +4 bp	*n*	Enrichment score $E_{(\Delta P50)}$	$E_{(\Delta P95)}$	$E_{(KS2)}$	Also linked to high / atypical overall %NR
.... A CCCC	*79*	0.490	0.478	0.478	x
.... A CCC.	*199*	0.479	0.468	0.475	x
.... A CC.C	*190*	0.472	0.462	0.467	x
.... A CC..	*515*	0.46	0.447	0.457	x
.... A C.CC	*184*	0.455	0.432	0.443	x
.... A C...	*1495*	0.390	0.357	0.373	x
.... A .CCC	*220*	0.108	0.145	0.091	x
.... A	*5124*	0.015	0.037	0.019	x*
.... T CCCC	*50*	0.430	0.413	0.400	x
.... T CCC.	*108*	0.402	0.39	0.367	x
.... T CC.C	*125*	0.377	0.362	0.348	x
.... T C.CC	*158*	0.352	0.339	0.315	x
.... T CC..	*361*	0.344	0.327	0.313	x
.... T C...	*1204*	0.249	0.244	0.221	x
.... T .CCC	*135*	0.167	0.203	0.136	x
.... T GGG.	*18*		0.317		x
.... T GG.G	*22*		0.255		x
.... T GG..	*100*		0.150		x
.... T G...	*513*	0.050	0.094		x
.... T AAAA	*46*	0.221	0.260	0.181	x
.... T AAA.	*129*	0.167	0.209	0.150	x
.... T .AAA	*133*	0.157	0.175	0.134	x
.... T A.AA	*110*	0.139	0.182	0.109	x
.... T AA.A	*149*	0.117	0.175		x
.... T AA..	*414*	0.075	0.137	0.062	x
.... T A...	*1373*		0.065		x
.... T	*4094*	0.073	0.104	0.059	x
.... C CCCC	*69*	0.195		0.21	
.... C CCC.	*224*	0.136		0.161	
.... C .CCC	*201*	0.123		0.151	
.... C CC.C	*205*	0.119		0.151	
.... C C.CC	*201*	0.101		0.124	
.... C CC..	*624*	0.068		0.093	
.... G TT..	*104*		0.117		

$E_{(\Delta P50)}$, $E_{(\Delta P95)}$ – Motif enrichment score among positions ranked by the difference between %NR strand-specific distribution medians or 95th percentiles; $E_{(KS2)}$ – Motif enrichment score among positions ranked by %NR strand-specific distribution shape overall differences as expressed by KS2 test statistic; n – number of instances motif is found in the mitochondrial genome; Wildcard '•' represents motif positions where base type is not restricted; Only motifs with statistically significant positive enrichment with p<0.05 after correction for multiple comparisons are listed, enrichment scores below significance threshold not shown; Motifs marked with 'x' are also significantly associated with elevated overall %NR levels of high %NR distribution shape atypicality, except where marked with asterisk, in which case motif is linked to high %NR atypicality only

Figure 3.4.2 F2: Sequence motifs linked to low strand-specific %NR distribution differences

Sequence motifs associated with high similarity between %NR strand-specific distributions

Upstream motifs -4 bp to +4 bp	*n*	E(ΔP50)	E(ΔP95)	E(KS2)
..TT **T**	*251*	-0.071		
...T **T**	*1004*	-0.055		-0.043
CCC. **C**	*205*	-0.198	-0.237	-0.172
CCCC **C**	*69*	-0.176	-0.235	-0.165
C.CC **C**	*201*	-0.167	-0.235	-0.157
.CCC **C**	*224*	-0.165	-0.206	-0.148
..CC **C**	*624*	-0.113	-0.175	-0.101
CC.C **C**	*201*		-0.120	
...C **C**	*1771*		-0.067	
...A **C**	*1495*	-0.072	-0.122	-0.044
...T **C**	*1204*	-0.044	-0.115	
...G **C**	*711*	-0.063	-0.126	-0.063
.... **C**	*5181*	-0.043	-0.102	-0.024
CCC. **G**	*61*			-0.172
..CC **G**	*141*			-0.107
...C **G**	*435*	-0.053		-0.088
...A **G**	*795*	-0.112	-0.059	-0.131
..AA **G**	*209*	-0.103		-0.127
.AAA **G**	*75*			-0.119
AA.A **G**	*97*			-0.115
TTT. **G**	*35*			-0.160
...T **G**	*513*	-0.053		-0.081
...G **G**	*426*			-0.065
.... **G**	*2169*	-0.070	-0.038	-0.098

Downstream motifs -4 bp to +4 bp	*n*	E(ΔP50)	E(ΔP95)	E(KS2)
.... **A** T.TT	*77*	-0.168		-0.175
.... **A** T...	*1230*	-0.132	-0.104	-0.128
.... **A** TT..	*330*	-0.089		-0.094
.... **A** A...	*1604*	-0.144	-0.082	-0.120
.... **A** AA..	*524*	-0.120	-0.078	-0.095
.... **A** A.AA	*187*	-0.110		-0.083
.... **A** AAA.	*182*	-0.103	-0.082	-0.089
.... **A** AA.A	*172*	-0.085		
.... **A** GG..	*174*	-0.124		-0.139
.... **A** G...	*795*	-0.140	-0.107	-0.135
.... **T** TTT.	*63*	-0.160		-0.148
.... **T** T.TT	*78*	-0.120		
.... **T** TT..	*251*	-0.087		
.... **T** T...	*1004*	-0.048		-0.041
.... **C** .AAA	*147*	-0.152	-0.127	-0.152
.... **C** AAA.	*145*	-0.121		-0.137
.... **C** AA.A	*150*			-0.108
.... **C** A.AA	*145*		-0.100	
.... **C** AA..	*465*	-0.083	-0.080	-0.106
.... **C** A...	*1534*	-0.029	-0.098	
.... **C** T...	*1440*	-0.126	-0.161	-0.099
.... **C** T.TT	*94*		-0.124	
.... **C** TT..	*318*	-0.079	-0.115	-0.063
.... **C** C...	*1771*		-0.068	
.... **C** G...	*435*		-0.062	
.... **C**	*5181*	-0.043	-0.102	-0.024
.... **G** G.GG	*26*		-0.193	
.... **G** GG..	*72*		-0.145	
.... **G** G...	*426*	-0.095	-0.122	-0.109
.... **G** A...	*613*	-0.075	-0.049	-0.104
.... **G** C...	*711*			-0.091
.... **G** T...	*419*	-0.094		-0.088
.... **G**	*2169*	-0.07	-0.038	-0.098

E(ΔP50), E(ΔP95) – Motif enrichment score among positions ranked by the difference between %NR strand-specific distribution medians or 95th percentiles; E(KS2) – Motif enrichment score among positions ranked by %NR strand-specific distribution shape overall differences as expressed by KS2 test statistic; n – number of instances motif is found in the mitochondrial genome; Wildcard '•' represents motif positions where base type is not restricted; Only motifs with statistically significant negative enrichment with p<0.05 after correction for multiple comparisons are listed; enrichment scores below significance not shown

Figures 3.4.2 F1 and **F2** point to a strong but complex relationship between the differences in the %NR behaviour between sequencing directions and the adjacent sequence context. Among the motifs associated with high %NR directional behaviour differences, bases A and T dominate at the position 0. The adjacent contexts tend to fit a handful of quite clear patterns, with variations of the upstream 'GGT'-type pattern and the downstream 'ACC' and 'TCC'-type patterns showing strongest links to %NR directional differences. While two of these motif families ('GGT' and 'ACC') are reverse complements of each other, not all complementary pattern pairs follow the same trends. For example, the 'TCC'-type motif family which itself is strongly linked to high %NR strand bias lacks its reverse complement among the high enrichment motif sets. And in some cases the motif families that are reverse complements of each other exhibit opposite association to %NR strand bias. An example of this is the upstream 'AAT'-type pattern which is linked to high %NR strand bias while its reverse complement downstream pattern 'ATT' displays significant negative association to directional %NR differences.

With large strand bias being generally considered a hallmark of sequencing errors (Nakamura *et al*., 2011; Allhoff *et al*., 2013), the patterns observed here strongly point to a presence of systematic sequencing errors linked to the specific adjacent sequence contexts within our datasets, with bases A or T within a C or G rich adjacent context identified as the main problematic contexts.

This raises a fundamental question of whether any of the %NR distribution behaviours we have observed represent real heteroplasmy at all, or whether the non-random patterns that characterise %NR landscape are simply a result of sequencing artefacts. To address this, we compared our identified set of sequence patterns significantly linked to %NR strand bias to the sets of motifs we found showing association to high %NR overall levels in the earlier **section 3.3.2**.

Our data shows a high degree of overlap between the motif set associated with direction-specific %NR behaviour differences and the motifs linked to high overall %NR levels or high %NR atypicality. In total, 33 of the upstream and 27 of the downstream patterns linked to %NR strand bias were also reported to correlate to high overall %NR levels or atypicality. This represents 66% of all motifs linked to high %NR median, 70% of all motifs linked to high %NR 95th percentile and 60% of the motifs linked to high %NR overall distribution atypicality, and includes variants of the upstream 'GGT', 'AAT', 'CCT', 'GGA', and 'CCA'-type patterns and the downstream 'ACC, 'TGG', 'TCC' and 'TAA'-type patterns. However, not all motif families linked to high %NR overall levels are also linked to strand bias. **Figure 3.4.2 F3** below highlights the subsets of motifs significantly linked to high overall %NR distribution median, 95th percentile or shape atypicality that did not exhibit significant association with %NR strand bias.

Figure 3.4.2 F3: Sequence patterns linked to high %NR level or atypicality but not to high %NR strand bias

Sequence patterns associated with high %NR level or atypicality but not linked to strand bias

Upstream motifs -4 bp to +4 bp	*n*	$E_{(P50)}$	$E_{(P95)}$	$E_{(KS2)}$	
CCCC **G**	*24*	0.244	0.242		
.CCC **G**	*47*	0.210	0.192		
C.CC **G**	*57*	0.202	0.188		
..CC **G**	*141*	0.182	0.145		
CC.C **G**	*53*	0.177	0.171		
...C **G**	*435*	0.162	0.111		
T.TT **G**	*29*	0.238	0.228		
..TT **G**	*116*	0.174	0.138		
...T **G**	*513*	0.128	0.069		
AAA. **G**	*59*	0.199	0.156		
...G **G**	*426*	0.141	0.121		
.... **G**	*2169*	0.098	0.074		
G.GG **A**	*15*	0.321			*
.GGG **A**	*17*	0.296			*
.AAA **A**	*182*			0.190	*
AAAA **A**	*70*			0.171	*
A.AA **A**	*175*			0.170	*
..AA **A**	*524*			0.166	*
AAA. **A**	*172*			0.155	*
AA.A **A**	*187*			0.134	*
T.TT **A**	*88*			0.161	
..TT **A**	*329*			0.152	
...T **A**	*1373*			0.152	
TT.T **A**	*96*			0.144	
CCC. **C**	*205*			0.115	

Downstream motifs -4 bp to +4 bp	*n*	$E_{(P50)}$	$E_{(P95)}$	$E_{(KS2)}$
.... **G** CCCC	*26*	0.367	0.344	0.244
.... **G** CCC.	*93*	0.304	0.289	
.... **G** CC.C	*79*	0.276	0.265	
.... **G** C.CC	*53*	0.268	0.256	
.... **G** CC..	*271*	0.261	0.243	
.... **G** .CCC	*68*	0.256	0.227	
.... **G** C...	*711*	0.190	0.164	
.... **G** AAAA	*27*	0.234		
.... **G** AA.A	*66*	0.221	0.169	
.... **G** A.AA	*69*	0.178		
.... **G** AA..	*201*	0.174	0.125	
.... **G** AAA.	*68*	0.161		
.... **G** A...	*613*	0.120	0.057	
.... **G**	*2169*	0.098	0.074	
.... **C** AA..	*465*	0.068		
.... **T** .GGG	*18*		0.311	
.... **A** AAAA	*70*			0.238
.... **A** AAA.	*182*			0.235
.... **A** AA.A	*172*			0.186
.... **A** A.AA	*187*			0.184
.... **A** AA..	*524*			0.180
.... **A** .AAA	*175*			0.169
.... **A** A...	*1604*			0.112

$E_{(P50)}$, $E_{(P95)}$ – Motif enrichment score among positions ranked by the overall %NR population distribution median or 95th percentile; $E_{(KS2)}$ – Motif enrichment among positions ranked by %NR population distribution shape atypicality expressed via KS2 test statistic; n – number of instances motif is found in the mitochondrial genome; Wildcard '•' represents motif positions where base type is not restricted; Only motifs with statistically significant negative enrichment with p<0.05 after correction for multiple comparisons listed; Motifs where related sequence patterns of the same family are linked to %NR strand bias marked with asterisk

The motif families representing variants of the 'CCT' and 'TTG'-type patterns and the downstream 'GCC' and 'GAA'-type patterns in particular are good candidates for sequence patterns potentially linked to genuine mitochondrial heteroplasmy with as large number of variants linked to high overall %NR and no variants found to show association with high NR strand bias. It should be noted that even though two 'GGA'-type motifs (GNGGA and NGGGA) linked to high overall %NR levels did not show statistically significant link to high to %NR strand bias, three other motifs of the same family (GGNGA, NNGGA and NNNGA) were significantly enrichment among positions ranked by strand-specific %NR behaviour differences. Therefore 'GGA'-type motif family most likely represents a pattern associated with sequencing artefacts rather than genuine heteroplasmy, and the instances where motifs did not reach statistical significance for strand bias link are due to the low number of instances these motifs are found within the MT genome. Finally, **figure 3.4.2 F3** includes multiple variants of sequence patterns where A at position of interest is found either preceded by upstream runs of As or Ts, or followed by a downstream run of As. However these motif families are linked to high %NR atypicality but low overall %NR median or 95th percentile levels. None of these patterns are linked to high %NR strand bias and may be potential indicators of MT genome sequences with lower than usual genuine heteroplasmy levels.

Overall, motif overlaps presented here demonstrate a major impact of what we assume to probably be systematic sequencing errors on the observable %NR landscapes in our analysed dataset. At the same time, these results also point to some sequence contexts that do not appear to be obviously linked to sequencing errors yet appear to be linked to higher %NR levels, which does suggest that among the false positives of technical artefacts there is potentially also signal we can detect that is due to genuine heteroplasmy peaks. However, we cannot prove this without a better understanding of the detailed error rate landscape of the mtDNA sequence.

3.5 Chapter conclusion

The data presented in this chapter provides an overview of the characteristics of the non-reference read frequency landscape found across the mitochondrial genomes of 2533 individuals, and demonstrates the highly sequence context dependent nature of this landscape. At the same time we also demonstrate the pervasive presence of what may be systematic sequencing errors, and show that many of the sequence contexts linked to the overall behaviours of the observed %NR landscape are in fact most likely sequences that induce mis-calls during sequencing.

However not every sequence context linked to specific %NR landscape behaviours fits the expected behaviour from an artefact, as we have identified some candidate sequence motifs which based on our data might be linked to higher true heteroplasmy levels on the population scale. Such examples are motif patterns 'CCG' and 'GGC'. It is not inconsistent with the literature that such motifs might be indeed linked to heteroplasmy, as for example 3 bp combination 'GGC' can be found within G4 quadruplex structures shown to be both capable of inducing sequencing errors as well as associated with mutations in vivo (Butler *et al*., 2020; Weissensteiner *et al*., 2022). Thus it is not unreasonable that amid the artefacts we are also detecting some true signal from a genuine heteroplasmy landscape.

The undeniable presence of what can only be described as an error landscape severely limits the ability to investigate the overall heteroplasmy landscape found within the population, and this must be handled in some way to separate what signal comes from a real biological phenomenon and what comes from technical artefacts introduced during the sequencing process. If we were able to quantify the landscape of just the errors excluding any signal that might be arising from true heteroplasmy, we might be able to devise a filter so that we could then identify sites where the %NR landscape does not match the expected error landscape, which would offer a way to identify the actual heteroplasmy levels. Whether this is a feasible possibility is however uncertain without a better understanding of what the error rate patterns across the mitochondrial genome are.

Understanding this sequencing error rate landscape within the mitochondrial genome context is the focus of the next chapter, where we systematically measure the precise error rates associated with different sequence contexts and analyse the variability and patterns of the error rate landscapes observed.

4 Results 2 – Context dependent DNA sequencing errors

4.1 Introduction

While the main research aim of this thesis is investigation of the mitochondrial heteroplasmy landscape, the results presented in the previous chapter shows that in order to do so we must first address an entirely different landscape – one of large, non-random presumably sequencing error patterns manifesting as DNA strand specific non-reference read frequency peaks found throughout the mitochondrial genome.

True heteroplasmy and sequencing errors both manifest as variant sequence reads and unless strand-specific differences are considered, the two may look indistinguishable leading to false positives being incorrectly interpreted as a real biological signal. Because of this, strand bias is commonly taken into account when analysing heteroplasmy, with positions exhibiting large allele frequency differences between sequencing directions often flagged up as artefacts. This limits one to being able to study sequence positions absent of error prone sequence contexts, potentially introducing bias in what heteroplasmies can be detected. For example, if a certain sequence pattern happened to be associated with higher mutation rates in vivo but also was more prone to inducing sequencing error, such pattern would be invisible to heteroplasmy identification approach where sites of suspected sequencing error are discarded, thus systematically biasing the results of such a study.

This is one of the major reasons why we need to address the error landscape first. Considering that the error component we have observed appears to be non-random showing links to specific sequence features, if we were able to systematically measure how much error is introduced by different sequence patterns across the entire population studied, we might be able to potentially build a model of the expected error landscape across the mitochondrial genome, which could provide a way to separate the true heteroplasmy component from the error component of an individual's %NR landscape.

Understanding sequence patterns that are associated with sequencing errors could also allow us to assess the likely validity of any sequence patterns we might later on find to be associated with genuine heteroplasmy, allowing us to ask how confident we can be that the particular motif is not in fact a result of misidentified systematic sequencing error masquerading as a true heteroplasmy.

Research questions addressed

The main questions that we address in this section are:

- What are the sequence patterns that show an association with significant levels of sequencing errors? Can sequence patterns be used to predict the error rate we should expect to find at a particular mtDNA site?
- How variable are the context specific error rates observed between the different samples? How much variability do individuals samples show in terms of which sequence patterns are found to be strongly linked to errors?

Motif error rate measurement approach

In order to address these questions, we must be able to quantitate the error rate associated with a particular sequence. In this section we do not address non-specific or random errors, but specifically focus on systematic context-dependent sequencing errors of the Illumina HiSeq platform.

As a sequencing-by-synthesis method, error rates associated with Illumina platform tend to be largely affected by the sequence context upstream of the base being added in that specific extension cycle, whereas the downstream sequence context should have a lesser impact unless some sort of back folding of the template occurs that could

interfere with the elongation step (Nakamura *et al*., 2011). Because of this we shall focus on analysing the context-dependent sequencing error rates (CSEs) associated with different upstream sequence motifs. The approach we use is similar to a method employed by a previous study (Allhoff *et al*., 2013), whereby motif error rate is estimated as the difference between the pooled read mismatch fractions in forward and reverse directions relative to the motif orientation across all identified motif instances within a sequence.

Unlike the study by Allhoff *et al*., 2013, which focused on data from a single individual, we assess error rates associated with sequence motifs both at a single individual level as well as across the entire population by pooling read match and mismatch counts across all motif instances identified in each of the analysed samples.

Motif error rate is calculated as the non-reference fraction of the pooled forward reads minus the non-reference fraction of the pooled reverse reads relative to motif orientation. On a single sample level we denote motif CSE rate as 'MER' and on the population level refer to motif error rate as 'ΣMER'. Motif strand bias statistical significance testing is done using Chi squared and Fisher's exact tests, and Bonferroni correction is applied to correct for multiple comparisons. For further motif error rate measurement specifics see **Materials & Methods 2.4**

4.2 Motifs associated with high level of context specific sequencing error

4.2.1 Motif spaces investigated

We report systematic measurement of context-dependent sequencing error rates associated with variable length mtDNA sequence motifs. In total we investigated 7 motif spaces encompassing all possible 2bp to 8bp sequence permutations. The number of motifs analysed within each motif space and the significance thresholds applied are summarised in **figure 4.2.1 F1** below.

The analysis was carried out using MT genome sequencing data of 2533 individuals. For calculation of motif error rates in each individual we used their respective mitochondrial genome consensus sequence generated by 1000 genomes consortium (1000 Genomes Project Consortium, 2015) as the reference (**Materials & methods 2.2**). For each motif space investigated we determined both the sample-level motif error rates in each individual as well as the population overall motif error rate for the pooled sequences across all samples.

We devised a strategy that uses both the calculated population-level and sample-level motif error rates and application of progressively more stringent filter criteria for shortlisting a set of motifs that we can be most confident about being strongly associated with context-dependent sequencing errors. The criteria applied include requirement that the motif is associated with statistically significant population-level error rate of at least 1%, as well as statistically significant sample-level motif error rates in at least 90% of the samples with that motif, and exclusion of any motifs only analysed in a single individual. **Figure 4.2.1 panel C** expands the characterisation of motif analysis described in Materials and Methods to include statistics of the number of motifs found meeting these sets of criteria.

The 1% threshold for population-level motif error rate was selected based on previous literature on sequencing error rates and background noise levels expected in high throughput sequencing systems. The minimum requirement that sample-level error rates reach statistical significance in 90% of samples with the motif was chosen to provide a robust filter removing motifs with large error rate variability between samples. Finally, the requirement that a motif was analysed in multiple samples is intended to reduce the impact of potential artefacts in the source data.

Using these criteria, we were able to identify the final sets of 1 2bp motif, 2 3bp motifs, 7 4bp motifs, 23 5bp motifs, 49 6bp motifs, 45 7bp motifs and 59 8bp motifs which we deem to be most strongly associated with significant context-dependent sequencing error rates. We analyse these shortlisted motifs in the next section.

A note on resolution limit of motif length

The strategy we use for measuring motif error rate relies on pooling sequence read data across many instances of a motif in order to mitigate the effect of sequence context on the opposite DNA strand. As motif length increases however, the probability of finding multiple instances of that motif in any given sequence length decreases. With fewer instances of a motif found the mitigation of the opposite sequencing direction context impact becomes less effective, effectively limiting the length of the motifs that can be usefully investigated. Therefore given the size of the human mitochondrial genome, we estimate 8bp as the upper practically useful motif length limit.

Figure 4.2.1 F1: Motif spaces investigated for association with context dependent DNA sequencing errors

Motif spaces investigated for association with context dependent DNA sequencing errors

(a) Seven motif spaces investigated for context dependent sequencing errors

Motif spaces of interest	2N	3N	4N	5N	6N	7N	8N
Motif length	2bp	3bp	4bp	5bp	6bp	7bp	8bp
Number of possible motifs	16	64	256	1,024	4,096	16,384	65,536

Motif spaces represents all possible DNA sequence permutations of a set length

(b) Significance thresholds used for assessing motif sequencing error rate statistical significance

Bonferroni adjusted thresholds	2N	3N	4N	5N	6N	7N	8N
For global population-level error rate ΣMER $\alpha_{\Sigma MER} = \alpha \,/\, motif\ space\ size$	3.13×10^{-3}	7.81×10^{-4}	1.95×10^{-4}	4.88×10^{-5}	1.22×10^{-5}	3.05×10^{-6}	7.63×10^{-7}
For individual sample-level error rate MER $\alpha_{MER} = \alpha \,/\, (motif\ space\ size \times population\ size)$	1.23×10^{-6}	3.08×10^{-7}	7.71×10^{-8}	1.93×10^{-8}	4.82×10^{-9}	1.20×10^{-9}	3.01×10^{-10}

**Significance testing performed using Fisher's exact and χ^2 tests at significance level $\alpha = 0.05$ adjusted for multiple comparisons*

(c) Number of motifs found meeting criteria for association with sequencing errors

Motifs found matching selected criteria	2N	3N	4N	5N	6N	7N	8N
Motifs found in at least 1 individual	16 *100%*	64 *100%*	256 *100%*	1,024 *100%*	4,088 *99.8%*	15,454 *94.3%*	42,375 *64.7%*
…and… Motif population error rate ΣMER > 0	9 *56.3%*	41 *64.1%*	156 *60.9%*	588 *57.4%*	2,244 *54.8%*	8,396 *51.2%*	22,593 *34.5%*
…and… Population ΣMER is statistically significant		39 *60.9%*	152 *59.4%*	573 *56.0%*	2,122 *51.8%*	6,503 *39.7%*	12,597 *19.2%*
…and… Population ΣMER ≥ 1%				27 *2.6%*	116 *2.8%*	460 *2.8%*	1,267 *1.9%*
…and… Sample-level MER is statistically significant in ≥ 90% of individuals with that motif	**1** *7.5%*	**2** *3.1%*	**7** *2.7%*	**23** *2.2%*	**49** *1.2%*	51 *0.31%*	116 *0.18%*
…and… Motif exists in more than 1 individual						**45** *0.27%*	**59** *0.09%*

Percentages indicate fraction of total motif space size; MER – sample-level motif error rate; ΣMER – population-level motif error rate

4.2.2 Motif sets with strongest association to sequencing errors

This section focuses on sets of 2bp to 8bp long sequence motifs deemed to be most strongly associated with significant context dependent sequencing errors.

The figures below list the sets of motifs with a statistically significant population ∑MER ≥ 1%, that were analysed in multiple samples and found to have statistically significant sample-level MER in at least 90% of those samples. The most error prone 2bp to 5bp motifs identified are listed in **figure 4.2.2 F1**, with the sets of the most significant 6bp, 7bp and 8bp motifs provided in **figures 4.2.2 F2**, **F3** and **F4** respectively.

Figure 4.2.2 F1: 2bp to 5bp mtDNA sequence motifs linked to most significant sequencing error rates

2bp to 5bp mtDNA sequence motifs with statistically significant global error rate ≥1% found in multiple individuals with ≥90% statistically significant sample-specific error rates

Motif space	Motif rank order	Motif sequence	Total motif instances found	Number of samples with motif	Average motifs per sample	Population global motif error rate **Σ*MER***	Median sample-specific motif error rate ***M MER***	Motif error rate variation in population **IQR*MER***	Sample % with statistically significant MER values
2N	*1*	GT	3,438,042	2,032	1691.9	1.43%	1.33%	0.77%	99.9%
3N	*1*	GGT	1,071,815	2,032	527.5	2.90%	2.73%	1.41%	99.9%
	2	AGT	1,039,377	2,032	511.5	1.08%	0.98%	0.65%	99.7%
4N	*1*	GGGT	379,982	2,030	187.2	4.43%	4.17%	2.09%	99.9%
	2	AGGT	289,435	2,030	142.6	2.86%	2.72%	1.46%	99.9%
	3	GAGT	269,055	2,032	132.4	1.88%	1.72%	1.19%	99.6%
	4	CGGT	102,709	2,031	50.6	1.71%	1.42%	1.09%	99.4%
	5	TGGT	299,689	2,032	147.5	1.47%	1.30%	0.80%	99.7%
	6	GGGA	215,003	2,030	105.9	1.12%	0.94%	0.84%	99.8%
	7	AAGT	217,185	2,031	106.9	1.04%	0.94%	0.61%	99.4%
5N	*1*	CGGGT	16,245	2,013	8.1	10.23%	9.98%	4.66%	99.1%
	2	GGGGT	139,819	2,027	69.0	5.54%	5.19%	2.50%	99.8%
	3	AGGGT	121,545	2,030	59.9	4.39%	4.15%	2.29%	99.9%
	4	CAGGT	51,558	2,026	25.4	3.84%	3.76%	1.94%	99.4%
	5	GAGGT	98,636	2,028	48.6	3.61%	3.33%	2.10%	99.7%
	6	CGAGT	20,013	2,021	9.9	2.67%	2.32%	1.78%	95.2%
	7	AAGGT	55,292	2,028	27.3	2.52%	2.30%	1.29%	99.8%
	8	AGAGT	58,708	2,029	28.9	2.46%	2.30%	1.47%	97.9%
	9	TGGGT	102,373	2,028	50.5	2.26%	2.02%	1.23%	99.8%
	10	GCGGT	44,209	2,025	21.8	2.22%	1.81%	1.43%	99.1%
	11	GGAGT	89,644	2,028	44.2	2.12%	1.94%	1.36%	98.0%
	12	TAGGT	83,949	2,029	41.4	1.69%	1.53%	0.92%	99.4%
	13	GTGGT	82,166	2,029	40.5	1.69%	1.44%	1.01%	99.4%
	14	ATGGT	73,548	2,031	36.2	1.65%	1.52%	0.85%	99.1%
	15	CTGGT	38,936	2,026	19.2	1.53%	1.37%	0.87%	98.8%
	16	CCGGT	15,435	2,024	7.6	1.49%	1.27%	0.88%	92.8%
	17	ACGGT	19,294	2,024	9.5	1.48%	1.25%	0.96%	93.9%
	18	GGGGA	90,074	2,029	44.4	1.38%	1.11%	1.01%	99.2%
	19	AAAGT	61,780	2,028	30.5	1.26%	1.14%	0.67%	98.3%
	20	GAAGT	70,088	2,030	34.5	1.21%	1.10%	0.82%	95.6%
	21	TGAGT	100,690	2,031	49.6	1.16%	1.05%	0.76%	95.9%
	22	TTGGT	105,039	2,032	51.7	1.14%	1.00%	0.64%	98.3%
	23	GGCGT	46,626	2,028	23.0	1.06%	0.91%	0.74%	94.8%

MER – Motif error rate, obtained as the non-reference base call frequency difference between two sequencing directions;
Σ MER – Population global motif error rate, determined from summed base call frequencies across all individuals;
M MER – Median person-specific motif error rate; IQR MER –Motif error rate population variation, expressed as interquartile range;
Significance testing: χ^2 and Fishers exact tests at significance level α = 0.05 after correction for multiple comparisons

Figure 4.2.2 F2: 6bp mtDNA sequence motifs linked to most significant sequencing error rates

6bp long mtDNA sequence motifs with statistically significant global error rate ≥1% found in multiple individuals with ≥90% statistically significant sample-specific error rates

Motif space	Motif rank order	Motif sequence	Total motif instances found	Number of samples with motif	Average motifs per sample	Population global motif error rate **Σ*MER***	Median sample-specific motif error rate ***M MER***	Motif error rate variation in population **IQR*MER***	Sample % with statistically significant MER values
	1	GCGGGT	3,546	1852	1.9	38.40%	37.75%	16.15%	99.9%
	2	GCGGGA	2,410	1449	1.7	12.50%	12.39%	15.19%	94.0%
	3	CGGGGT	14,894	2023	7.4	8.57%	8.12%	3.87%	97.5%
	4	GCAGGT	14,833	2008	7.4	8.11%	7.93%	4.29%	98.8%
	5	GAGGGT	37,836	2028	18.7	6.12%	5.69%	3.17%	99.8%
	6	GGGGGT	54,496	2025	26.9	6.03%	5.69%	2.69%	99.6%
	7	AAGGGT	17,798	2021	8.8	5.46%	5.36%	2.69%	98.8%
	8	AGGGGT	30,897	2025	15.3	5.19%	4.97%	2.73%	99.6%
	9	TCGGGT	3,801	1984	1.9	5.17%	4.49%	3.07%	94.9%
	10	GGAGGT	38,588	2019	19.1	4.48%	4.13%	2.66%	99.1%
	11	AGAGGT	22,582	2025	11.2	4.19%	3.94%	2.31%	98.4%
	12	TGGGGT	39,532	2024	19.5	3.97%	3.69%	2.19%	99.3%
	13	ACGGGT	5,469	1978	2.8	3.71%	3.38%	2.48%	93.6%
	14	GAAGGT	16,100	2027	7.9	3.64%	3.39%	1.90%	99.2%
	15	TAGGGT	49,436	2028	24.4	3.23%	3.04%	1.78%	99.5%
	16	CAAGGT	7,506	2008	3.7	2.99%	2.62%	1.96%	95.2%
	17	GGCGGT	13,060	2006	6.5	2.97%	2.35%	2.07%	97.8%
	18	CAGGGT	16,475	2023	8.1	2.86%	2.67%	1.46%	97.5%
	19	GAGAGT	19,185	2026	9.5	2.71%	2.46%	1.67%	91.9%
	20	GGGAGT	17,697	2019	8.8	2.70%	2.48%	1.80%	91.6%
	21	ACAGGT	11,986	2004	6.0	2.68%	2.41%	1.55%	95.3%
	22	ATGGGT	15,446	2015	7.7	2.60%	2.23%	1.51%	97.9%
	23	CTAGGT	13,039	2014	6.5	2.57%	2.42%	1.39%	95.4%
	24	GTGGGT	34,400	2027	17.0	2.47%	2.14%	1.42%	99.3%
6N	*25*	TGCGGT	11,392	2017	5.6	2.46%	2.10%	1.47%	91.7%
	26	CGAGGT	11,130	2012	5.5	2.41%	2.25%	1.53%	95.6%
	27	TGAGGT	26,336	2027	13.0	2.39%	2.18%	1.42%	97.1%
	28	GGTGGT	24,847	2027	12.3	2.27%	1.94%	1.37%	97.2%
	29	AGGAGT	35,001	2028	17.3	2.24%	2.07%	1.43%	94.9%
	30	TTGGGT	39,497	2026	19.5	2.13%	1.95%	1.17%	98.9%
	31	TCTGGT	11,200	2012	5.6	2.13%	1.94%	1.28%	93.3%
	32	AAAGGT	22,651	2025	11.2	2.05%	1.83%	1.04%	97.7%
	33	TCAGGT	17,123	2025	8.5	1.98%	1.78%	1.19%	96.6%
	34	ATAGGT	20,670	2027	10.2	1.84%	1.68%	1.02%	94.5%
	35	GATGGT	24,457	2029	12.1	1.77%	1.56%	1.08%	95.8%
	36	TGGAGT	33,001	2026	16.3	1.76%	1.58%	1.16%	92.0%
	37	CTGGGT	13,030	2021	6.4	1.70%	1.51%	1.01%	90.6%
	38	GTAGGT	22,700	2025	11.2	1.68%	1.45%	1.06%	96.4%
	39	GAAAGT	16,632	2021	8.2	1.68%	1.48%	0.99%	90.9%
	40	GGGGGA	24,733	2024	12.2	1.54%	1.24%	1.13%	93.4%
	41	AATGGT	17,603	2027	8.7	1.50%	1.31%	0.75%	95.4%
	42	TGGGGA	34,463	2028	17.0	1.47%	1.12%	1.19%	93.5%
	43	TTTGGT	35,531	2031	17.5	1.34%	1.20%	0.88%	92.3%
	44	TGTGGT	32,533	2028	16.0	1.29%	1.11%	0.78%	95.3%
	45	AGGGGA	27,562	2024	13.6	1.19%	0.97%	0.77%	90.5%
	46	ATGAGT	39,835	2031	19.6	1.18%	1.07%	0.74%	91.2%
	47	TATGGT	22,105	2024	10.9	1.17%	1.01%	0.66%	90.2%
	48	TTAGGT	27,540	2026	13.6	1.17%	1.01%	0.66%	91.5%
	49	ATTGGT	29,232	2029	14.4	1.11%	0.95%	0.59%	93.1%

MER – Motif error rate, obtained as the non-reference base call frequency difference between two sequencing directions;
Σ MER – Population global motif error rate, determined from summed base call frequencies across all individuals;
M MER – Median person-specific motif error rate; IQR MER –Motif error rate population variation, expressed as interquartile range;
Significance testing: χ^2 and Fishers exact tests at significance level $\alpha = 0.05$ after correction for multiple comparisons

Figure 4.2.2 F3: 7bp mtDNA sequence motifs linked to most significant sequencing error rates

7bp long mtDNA sequence motifs with statistically significant global error rate ≥1% found in multiple individuals with ≥90% statistically significant sample-specific error rates

Motif space	Motif rank order	Motif sequence	Total motif instances found	Number of samples with motif	Average motifs per sample	Population global motif error rate Σ_{MER}	Median sample-specific motif error rate M_{MER}	Motif error rate variation in population IQR_{MER}	Sample % with statistically significant MER values
	1	GGCGGGT	3,534	1,851	1.9	38.56%	37.89%	16.04%	99.9%
	2	CCATGGT	5	5	1	18.88%	19.02%	3.82%	100%
	3	GCGGGGT	5,711	1,987	2.9	17.41%	16.85%	7.44%	97.2%
	4	GGCAGGT	7,096	1,991	3.6	15.70%	15.14%	7.89%	97.8%
	5	GGCGGGA	2,254	1,411	1.6	13.97%	12.87%	15.51%	96.7%
	6	GGAGGGT	13,131	2,021	6.5	9.48%	8.69%	5.19%	99.1%
	7	GTGCGGT	1,783	1,780	1.0	9.36%	8.42%	5.97%	94.6%
	8	TAGGGGT	7,668	2,023	3.8	8.08%	8.40%	6.66%	91.9%
	9	CGGAGGT	5,554	1,979	2.8	7.45%	6.53%	4.76%	93.5%
	10	CCAAGGT	1,958	1,934	1.0	7.22%	6.34%	4.60%	94.0%
	11	AGGGGGT	23,331	2,024	11.5	6.86%	6.67%	2.41%	98.7%
	12	GAGGGGT	7,815	2,005	3.9	6.09%	5.77%	3.70%	95.6%
	13	GGGGGGT	13,135	2,007	6.5	6.08%	5.42%	3.51%	97.5%
	14	AAGAGGT	5,707	2,017	2.8	5.18%	4.66%	2.91%	93.3%
	15	TAAGGGT	9,650	2,011	4.8	5.11%	4.85%	2.74%	98.1%
	16	CTCGGGT	2	2	1	5.09%	4.99%	0.35%	100%
	17	GTGGGGT	11,355	2,002	5.7	5.01%	4.73%	2.70%	98.0%
	18	TGGGGGT	10,957	2,021	5.4	4.85%	4.52%	3.17%	95.3%
	19	AGAGGGT	5,397	1,988	2.7	4.81%	4.34%	2.72%	91.4%
	20	CGGGGGT	7,073	1,986	3.6	4.74%	4.22%	2.81%	94.3%
	21	CGAGGGT	5,573	2,014	2.8	4.65%	4.15%	2.47%	93.1%
	22	TAGAGGT	5,577	1,990	2.8	4.63%	4.06%	2.71%	92.6%
7N	*23*	GACGGGT	3,815	1,974	1.9	4.44%	3.99%	2.82%	90.4%
	24	ATGGGGT	18,080	2,017	9.0	4.19%	3.84%	2.30%	98.9%
	25	GGGAGGT	14,852	2,015	7.4	4.18%	3.83%	2.58%	96.8%
	26	TGAGGGT	13,735	2,026	6.8	4.14%	3.77%	2.19%	98.5%
	27	CTAGGGT	18,809	2,026	9.3	3.96%	3.67%	2.38%	98.8%
	28	GAGAGGT	5,800	2,011	2.9	3.68%	3.36%	2.43%	90.7%
	29	AGAAGGT	7,546	2,019	3.7	3.66%	3.49%	1.98%	92.7%
	30	AGGAGGT	12,781	2,005	6.4	3.60%	3.31%	2.28%	94.1%
	31	AAGGGGT	9,904	2,022	4.9	3.47%	3.05%	2.02%	95.2%
	32	GTAGGGT	8,853	2,010	4.4	3.45%	3.19%	1.97%	93.3%
	33	CCCGAGT	3	3	1	3.42%	3.65%	3.09%	100%
	34	TCGGGGT	9,175	2,018	4.5	3.35%	3.10%	1.97%	90.6%
	35	TTAGGGT	12,710	2,025	6.3	3.15%	3.04%	2.08%	92.6%
	36	GCTAGGT	7,620	2,003	3.8	3.08%	2.93%	1.57%	93.4%
	37	GTTGGGT	13,180	2,010	6.6	3.02%	2.79%	1.71%	96.4%
	38	AGGTGGT	9,129	2,016	4.5	3.01%	2.66%	1.83%	91.2%
	39	GGTGGGT	8,254	2,021	4.1	2.90%	2.53%	1.89%	91.7%
	40	GTCTGGT	7,710	2,009	3.8	2.76%	2.49%	1.69%	92.6%
	41	CTTGGGT	7,486	2,000	3.7	2.76%	2.33%	1.84%	90.6%
	42	TCAGGGT	11,117	2,022	5.5	2.67%	2.47%	1.52%	93.9%
	43	TTGAGGT	16,871	2,024	8.3	2.25%	2.05%	1.40%	92.8%
	44	TGATGGT	13,410	2,012	6.7	2.22%	1.91%	1.44%	93.6%
	45	TGTGGGT	15,530	2,024	7.7	2.20%	1.86%	1.29%	93.9%

MER – Motif error rate, obtained as the non-reference base call frequency difference between two sequencing directions;
Σ MER – Population global motif error rate, determined from summed base call frequencies across all individuals;
M MER – Median person-specific motif error rate; IQR MER –Motif error rate population variation, expressed as interquartile range;
Significance testing: χ^2 and Fishers exact tests at significance level $\alpha = 0.05$ after correction for multiple comparisons

Figure 4.2.2 F4: 8bp mtDNA sequence motifs linked to most significant sequencing error rates

8bp long mtDNA sequence motifs with statistically significant global error rate ≥1% found in multiple individuals with ≥90% statistically significant sample-specific error rates

Motif space	Motif rank order	Motif sequence	Total motif instances found	Number of samples with motif	Average motifs per sample	Population global motif error rate **Σ*MER***	Median sample-specific motif error rate ***M MER***	Motif error rate variation in population **IQR*MER***	Sample % with statistically significant MER values
	1	TGGCAGGT	1,702	1,701	1.0	56.54%	55.93%	21.26%	100%
	2	TGGCGGGT	1,831	1,752	1.0	50.66%	50.09%	13.77%	100%
	3	CGGCGGGT	1,661	1,661	1	25.22%	22.49%	15.28%	97.8%
	4	GGCGGGGT	3,795	1,969	1.9	23.91%	23.30%	10.42%	95.8%
	5	CGGCGGGA	1,378	1,377	1.0	22.98%	22.31%	11.60%	99.4%
	6	GCAGGGGT	2	2	1	21.12%	25.07%	40.65%	100%
	7	CTCCCCCG	84	83	1.0	20.37%	22.39%	15.70%	92.8%
	8	CCCATGGT	5	5	1	18.88%	19.02%	3.82%	100%
	9	GGGGGGGT	490	458	1.1	18.66%	18.88%	12.97%	95.2%
	10	CTCATGGT	1,478	1,471	1.0	17.88%	18.48%	6.77%	96.1%
	11	CGAAGGGT	1,616	1,616	1	17.88%	16.85%	9.08%	98.8%
	12	GCGGAGGT	1,740	1,739	1.0	15.10%	13.92%	8.86%	99.1%
	13	CGGGGGGT	3	3	1	14.65%	8.03%	15.47%	100%
	14	CAGGGGGT	7,470	2,010	3.7	14.46%	14.16%	5.20%	98.0%
	15	GCTTGGGT	50	50	1	13.66%	13.94%	9.95%	94.0%
	16	GTGGCAGT	2	2	1	13.10%	16.00%	22.13%	100%
	17	CGAAAGGT	69	69	1	12.74%	11.40%	6.32%	98.6%
	18	TCTTTGGT	1,511	1,510	1.0	12.54%	11.74%	5.43%	97.5%
	19	CAAGGGGT	18	18	1	11.70%	12.32%	6.72%	100%
	20	ATGGCTTG	2	2	1	11.31%	11.32%	0.19%	100%
	21	AGGAGGGT	9,422	2,020	4.7	11.16%	10.09%	5.98%	98.8%
	22	GCAGGTTA	1,404	1,400	1.0	10.44%	8.12%	7.72%	91.1%
	23	TGCGCGGT	3	3	1	10.03%	10.47%	4.10%	100%
	24	CTCGGGGT	3	3	1	9.81%	8.31%	5.71%	100%
	25	TGTGCGGT	1,779	1,779	1	9.37%	8.42%	5.99%	93.9%
	26	CAAGAGGT	1,843	1,843	1	9.30%	8.13%	5.82%	90.7%
	27	CTGGAGGT	2	2	1	9.19%	9.19%	1.72%	100%
	28	AACGGGGT	2	2	1	8.57%	9.96%	12.91%	100%
	29	GTGGGGGT	4,031	1,951	2.1	7.82%	7.35%	5.19%	90.3%
N8	*30*	ACGGAGGT	23	23	1	7.36%	7.30%	3.78%	95.7%
	31	GCAAGGGT	2	2	1	7.27%	7.08%	2.07%	100.0%
	32	GCCAAGGT	1,954	1,932	1.0	7.23%	6.34%	4.60%	93.0%
	33	TAGAGAGT	3	3	1	6.70%	6.72%	3.50%	100%
	34	CTAGAGGT	3,629	1,943	1.9	6.62%	5.98%	3.58%	94.4%
	35	ACCAAGGT	3	3	1	6.18%	6.39%	0.67%	100%
	36	AAGAAGGT	3,696	1,969	1.9	6.17%	6.00%	3.29%	92.3%
	37	GATGGGGT	5,645	1,994	2.8	6.10%	5.67%	3.39%	97.0%
	38	TGGGGGGT	8,647	2,001	4.3	5.91%	5.45%	3.41%	95.3%
	39	AGTGGGGT	7,670	1,994	3.8	5.74%	5.49%	3.22%	97.1%
	40	CGCAGGGT	7	7	1	5.72%	6.07%	4.15%	100%
	41	ACGGGAGT	2	2	1	5.51%	5.62%	3.44%	100%
	42	GTGGCGGT	3	3	1	5.46%	6.01%	4.13%	100%
	43	GCTAGGGT	9,074	2,011	4.5	5.43%	5.04%	3.22%	98.4%
	44	TCTCGGGT	2	2	1	5.09%	4.99%	0.35%	100%
	45	TTAAGGGT	4,285	2,005	2.1	5.01%	4.62%	3.13%	90.9%
	46	GTGAGGGT	5,440	2,009	2.7	4.92%	4.53%	2.60%	91.6%
	47	TAGAGGGT	5,363	1,988	2.7	4.81%	4.34%	2.72%	90.8%
	48	GTAAGGGT	3,707	1,957	1.9	4.79%	4.46%	2.45%	91.3%
	49	TGCCAGGT	2	2	1	4.63%	4.63%	0.10%	100%
	50	ACGAGGGT	5,496	2,014	2.7	4.63%	4.14%	2.45%	92.5%
	51	TGCTAGGT	3,913	1,987	2.0	4.35%	4.23%	2.37%	90.5%
	52	AATGGGGT	7,545	2,013	3.7	4.06%	3.60%	2.46%	95.7%
	53	CCTTTGGT	2	2	1	3.96%	4.49%	3.75%	100%
	54	AGGGAGGT	9,164	2,008	4.6	3.92%	3.61%	2.57%	91.9%
	55	ATGAGGGT	6,047	2,022	3.0	3.57%	3.21%	1.95%	90.6%
	56	TAAGGGGT	7,759	2,018	3.8	3.53%	3.09%	2.12%	92.6%
	57	AGTAGGGT	7,492	2,010	3.7	3.24%	2.97%	1.91%	91.1%
	58	GCCCGAGT	2	2	1	2.78%	2.92%	1.47%	100%
	59	GTTAGGGT	11,168	2,025	5.5	2.25%	1.97%	1.25%	90.4%

MER – Motif error rate, obtained as the non-reference base call frequency difference between two sequencing directions;
Σ MER – Population global motif error rate, determined from summed base call frequencies across all individuals;
M MER – Median person-specific motif error rate; IQR MER –Motif error rate population variation, expressed as interquartile range;
Significance testing: χ^2 and Fishers exact tests at significance level $\alpha = 0.05$ after correction for multiple comparisons

The first and most obvious observations from **figure 4.2.2 F1**-**F4** is the similarity between the motifs, with the base at the error prone position (position 0) in the vast majority of cases being 'T' that is nearly always preceded by 'G' at positions -1 and -2. Positions further upstream show slightly more variation, but generally the closer an upstream base is to position 0, the more biased towards 'G' it appears to be.

Secondly, the longer the investigated motifs become, the higher the maximum observed error rate associated with the most extreme motif tends to be, and among the longer motif sets we see some very extreme values such as 38.5% ∑MER associated with 7bp motif GGCGGGT or 56.5% and 50.7% ∑MER associated with 8bp motifs TGGCAGGT and TGGCGGGT respectively. Generally, the most extreme motifs of different lengths tend to be closely related, but we see large ∑MER differences even between very similar motifs. For example, the 2nd and 3rd highest ∑MER 8bp motifs (TGGCGGGT & CGGCGGGT) are both derivatives of the highest ∑MER 7bp motif GGCGGGT, but the ∑MER associated with motif where the base at position -7 is 'T' is twice the ∑MER of the motif where position -7 is 'C'. This is suggestive that at least in some cases bases even quite far upstream can have a noticeable impact on the error rate at the position of interest.

Evidence that bases as far upstream as -7 might be impacting sequencing error rates highlights the importance of looking at the sequence contexts further upstream than just the adjacent few positions. However we are limited by the length of the mitochondrial genome. Figures above show that as the motif length increases, we observe fewer instances of each motif found per sample, and we also start seeing motifs found only at a single instance only in and in a very small number of samples. While all highly error prone 6bp motifs were found in 1400 samples minimum at minimum of 1.7 instances per sample on average, at 7bp length there are 3 motifs which we found exactly once in 5 or fewer individuals. This is even more extreme among the 8bp motifs, where 24 of the 59 most error prone motifs were analysed in less than 100 samples with exactly one instance found per sample. Of these 19 were analysed in less than 10 samples. This is a strong indicator that 8pb motifs are approaching the upper limit of the sequence pattern length that can be usefully analysed with our dataset.

All of the 2bp to 6bp motifs and most of the 7bp motifs show higher population-level ∑MER compared to the median sample-specific MER. Among the 8bp motifs the population ∑MER tends to be higher than the sample level MER median for motifs that are analysed in a large number of samples, but not when a motif is found in only a few samples. The population global ∑MER is calculated from the summed direction read match and mismatch counts across all samples, therefore it is in principle is similar to an average of all the sample MER values, except that it is weighted such that samples with fewer analysed reads contribute less to the overall value than samples with higher read depth do. But this also means that, as a type of an average, ∑MER would be more affected by outlier individuals compared to sample-level MER median. So the observation that median MER tends to be lower than the global ∑MER hints to ubiquitous presence of outlier individuals. The implications of this will be addressed in a later section.

4.2.3 Generalising sequence patterns associated with high error rates

As noted in the previous sections, there is an obvious link between increased sequencing error rate and upstream DNA sequence, especially 'GGT'-type patterns. Such patterns make up a far larger proportion of motifs associated with highest error rates than they do across the analysed motif spaces in general. This is true regardless of motif length investigated. And when we look at what proportion of the analysed positions upstream of the 'T' are Gs, this is much higher among highest error rate motif subsets than it is across all analysed motifs. For more detailed breakdown of the highly error prone sequence motif upstream region G-content data see **Supplementary figure 4.2.3 S1.**

The characteristics of the sequence patterns associated with context dependent sequencing errors can be visualised via the use of sequence logos (Schneider *et al.*, 1990), shown in **figure 4.2.3 F1**

Figure 4.2.3 F1 shows 'GT' (and 'GGT' among the longer motif sets) as a core component of the vast majority of the sequence patterns associated with elevated sequencing error rates, with the base compositions further upstream becoming progressively less skewed, as visualised by the big drop-off in the sequence logo base stack heights representing a steep decrease in the information content of the further upstream positions. Nevertheless, noticeable 'G' enrichment remains to be present throughout the entire upstream region, indicating that even though G positioning may be more flexible further upstream, increasing upstream G content must have an effect on the motif associated error rate. It is also of note that despite being enriched in G, the error prone motif sequences as far upstream as -5 tend to be very much depleted of C as far upstream as position -5, indicating that the reasons behind error proneness of a particular position goes beyond simply it's immediate upstream GC content.

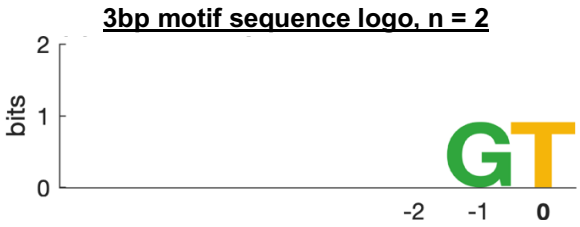
Addressing the few cases of highly error prone motifs that representing deviations from a 'GGT' type pattern, firstly there are three highly error prone 8bp motifs 'CTCCCCCG', 'GCAGGTTA', 'ATGGCTTG' which do not appear to share any patterns in common with other error prone sequences. Secondly, unlike the aforementioned unique 8bp motifs, a second group of 9 non-GGT motifs of variable lengths exists which do share a common pattern "GGA". This includes 4bp motif 'GGGA', 5bp motif 'GGGGA', four 6bp motifs 'GGGGGA', 'TGGGGA', 'AGGGGA' and 'GCGGGA', 7bp motif 'GGCGGGA' and an 8bp motif 'CGGCGGGA'. Interestingly, for every one of these GGA-type motifs there is an equivalent GGT-type motif with otherwise identical upstream sequence found within the most error prone motif set, and in every one of these motif pairs the error rate associated with the GGA-type motif is lower than that associated with its GGT-type equivalent. This is interesting as this shows that not only can 'A' also be significantly vulnerable to sequencing errors in certain contexts, but also that there might be a potentially consistent vulnerability difference between T and A when found with the same upstream sequence context.

Figure 4.2.3 F1: The common features of the most error prone sequence motif sets of different lengths

Base composition and sequence logos of the most error prone 3bp to 8bp motifs found in multiple samples with global ΣMER ≥1% and sample-level MER significance fraction ≥90%

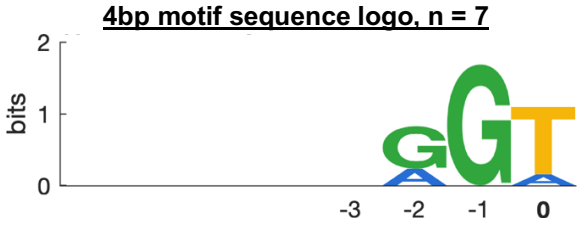
(a) Motif space N3: 2 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A						50%	0%	0%
C						0%	0%	0%
G						50%	100%	0%
T						0%	0%	100%



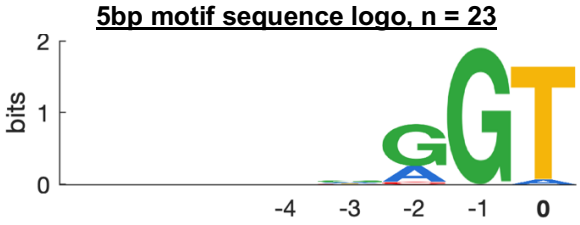
(b) Motif space N4: 7 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A					28.6%	28.6%	0%	14.3%
C					14.3%	0%	0%	0%
G					42.9%	71.4%	100%	0%
T					14.3%	0%	0%	85.7%



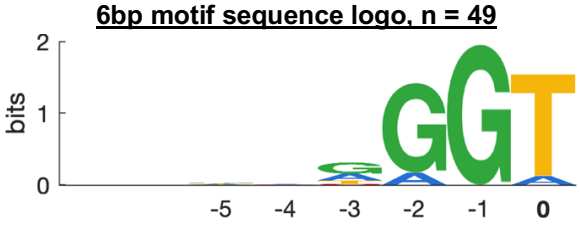
(c) Motif space N5: 23 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A				26.1%	26.1%	26.1%	0%	4.3%
C				21.7%	13.0%	4.3%	0%	0%
G				34.8%	43.5%	69.6%	100%	0%
T				17.4%	17.4%	0%	0%	95.7%



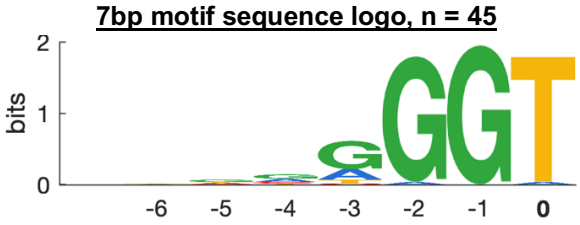
(d) Motif space N6: 49 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A			26.5%	24.5%	30.6%	12.2%	0%	8.2%
C			12.2%	16.3%	4.1%	0%	0%	0%
G			32.7%	36.7%	49.0%	87.8%	100%	0%
T			28.6%	22.4%	16.3%	0%	0%	91.8%



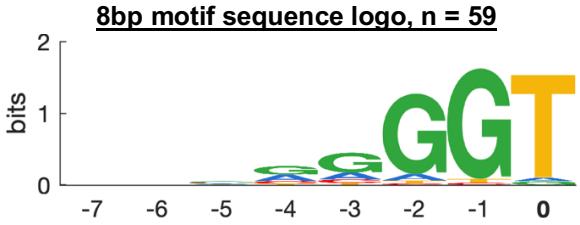
(e) Motif space N7: 45 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A		17.8%	17.8%	28.9%	24.4%	2.2%	0%	2.2%
C		20.0%	15.6%	15.6%	2.2%	0%	0%	0%
G		37.8%	42.2%	44.4%	64.4%	97.8%	100%	0%
T		24.4%	24.4%	11.1%	8.9%	0%	0%	97.8%



(f) Motif space N8: 59 motifs

Base type	Base composition by motif position							
	-7	-6	-5	-4	-3	-2	-1	0
A	22.0%	16.9%	18.6%	33.9%	22.0%	6.8%	0%	3.4%
C	27.1%	27.1%	22.0%	13.6%	10.2%	1.7%	1.7%	0%
G	28.8%	32.2%	40.7%	45.8%	61.0%	88.1%	94.9%	3.4%
T	22.0%	23.7%	18.6%	6.8%	6.8%	3.4%	3.4%	93.2%



Colour scale represents base frequency increase (red) or decrease (blue) from the expected baseline

4.2.4 Sequencing error rate differences between closely related motifs

While motifs linked to the highest sequencing error rates largely share common sequence patterns, there are obvious differences in the sequencing error rates associated with even very similar motifs. Here we investigate how the error rates differ between near-identical motifs depending on the location of the divergent base position.

To address this we shall focus on the top 10 8bp, 7bp and 6bp most error prone motifs with highest population-level ΣMER values. In the figures we compare the population ΣMER values associated with each of these top 10 most error prone motifs with the ΣMER of every possible motif that differs from the top motif by one single base. The data for the top 10 8bp motifs is shown in **Figure 4.2.4 F1**, with **figures 4.2.4 F2** and **figures 4.2.4 F3** focusing on 7bp and 6bp motifs respectively. In each figure **panel A** shows the 10 top error prone motifs of interest with their respective ΣMER values. **Panel B** lists ΣMER values of every possible related motif that differs by a single base from the corresponding motif of interest listed in panel A directly above. For equivalent analyses of the most error prone 5, 4, 3 and 2bp motifs see **supplementary figures 4.2.4 S1** and **S2**

It should be noted that any negative ΣMER values shown in the figures below are an artefact of MER calculation where the motif-specific error rate is below the combined background noise of reverse strand mismatch rates across all instances of the motif, with the magnitude of sub-zero ΣMER values thus essentially meaningless and functionally equivalent to zero.

Figure 4.2.4 F1: Sequencing error rate differences between motifs closely related to the 10 highest CSE 8bp motifs

Population level sequencing error rates associated with every possible 8bp long motif that differs only by a single base from the top 10 most error prone 8bp sequences

(a) Top 10 most error prone 8bp motifs associated with highest population level sequencing error rate

Top 8bp motifs	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 8	Motif 9	Motif 10
Motif sequence	TGGCAGGT	TGGCGGGT	CGGCGGGT	GGCGGGGT	CGGCGGGA	GCAGGGGT	CTCCCCCG	CCCATGGT	GGGGGGGT	CTCATGGT
Population ΣMER	56.54% **	50.66% **	25.22% **	23.91% **	22.98% **	21.12% **	20.37% **	18.88% **	18.66% **	17.88% **

(b) ΣMER across motifs most closely related to each of the 10 most error prone 8bp sequences

Related motif variant position		Motif 1 TGGCAGGT	Motif 2 TGGCGGGT	Motif 3 CGGCGGGT	Motif 4 GGCGGGGT	Motif 5 CGGCGGGA	Motif 6 GCAGGGGT	Motif 7 CTCCCCCG	Motif 8 CCCATGGT	Motif 9 GGGGGGGT	Motif 10 CTCATGGT
Position 0	-------A	1.42%	none found	22.98% **	3.33%		0.24%	0.12%	0.08%	2.74%	none found
	-------C	0.31%	0.58%	11.23%	none found	11.23%	0.94%	0.06%	0.13%	none found	0.26%
	-------G	-0.20%	-0.17%	-0.05%	-0.41%	-0.05%	-0.21%		none found	-0.32%	-0.06%
	-------T					25.22% **		0.10%			
Position -1	------A-	0.2%	0.11%	none found	-0.05%	none found	0.17%	none found	0.17%	0.17%	0.1%
	------C-	0.07%	0.18%	0.12%	2.87%	none found	0.29%		0.01%	-0.04%	-0.01%
	------G-							none found			
	------T-	none found	0.15%	none found	1.07%	none found	-0.07%	0.00%	none found	0.16%	-0.05%
Position -2	-----A--	0.62%	1.83%	none found	none found	none found	3.56% *	-0.25%	0.63%	1.38%	0.43%
	-----C--	none found	none found	none found	none found	none found	-0.31%		0.49%	0.64%	0.45%
	-----G--							none found			
	-----T--	0.64%	none found	none found	0.89%	none found	-0.13%	-0.04%	0.97%	0.98%	0.5%
Position -3	----A---		56.54% **	4.22%	1.82%	none found	none found	none found	2.43% **	3.06%	none found
	----C---	none found	none found	none found	1.83%	-0.05%	2.11%		none found	6.14%	13.95% **
	----G---	50.66% **						none found	1.44%		6.77% *
	----T---	1.40%	1.40%	0.80%	3.55%	none found	2.05%	0.02%		3.75%	
Position -4	---A----	5.85% *	3.44%	none found	1.69%	none found	7.27% **	0.17%		6.67% *	
	---C----				none found		none found		1.07%	9.54% *	0.58%
	---G----	6.97% *	5.91% **	14.65% **		1.68%		-0.11%	none found		none found
	---T----	5.42%	3.42% *	none found	3.36%	none found	none found	none found	1.09%	2.42%	0.92%
Position -5	--A-----	none found	1.04%	1.97%	10.91% *	none found		0.06%	2.02%	10.91% *	0.99%
	--C-----	4.63% **	none found	2.32%		0.46%	none found			23.91% **	
	--G-----				18.66% **		6.2% *	-0.06%	1.29%		1.86%
	--T-----	0.67%	4.77% **	none found	4.55%	none found	6.43% *	0.03%	1.01%	4.55%	none found
Position -6	-A------	1.77%	4.99%	none found	3.63%	none found	3.12%	-0.16%	1.49%	3.81% *	1.49%
	-C------	2.32%	3.29%	none found	none found	none found		26.06% *		6.2% *	18.88% **
	-G------						10.91% *	-0.17%	none found		none found
	-T------	1.68%	none found	none found	2.23%	none found	16.58% *		17.88% **	7.82% **	
Position -7	A-------	1.20%	none found	none found	4.26%	0.05%	2.08%	1.13%	none found	5.18% *	2.09%
	C-------	4.22%	25.22% **		none found		3.79%			14.65% **	
	G-------	4.46% *	9.54% *	9.54% *		none found		0.28%	none found		0.53%
	T-------			50.66% **	none found	none found	3.82% *	0.17%	none found	5.91% **	1.3%

*Colour scale: ΣMER above 1% in red, ΣMER below 1% in blue; 'none found' – no instances of motif found in any of the samples; Motifs marked with ** and * show statistically significant population ΣMER plus either statistically significant MER in ≥ 90% of samples (**) or ≥ 75% of samples (*) with that motif*

Figure 4.2.4 F2: Sequencing error rate differences between motifs closely related to the 10 highest CSE 7bp motifs

Population level sequencing error rates associated with every possible 7bp long motif that differs only by a single base from the top 10 most error prone 7bp sequences

(a) Top 10 most error prone 7bp motifs associated with highest population level sequencing error rates

Top 7bp motifs	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 8	Motif 9	Motif 10
Motif sequence	GGCGGGT	CCATGGT	GCGGGGT	GGCAGGT	GGCGGGA	GGAGGGT	GTGCGGT	TAGGGGT	CGGAGGT	CCAAGGT
Population ΣMER	38.56% **	18.88% **	17.41% **	15.70% **	13.97% **	9.48% **	9.36% **	8.08% **	7.45% **	7.22% **

(b) ΣMER across motifs most closely related to each of the 10 most error prone 7bp sequences

Related motif variant position		Motif 1 GGCGGGT	Motif 2 CCATGGT	Motif 3 GCGGGGT	Motif 4 GGCAGGT	Motif 5 GGCGGGA	Motif 6 GGAGGGT	Motif 7 GTGCGGT	Motif 8 TAGGGGT	Motif 9 CGGAGGT	Motif 10 CCAAGGT
Position 0	------A	13.97% **	0.10%	0.39%	1.50%		0.61%	0.07%	1.27%	0.63%	0.31%
	------C	4.87% *	0.03%	*none found*	0.38%	4.87% *	0.90%	-0.03%	2.62%	0.09%	0.16%
	------G	-0.17%	-0.10%	-0.30%	-0.19%	-0.17%	-0.30%	-0.09%	-0.17%	-0.21%	-0.03%
	------T					38.56% **					
Position -1	-----A-	0.11%	-0.23%	0.21%	0.15%	0.18%	0.15%	-0.10%	0.09%	0.17%	0.03%
	-----C-	0.13%	0.00%	0.37%	0.07%	0.78%	0.06%	-0.01%	0.20%	0.10%	-0.19%
	-----G-										
	-----T-	0.06%	-0.01%	1.07%	0.03%	0.12%	0.01%	0.01%	-0.07%	-0.26%	*none found*
Position -2	----A--	1.83%	0.45%	*none found*	0.80%	0.57%	3.60% *	-0.04%	1.96% *	0.77%	3.57%
	----C--	*none found*	0.30%	1.72%	*none found*	0.07%	0.36%	0.21%	0.70%	0.45%	1.01%
	----G--										
	----T--	1.55%	0.54%	1.64%	0.73%	0.35%	0.56%	0.74%	0.94%	0.95%	0.59%
Position -3	---A---	15.70% **	7.22% **	1.51%		1.50%	3.66% *	3.64% *	4.63% **		
	---C---	1.90%	0.00%	7.85% *	1.90%	1.38%	*none found*		1.48%	2.55%	0.00%
	---G---		3.76%		38.56% **			5.01% **		4.74% **	3.76% *
	---T---	1.92%		2.07%	1.92%	3.73%	2.26%	1.14%	1.07%	2.79%	18.88% *
Position -4	--A----	9.48% **		2.76%	3.66% *	0.61%		1.33%	5.11% **	4.12%	
	--C----		1.07%	2.32%			38.56% **	2.64% *	1.89%	2.34%	3.66%
	--G----	6.08% **	2.13%		4.18% **	2.57%	6.08% **				3.23% *
	--T----	2.90% **	1.29%	2.03% *	3.25% *	0.44%	2.90% **	0.86%	2.35%	1.33%	4.50%
Position -5	-A-----	4.44% **	2.02%	6.09% **	3.72%	0.90%	8.44% *	0.76%		3.32% *	3.41%
	-C-----	2.32%			1.92%	0.46%	2.76%	7.85% *	3.35% **	3.23% *	
	-G-----		1.54%	6.08% **				3.84% *	4.85% **		4.12%
	-T-----	5.22%	1.67%	5.01% **	2.58%	-0.19%	3.45% **		1.84%	2.04%	0.90%
Position -6	A------	5.14%	0.62%	5.91%	1.77%	0.25%	4.81% **	1.83%	3.47% **	3.60% **	1.45%
	C------	3.29%		6.24% **	2.34%	*none found*	4.65% **	0.67%	3.32% *		
	G------		1.03%						6.09% **	4.18% **	1.89%
	T------	*none found*	6.97% *	3.35% **	1.71%	0.57%	4.14% **	1.22%		4.22% *	1.50%

*Colour scale: ΣMER above 1% in red, ΣMER below 1% in blue; 'none found' – no instances of motif found in any of the samples; Motifs marked with ** and * show statistically significant population ΣMER plus either statistically significant MER in ≥ 90% of samples (**) or ≥ 75% of samples (*) with that motif*

Figure 4.2.4 F3: Sequencing error rate differences between motifs closely related to the 10 highest CSE 6bp motifs

Population level sequencing error rates associated with every possible 6bp long motif that differs by only a single base from the top 10 most error prone 6bp sequences

(a) Top 10 most error prone 6bp motifs associated with highest population level sequencing error rates

Top 6bp motifs	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 8	Motif 9	Motif 10
Motif sequence	GCGGGT	GCGGGA	CGGGGT	GCAGGT	GAGGGT	GGGGGT	AAGGGT	AGGGGT	TCGGGT	GGAGGT
Population ΣMER	38.4% **	12.5% **	8.57% **	8.11% **	6.12% **	6.03% **	5.46% **	5.19% **	5.17% **	4.48% **

(b) ΣMER across motifs most closely related to each of the 10 most error prone 6bp sequences

Related motif variant position		Motif 1 GCGGGT	Motif 2 GCGGGA	Motif 3 CGGGGT	Motif 4 GCAGGT	Motif 5 GAGGGT	Motif 6 GGGGGT	Motif 7 AAGGGT	Motif 8 AGGGGT	Motif 9 TCGGGT	Motif 10 GGAGGT
Position 0	-----A	12.5% **		0.62%	0.83%	1.03% *	1.54% **	0.54%	1.19% **	1.02%	0.74% *
	-----C	2.95% *	2.95% *	0.22%	0.14%	0.44%	0.39%	0.12%	1.26%	0.39%	0.32%
	-----G	-0.18%	-0.18%	-0.18%	-0.23%	-0.18%	-0.25%	-0.12%	-0.19%	-0.15%	-0.17%
	-----T		38.4% **								
Position -1	----A-	0.13%	0.06%	0.30%	0.07%	0.08%	0.19%	0.07%	0.11%	0.02%	0.03%
	----C-	0.09%	0.37%	0.02%	-0.05%	0.02%	0.22%	0.06%	0.16%	0.01%	0.17%
	----G-										
	----T-	0.02%	0.02%	0.19%	-0.01%	-0.03%	0.04%	-0.03%	-0.01%	0.03%	-0.1%
Position -2	---A--	1.58%	0.46%	1.57%	0.74%	2.71% **	2.7% **	2.93% *	2.24% **	2.43% *	1.17% *
	---C--	0.63%	0.07%	0.77%	0.49%	0.6%	1.45% *	0.86%	1.08% *	0.76%	0.45%
	---G--										
	---T--	1.15%	0.53%	1.38%	0.65%	0.71%	0.84% *	0.88%	0.78%	0.66%	0.67%
Position -3	--A---	8.11% **	0.83%	2.41% **		3.64% **	4.48% **	2.05% **	4.19% **	1.98% **	
	--C---	1.95% *	0.88%	1.95% *	1.95% *	4.17%	2.97% **	1.07%	1.49% *	0.73%	2.97% **
	--G---				38.4% **						6.03% **
	--T---	1.38%	1.61%	1.73% *	1.38%	1.77% **	2.27% **	1.5% **	1.55% *	2.13% **	2.27% **
Position -4	-A----	6.12% **	1.03% *	2.86% **	3.64% **		6.12% **		5.46% **	3.23% **	3.64% **
	-C----			1.06%		38.4% **	38.4% **	3.71% **	3.71% **		8.11% **
	-G----	6.03% **	1.54% **		4.48% **	6.03% **		5.19% **		3.97% **	
	-T----	2.47% **	0.79%	1.7% **	1.68% **	2.47% **	2.47% **	2.60% **	2.60% **	2.13% **	1.68% **
Position -5	A-----	3.71% **	0.31%	5.19% **	2.68% **	5.46% **	5.19% **			3.71% **	4.19% **
	C-----	1.06%	0.91%		1.61%	2.86% **	8.57% **	2.86% **	8.57% **	1.06%	2.41% **
	G-----			6.03% **				6.12% **	6.03% **	38.4% **	
	T-----	5.17% **	1.02%	3.97% **	1.98% **	3.23% **	3.97% **	3.23% **	3.97% **		2.39% **

*Colour scale: ΣMER above 1% in red, ΣMER below 1% in blue; Motifs marked with ** and * show statistically significant population ΣMER plus either statistically significant MER in ≥ 90% of samples (**) or ≥ 75% of samples (*) with that motif*

While the 6bp motif space is nearly fully represented in our data, several of the 7bp and to an even greater extent 8bp motifs were not present in any of our analysed mtDNA sequences (corresponding cells in the figures are marked 'none found'). Nevertheless, all three figures visibly demonstrate several patterns of variable constraint experienced across different motif positions.

Generally, the closer we are to position 0, the less interchangeable different base types appear to be. For any of the GGT-type motifs a T>G change at the position 0 always fully eliminates any error rate association. T>A or T>C changes are less consistent and can sometimes eliminate and sometimes merely reduce the motif ΣMER.

Across the GGT and GGA type motifs, position -1 appears to be even more constrained compared to position 0, with a near universal loss of error association when G is changed to any other base. G at position -1 is however not a universal requirement for high ΣMER, as evidenced by the 8bp motif CTCCCCCG where changing C>G at position 0 also fully eliminated sequencing error association.

Compared to position -1, the base at position -2 shows a slightly lower constraint, with G>A change in GGT-type and GGA-type motifs being the least impactful on motif ΣMER. Beyond position -2, the base type constraint for 'GGT' type motifs generally tends to decrease the further upstream we go with a larger proportion of variant motifs generally showing higher ΣMER values. However this is not universal, and with many instances where the same type

of base change results in different levels and direction of ΣMER change between motif sets. Position -3 is interesting in that we see possibly the biggest variability in the level of the constraint between motifs and between base types here, with some motifs being far more permissive to base changes compared to others. Generally from positions -3 onwards, there appears to be little consistency in the effects of different base changes between motifs.

Considering that nearly all of the high ΣMER motifs addressed in the figures above represent variations of a 'GGT' pattern, it is harder to speculate about the characteristics of the sequence constraint in the few non-GGT type motifs present beyond noting that compared to motifs ending in 'GGT' both the three motifs representing a 'GGA' pattern as well as the single 8bp motif ending in a 'G' appear to show a higher level of base constraint throughout.

The presence of these rarer non-GGT pattern motifs with strong error association found in our analyses does suggest that while clearly less ubiquitous, other sequencing error inducing sequences may exist. If such sequences exist, but are longer or more tightly constrained as our data suggests, then the limit of maximum 8bp investigated here is insufficient to identify these with confidence in an mtDNA system. Instead, much longer nDNA would need to be analysed, with any conclusions made presently regarding high ΣMER motif positional constraints generally only applicable to GGT-type and to a lesser extent GGA-type patterns.

4.2.5 Predictive power of motif patterns for identification of error sites

The large number of 'GGT'-type sequence patterns found among the motif sets linked to highest sequencing errors does not automatically imply that this pattern is always a strong predictor of high error rate. Therefore in **figure 4.2.5 F1** we present the results of analyses assessing the robustness of different 'GGT' type patterns as predictors of 6, 6 and 8bp motifs associated with high sequencing error rate, defined as ΣMER >1% with statistical significance in >90% samples.

Figure 4.2.5 F1: GGT-like patterns as predictors for high motif associated sequencing error rate

Usefulness of 'GGT'-like patterns as predictors for 6bp to 8bp high error rate motifs found in multiple samples with statistical significance in ≥90% of samples and global ΣMER ≥1%

(a) Predictive power for high error rates associated with 8bp upstream sequences

8bp PATTERNS		Pattern match	Actual positive	True positive	False negative	False positive	True negative	PRECISION	SENSITIVITY	SPECIFICITY
Total Gs	Defined positions							Positive predictive value	True positive rate	True negative rate
≥4	---**GGGGT**	51	59	14	45	37	36,924	27.5%	23.7%	99.9%
	----**GGGT**	123		29	30	94	36,867	23.6%	49.2%	99.7%
	-----**GGT**	261		37	22	224	36,737	14.2%	62.7%	99.4%
	------**GT**	455		39	20	416	36,545	8.6%	66.1%	98.9%
	-------**T**	797		39	20	758	36,203	4.9%	66.1%	97.9%
≥3	----**GGGT**	175		31	28	144	36,817	17.7%	52.5%	99.6%
	-----**GGT**	485		46	13	439	36,522	9.5%	78.0%	98.8%
	------**GT**	1,072		50	9	1,022	35,939	4.7%	84.7%	97.2%
	-------**T**	2,354		50	9	2,304	34,657	2.1%	84.7%	93.8%
≥2	-----**GGT**	619		51	8	568	36,393	8.2%	86.4%	98.5%
	------**GT**	1,799		55	4	1,744	35,217	3.1%	93.2%	95.3%
	-------**T**	4,979		55	4	4,924	32,037	1.1%	93.2%	86.7%

(b) Predictive power for high error rates associated with 7bp upstream sequences

7bp PATTERNS		Pattern match	Actual positive	True positive	False negative	False positive	True negative	PRECISION	SENSITIVITY	SPECIFICITY
Total Gs	Defined positions							Positive predictive value	True positive rate	True negative rate
≥4	--**GGGGT**	15	45	11	34	4	14,881	73.3%	24.4%	100%
	---**GGGT**	35		21	24	14	14,871	60.0%	46.7%	99.9%
	----**GGT**	63		28	17	35	14,850	44.4%	62.2%	99.8%
	-----**GT**	99		28	17	71	14,814	28.3%	62.2%	99.5%
	------**T**	146		28	17	118	14,767	19.2%	62.2%	99.2%
≥3	---**GGGT**	60		27	18	33	14,852	45.0%	60.0%	99.8%
	----**GGT**	161		41	4	120	14,765	25.5%	91.1%	99.2%
	-----**GT**	330		41	4	289	14,596	12.4%	91.1%	98.1%
	------**T**	628		41	4	587	14,298	6.5%	91.1%	96.1%
≥2	----**GGT**	238		43	2	195	14,690	18.1%	95.6%	98.7%
	-----**GT**	684		44	1	640	14,245	6.4%	97.8%	95.7%
	------**T**	1,696		44	1	1,652	13,233	2.6%	97.8%	88.9%

(c) Predictive power for high error rates associated with 6bp upstream sequences

6bp PATTERNS		Pattern match	Actual positive	True positive	False negative	False positive	True negative	PRECISION	SENSITIVITY	SPECIFICITY
Total Gs	Defined positions							Positive predictive value	True positive rate	True negative rate
≥4	-**GGGGT**	4	49	4	45	0	4,031	100%	8.2%	100%
	--**GGGT**	7		7	42	0	4,031	100%	14.3%	100%
	---**GGT**	10		10	39	0	4,031	100%	20.4%	100%
	----**GT**	13		11	38	2	4,029	84.6%	22.4%	100%
	-----**T**	16		11	38	5	4,026	68.8%	22.4%	99.9%
≥3	--**GGGT**	16		15	34	1	4,030	94%	30.6%	100%
	---**GGT**	37		27	22	10	4,021	73.0%	55.1%	100%
	----**GT**	67		31	18	36	3,995	46.3%	63.3%	99.1%
	-----**T**	106		31	18	75	3,956	29.2%	63.3%	98.1%
≥2	---**GGT**	64		39	10	25	4,006	60.9%	79.6%	99.4%
	----**GT**	174		45	4	129	3,902	25.9%	91.8%	96.8%
	-----**T**	374		45	4	329	3,702	12.0%	91.8%	91.8%

ΣMER – Population global motif error rate, determined from summed base call frequencies across all individuals;
Statistical significance fraction – Individuals with a statistically significant MER as a fraction of all individuals with that motif;
Sequence pattern defined positions: 'T' – thymine at position 0, 'G' – guanines upstream of position 0, '-' – any base type;
Sequence pattern total Gs – minimum total number of guanines within the pattern, including any within the defined positions;
Pattern match – number of motifs within the motif subset analysed that match the pattern of interest;
Actual positive – number of motifs within the subset analysed that match the ΣMER and population significance fraction criteria

Generally, the 'GGT' type patterns investigated tend to show a high specificity, with the true negative rate ranging between 86.7% to 100% depending on motif space and pattern complexity. Across all three motif spaces investigated, while less tightly defined patterns with lower minimum G content high a higher sensitivity matching a larger proportion of the actual highly error prone motifs, such patterns also match in a high number of false positives and thus have very low predictive value. Patterns with higher minimum G content and with more consecutive 'G' required to immediately precede the 'T' result in higher specificity with fewer false positives. However, such patterns with more consecutive upstream Gs have lower sensitivity. It is worth noting, that the more tightly defined and longer the pattern is, the fewer possible sequence combinations matching it can exist. Motif 'GGGGT' is the most extreme case of this, with only 4 matches possible among 6bp sequences, requiring at minimum 45 out of the 49 identified highly error prone 6bp sequences to be classed as false negatives. If we look at 7bp motifs, 'GGGGT' pattern has 16 theoretically possible matching motifs, and even if all 16 were true positives (actually 11 are), this would still leave more than 2/3 of the most error prone 7bp motifs unaccounted for. Thus even though 'GGGGT' is the most precise predictor of high motif error rate, it is by necessity the least sensitive one, with true positive rates of only 8.2% for 6bp motifs and 24.7% for 7bp motifs. Interestingly for 8bp motifs 'GGGGT' pattern sensitivity does not increase any further and is just 23.7%, suggesting that the number of possible combinations is no longer the limiting factor when it comes to the longest motif set.

When looking at slightly more complex sequence patterns with fixed certain overall minimum upstream G content but only some of these Gs defined as immediately adjacent to T, the sensitivity of the pattern increases when one or two G positions are allowed to vary, while allowing further variability in G positioning leads to a drop in precision.

Overall this data indicates that while strong commonalities exist between many of the most error prone motifs, and the pattern 'GGT' offers perhaps one of the best compromises between high error-proneness prediction precision and sensitivity, there does not appear to be a single simple definitive pattern fully capable of describing all of our reported highly error prone sequences.

An important limitation of the above analyses that we have only classed as true positives the 'GGT' type motifs that passes all the chosen criteria for inclusion among our reported lists of the most significantly error prone sequences (i.e. population ΣMER ≥ 1% supported by evidence from multiple samples with ≥90% of which show statistically significant sample-level MER), while every other GGT type motif is classed as a false positive. This means we do not distinguish between the false positives classed as such due to genuinely low sequencing error rates across the population from motifs that are either very rare or for whatever reason show high error association in some but not all samples thus failing to reach our chosen 90% statistically significant sample threshold.

Distinguishing between these categories of 'GGT' type pattern match false positives is important due to the implications of a lack of GGT-type motifs with genuinely low ΣMER for the interpretation of our results. This is however not the case, and among all our analysed motifs that match a 'GGT' pattern, 22.7% of the 8bp motifs, 18.2% of the 7bp motifs and 6.3% of the 6bp motifs are associated with ΣMER below 1%. Even extending the pattern to 'GGGT', we still see 6.3% of the matching 8bp motifs and 3.3% of the 7bp with ΣMER <1%, although in this case all 'GGGT' matching 6bp motifs do show ΣMER >1% (See **supplementary figure 4.2.5 S1**)

The existence of a non-trivial proportion of 6, 7 and 8bp motifs matching GGT patterns exhibiting genuinely low sequencing error-rate association confirms that despite 'GGT' being a common feature, this pattern does not fully account for the sequence context dependent error rates we observe. This points to significant CSE-modifying impact from the further upstream positions, with the noticeable increase in low ΣMER motif proportion among 7bp and 8bp 'GGT'-type pattern motifs compared to 6bp motifs again highlighting the effects that bases as far upstream as positions -6 and -7 can have on the sequencing error rate.

4.2.6 Comparison of identified error prone contexts to previous literature

Unlike the present study, most literature on context dependent sequencing errors focuses the nuclear DNA and fairly short sequence motifs, with 'GGT' ubiquitously reported as the most error prone sequence motif identified across various sequencing-by-synthesis platforms, with motifs 'AGT' and 'CGT' being reported to a slightly lesser extent (Meacham *et al*., 2011; Abnizova *et al*., 2012; Stoler *et al*., 2021). A 2016 study of base triplets upstream of sequencing error sites reported 'GGG','CGG' and 'AGG' as the two three most error prone contexts for both Illumina HiSeq and MiSeq platforms, with the third highest HiSeq error rate reported for triplet 'AGG', and highest substitution rate at position 0 found for Ts (Schirmer *et al*., 2016). For MiSeq platform the next most error prone contexts after 'GGG' and 'CGG' reported by Schirmer *et al*., 2016, were 'TGG' and 'GCG', with A and T at position 0 showing higher substitution rates compared to C or G. Some studies report elevated sequencing error rates associated with DNA structures capable of forming non B-DNA, such as G quadruplexes (G4), with Weissensteiner *et al*, 2022, reporting 2.4 times increased base mismatch rate increase at G4 motifs for Illumina system with bases in loops (non-G spacers) 5 times more likely to be miscalled compared to stems (G rich strings) (Weissensteiner *et al*., 2022). However little published literature is available that would systematically quantify the sequencing error profiles associated with such longer sequence motifs, whether G4 forming or otherwise.

The one exception to this is a 2013 study by Allhoff *et al.*, which reports error rates associated with 4 and 8 bp sequence contexts in chromosome 1 across several sequencing platforms (Allhoff *et al.*, 2013). Error rates associated with sequence contexts reported in Allhoff *et al*., 2013, are expressed as 'error rate difference' or ERD, which is calculated as the difference between forward and reverse mismatch rates across all motif instances, and is equivalent to sample-level motif error rates reported in this study. Of particular interest to us are error rates reported by Allhoff *et al*. for Illumina HiSeq platform measured on chromosome 1 sequence of the 1000 genomes project phase 1 sample HG00108. Unlike Allhoff *et al*., we analyse mitochondrial sequences of the entire 1000 genomes project phase 3 dataset focusing mainly on population level error rates measured across all 2533 analysed samples, our dataset does also include the same sample GH00108, allowing for a direct comparison between context specific error rates detected in mtDNA to those reported in the nuclear genome of the same individual.

Figures 4.2.6 F1 and F2 below compare the most error prone 4bp and 8bp motifs respectively in our dataset to sequence contexts with strongest strand bias reported in Allhoff *et al*., 2013. Their reported motif ERD values are compared to our measured population ΣMER values as well as with the sample-level MER specific to sample HG00108. Allhoff *et al*. reported 4 significant 4bp-long sequence contexts that are associated with ERD≥1%. These are compared to our observed ΣMER and $MER_{HG00108}$ values in **Figure 4.2.6 F1 Panel A** with a scatter plot visualisation provided in **Panel B.**

Figure 4.2.6 F1: Overlap between our identified error-prone 4bp contexts with those reported in an earlier study

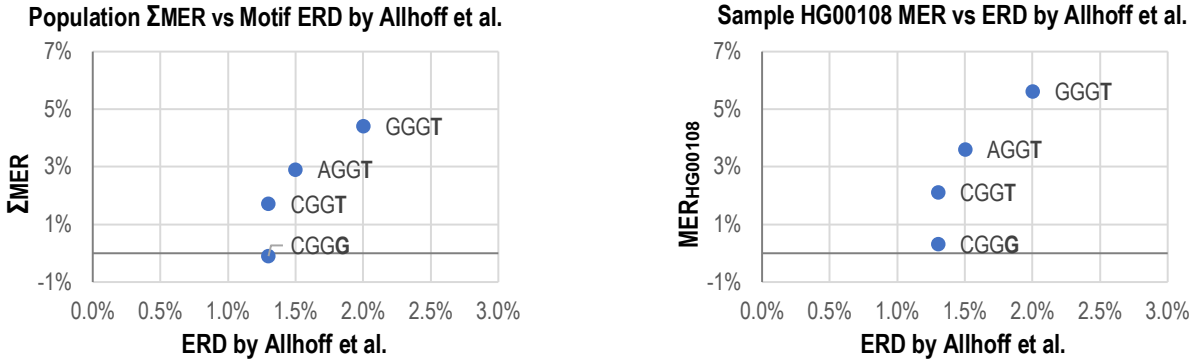
Overlap of the most error prone 4bp motifs discovered in this study with sequence contexts reported as strongly associated with sequencing errors by Allhoff et al., 2013

(a) Error rates measured in this study for all 4bp motifs reported as error prone* by Allhoff et al.

Motif sequence	Reported by Allhoff et al. Chr.1, HG00108		Found in this study mtDNA, 1000 Genomes phase 3 population (n = 2533)				mtDNA, HG00108 only	
	Rank[1]	ERD	Rank[2]	ΣMER	In samples	% Significance	Rank[3]	$MER_{HG00108}$
GGGT	1	2.0%	1	4.4%	2030	99.9%	1	5.6% ‡
AGGT	2	1.5%	2	2.9%	2030	99.9%	2	3.6% ‡
CGGG §	3	1.3%	240	-0.1%	2026	0.8%	26	0.3% ‡
CGGT	4	1.3%	4	1.7%	2031	99.4%	3	2.1% ‡

** Significant ERD ≥1% reported by Allhoff et al.; § 'CGGG' not linked to errors in our dataset (ΣMER <1%); ‡ Significant $MER_{HG00108}$*

(b) Most error prone 4bp motif error rate comparison between studies: a visualisation



***MER** – sample level motif error rate; **ΣMER** – population motif error rate; **ERD** – reported error rate difference, equivalent to MER;*
***Motif rank: 1** – by ERD reported in sample HG00108 excluding any contexts with wildcards, **2** – by population level ΣMER among all analysed motifs, with rank among most error prone motif subsets given in brackets, **3** – by sample-level MER measured in the sample HG00108; **% Significance** – samples with statistically significant MER as a fraction of all samples with the motif*

When analysing error rates associated with the longer 8bp contexts, Allhoff *et al*. used ERD threshold ≥10%, reporting 42 error prone 8bp sequence contexts above this threshold. 10 of these 42 contexts were not represented in mtDNA. Of the remaining 32 reported 8bp contexts present in the mitochondrial genome, 7 motifs, listed in **Figure 4.2.6 F2 panel A**, overlap with the 8bp motifs we also report as the most error prone (i.e. found in multiple samples, with significant ΣMER ≥1%, and >90% sample significance fraction). **Figure 4.2.6 F2 Panel B** shows a visual comparison between our observed error rates with those reported by Allhoff *et al*. for these 8bp motifs. **Panel C** lists further 13 reported 8bp contexts with ERD ≥10% which in our analysis did not pass the necessary significance thresholds, but either showed ΣMER ≥1% (11 motifs) or sample-level MER ≥1% (2 motifs) specifically in the same individual HG00108 as analysed by Allhoff *et al*. Lastly, 12 8bp contexts reported by *Allhoff et al*. showed no error association in our dataset having both ΣMER and $MER_{HG00108}$ values below 1%. These contexts, as well as the 10 motifs not found in mtDNA at all are listed in **panel D**.

Figure 4.2.6 F2: Overlap between our identified error-prone 8bp contexts with those reported in an earlier study

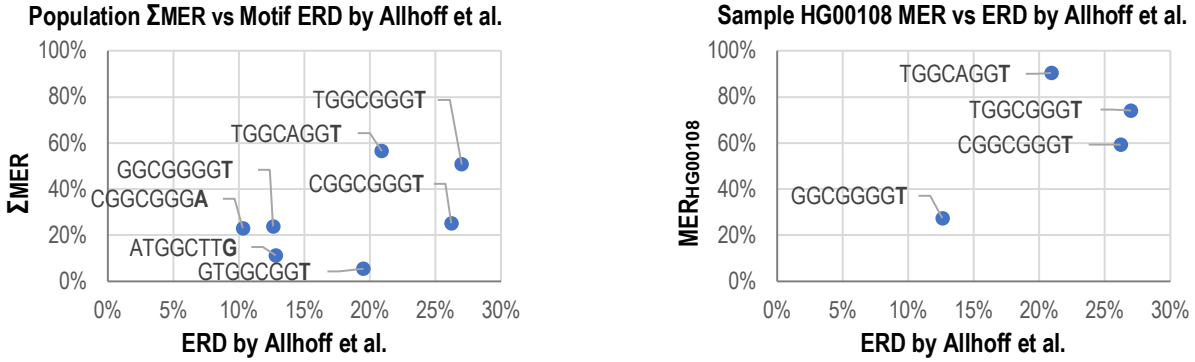
Overlap of the most error prone 8bp motifs discovered in this study with sequence contexts reported as strongly associated with sequencing errors by Allhoff et al., 2013

(a) Most error prone 8bp motifs* found in this study also reported as error prone by Allhoff et al**

Motif sequence	Reported by Allhoff et al. Chr.1, HG00108		Found in this study mtDNA, 1000 Genomes phase 3 population (n = 2533)				mtDNA, HG00108 only	
	Rank[1]	ERD	Rank[2]	ΣMER	In samples	% Significance	Rank[3]	$MER_{HG00108}$
TGGCGGGT	1	27.0%	2 (2)	50.7%	1752	100%	2	74.1% ‡
CGGCGGGT	2	26.2%	4 (3)	25.2%	1661	97.8%	3	59.3% ‡
TGGCAGGT	6	20.9%	1 (1)	56.5%	1701	100%	1	90.4% ‡
GTGGCGGT	10	19.5%	107 (42)	5.5%	3	100%	-	-
ATGGCTTG	25	12.8%	33 (20)	11.3%	2	100%	-	-
GGCGGGGT	28	12.6%	5 (4)	23.9%	1969	95.8%	9	27.4% ‡
CGGCGGGA	39	10.3%	6 (5)	23.0%	1377	99.4%	-	-

** Significant ΣMER≥1%, sample significance ≥90%; **Significant ERD ≥10% reported by Allhoff et al.; ‡ Significant MER $_{HG00108}$*

(b) Most error prone 8bp motif error rate comparison between studies: a visualisation



(c) 8bp motifs below 90% population significance with ΣMER ≥1% or with MER ≥1% in HG00108

Motif sequence	Reported by Allhoff et al. Chr.1, HG00108		Found in this study mtDNA, 1000 Genomes phase 3 population (n = 2533)				mtDNA, HG00108 only	
	Rank[1]	ERD	Rank[2]	ΣMER	In samples	% Significance	Rank[3]	$MER_{HG00108}$
GTGGCTCG	3	22.0%	38	10.8%	1843	83.8%	4	52.1% ‡
CGGCAGGT	4	21.9%	190	4.2%	1538	57.0%	-	-
TGGCTGGT	5	21.1%	930	1.4%	1853	11.8%	161	2.8%
ATGGCGGT	17	15.0%	641	1.8%	1972	19.5%	514	1.0%
ATGGCTGC	18	14.6%	1055	1.3%	1915	16.3%	19	14.3% ‡
GTGGCTGT	19	14.4%	151	4.7%	1972	30.3%	5	34.1% ‡
ACGGCGGT	21	13.7%	442	2.5%	2	0%	-	-
GTGGCTAC	27	12.6%	2143	0.7%	1803	11.0%	46	7.5% ‡
GTGGCTGC	30	12.3%	5133	0.3%	1938	2.5%	335	1.5%
AGGCAGGT	35	11.0%	1132	1.2%	2	0%	-	-
TGGCAGGA	36	10.7%	900	1.4%	1868	12.4%	66	5.6% ‡
GGCAGGGT	38	10.5%	732	1.7%	25	16.0%	-	-
CTGGCGGT	42	10.1%	547	2.1%	1807	15.6%	225	2.1%

‡ Statistically significant MER in sample HG00108

(d) 8bp motifs with reported ERD ≥10% but no error association detected or not found in mtDNA

No error association:	***Population ΣMER < 1% and sample-level $MER_{HG00108}$ < 1%***					
(n = 12)	*CGGCTGGT*	*ATGGCTAG*	*ACGGCGGG*	*GAGGCGGG*	*TATGGCGG*	*TATGGCTG*
	GTGGCTGG	*GTGGCTAG*	*ATGGCGGG*	*GTGGCGGG*	*GTGGCAGG*	*GTGGCTTG*
Not found in mtDNA:	***Not instances found in any of the 2533 analysed MT genomes***					
(n = 10)	*GCGGCGGT*	*TCGGCGGT*	*AGGCGGGT*	*ATGGCTCG*	*GCGGCTCG*	
	CCGGCGGT	*ATCGGCGG*	*TGGCGGGA*	*ACGGCTCG*	*ACGGCTTG*	

***MER** – sample level motif error rate; **ΣMER** – population motif error rate; **ERD** – reported error rate difference, equivalent to MER;*
***Motif rank: 1** – by ERD reported in sample HG00108 excluding any contexts with wildcards, **2** – by population level ΣMER among all analysed motifs, with rank among most error prone motif subsets given in brackets, **3** – by sample-level MER measured in the sample HG00108; **% Significance** – samples with statistically significant MER as a fraction of all samples with the motif*

In terms of 8bp contexts linked to sequencing errors, the 7 motifs listed in panel B represent 12% of all highly error prone 8bp motifs identified in our study and 22% of all such contexts reported by Allhoff *et al*. (excluding contexts with wildcards or motifs not present in the mitochondrial genome). These 7 overlapping motifs discovered by both

studies include all of the top 5 highest ΣMER motifs we have identified, as well as motifs ranked 20th and 42nd within our 59 most error prone set. If we also consider any other 8bp motifs with significant population ΣMER ≥ 1% which did not reach the 90% threshold sample significance fraction, this adds 11 further motifs which overlap between our dataset and sequence contexts reported by Allhoff *et al*., indicating a considerable overlap between the results of these two studies.

The differences however are very interesting. Excluding any sequence contexts not represented in the mitochondrial genomes within our dataset, Allhoff *et al*. report in total 32 error prone 8bp long motifs, of which base at position 0 most frequently is 'T' in 14 cases (43.8%) or 'G' in 13 cases (40.6%), with 'A' only found in 2 motifs and 'C' in 3 motifs. Of the two most common reported motif groups ending in T or G, our results support error association for 13 out of the 14 motifs ending in 'T' but only 2 of the 13 motifs ending in 'G'. Overall we found no population level error association for 11 reported motifs ending with 'G', two motifs ending with 'C' and one reported motif ending with 'T'. We also see similar difference between 4bp context reported in Allhoff *et al*., 2013, and our results, with our data supporting the reported error association of three 'GGT' type motifs 'GGGT', 'AGGT', and 'CGGT', but not for the reported motif 'CGGG'.

The results reported by Allhoff *et al*. come from a single individual, whereas our analysis was focused on measuring the aggregate motif error rates across a population. Because of this we also extracted only the error rates measured in the same individual as studied by Allhoff *et al*.. Inclusion of any motifs that match contexts reported by Allhoff *et al*. and for which we find MER>1% in the same individual, only results in addition of two new 8bp motifs, both ending with 'C'. We found no additional reported contexts ending in 'G' to be error prone in this specific individual based on our data.

It is interesting that even in the same individual our data selectively replicate error association findings by Allhoff *et al*. for 'GGT' and 'GGA'-type patterns but not for motifs ending in 'G'. This suggests the differences between our observed population level motif error rates and those reported by Allhoff *et al*. may not be a result of different data source (one individual vs population of 2533) or error rate calculation approach (over motif instances in one individual vs over aggregate motif instances across all individuals). Instead, these differences may stem from intrinsic characteristics of the different genomes being studied (nDNA Chr1 vs mtDNA) such as different GC content. Equally, hey could also be due to different data processing and sequence alignment steps (aligning raw sequence reads vs re-using alignment produced by 1000 genomes consortium, alignment to linear versus circular genome, different parameters used etc.) and potentially inclusion of reads from different sequencing runs (1000 genomes phase 1 vs phase 3 data).

It is worth noting that for all error prone 8bp contexts reported in Allhoff *et al*., 2013, and found in mtDNA that end in either 'T' (n=14) or 'A' (n=3) the base at position 0 is preceded invariably by at least two 'G's at positions -1 and -2. This is very similar to patterns we have observed in our datasets. This is not a case for motifs reported by Allhoff *et al*. that end in 'G' or 'C', where we see more variability at positions -1 and -2.

It should also be noted that even though our data does not replicate error rate findings for many of the reported 8bp motifs ending in 'G', for motif 'ATGGCTTG' for which Allhoff *et al*. report ERD of 12.8%, we did measure a comparable population error rate of 11.3% across two individuals, none of whom was the same individual analysed by Allhoff *et al*.. This suggests that even though this specific motif is not found in many samples, it is a highly error prone context in at least another study in a completely different sequence segment, all of which speak to this non-'GGT' type motif probably not being an artefact of our experiential approach.

Differences between error rates measured in the same individual

So far we have mostly focused on whether the same motifs were identified by the two studies as associated with sequencing errors, but we have not addressed the differences in the actual error rates measured.

Allhoff *et al*. reported a total of 42 error prone 8bp contexts associated with ERD≥10% that were identified in the individual HG00108. In the mitochondrial genome of the same individual we found 33 8bp motifs with statistically significant MER ≥10%. 7 motifs with motif error rates above 10% were found in this individual by both studies. The highest ERD that Allhoff *et al*. report is 27.0% for motif 'TGGCGGGT', for which we report 74.1% MER. The highest MER we found in this individual is 90.4% for 'TGGCAGGT', for which Allhoff *et al*. report 20.9% ERD. However MER reported in this study are not always higher than the ERD reported for the same motifs by Allhoff *et al*..

Of the 42 reported error prone 8bp contexts, we identified 21 in the mtDNA of individual HG00108, 9 of which showed statistically significant MER . For 6 of these motifs our measured motif error rates exceeded those reported by Allhoff *et al*., whereas for the other 3 significant motifs we measured sequencing error rates below those reported by Allhoff *et al*.. And for 12 motifs reported as error prone by Allhoff *et al*. we found no significant error association in sample HG00108 mitochondrial sequence in our data. Three of these were motifs ending in 'GGT' with MER between 1% and 2.8% but not significant. One motif, GTGGCTGC, showed 1.5% MER but also was non-significant. The remaining 8 motifs with motif error rates below 0.7% were representative of patterns ending in 'G'.

It must be noted that in this comparison between the reported motif error rates in sample HG00108, Allhoff *et al*. report ERD values measured across hundreds to thousands of motif occurrences found in Chromosome 1, whereas our reported MER values are calculated based on between 1 and 5 motif occurrences in the much shorter mitochondrial genome.

When we measure population level motif error rates across all motif instances found in any individual, we see lower maximum ΣMER values. Of the 7 most highly error prone motifs we identified across the population that were also reported by Allhoff *et al*., the error rates reported are below those measured by us for 4 motifs, whereas for the other 3 motifs ERD values reported by Allhoff *et al*. were above our measured ΣMER levels.

Overall the differences in motif error rates reported in Allhoff *et al*., 2013, or measured in this study across the population versus in just the individual HG00108 point to a substantial error rate variability for the same sequence contexts between samples. This is also demonstrated by our data presented earlier in this chapter where for the most error prone motifs identified we not only list the population aggregate ΣMER error rate levels but also include the sample level MER median and interquartile range values across the population which show substantial variation in sample-level motif error rates across the population as well as a noticeable differences in inter-sample variability levels between different motifs. This context dependent sequencing error rate variability between samples and between motifs is the focus of the next section of this chapter.

4.3 Motif specific sequencing error rate variation

We observed in the previous sections that the motifs identified as most significantly linked to sequencing errors at a population level show a substantial variation in the error rates detected between individuals. For example, for the three most error prone 8bp motifs identified with population ΣMER levels of 56.54%, 50.66% and 25.22%, we observe the MER IQR between samples of 21.26%, 13.77% and 15.28% respectively (see **figure 4.2.2 F4**).

High levels of sequencing error rate variability between samples could potentially complicate any attempts to systematically account for such context-dependent sequencing errors in experiments and therefore warrant addressing in more detail.

Two aspects of sequencing error rate variability need to be distinguished. One aspect is the variability in the error rates observed between different sequenced individuals for the same sequence motif. I.e. how much variation is there between samples, how does the level of inter-sample variability scale relative to the observed overall motif error rate, and is this consistent across the motif spaces or are there some outlier sequence patterns with atypically low or high levels of error rate variation between samples.

Another aspect of sequencing error rate variability is how much variation exists between different motif error rates within the same sample, and how this differs from sample to sample. I.e. do samples exhibit different levels of 'error prone-ness' and is this something across all or most analysed sequence motifs, or are there motifs which behave as outliers exhibiting atypically high or low error rates is some but not all individuals.

We attempt to address these questions in this section. Rather than focusing on just the subsets of highest ΣMER motifs, we address the wider motif error rate variability for entire motif spaces assessing what the MER variation behaviour characteristics are more generally.

In the following analyses we only included motifs with ΣMER >0 that were analysed in a minimum of two samples. We shall also focus mainly on 6bp motifs as the longest set of sequences for which the motif space is nearly fully represented on the mitochondrial genome.

4.3.1 Relationship between motif population ΣMER level and the MER variation between samples

When measuring sequencing error rates associated with different motifs we recorded both the motif overall population-level ΣMER as well as the sample-level MER values in each analysed sample. The interquartile range of the sample-level MER (MER IQR) provides a convenient measure of the level of error rate variability observed between samples for any given motif.

Supplementary figures available in **Appendix 2.2** provide an overview of motif MER IQR value distributions observed for each motif space N2 to N8. While the majority of the analysed motifs of all lengths have population MER IQR below 0.1%, as the investigated motif length increases, so does both the maximum observed global MER levels as well as the maximum observed inter-sample MER variability (See **supplementary figure 4.3.1 S1**).

When we look at individual motifs, we find that the inter-sample MER variability as measured by MER IQR correlates strongly with the motif global population error rate ΣMER across all motif spaces, with Pearson correlation coefficient ranging from 0.98 for 2bp motifs to 0.80 for 8bp motifs (**Supplementary figure 4.3.1 S2**). Interestingly, when the relationship between motif global error rate and its inter-sample MER variability is visualised (**Supplementary figure 4.3.1 S3**), the correlation between the two measures is very evident for motifs with ΣMER above around 0.1%-0.2%, but disappears for lowers ΣMER motifs, suggesting that below this threshold any MER variation between individuals may largely be just noise.

While MER IQR can be a very useful measure of motif error rate variability between samples, it is worthwhile to consider other descriptive statistics that describe motif MER distribution behaviours across the analysed motifs. For this we chose 6bp motif space as it is the longest mostly completely represented motif set in our data, and **Supplementary figure 4.3.1 S4** provides a visualisation of a range of motif MER distribution descriptive statistic distributions besides IQR for this 6bp motif space. It shows that the observed MER values 5^{th} to 95^{th} percentile range as well as the total range across the samples varies widely from motif to motif and produces highly skewed distributions, with an order of magnitude more variation found between motifs in terms of MER total range compared to MER IQR. Furthermore, the motif MER distribution skewness and kurtosis parameter distributions highlight that the vast majority motif MER distributions tend to be generally tail-heavy with presence of high MER outlier individuals for most motifs.

4.3.2 Variation in the sequencing error rate profiles observed between different samples

As discussed, on a single motif level the distributions of motif MER values from our analysed population tend to be positively skewed, albeit to variable extent. Similarly, when we look at the error rates measured for every analysed motif in the same sample, such individual-specific MER distributions also tend to be positively skewed. This is consistent with the vast majority of motifs in any given individual generally being associated with low error rates, with high MER values only observed in a relatively smaller proportion of motifs.

The individual-specific MER distribution behaviours across our analysed population are visualised in **Supplementary figure 4.3.2 S2** for 6bp motifs via box-and-whiskers plots depicting the distributions of various descriptive statistics describing all analysed motif MER value distributions observed in each sample. This figure demonstrates that when we compare how different MER destitution quantiles vary between individuals, much more variability is evident between individuals in terms of their MER distribution 75^{th} and 95^{th} quantiles compared to the respective MER distribution medians. This indicates that the majority of the variation in the sequencing error rate profiles between individuals must be arising from motifs generally associated with highest error rates and that at least some motifs with very high MER levels are found in nearly all individuals.

Considering our earlier observations on differences between the motif error rates measured on a population level to those measured in a single individual both in this study as well as those reported in previous literature (see **section 4.2.6**), the present observations of the high level of sample MER variability raises questions about how consistent the context-dependent sequencing error rate profiles are between different samples. Specifically, whether it is the same set of most 'error prone' samples producing highest MER values across all motifs, or are the unusually high outlier MER values for each motif observed in different sets of samples.

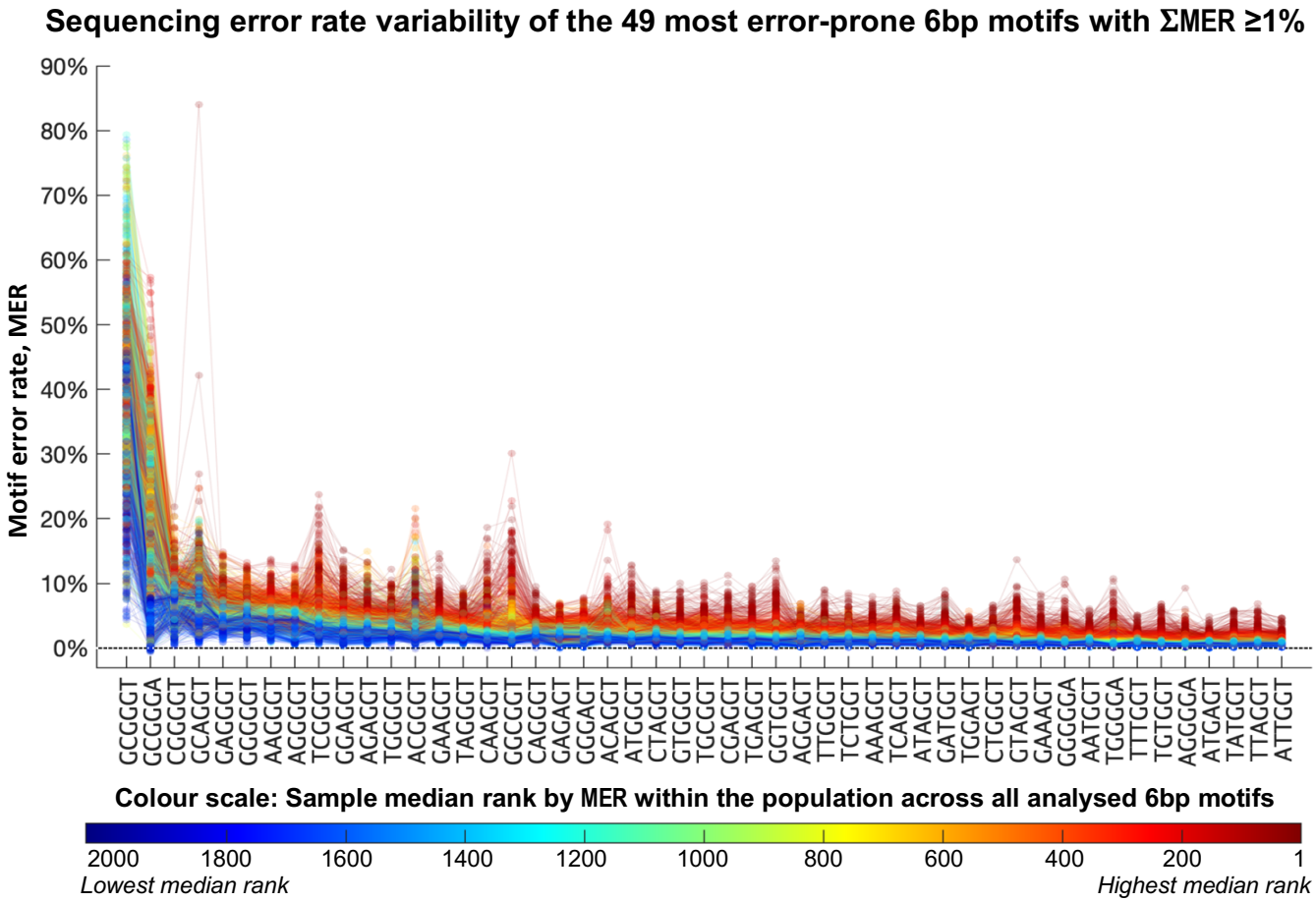
To address this we must look at how the measured MER values from a specific sample compare to the rest of the population across different motifs, and assess whether the patterns of which samples tend to have highest MER values across different motifs are the same.

Assuming that there is some level of error profile consistency (i.e. the same motifs show highest MER levels across many samples) between different sample, it is reasonable to expect that even if the actual error-proneness varies widely between samples, we should generally expect to see individuals with higher than average MER for one motif to also have higher than average MER for most other motifs. If indeed there is some level of consistency between motifs in terms of where within the population each sample ranks by its MER for that particular motif, we can potentially use the information of where within the population MER distribution on average the sample tends to rank as an approximation of the error-proneness of that sample. This principle is employed in the following figures.

Focusing first on the subsets of our identified most error prone motifs, **Figure 4.3.2 F1** presents a visualisation of the sample-specific motif error rate measurements for each of the 49 most significant 6bp motifs with ΣMER >1%. Each data point represent MER value of the specific motif in one sample, with the colour indicating the sample average MER rank within the population across all 6bp motifs, and aiming to highlight whether the samples with highest error rates on average do indeed consistently rank at the top of the population for all motifs.

Interestingly, **Figure 4.3.2 F1** demonstrates that this is not always the case, with what appears to be a fairly different sample order emerging for some motifs compared to others (i.e. many instances of data points in blue and green found at higher MER values compared to those in red). This is especially true for the motifs with highest ΣMER levels, while sample order for the slightly lower ΣMER motifs appears to be more consistent.

Figure 4.3.2 F1: Most error prone 6bp motifs: sequencing error rate variability between samples

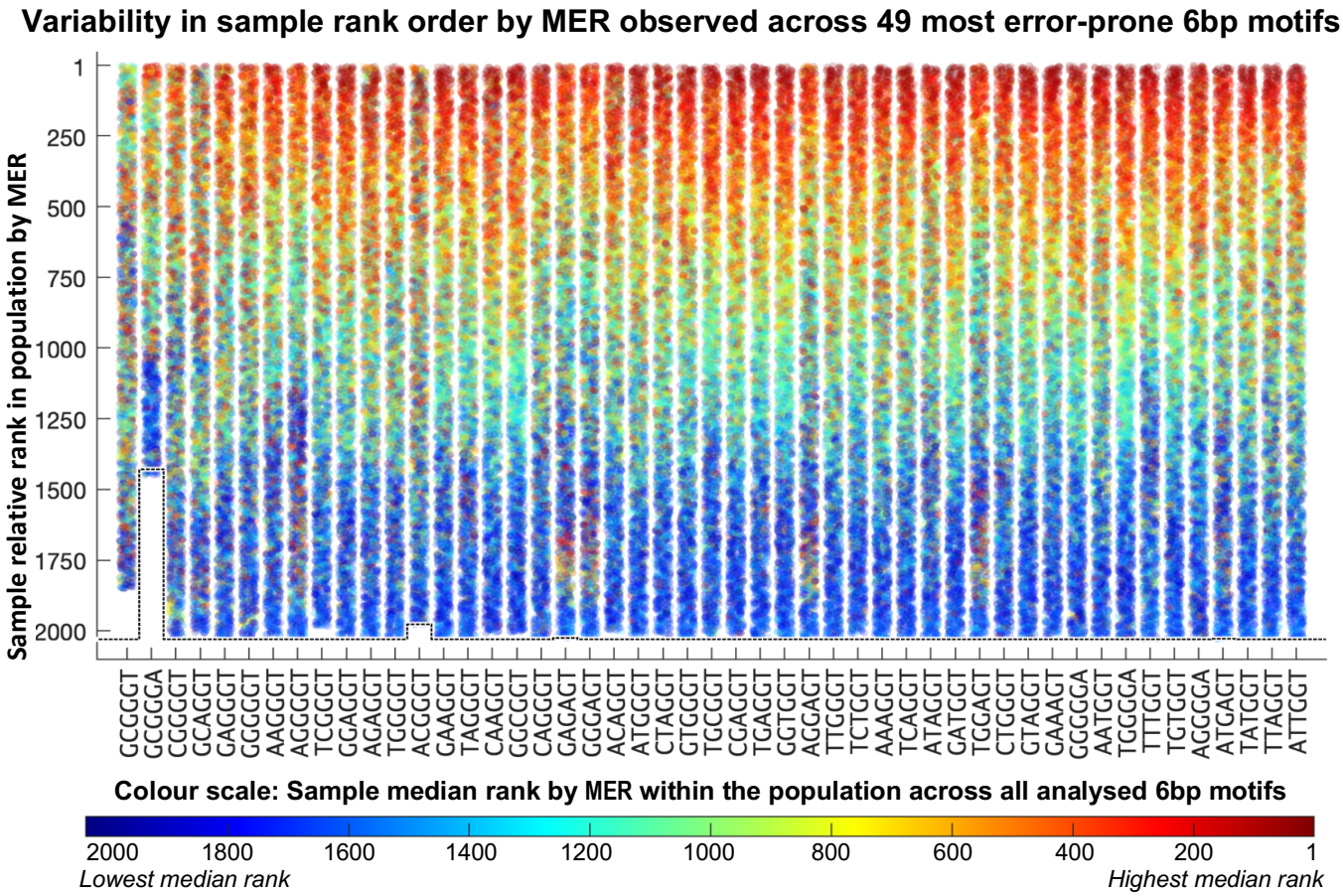


Motif set includes 49 most error prone 6bp motifs with ΣMER ≥ 1% and statistically significant MER in ≥ 90% of analysed samples. Each data point represents motif error rate observed in a specific sample, with data points from the same individual linked by lines. Colours represent individual sample median rank within the population by MER across all analysed 6bp motifs. Only data from 1959 highest coverage samples shown, samples with <50% mtDNA reaching minimum 500 reads/sequencing direction excluded

Some other important observations from **Figure 4.3.2 F1** include the fact that for several motifs the outlier individuals with highest MER levels for that particular motif are not necessarily the individuals with highest MER levels on average, as well as the observation that even though these are 49 motifs with highest ΣMER levels, for nearly every motif there are some samples in which that motif MER is in fact very low.

The differences between motifs in terms of the sample MER spread hinders a more detailed visual comparison between how the samples stratify based on their respective error rates for different motifs. To circumvent this, **figure 4.3.2 F2** presents the same data, except instead of showing the absolute MER values on Y axis, in this case the vertical axis is used to represent each sample relative rank within the population by its MER for that specific motif.

Figure 4.3.2 F2: Most error prone 6bp motifs: variation of sample relative ranking by MER at different motifs



Motif set includes 49 most error prone 6bp motifs with ΣMER ≥ 1% and statistically significant MER in ≥ 90% of analysed samples. Each data point represents the relative rank of a sample within the population based on the specific motif error rate observed in that sample. Colour represents sample median rank within the population by MER across all analysed 6bp motifs. Only data from 1959 highest coverage samples shown, with samples with <50% mtDNA reaching minimum 500 reads/sequencing direction excluded

Figure 4.3.2 F2 confirms the observation that the top handful of high ΣMER motifs are outliers in their behaviour. Specifically, their sample rank order deviates most sharply from the sample rank order averaged for all motifs. As the motif ΣMER decreases, the sample rank order becomes more consistent between motifs, however some noise remains, with a number of samples found for each motif that are ranked uncharacteristically high or low. It should be noted that the sample median MER rank is going to be affected more by the much larger number of motifs with generally low error rates compared to motifs with highest error rates, which could explain why sample rank order for high ΣMER motifs may appear more jumbled compared to lower ΣMER motifs. Thus perhaps using this type of sample average sample rank order by MER as an approximation of the overall error proneness of the particular sample may not be as straightforward.

For a more in depth discussion of the possible underlying models that can explain the differences between samples in their sequencing error rate profiles, and proposals for more suitable measures for sample sequencing error-proneness, see supplementary data and analysis in **Appendix 2.3**.

One observation of note emerging from **figure 4.3.2 F2** that was not necessarily apparent from **figure 4.3.2 F1** is the subset of five motifs, all representing pattern 'AGT' at 3' end (GAGAGT, GGGAGT, AGGAGT, TGGAGT, ATGAGT), that exhibit a noticeably different sample rank order pattern compared to the rest, with a visible cluster of samples with generally higher error rates (top 1/3 in population ranking, orange to red colour) found ranked uncharacteristically low between positions 1600-1800. This is not fully consistent with an assumption that some universal context dependent sequencing error profile is shared across all samples, with all error rate variation between samples being entirely explained by the differences in some intrinsic 'error-proneness' between samples. Instead, this could

indicate that different sequence contexts perhaps exhibit differential sensitivity to some external factors, resulting in high error rates in some but not all samples.

Identification of such candidate sequence patterns with unusual sequencing error rate levels in some but not all samples requires a way for quantitatively assessing how commonly a particular motif exhibits outlier behaviour in terms of its MER levels relative to the overall error rates of the sample. The detailed results on an example of such analyses is presented in **Appendix 2.3** where after scoring all 6bp motifs with ΣMER >0.2% by how often the motif results in an unusual error rate for the sample (**Supplementary figure 4.3.3 S9**), we clearly see an overrepresentation of 3' patterns 'AGT' and 'GGC' among the motifs with highest outlier scores. Such overrepresentation of the 'AGT' pattern among motifs more often acting as error rate outliers in samples is consistent with the visual observations made from **figure 4.3.2 F2**, lending more support to the idea that sequence context dependent error rate variability between samples may in itself also be context dependent.

4.4 Chapter discussion

The data presented in this chapter reveals complex patterns of context-dependent sequencing errors (CSE) present in the mitochondrial genome sequences, demonstrating a strong correlation between sequence composition upstream of the error prone site and the associated CSE level. The common thread among most of our reported error prone sequence contexts is very clear and can be approximated to "an error prone T or sometimes A preceded directly by several Gs". It is important to note however that this simplification does not nearly capture the level of complexity emerging from our data, with evidence of even single base changes within the upstream context as far as 7 nucleotides away from the site of interest resulting in dramatic differences in the observed error rates.

The major error prone motif patterns we have observed are largely consistent with previous literature and match well with the most commonly CSE-linked 3-4 bp contexts reported in several previous (Schirmer *et al*., 2016; Stoler *et al*., 2021). Only one study we found systematically reports CSEs in nuclear DNA for up to 8bp long sequence contexts (Allhoff *et al*., 2013) with results broadly similar in terms of the most error prone contexts to those reported here, although there are differences in the precise error prone contexts reported as well as substantial variability between reported motif error rates.

Our dataset of mtDNA sequences from over 2000 individual samples provides a near ideal system for addressing the variability of the error prone sequence contexts identified and the CSE levels measured between different individuals. The results presented in this thesis are to the best of our knowledge the first systematic attempt to quantify the context-dependent sequencing error rate variability between samples and how this differs between motifs.

Our analysis points to variable CSE profiles between different samples, meaning that the exact CSE landscape of a specific sample is unlikely to be a result of simple scaling of some generic baseline 'CSE pattern' to that individual sample's error proneness level. With evidence linking the motif sequence pattern to its likelihood of emerging as an outlier in terms of its error rate variability between samples, these findings are highly relevant and may provide an insight into the differences versus similarities that we found between the most error prone sequence contexts reported in this study compared to those previously reported by Allhoff *et al*., 2013.

The vast majority of our reported highly error prone sequence motifs fit the general 'GGT' pattern, whereas the most error prone contexts reported by Allhoff *et al*., 2013, tend to mostly fit one of two major pattern groups – the aforementioned 'GGT' as well as 'CGG' patterns, with the latter being largely absent from our results. Intriguingly, our analyses points to the 'CGG' patterns motifs exhibiting more outlier MER behaviours, suggesting that at least to some extent the disagreement between our population-level CSE rate data to the sample-level results by Allhoff *et al*. in relation to non-GGT type motifs could be related to higher CSE inter-sample variability of 'CGG' pattern sequence contexts.

Approach limitations

One of the major limitations to the CSE detection method employed in this study is presented by the small size of the mitochondrial genome. It is important to consider in how many instances across how many samples our analysed sequence motifs were found in. The fewer instances of a motif are encountered, the higher the risk that error rate measured may be affected by something other than the sequence context. Finding a motif in a large number of samples, even if that motif is only present once per sample, mitigates this to some extent as far as impacts from random chance events go, as we would not expect the same position to be affected by the same chance event across thousands of independently sequenced samples. This does not however mitigate against systematic events that for whatever reason could be introducing sequencing error at a particular position without necessarily being linked to

the immediate sequence context. If for example a specific region was for whatever reason more predisposed to experience alignment artefacts due to for example NUMTs (Maude *et al*., 2019; Nandakumar *et al*., 2021) such artefacts might manifesting as strand bias and thus get picked up by our motif error rate discovery algorithm.

This highlights the importance of carefully choosing such motif length that would provide multiple hits across the analysed sequence in each sample. In our dataset of 16.5 kb mitochondrial sequences, median instances that 8bp motifs are found in per sample is 1 with interquartile range between 1 and 1.003. For 7bp motifs the median rises to 1.4 motifs per sample on average with IQR of 1.0 to 2.8, suggesting that even 7bp might be pushing the upper limit of a motif length that can be reliably investigated in the mitochondrial context. Shortening the motif to 6bp leads to median 5.8 hits per sample with IQR of 3.1 to 9.8, which while still not a large number of instances, certainly should provide much more robust error rate data for the available genome size. Limiting motif length to for example 6bp presents a trade-off however in terms of how many upstream positions we are accounting for, which is very significant given the major effects on error rates observed with just single base changes at positions even as far upstream as -6 and -7. The likelihood of error rates being significantly modified by positions at and possibly even beyond our available resolution limit, as demonstrated in **Figures 4.2.4 F1, F2 and F3**, means that defining an exact 'error-proneness' landscape of the mitochondrial genome is not possible from this analysis alone. Furthermore, the complexity of the SCE associated sequence patterns and the CSE variability that we have observed means that based on our data alone we cannot yet fully tease apart the relationships between the sequence patterns and their associated error rates that determine which motifs and which positions will be more prone to an error than others in the particular sample.

It may well be that even the 'GGT' type patterns, by far the most common feature of our reported highly error prone sequence contexts, may not represent a single pattern, but rather might be representative of several distinct pattern groups, each with its own set of rules determining what positions are the most important and why. If one were to visualise error proneness associated with different sequence motifs as a landscape, our most error prone motifs identified would be at the top of distinct maximum error proneness peaks surrounded by most closely related sequences with most similar error rates, and separated from other peaks by valleys of sequences patterns least prone to errors. One could imagine how mapping the distance between related sequences in terms of error proneness versus sequence differences could allow us to start identifying the landscape peaks of motif clusters governed by the same mechanism, where similar base changes produce similar effect on the error rates. An approach where similar single base substitutions were scored for their effect on error rates across many motifs could potentially circumvent the resolution limit imposed in the mitochondrial context of how long a single motif can be that we can usefully interrogate given the available sequence length.

Another complementary approach for interrogating the mechanisms behind context dependent sequencing errors further would be to move away from mitochondrial sequences to the nuclear genome. Several orders of magnitude more available sequence means that even long motifs will be found often enough to ensure robust measurement of error rates. Considering the vast amounts of easily accessible very high depth whole genome and whole exome sequencing that have been made public in the recent years, including since the work presented here was carried out, makes this the next logical step in obtaining more precise measurements of the systematic error component that depends solely on the upstream sequence context. If higher resolution provided by longer nuclear sequences enabled us to draw the threshold for how far a base must be to be able to impact sequencing error rate, this would be highly informative about the underlying mechanisms and therefore relevant to any sequencing context, including in the mtDNA.

Possible role of G quadruplex structures in context dependent sequencing errors

So far we have not addressed the potential causes of the context dependent sequencing error reported in this chapter. Systematic sequencing errors associated with Illumina platform are thought to usually arise from phasing issues whereby elongation of different molecules within a cluster falls out of sync, leading to increased probability of a miscall (Fuller *et al*., 2009; Nakamura *et al*., 2011). Possible processes leading to elongation delays and causing such phasing issues could be related to transient DNA secondary structure formation capable of temporarily blocking nucleotide addition, or to ability of certain sequence patterns to potentially cause temporary arresting of polymerase though alteration of enzyme affinity (Nakamura *et al*., 2011).

One particular possible mechanism of note, which could involve both these processes is G quadruplex formation. G quadruplex (G4) is a DNA secondary structure that can form intramolecularly when 4 or more G tracts (3-7Gs) are separated by 1-7nt spacers or extra molecularly from G tracts of different molecules (Falabella *et al*., 2021). G4 motifs have been shown to affect polymerase kinetics reducing polymerisation speed both in vivo and in vitro, resulting in an increased genuine mutation rates as well as elevated error rates in sequencing experiments (Guiblet *et al*., 2018). Within G4 motifs over 5 fold increase in mismatch error rates are reported in the non-G spacer regions compared to G tracts (Weissensteiner *et al*., 2022). Some non-canonical G rich sequences such as [GGT]n repeats, have also been shown to form G4-like structures with similar kinetics to canonical G4 complexes (Guiblet *et al*., 2018).

Human mtDNA contains many G-rich sequences with predicted quadruplex forming potential, of which 134 have been experimentally validated (Bedrat *et al*., 2016). Many of our identified highly error prone sequence motifs coincide with such reported sequence regions capable of forming G quadruplexes, with the error prone base nearly always being non-G positioned immediately after a run of Gs. These observations are strongly suggestive of a potential G quadruplex role in the context dependent sequencing errors investigated in this study. The error prone base usually being non-G immediately downstream of several Gs fits the model where the most vulnerable position is within the G4 complex linker region. This is fully consistent with findings by Weissensteiner *et al*., 2022, that showed significantly higher vulnerability to sequencing errors in the G4 linkers compared to stems on Illumina platforms.

A G4 role would also fit our observations where single base changes much further upstream than perhaps expected (e.g. position -6 or -7 and beyond) appear to have a large impact on the error rate. Relatively distant positions can affect G quadruplex kinetics and differences in complex stability would affect the error probability, so it is not unreasonable that even positions much further upstream than the motifs we have investigated could have strong impact on the measured error rates if G quadruplex formation was indeed the underlying reason for increased error probability.

G quadruplex ability to form through interaction of G tracts of different DNA molecules means that a canonical full length G4 sequence may not even be required in order for several DNA molecules, each containing just a few Gs, within the same flow cell cluster to potentially assemble into a G4 structure and impact the sequencing process.

Differences in G quadruplex folding/unfolding kinetics (i.e. the speed and strength of G quadruplex formation) could also contribute to the observed inter-sample variability of motif error rates. Nominally the same motifs could seem to have different error proneness in different samples if these individuals had some sequence differences within the G4 forming sequence outside our investigated motif window. One would expect longer motifs, which we found only in very few instances per sample in our dataset, to be more affected by this, and this is consistent with our observations.

Considering the multiple lines of evidence pointing towards G-quadruplex as a very possible explanation of the types of sequencing error rats we are observing, testing this by specifically looking at sequences known to be G4 formation capable and comparing CSEs to non-G4 forming sequences would be a logical next step in this work.

Outlook for context dependent sequencing error separation from the true heteroplasmy

In this study we shall not pursue the angle of context dependent sequencing error rate variability analysis further. Rather, our primary focus is to apply the new insights in CSE patterns and dynamics described in this chapter to mitochondrial DNA heteroplasmy analysis.

Our results have presented a more complex picture of widespread yet variable context dependent sequencing error profiles within our data that we must account for when attempting to analyse mitochondrial heteroplasmy landscapes. The variable nature of error profiles detected between samples means that it is unlikely that we can distinguish non reference base frequency proportion rising from CSEs from non-reference calls due to true heteroplasmy by simply applying some 'expected' population error rate threshold at a given site. Instead, any error correction has to be applied on a sample by sample basis.

An intriguing possibility related to G quadruplex formation discussed above is that G4 structures could not only be one of a major cause of sequencing errors, but also may be a contributor to genuine mutation in the mitochondrial genome in vivo. Guiblet *et al*., 2018, demonstrate evidence that the G4 motif effect on polymerase kinetics can be observed in vivo as well, resulting in increase in genuine mutation rates. This, in combination with evidence that mitochondrial DNA heavy strand is prone to G4 complex formation, suggests that G4 structures might play a role not only in sequencing errors that we are trying to remove from our data, but also could have a role in the true heteroplasmy landscape which we are aiming to study.

This consideration means that an approach where, armed with the error profile data described in this chapter we could flag all potentially error prone positions within individual MT genomes and simply exclude these from analysis, may not be appropriate for heteroplasmy investigation. If both the probability of genuine mutation as well as that of a sequencing error is modulated by the same mechanism, such an approach of blanket exclusion of potentially CSE prone sites could easily eliminate true increased heteroplasmy sites from the analysis and thus bias any findings about the heteroplasmy landscapes found across the mitochondrial genome.

Therefore in the next chapter we propose and apply sequence data clean-up method where mitochondrial sequences are corrected for suspected context dependent errors in what we hope to be more nuanced way based on assessing whether there is evidence for any read in both sequencing directions and excluding the read fractions without bi-directional support.

5 Results 3 – Heteroplasmy analysis in a high coverage dataset

5.1 Introduction

Context-dependent sequencing error (CSE) rates analysis addressed in the previous chapter was based on the principle that any variants only supported by reads originating from one DNA strand are suspect of being sequencing artefacts introduced during the sequencing-by-synthesis process. The difference between variant allele frequencies observed in each sequencing direction is a direct measure of how error-prone a particular site or motif is.

With this in mind, the present chapter focuses on applying our insights gained about CSEs in the form of as rigorous a data filtering strategy as possible to the identification and analysis of what we hope is a close approximation of the true low level human mitochondrial heteroplasmy landscape. Using principles similar to those described in the previous chapter for assessment of the CSE patterns, here we propose a sequence data filtering approach which aims to exclude some of the CSE component from the detected heteroplasmy landscapes via an additional step of selective removal of sequence read fraction found lacking sufficient support from both sequencing directions.

The overall research aims of this chapter are two-fold. First, we present an assessment of the validity and usefulness of our proposed heteroplasmy measurement approach, followed by an application of this method with the goal of identification and analysis of low level mitochondrial heteroplasmy landscape of the human population.

In section 5.1 and 5.2 we present our proposed heteroplasmy measurement approach and compare its results with the wider heteroplasmy literature. Following this, sections 5.3 to 5.5 focus slicing the observed heteroplasmy landscapes themselves in differ ways with the aim to glean insights into the low level heteroplasmy phenomenon on the population scale.

The challenges of dissecting the population-level mitochondrial heteroplasmy landscape lie in the complexity of the system. Not only are we likely looking at highly variable heteroplasmic sample counts at each on the 16568 mtDNA positions, but also the heteroplasmies found at any given locus may vary dramatically in terms of their minor allele frequencies (MAF). To address these challenges, we approach the problem from multiple angles. In section 5.4 we look at heteroplasmy through the lens of its incidence in the population, while the focus of section 5.5 is the heteroplasmy levels themselves in terms of MAF. Finally in section 5.6 we address the links between heteroplasmy and the functional genomic context.

All analysis presented in this chapter is carried out using a different, more recent sequence dataset compared to that studied in the previous two chapters. The dataset we employ here was made publicly available in 2020 by New York Genome Center (https://www.nygenome.org/) as part of the study by Byrska-Bishop and colleagues (Byrska-Bishop *et al*., 2022) and comprises 3202 high coverage WGS sequences sequenced to a 30X target nuclear coverage using Illumina NovaSeq 6000 platform. This dataset combines sequences of the original 2504 samples forming the core of the 1000 genomes (G1K) project phase 3 release with sequences of 698 new samples of individuals related to the original G1K cohort. The considerably higher read depth compared to the original G1K phase 3 dataset analysed in the previous chapters presents a better opportunity for detecting low level heteroplasmies while applying stringent read quality filtration.

5.2 Heteroplasmy detection using consensus read frequencies

The proposed approach for minimising CSE contribution in heteroplasmy detection

A wide range of NGS data clean-up practices can be found in heteroplasmy literature (González *et al*., 2020). Most approaches rely on some form of read quality filtration, exclusion of certain patterns, such as low complexity regions (Goto *et al*., 2011) and known artefact-prone sites (Seneca *et al*., 2015; Laricchia *et al*., 2020; Liu *et al*., 2022). Some studies apply requirements for minor variants to be supported by reads in both sequencing directions (Li *et al*., 2010) or apply strand bias thresholds (González *et al*., 2020; Korolija *et al*., 2024), while others do not (He *et al*., 2010). Generally, only alleles above some minimum frequency threshold are treated as genuine variants, although what that threshold is varies widely, ranging from as low as 0.1% to over 10% depending on the study (González *et al*., 2020; Laricchia *et al*., 2022; Korolija *et al*., 2024).

Sufficiently stringent quality controls aim to eliminate as many of the artefact-derived variants from the final dataset as possible. However, among the variants passing all filters, the actual observed MAF data may still contain error components small enough to be within the tolerance limits applied, and this will likely be recorded as heteroplasmy on top of the true signal. The lower the minimum heteroplasmy thresholds one may wish to study, the more significant could the impact of such small CSE components within the MAF data become.

We propose a method which addresses the potential error component by selectively removing read fractions we identify as more likely to have arisen from sequencing errors, thus allowing for a more conservative heteroplasmy level estimation. The proposed approach is an extension of double strand verification principle but goes beyond a binary filter. Rather than interpreting the entire raw MAF as heteroplasmy, we extract only the consensus MAF component supported by reads in both sequencing directions, whereas the excess MAF component supported only by the read fraction in one sequencing direction is excluded. This differs from simply requiring variants to be supported by reads in both sequencing directions in that we require the number of supporting reads to be proportional to overall read count in each direction, and any excess reads in one direction only are disregarded for calculation of MAF. Thus only the consensus MAF component forms the basis of the heteroplasmy level estimate. Theoretically, such an approach based on only the consensus allele frequencies should in most cases reduce the impact of direction specific sequencing errors, resulting in a more conservative heteroplasmy level estimate compared to a more standard approach.

In this section we present a comparison between our proposed heteroplasmy identification method, which incorporates our additional data clean-up step of excess read removal defining heteroplasmy as only the consensus component of MAF, to a more standard approach where the full observed MAF as heteroplasmy.

Throughout the chapter, the heteroplasmy measurement approach using full MAF is referred to as ‘standard’ method whereas the approach excluding excess reads and measuring only the MAF component based on consensus read fraction supported in both sequencing directions is referred to as ‘consensus’ method.

5.2.1 Putative heteroplasmy counts at different allele filtration stringency levels

For both heteroplasmy detection methods we applied the same initial minimum read depth and indel filtration to the dataset of 3202 MT genome sequences (see **Materials & Methods 2.5**) removing any insertion or deletion sites as well as any positions without at least <500 reads in each sequencing direction, leaving 97.8% of the original dataset (**Supplementary figure 5.2.1 S1**). For consensus method we also applied removal of excess read fractions not supported in both sequencing directions, after which any additional sites with now less than 500 consensus reads per direction were excluded. This resulted in removal of further 593 data points due to a now insufficient read depth. The 500 read /direction minimum threshold was chosen to theoretically allow detection of minor reads at the frequency of 0.1% in even the lowest read depth sites.

The resulting dataset shows good level of coverage with data from ≥90% of population being analysed at 97.4% of mtDNA sites (**Supplementary figure 5.2.1 S2**), with only areas around positions 19569/1, 310 and 3107, which are known to be prone to alignment artefacts (Laricchia *et al*., 2020), experiencing reduced data coverage (**Supplementary figure 5.2.1 S3**). No samples were found with <89% of mtDNA sites analysed, with the vast majority of samples having between 96% to 99% of mtDNA site data included in the analyses (**Supplementary figure 5.2.1 S4**).

After the initial dataset filtration, we analysed the dataset identifying all sites with reads supporting more than one allele. We applied 7 filtration steps to any such identified putative heteroplasmies, intended to exclude lower confidence variants with progressively increasing stringency. The criteria used at each filtration step are summarised in **figure 5.2.1 F1**.

Figure 5.2.1 F1: Criteria used for putative heteroplasmy filtration

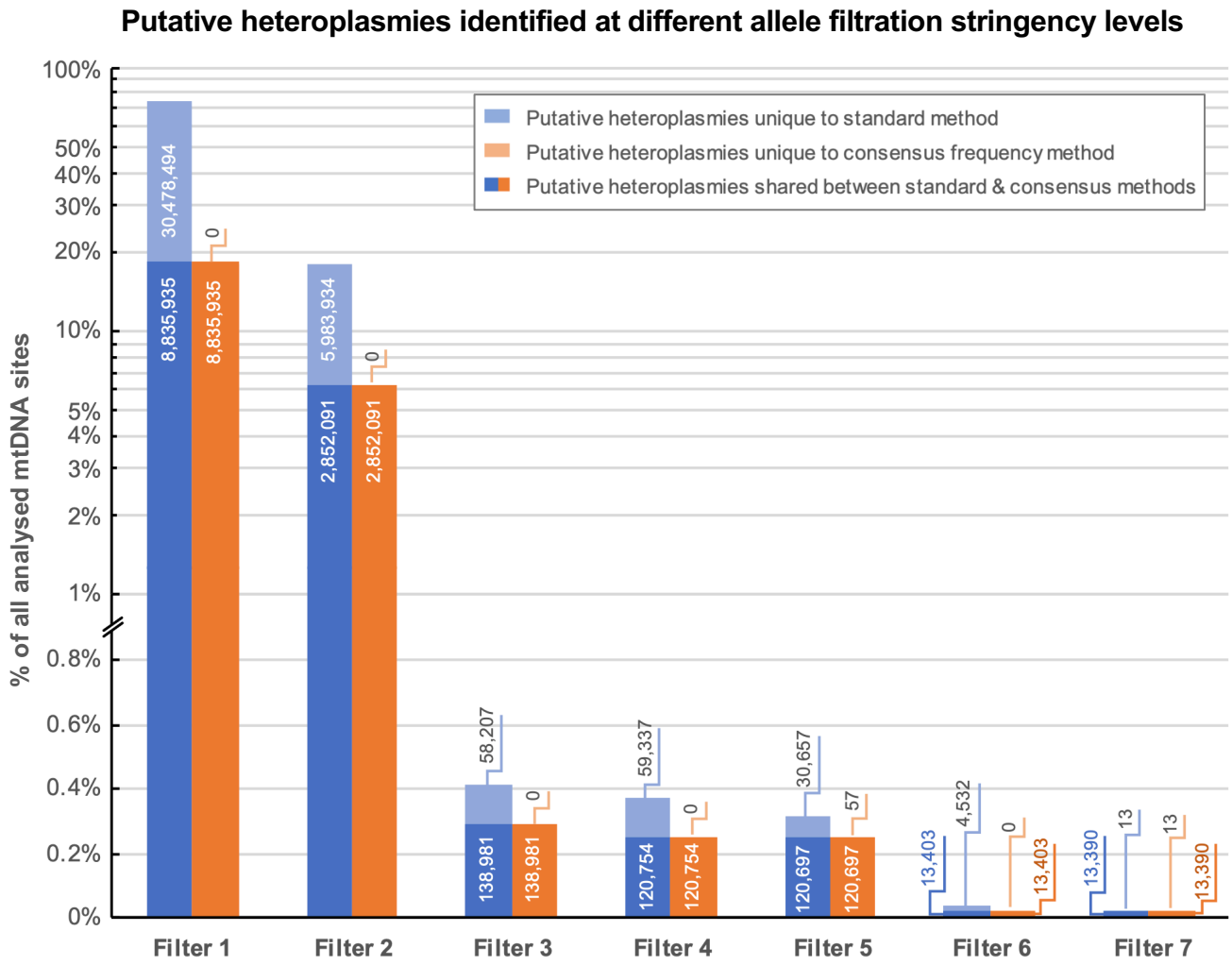
Criteria used for putative heteroplasmy filtration

Filter	Minimum allele read count threshold	Minimum allele frequency threshold
Filter 1	*1 read total*	*none*
Filter 2	*1 read in each sequencing direction*	*none*
Filter 3	*5 reads in each sequencing direction*	*none*
Filter 4	*5 reads in each sequencing direction*	*0.1% total*
Filter 5	*5 reads in each sequencing direction*	*0.1% in each sequencing direction*
Filter 6	*5 reads in each sequencing direction*	*1% total*
Filter 7	*5 reads in each sequencing direction*	*1% in each sequencing direction*

The progressively increasing filtration stringency approach was chosen in order to allow assessment of the effectiveness of the excess read fraction removal on the putative heteroplasmy detection at each stage. Thus we do not expect all putative heteroplasmies passing the lower stringency filters to be truly representative of the actual heteroplasmy landscape of the population, whereas the smaller number of variants passing the more stringent filtration steps should represent a closer approximation of the true heteroplasmy levels in our dataset. The caveat is false negatives, with the potential risk of removing true heteroplasmic loci if filtration stringency is too high. To avoid this, highest MAF minimum threshold we use is 1%. This is well within the range of MAF minimum thresholds employed in the existing heteroplasmy literature (González *et al.*, 2020).

Figure 5.2.1 F2 below presents a comparison between the number of putative heteroplasmies identified in our dataset with the standard and consensus methods that pass each allele filtration step. This figure distinguishes heteroplasmies that pass filter step with both methods (shared heteroplasmies) from those that only pass the respective filter with one of the methods (unique heteroplasmies). For quantification of unique heteroplasmy proportion and how this changes with increasing allele filtration stringency see **Supplementary figures 5.2.1 S5** and **S6**.

Figure 5.2.1 F2: Heteroplasmies identified at different allele filtration stringency levels



As expected, lower stringency filters result in higher total number of putative heteroplasmies with both methods, but more so for standard method. The large difference in the number of putative heteroplasmies found at filter 1 with and without excess read removal (17.3% vs 75.8% of all analysed sites) hints to potentially some sequencing artefact presence at over half of the analysed sites, highlighting the importance of double strand verification for heteroplasmy analyses. At the higher filtration stringency levels we see the number of putative heteroplasmies identified by both methods becoming more similar. Nevertheless, even at filter 5 where minimum 5 reads and 0.1% minimum minor allele frequency in each sequencing direction is required, 20% of putative heteroplasmies identified by the standard method are not supported by the consensus method. It is only with the minimum MAF threshold of 1% where the number of unique heteroplasmies becomes negligible, which strongly points to excess read filtration being a potentially valuable tool for attempting to analyse extremely low level heteroplasmies with MAF below the 1% threshold.

The large fraction of unique heteroplasmies identified by the standard method but not by the consensus method is consistent with the premise that using consensus reads for analysis allows better exclusion of reads more likely to be arising from sequencing errors than standard method does. Based on this, we would assume that unique putative heteroplasmies identified only by the standard method are more likely to be arising from errors, and therefore should show some correlation with locations of known error-prone motifs discussed in **chapter 4**.

Number of mtDNA positions with detected heteroplasmies

In terms of the putative heteroplasmic sample count distribution across the MT genome, addressed in **Supplementary figure 5.2.1 S7**, when no additional allele filtration is applied, putative heteroplasmic individuals are found in nearly every MT genome position. As filtration stringency increases, the number of MT genome positions with any detected putative heteroplasmies drops, with the standard method across filter levels 2-6 resulting in a slightly higher number of mtDNA positions with detected putative heteroplasmies, whereas at filter 7 both methods result in a near identical number of such sites. We observe at least 1 putative heteroplasmy with MAF>0.1% per sequencing direction at approximately 65.9% of analysed MT genome positions with the consensus method and 66.8% of analysed positions with the standard method, which decreases to approximately 35.6% with both methods when minimum MAF threshold set as 1%.

While the differences between methods in terms of the positions with any detected heteroplasmies are fairly small, the average number of putative heteroplasmies found per site differs drastically between consensus and standard methods, especially at the lower stringency filtration steps. By filter 5 on average at each analysed MT genome site method produces 9.1 putative heteroplasmies compared to 7.3 with consensus method, with this dropping to 0.8 with both methods at filter 7.

A summary of the putative heteroplasmic sample count per position central tendency and data dispersion measures, presented in **supplementary figure 5.2.1 S8**, shows that for all bar one filter level the distribution of heteroplasmy counts per position is highly skewed and a high level of variability is present across all filter levels, which is generally consistent with observations from literature, whereby heteroplasmies have been reported at large number of mitochondrial genome sites, but with few heteroplasmic individuals usually reported at any one site, with some notable exceptions such as site 16593 and some others.

Number of individuals with detected heteroplasmies

Addressing the identified putative heteroplasmy distribution between individuals, there is no difference between methods in terms of how many individuals are identified at each filtration stringency level with at least some putative heteroplasmies. At the lower filtration stringency levels putative heteroplasmies are present in all analysed samples, including up to filter 5 (MAF≥0.1% in both sequencing directions). Only with MAF threshold increasing to 1% we see samples without any detected heteroplasmies emerging, however even at filter 7 at least some putative heteroplasmies are detected in 94.9% of all samples (**Supplementary figure 5.2.1 S9**).

The mean number of putative heteroplasmies detected per sample decreases from 12,278.1 per sample (74.1% of all mtDNA sites) with standard method and 2,759.5 per sample (16.6% of mtDNA sites) with consensus method at filter level 1 to 47.7 and 37.7 per sample with standard and consensus methods respectively at filter 5, which then drops even further to just 4.2 per sample with both methods at filter level 7.

Unlike the central tendency measures for number of putative heteroplasmic individuals per site discussed earlier, the mean and median putative heteroplasmy counts detected per sample remain much more similar, with the median to mean ratio remaining between approximately 1 and 0.86 across all filter levels with both methods (**Supplementary figure 5.2.1 S10**). This indicates relatively symmetric distribution of heteroplasmy counts per sample across the population with few outlier individuals. The lack of outliers with large heteroplasmy counts is reasonably consistent with previous literature, where typically a small number of heteroplasmies are usually reported in any one individual. However, very variable heteroplasmy identification and filtration criteria employed by different authors complicate generalisations across the heteroplasmy literature.

Overall, the data presented here confirms that the more stringent variant filters that are applied to both sequencing directions result in significant reduction in the number of observed putative heteroplasmies, with the excess read exclusion employed by the consensus frequency based heteroplasmy detection method resulting in lower final heteroplasmy counts compared to the more standard approach.

As we fundamentally are interested in studying low level heteroplasmy landscapes in the human population, we ideally wish to focus on the heteroplasmies supported by a minimum of 5 consensus reads in each sequencing direction with at least 0.1% MAF in both sequencing directions (i.e. variants passing filter level 5). These filtration criteria result in a dramatic reduction of putative variants. The resulting 120,754 heteroplasmies identified by consensus method represent 0.3% of all originally identified variants prior to excess read exclusion and 1.36% of the putative variants remaining after excess read filtration.

Nevertheless, the possibility that this set of the very low MAF heteroplasmies identified at filter level 5 may still contain a significant proportion of artefacts cannot be dismissed. In the existing literature setting higher MAF thresholds is considered to increase the confidence of the detected heteroplasmies (González *et al*., 2020; Laricchia *et al.*, 2020), therefore we shall also focus on a secondary subset of presumed higher confidence heteroplasmies supported by ≥1% MAF in both sequencing directions. Treating the heteroplasmy subset passing this ten-fold higher MAF threshold as being higher confidence does not equal dismissing possibility of artefacts at either of these MAF thresholds.

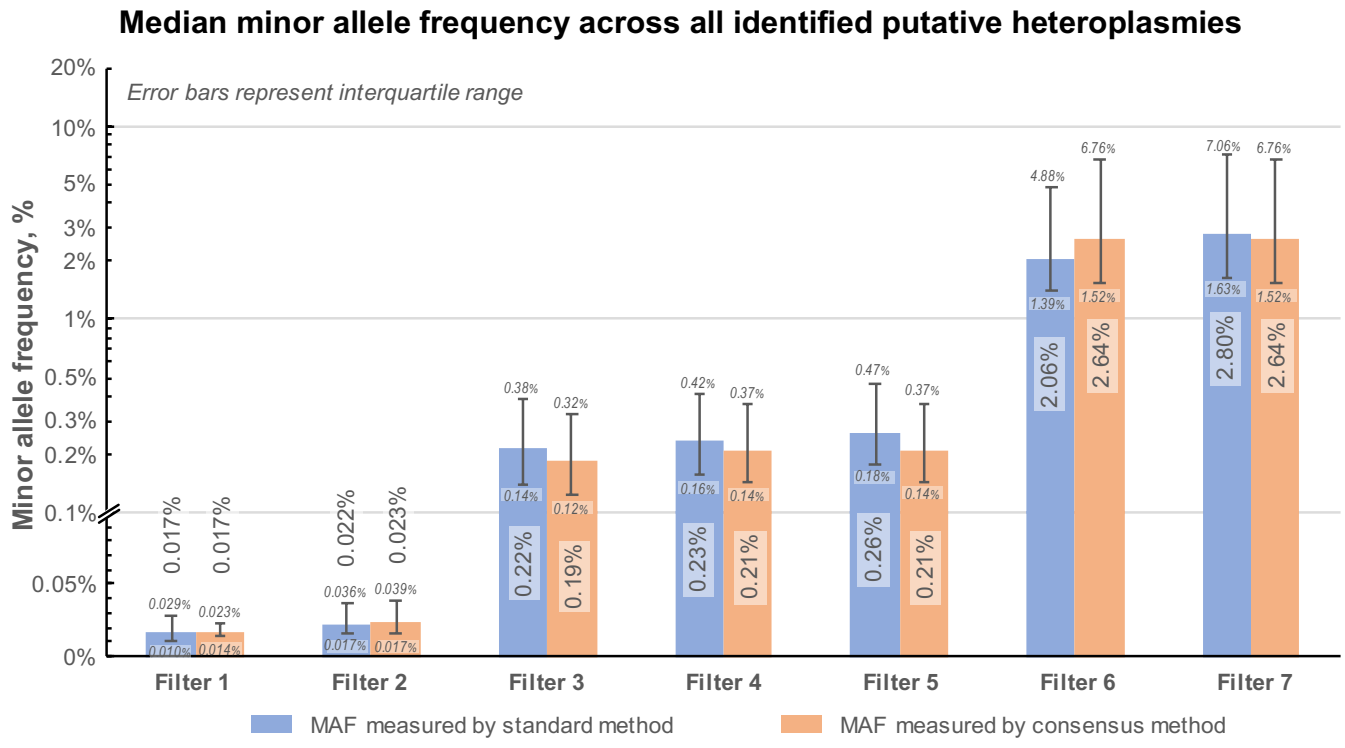
5.2.2 Putative heteroplasmy MAF at different filtration stringencies

Assuming that the excess read filtration approach used by our proposed heteroplasmy identification method does successfully eliminate some of the CSE-derived variant read fraction component, using consensus rather than raw allele frequencies should lead to lower putative heteroplasmy levels estimates. Here we address whether that is the case in our data. Unless specified otherwise, here and further we define heteroplasmy level as the total minor allele frequency found at a site, which means in cases where more than one minor allele is present, we consider the sum total frequency of all these minor alleles as the putative heteroplasmy level.

Figure 5.2.2 F1 visualises the median MAF detected per site across all analysed putative heteroplasmies found at different allele filtration stringency levels, showing that the MAF increases with increasing allele filtration stringency increases. This matches what we expect, especially at filtration steps where by definition a higher minimum MAF threshold is imposed.

What is however interesting is the comparison between the average MAF measured by the two different heteroplasmy detection methods. Even though we would expect the consensus method to result in lower MAF estimates for the same heteroplasmies due to the removal of some variant read fraction that lacks bi-directional support, at filter levels 2 and 6 the actual median MAF measured by the consensus method marginally exceeds that produced by the standard approach. This is even more noticeable when considering mean MAF (**Supplementary figure 5.2.2 S1**), where the consensus method results in a higher mean compared to the standard approach at all filter levels except filter 7. This discrepancy however can be explained by MAF being averaged across different sets of heteroplasmies with each method, as the consensus method consistently identifies fewer heteroplasmies compared to the standard method at 6 of the 7 filtration levels. This is also confirmed in **Supplementary figure 5.2.2 S2**, which shows the average MAF measured across specifically only the shared putative heteroplasmy subsets identified by both standard and consensus methods. When only shared putative heteroplasmies are considered, the consensus method does in fact result in slightly lower mean and median MAF compared to standard method. The difference is very small but consistent across all filter levels .

Figure 5.2.2 F1: Median minor allele frequency of the identified heteroplasmies



The difference between the shared putative heteroplasmy average MAF measured using standard and consensus methods does not decrease with increasing allele filtration stringency. It shows that even with stricter minimum MAF or minimum allele-supporting read count filter criteria there still remains a component of MAF that comes from reads lacking double-strand support. This further justifies the usefulness of our proposed heteroplasmy measurement approach where such reads are selectively excluded prior to MAF calculation.

Assuming that use of the consensus method allows exclusion of a higher proportion of artefact sites, it stands to reason that there should be a higher proportion of artefacts among the putative heteroplasmies uniquely identified by the standard method. Thus it is interesting to consider how the average MAF measured by the standard method differ between the subsets of heteroplasmies supported by the consensus method versus those unique to the standard method alone, as this could inform us about the background levels of errors that will manage to pass different allele filtration stringency levels unless consensus read frequency levels are used for heteroplasmy estimates.

The differences between median MAF measured by the standard method across shared versus unique heteroplasmy subsets are visualised in **Figure 5.2.2 F2** below**.**

Figure 5.2.2 F2: Median MAF of heteroplasmies identified by both methods versus standard method only

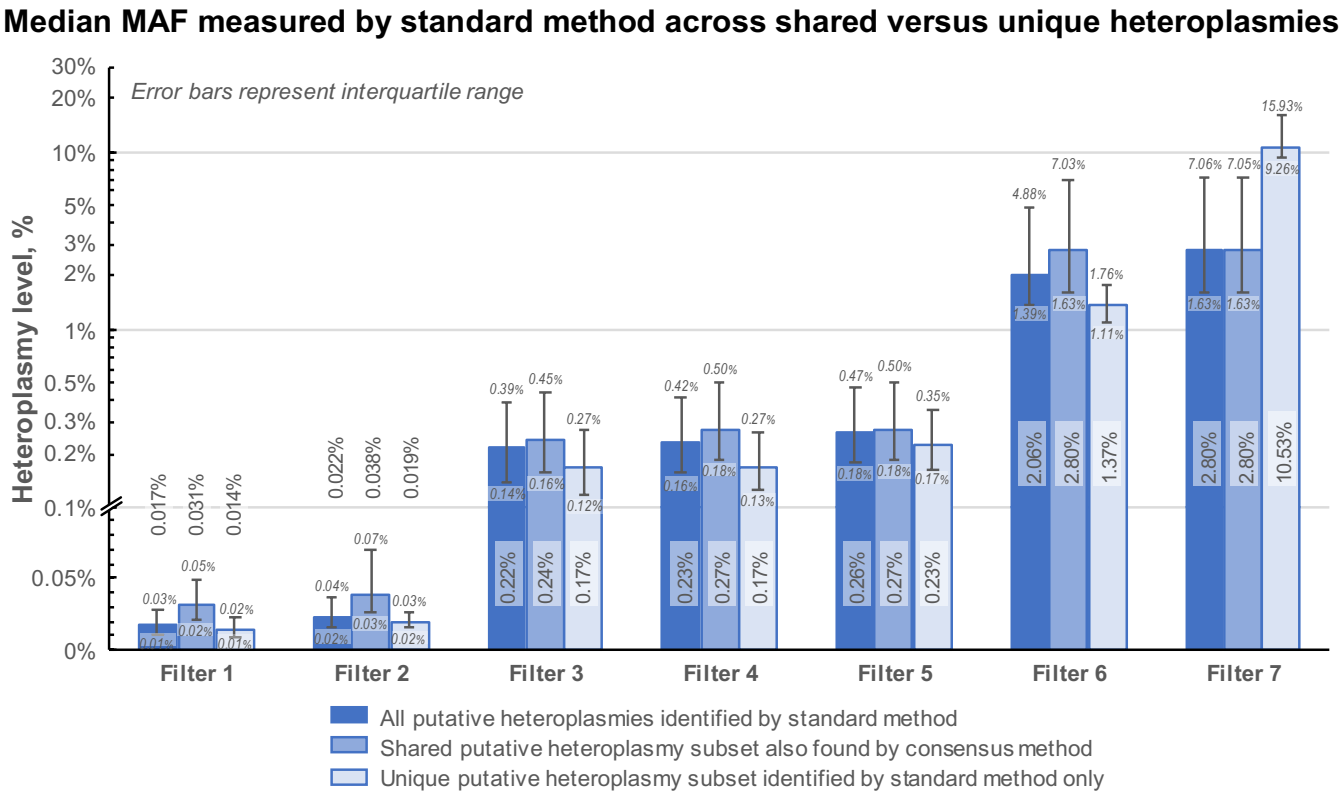


Figure 5.2.2 F2 shows that even though there is significant overlap, across filter levels 1-6 MAF of putative heteroplasmies unique to the standard method tend to be on average lower compared to heteroplasmies confirmed by both methods. Mean MAF, addressed in **Supplementary figure 5.2.2 S3**, follow the same trend.

The one notable exception is filter level 7, where the unusually high MAF observed across the 13 heteroplasmies unique to the standard method exceed MAF on average observed among heteroplasmies corroborated by the consensus method. The fact that these 13 unique putative heteroplasmies, listed in **Supplementary figure 5.2.2 S4** heteroplasmies, all pass the minimum filtration threshold of MAF>1% in both sequencing directions yet are nevertheless not corroborated by the consensus method means that the high MAF must be arising from a large number of variant reads of two different base types, one in each direction. Thus these are very likely to be sites with particularly high error rates affecting both sequencing directions, either due to alignment issues or due to

coincidence of highly error prone CSE-linked motifs on both H and L strands. When locations of the 13 unique heteroplasmies are checked, 9 do indeed fall within the low coverage genome regions. Of these, 5 are at the known problem position 310, with 2 more found adjacent to this at 309, thus lending support to the alignment artefact hypothesis.

Overall the data presented in figures 5.2.2 F1-F2 and S1-S4 confirms that not only does the consensus method result in fewer identified heteroplasmies overall compared to standard method, it also produces a slightly lower MAF estimate across the subset of heteroplasmies found by both methods. Yes, these differences are very small. They are however consistent across different allele filtration stringency levels, indicating a potential benefit of our approach for estimating heteroplasmy from only the consensus read fraction with full bi-directional support.

Comparing the two heteroplasmy identification method outcomes at the two allele filtration stringency levels that we are most relevant for our analyses of the low level mitochondrial heteroplasmy landscapes, with the minimum MAF threshold set to 0.1% (filter 5) the consensus frequency based method identifies on average 37.7 putative heteroplasmies per person with the median MAF of 0.21%, compared to 47.3 heteroplasmies per sample with median MAF 0.26% found by standard method. At filter level 7 with the minimum MAF threshold set to 1%, consensus method identified on average 4.2 heteroplasmies per sample with median MAF of 2.6%. While the standard method in this case also results in the same number of identified heteroplasmies, the detected median MAF with the standard method are higher at 2.8%, thus supporting the use of consensus method as the more conservative heteroplasmy measurement approach.

5.2.3 Putative heteroplasmy set major allele base composition

We know from our earlier context-dependent sequencing error (CSE) rate analysis that different base types can exhibit different levels of vulnerability to artefacts. Variant reads arising from such from sequencing artefacts may be misinterpreted as evidence of genuine heteroplasmies, especially at lower putative heteroplasmic variant filtration stringency levels.

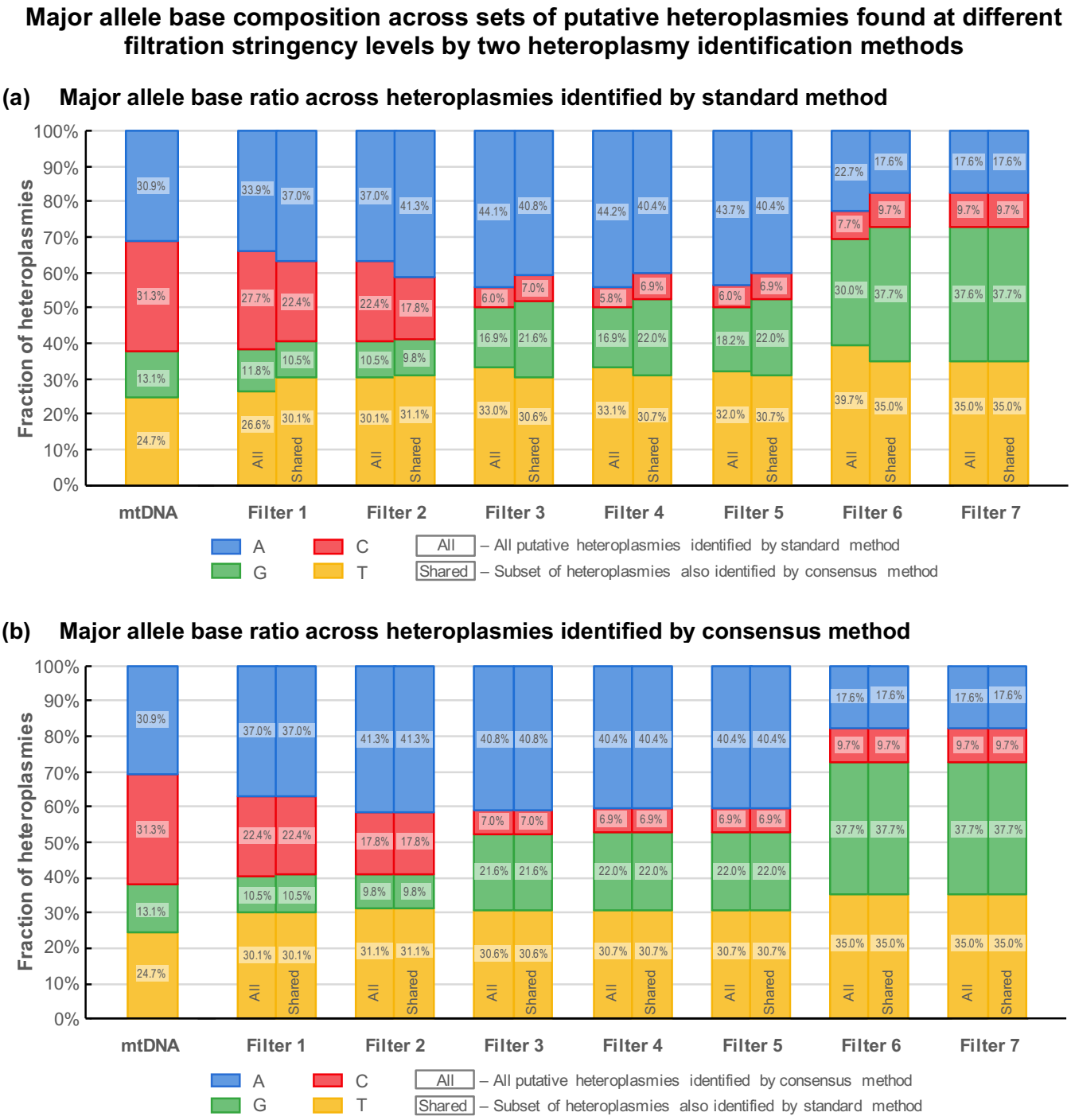
Our previous work on sequencing error profiles highlights base 'T' as by far the most vulnerable to CSE, with 'A' being the second most affected base. If we are misidentifying CSEs as heteroplasmy, we would expect variants with major alleles T and A to be overrepresented in our sets of putative heteroplasmies. Assuming that the progressively more stringent allele filtration steps we apply are reducing CSE misidentification, we should expect the T and A overrepresentation to decrease.

Genuine heteroplasmy however could also be sequence context dependent, thus also biasing our detected putative heteroplasmy major allele base composition. This complicates interpretation of whether overrepresentation of certain bases comes from CSEs or from genuine heteroplasmy simply being more common in some sequence contexts.

A potential solution lies in progressive application of increasingly more stringent allele filtration and evaluation of how the putative heteroplasmy set base composition shifts. Assuming there CSE artefacts are less common among putative heteroplasmies passing the most stringent allele filter 7 compared to filter 1, the base composition of this smallest putative heteroplasmy subset should also be most representative of any sequence context dependence of true heteroplasmy. In contrast, bases found more overrepresented among putative heteroplasmies at filter 1 would be suspect as misidentified CSEs. Thus, as long as the true heteroplasmy context dependence is not exactly the same as for CSEs, we should see informative shifts in the putative heteroplasmy base composition between filters 1 and 7.

This is explored in **figure 5.2.3 F1**, which illustrates the major allele base composition of the sets of putative heteroplasmies identified at different filtration stringency levels, and how this compares to base composition of MT genome overall. **Panel A** shows the major allele base composition of putative heteroplasmies identified by the standard method, whereas **panel B** shows putative heteroplasmies identified with our proposed consensus frequency based method. In both panels shown side by side at each filtration stringency level are the overall base composition of all heteroplasmies identified by that method as well as the base composition of only the shared heteroplasmy subset corroborated by both methods. A more detailed illustration the putative heteroplasmy major allele base composition change dynamic is provided in **Supplementary figure 5.2.3 S1**, which visualises the differences between base type frequencies among putative heteroplasmy sets compared to MT genome baseline and visualises each base type frequency change trends across all 7 filtration stringency levels separately with a breakdown of differences between total, shared and unique putative heteroplasmy sets identified by standard and consensus methods.

Figure 5.2.3 F1: Major allele base composition of heteroplasmies passing different filtration stringency levels



As there are only very few unique heteroplasmies detected by the consensus method alone, the major allele base composition of the shared heteroplasmies is functionally identical to that of all heteroplasmies found by the consensus method, confirmed by **figure 5.2.3 F1 panel B**. Thus we can largely focus just on data presented in **panel A**. There are several important observations.

Firstly, the frequency changes of different base types do not follow a linear or necessarily even a unidirectional trend with increasing allele filtration stringency. 'A' is perhaps the most extreme example of this, showing a slight initial overrepresentation at filter level 1 which increases over the next few filtration stringency increase steps, before experiencing a dramatic drop in frequency leading to being underrepresented among heteroplasmies with MAF >1% passing filter 7. The fact that the frequencies of different bases do not change in a unidirectional way could be an indicator of the observed base composition changes being a result of multiple interacting processes, such as both CSEs plus random noise plus some level of true heteroplasmy, some of which could be context-dependent.

Secondly, in terms of artefacts arising from misidentified context dependent sequencing errors if 'T' was the base most significantly affected by CSEs in this dataset and thus misinterpreted as heteroplasmy most often, then we should be seeing overrepresentation of T decreasing with increasing filtration stringency. Interestingly this is not the case, and 'T' is more overrepresented at filter 7 compared to filter 1. At the same time however, if we compare consensus and standard methods, 'T' tends to be more overrepresented among heteroplasmies identified uniquely by the standard method compared to those corroborated by the consensus method, which could be consistent with higher proportion of putative heteroplasmies at 'T' positions being CSE artefacts that are successfully removed by consensus-frequency based heteroplasmy detection method.

It is however interesting that at filter level 1 the proportion of putative heteroplasmies with major alleles A and T reported with the standard method is actually lower than that reported with the consensus method, which is opposite to what we were expecting if T and to a lesser extent A are more prone to sequencing errors. However, one possibility is that at the very low filtration levels we are misidentifying so many sites with random errors that the signal from both the genuine heteroplasmy as well as from CSE artefacts (which are by definition sequence context dependent and therefore not random) becomes swamped by the high levels of random noise. If so, we might expect the composition of putative heteroplasmies at filter 1 to be more similar to MT genome baseline, which is consistent with what we see in the data.

The differences between major allele base composition of putative heteroplasmies found passing filter levels 5 and 7 are also very interesting and point to genuine heteroplasmy likely being linked to the DNA base type. The increase in putative heteroplasmy MAF threshold from 0.1% to 1% results in a significant increase in the proportion of the remaining putative heteroplasmies with major allele G, suggesting that this base is the most likely candidate for being linked to genuine heteroplasmy. This is further supported by G frequency showing a steady increase through all of the earlier filtration steps, as well as our earlier work on sequencing error profiles, where 'G' is generally shown to be one of the bases least vulnerable to sequencing errors.

Importance of base pair orientation

Individual base types should not however be analysed in isolation, as they exist in complementary pairs in the sequence, and thus mechanisms affecting a particular base should affect all locations of that complementary base pair. It is therefore very interesting that while base 'G' is overrepresented among our identified putative heteroplasmies, the same is not true for 'C', suggesting some level of non-randomness affecting which orientation of CG base pairs is more likely to be flagged as putative heteroplasmy. In fact, we observe a very strong asymmetry affecting both CG and AT pairs, as shown by differences at which complementary bases are found among the putative heteroplasmies compared to MT genome as a whole. At filter level 7 C to G ratio among our putative heteroplasmies is just under 0.3 whereas in the MT genome overall it is approximately 2.4. Similarly, the A to T ratio of putative heteroplasmies is approximately 0.5 compared to 1.25 in the MT genome overall, indicating that for both complementary bases the heteroplasmy probability is affected by the base pair orientation.

Supplementary figure 5.2.3 S2 provides a visualisation comparing the actual observed putative heteroplasmy major allele base composition across different filtrations stringency levels to that expected if complementary bases were found in the same ratios as in the MT genome overall. It is interesting to note that across the different levels of putative heteroplasmy filtration stringency, the two complementary base pair ratios diverge from the MT genome baseline at different times. The primary C to G ratio shift away from MT genome baseline takes place at filter level 3 upon introduction of the requirement for minor allele to be supported by at least 5 reads in each sequencing direction. In contrast the largest shift of the putative heteroplasmy A to T ratio away from MT genome baseline happens at filter level 6, when the minimum MAF threshold is raised to 1%. While the significance of this is unclear,

it could perhaps be a result of the signal interpreted as putative heteroplasmy arising through different mechanisms at AT versus CG base pair sites.

Overall, the clear asymmetry between complementary base pair orientations in terms of probability of being found among our sets of identified heteroplasmies hints to the possible adjacent sequence context effect. Considering the different base composition of MT genome strands, it is not unreasonable that sequence contexts capable of modulating the chance of a site being flagged as heteroplasmy could be more common on one strand compared to the other. Consequently, we should see overrepresentation of one complementary base pair orientation among the identified putative heteroplasmies.

Are the remaining putative heteroplasmies at AT base pairs an artefact of context dependent sequencing errors?

Data presented here suggests that positions with base G and to a lesser extent T may be more likely to be heteroplasmic, with good evidence for the adjacent sequence context having a heteroplasmy modulating effect. However, a possibility needs to be addressed that even at the most stringent allele filtration the signal we are seeing is not in fact genuine heteroplasmy but rather sequencing artefacts. While our data on sequencing error rates does not raise many suspicions about heteroplasmies detected at positions with base G, the heteroplasmies detected at sites with reference base T are far more suspicious, especially where the T is found within G-rich contexts.

We explored the likelihood of our observed skewing in A/T and C/G ratios among the detected putative heteroplasmies arising entirely due to errors at GGT-rich sequence positions (**Supplementary figure 5.2.4 S1**) via a simplified model of how the combination of random heteroplasmy rates and context-dependent error rates would impact the resulting A/T ratios on putative heteroplasmies, described in detail in **Appendix 3.1**. While the model used is highly simplified and relies on multiple assumptions, it is informative about the general expected interaction and effects of two simultaneous processes, one random and one context-dependent, that both introduce putative heteroplasmies, and how the balance between the two processes should affect the A/T ratio of the resulting putative heteroplasmy set.

Our data suggests that the A/T ratios we see among our observed putative heteroplasmy sets are not consistent with high levels of systematic sequencing errors at GGT-type contexts being misinterpreted as heteroplasmy after filtration. If we were failing to remove a large proportion of context dependent errors at such motifs and were consistently classing these as heteroplasmies, we should see much higher A to T ratios among the resulting putative heteroplasmies. Instead our data is more consistent with true heteroplasmy rates also being context dependent (but not associated with the GGT-type contexts tested), with any contribution from misinterpreted CSEs at GGT-type motifs therefore likely being minor.

Data supports a link between heteroplasmy and sequence context

The data presented here supports the existence of a link between heteroplasmy and the major allele base type at the heteroplasmic site. Major allele base composition of our reported heteroplasmy sets is not consistent with being caused by the presence of misidentified sequencing errors alone, nor can it be attributed to any single mechanism, and is more suggestive of there being several underlying mechanisms interacting to produce the base frequency changes we observe across different allele filtration stringency levels.

Deviation in the complementary base frequencies from their expected ratios support there being some sequence context dependent mechanisms affecting the likelihood of observing heteroplasmy. But is this due to genuine heteroplasmy or is this CSEs being misinterpreted? We know that GGT-type patterns are highly prone to CSEs (see Chapter 4), and indeed our data is not inconsistent with some contribution from CSEs at such sequences when using

the standard method and little allele filtration. However when excess reads are excluded and more stringent allele filtration applied, the heteroplasmy A to T ratios we observe are not consistent with CSEs at GGT-type sequencing being a major contributor.

(For in depth discussion of how the GGT-type pattern asymmetric distribution between mtDNA L and H strands could lead to high 'apparent' heteroplasmy A to T ratio see Appendix 3.1)

Instead, at the most stringent allele filtration, we see strong base pair orientation bias for CG base pairs, indicating significant impact from the adjacent context affecting how likely heteroplasmy is be observed at sites where the base is C or G. In chapter 4 however we showed that Cs and Gs are far less affected by CSEs compared to As and Ts. It is unlikely that we are somehow successfully filtering out CSEs affecting As and Ts yet failing to do the same for Cs and Gs. It is far more likely that there is in fact some other context dependent mechanism besides CSEs that is contributing significantly to our observed putative heteroplasmies.

In particular it is worth noting that the heteroplasmy major allele base pair orientation bias for CG base pairs holds true across all allele filtration stringencies, with the effect increasing in magnitude as the allele filtration stringency increases. This is what we would expect, if genuine heteroplasmy was affecting CG pairs in a context dependent manner. Alternatively it could be some other mechanism besides true heteroplasmy, however it is unlikely to be purely CSEs.

5.3 Comparison with literature

5.3.1 Heteroplasmy frequency reported in literature

We reviewed 20 mitochondrial heteroplasmy studies spanning the past 30 years with available datasets of identified heteroplasmic positions with the aim to identify a list of commonly reported mitochondrial heteroplasmies. Studies reviewed are summarised in **Figure 5.3.1 F1.** The majority of studies used Illumina NGS sequencing approach either exclusively (14) or in combination with another method such as 454 sequencing system (2). Two studies used 454 system exclusively, one used Sanger sequencing, and one study measured heteroplasmy levels at a set of known mitochondrial SNPs using a genotyping microarray. Minimum minor allele frequency thresholds used by different teams for heteroplasmy detection vary from 0.5% to over 10%. Data clean-up strategies used between studies also vary, with only 12 out of the 20 studies using double strand heteroplasmic allele validation of some form.

Source DNA used is a mix of PCR amplified mtDNA (8 studies), whole genome (9 studies) and whole exome (3 studies) DNA. Source tissues used are blood in 5 studies, lymphoblastoid cell lines exclusively in 7 studies, and other tissues or mix of tissues or cell lines in the remaining 8 studies. Sample sizes vary across different studies by 6 orders of magnitude. In order to compare studies, we calculated heteroplasmy discovery rate as heteroplasmies discovered per sample per 1kb of sequence analysed. Even after correcting for sample size and amount of sequence studied, large variation in heteroplasmies discovered remains between studies. As would be expected, the studies with lower detection MAF threshold generally tend to detect more heteroplasmies, although the relationship is quite noisy.

Figure 5.3.1 F1 summarises the combined extent of reported heteroplasmy from the following 20 publications: Stoneking, 2000; Abecasis *et al*., 2010; He *et al*., 2010; Li *et al*., 2010; Goto *et al*., 2011; Sondheimer *et al*., 2011; Abecasis *et al*., 2012; Sosa *et al*., 2012; Calabrese *et al*., 2014; Diroma *et al*., 2014; Ding *et al*., 2015; Skonieczna *et al*., 2015; Liu *et al*., 2018; Duan *et al*., 2019; Wei *et al*., 2019; Bolze *et al*., 2020; Calabrese *et al*., 2022; Laricchia *et al*., 2022.

Figure 5.3.1 F1: Heteroplasmy incidence reported in literature

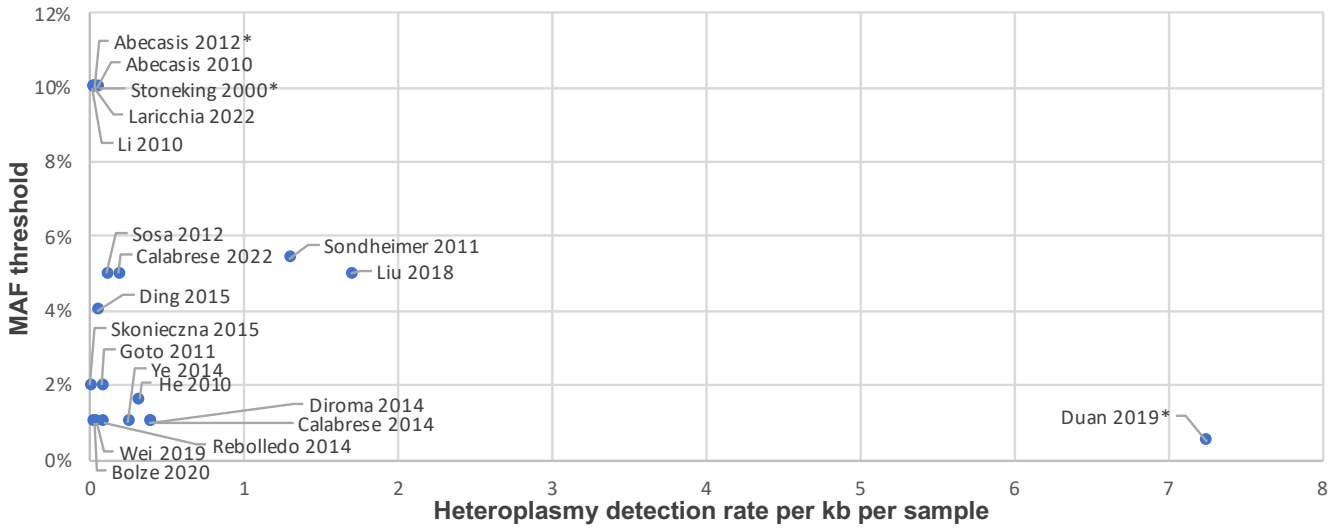
Heteroplasmy incidence reported in literature

(a) Overview of 20 published mitochondrial heteroplasmy studies

Study	Platform	Source DNA	Source tissue	MAF threshold	Sample size	Heteroplasmies		Incidence in mtDNA		Incidence in population	
						Total found	HDR* per kb	Positions	%	Samples	%
Stoneking 2000[1]	Sanger	MT[4]	B	5% - 40%[5]	1,152	56	0.051	24	2.50%	53	4.60%
He 2010‡	Illumina	MT	L,C	1.6%	20	103	0.311	70	0.40%	19	95%
Li 2010‡	Illumina	MT	B,O	10%	131	37	0.017	34	0.20%	32	24.40%
Abecasis 2010†	Illumina, 454	WG	L	10%	163	108	0.040	99	0.60%	74	45.40%
Goto 2011‡	Illumina	MT	B,O	2%	9	11	0.074	4	0.02%	7	77.80%
Sondheimer 2011	SNP array[3]	WG	B	5.40%	1,250	223	1.293	45	32.60%	210	16.80%
Abecasis 2012†‡	Illumina, 454	WG	L	unknown[6]	1,074	388	0.022	339	2.10%	299	27.80%
Sosa 2012†‡	454	MT	L	5%	40	71	0.107	68	0.40%	26	65%
Ye 2014†‡	Illumina	WG	L	1%	1,085	4,342	0.242	2,531	15.30%	973	89.70%
Rebolledo 2014‡	Illumina	MT	B,O	1%	78	111	0.086	87	0.50%	57	73.10%
Diroma 2014†	Illumina	WE	L	1%	1,242	8,012	0.389	2,223	13.40%	1017	81.90%
Calabrese 2014†	Illumina	WE	L	1%	2,419	15,478	0.386	3,519	21.20%	2101	86.90%
Ding 2015‡	Illumina	WG	B	4%	2,077	1,476	0.043	687	4.20%	878	42.30%
Skonieczna 2015‡	454	MT	O	2%	3,052	23	0.0005	23	0.10%	16	0.50%
Liu 2018	Illumina	MT	L	5%	356	9,986	1.693	2,405	14.50%	356	100%
Wei 2019‡	Illumina	WG	B	1%	3,052	1,936	0.038	1,140	6.90%	1376	45.10%
Duan 2019	Illumina	WG	L,O	0.2%, 0.5%[7]	2	240	7.243	223	1.40%	2	100%
Bolze 2020[2]	Illumina	WE	O	1%	195,983	78,258	0.024	8,413	50.80%	unknown[8]	< 40%
Laricchia 2022†‡	Illumina	WG	L,B,O	10%	56,434	22,575	0.024	5,150	31.10%	unknown[8]	< 40%
Calabrese 2022‡	Illumina	WG	B	5%	6,214	18,941	0.184	2,334	14.10%	4135	66.50%

*Source DNA: MT - PCR amplified mtDNA, WG - whole genome, WE - whole exome. Source tissues: L - lymphoblastoid cell lines, B - blood, O – other. * HDR – Heteroplasmy detection rate, expressed as number of heteroplasmies reported per sample per kilobase of sequence analysed. Colour scale for heteroplasmy detection rate / incidence values: lowest (white) to highest (red). † Reported populations have some overlap with population used in this study; ‡ Double strand heteroplasmy validation used; [1] Review; [2] Pre-print; [3] HumanHap550 & HumanHap610 SNP arrays with 138 mitochondrial SNPs; [4] PCR amplified 955bp CR region fragment only; [5] MAF estimated from chromatograms with thresholds between 5% and 40%; [6] MAF threshold not stated, median MAF stated 33%; [7] MAF thresholds 0.2% & 0.5% used for different samples; [8] Heteroplasmic sample count not stated.*

(b) Relationship between MAF threshold and the heteroplasmy detection rate across 20 studies



**Estimated 10% MAF plotted for Abecasis 2012 (threshold unknown) and Stoneking 2000 (variable thresholds reported); 0.5% MAF threshold plotted for Duan 2019 (variable thresholds reported)*

In total, 11180 positions are found to have heteroplasmy reported in at least one study, which is 67.5% of mtDNA positions. Of these, 4011 positions have been reported in only one study, which is 24.2% of all mtDNA positions and 35.9% of known heteroplasmic positions. Another 4346 (26.2% of mtDNA, 38.9% of known heteroplasmies) are reported by 2 or 3 papers, and a further 2825 (17% of mtDNA, 25.3% of known heteroplasmies) are reported by 4 or more papers. For more detailed breakdown see **figure 5.3.1 F2**.

The number of actual heteroplasmic individuals reported also varies widely. Of all reported heteroplasmic positions 2751 are only reported in a single individual, this is 16% of mtDNA. 6016 is reported in between 2 and 9 individuals, 2207 are reported in between 10 and 99 individuals and a further 206 are reported in more than 100 individuals, of which 13 are found in over 1000 individuals each.

Generally, the sites with reported heteroplasmies are distributed relatively evenly across the mitochondrial genome, with the only obvious hotspot of higher number of heteroplasmies reported in a larger number of studies roughly matching the locations of hypervariable sequence regions within the D-loop. It is also within these areas where the largest number of total heteroplasmic individuals are reported. For a visualisation of this see **Supplementary figure 5.3.1 S1**.

Figure 5.3.1 F2: Mitochondrial genome positions with previously reported heteroplasmies found in literature

Mitochondrial genome positions with heteroplasmies previously reported in literature

(a) Positions with and without known previously reported heteroplasmies

Site reported in literature as	Number of positions	Fraction of mtDNA
Heteroplasmic	11180	67.5%
Not heteroplasmic	5388	32.5%

(b) Known heteroplasmies grouped by number of studies position has been reported in

Group	Studies reported in	Number of positions	Fraction of mtDNA
Unique	1 study only	4011	24.2%
Rare	2 to 3 studies	4346	26.2%
Intermediate	4 to 5 studies	1783	10.8%
Common	6 to 9 studies	992	6.0%
Very common	10 or more studies	48	0.3%

(c) Known heteroplasmies grouped by number of individual samples position has been reported in

Group	Heteroplasmic individuals	Number of positions	Fraction of mtDNA
Unique	1 individual only	2751	16.6%
Very rare	2 to 9 individuals	6016	36.3%
Rare	10 to 99 individuals	2207	13.3%
Intermediate	100 to 999 individuals	193	1.2%
Most common	1000 or more individuals	13	0.1%

5.3.2 Most common heteroplasmic positions

The positions reported as heteroplasmic in previous literature can provide a test case for our heteroplasmy detection approaches, with the most commonly reported heteroplasmies across multiple studies representing the set of highest confidence common heteroplasmies that we should also be detecting with our heteroplasmy detection approaches. Sites reported in literature vary widely both in number of studies and both the number of samples in which heteroplasmies are reported, presenting a difficulty for choosing the best criteria for selecting the highest confidence common heteroplasmy set to test against our data. There are pros and cons for both ways of looking at what is the most common heteroplasmic positions set.

We know sequencing errors are a problem for any study, and different authors are using different methods to remove sequencing errors. Thus if heteroplasmy is reported across multiple studies, it should be reasonable to treat such sites with more confidence compared to sites reported in a high number of samples but only a single study for example. However, as the sample sizes vary widely between the reviewed studies, and so do the source populations and tissues studied, some of the smaller studies are likely not as representative of heteroplasmy frequencies found in the population overall. Therefore, if we primarily use the number of reporting papers as the criterion for identifying most common heteroplasmies, some actually quite rare heteroplasmies reported in smaller studies could appear more common than they really actually are. It is a problem that could be addressed through a statistical model, however this is beyond the scope of the study.

Looking at the total number of heteroplasmic samples reported across all studies could be more representative of the real heteroplasmy frequencies in a population, but this also has some issues. Firstly, not all populations studied consist of unrelated individuals. Several studies reviewed were either partially or entirely investigating related individuals (Goto *et al*., 2011; Sondheimer *et al*., 2011; Wei *et al*., 2012; Rebolledo-Jaramillo *et al*., 2014; Ding *et al*., 2015). Therefore frequencies of some heteroplasmies reported from these datasets would not be representative of the general population. Most of these are smaller studies however, and thus should have a smaller overall impact on the total heteroplasmy numbers compared to papers reporting heteroplasmies from larger, population scale datasets.

Secondly, not all population scale datasets are fully independent from one another, with some of the larger papers with over a 1000 samples have included partially overlapping source populations of subsets of samples from the 1000 genomes and HapMap projects (Abecasis *et al*., 2012; Calabrese *et al*., 2014; Diroma *et al*., 2014; Ye *et al*., 2014). Some level of duplicate heteroplasmies reported between these studies are therefore likely to be reported from the same individuals, and could potentially artificially inflate the overall counts of reported heteroplasmies at some of the mtDNA positions

Due to variable dataset availability and different approaches used by the studies for reporting result breakdowns, tracing which reported heteroplasmies arose from which samples was not feasible. As a result our produced aggregate dataset of heteroplasmy reports in literature does not account for such possible duplicates.

Finally, if we look only at the raw number of heteroplasmies reported across studies, then the bigger a study is, the larger its impact on heteroplasmy counts will be. This means any systematic errors or bias present in one of the very large studies has much bigger impact on total heteroplasmy counts reported and consequently what we would consider as the most common heteroplasmies in the population.

Both measures of how commonly reported the heteroplasmy is thus have pros and cons, that should be considered when interpreting the data. **Figure 5.3.2 F1** below lists the sets of most commonly reported heteroplasmies by both measures, with positions with heteroplasmies reported in ≥1000 individuals shown in **panel A** and the positions with heteroplasmies reported in ≥10 studies listed in **panel B**.

Figure 5.3.2 F1: Mitochondrial genome positions most frequently reported in literature as heteroplasmic

Mitochondrial genome positions most commonly reported in literature as heteroplasmic

(a) 13 positions with ≥ 10000 reported heteroplasmies

Position	rCRS base	Upstream context	Downstream context (RC)	Studies reported in	Individuals reported in	Het. sample fraction	Genomic context
16093	T	TATGTATT	TGTACGAA	15	6656	2.4%	CR: HV1
16183	A	CATCAAAA	GAGGGGGT	9	5617	2.0%	CR: HV1
318	T	CCCCCGCT	GGCCAGAA	5	3941	1.4%	CR: HV2
319	T	CCCCGCTT	TGGCCAGA	4	2662	1.0%	CR: HV2
152*	T	CTCATCCT	AAATAATA	11	1915	0.7%	CR: HV2
195	T	AACATACT	TTTAGTAA	14	1348	0.5%	CR: HV2
7521	G	CAAAAAGG	TCTAATAC	9	1254	0.5%	tRNA: Asp
215	A	TTAATTAA	GCATTAAT	10	1170	0.4%	CR: HV2
310*	T	CCCCCCCT	GCGGGGGA	4	1120	0.4%	CR: HV2
204	T	ACTAAAGT	TTAACACA	14	1117	0.4%	CR: HV2
323	G	GCTTCTGG	GCTGTGGC	2	1085	0.4%	CR: HV2
16189	T	AACCCCCT	CATGGGGA	6	1082	0.4%	CR: HV1
146*	T	TCCTGCCT	TAGGATGA	10	1034	0.4%	CR: HV2

(b) 48 positions with heteroplasmies reported in ≥ 10 studies

Position	rCRS base	Upstream context	Downstream context (RC)	Studies reported in	Individuals reported in	Het. sample fraction	Genomic context
16093	T	TATGTATT	TGTACGAA	15	6656	2.42%	CR: HV1
195	T	AACATACT	TTTAGTAA	14	1348	0.49%	CR: HV2
204	T	ACTAAAGT	TTAACACA	14	1117	0.41%	CR: HV2
10398	A	AGACTGAA	AATTCGGT	13	466	0.17%	protein: ND3
16311	T	TACATAGT	TTTATGTA	12	788	0.29%	CR: HV1
16129	G	ATTGTACG	ATGGTACC	12	730	0.26%	CR: HV1
189	A	CAGGCGAA	AAGTATGT	12	633	0.23%	CR: HV2
16362	T	CTTCTCGT	CATGGGGA	12	484	0.18%	CR: HV1
16256	C	TCCAAAGC	AGGGGTGG	12	481	0.18%	CR: HV1
16172	T	AACCCAAT	GATGTGGA	12	369	0.13%	CR: HV1
152*	T	CTCATCCT	AAATAATA	11	1915	0.70%	CR: HV2
16519	T	TTCAGGGT	CTTTATGA	11	699	0.25%	CR
234	A	AGGACATA	TTATTATT	11	601	0.22%	CR: HV2
16223	C	ATCAACCC	TAGTTGAG	11	480	0.17%	CR: HV1
64**	C	TTTTCGTC	CCCCCCAG	11	474	0.17%	CR: HV2
16390	G	AGATAGGG	AAGGGACC	11	395	0.14%	CR
16092	T	CTATGTAT	GTACGAAA	11	367	0.13%	CR: HV1
228	G	GCTTGTAG	TTATGTCC	11	254	0.09%	CR: HV2
14569	G	CCACACCG	TTGTTAGC	11	163	0.06%	protein: ND6 [w]
16234	C	ACTATCAC	TTGATGTG	11	160	0.06%	CR: HV1
593	T	GCTTACCT	TTTGAGGA	11	79	0.03%	tRNA: Phe
14869	G	CCTGCCTG	GGAGGATC	11	67	0.02%	protein: CYB [w]
5054	G	TTCTACCG	GGTTGTAC	11	61	0.02%	protein: ND2 [w]
16324	T	AAGCCATT	TACGGTAA	11	35	0.01%	CR: HV1
215	A	TTAATTAA	GCATTAAT	10	1170	0.43%	CR: HV2
146*	T	TCCTGCCT	TAGGATGA	10	1034	0.38%	CR: HV2
214	A	GTTAATTA	CATTAATT	10	844	0.31%	CR: HV2
5147	G	GCACCACG	GTAGGGTC	10	737	0.27%	protein: ND2 [w]
16296	C	ACCCACCC	CTGTTAAG	10	594	0.22%	CR: HV1
15043	G	ACATCGGG	GGCCTCGC	10	423	0.15%	protein: CYB [w]
16126	T	AATATTGT	GTACCGTA	10	396	0.14%	CR: HV1
5460	G	ACTCATCG	GTAAGGGC	10	380	0.14%	protein: ND2
7028	C	TTGTAGCC	GGAAGTGG	10	325	0.12%	protein: CO1 [w]
150*	C	GCCTCATC	ATAATAGG	10	299	0.11%	CR: HV2
16292	C	ACCTACCC	TAAGGGTG	10	262	0.10%	CR: HV1
194	C	GAACATAC	TTAGTAAG	10	212	0.08%	CR: HV2
16086	T	CAACCGCT	AATACATA	10	204	0.07%	CR: HV1
16291	C	AACCTACC	AAGGGTGG	10	200	0.07%	CR: HV1
16148	C	ACTTGACC	TACAGGTG	10	183	0.07%	CR: HV1
2706	A	CGGGCATA	GCTGTGTT	10	162	0.06%	rRNA: RNR2
13708	G	ACGCCTGG	CCGGCTGC	10	156	0.06%	protein: ND5
14798	T	CCACTCAT	TCGATGAA	10	130	0.05%	protein: CYB
14040	G	TTTCACAG	TTTGGTGC	10	78	0.03%	protein: ND5 [w]
2623	A	AATAGGGA	ATACAGGT	10	65	0.02%	rRNA: RNR2
16131	T	TGTACGGT	TTATGGTA	10	52	0.02%	CR: HV1
12775	G	AGAGGGCG	ATTCCTAC	10	34	0.01%	protein: ND5
8577	A	TAGGCCTA	CGGCGGGT	10	18	0.01%	protein: ATP6 [w]
988	G	CACCTGAG	TTTACAAC	10	18	0.01%	rRNA: RNR1

** Positions in regions of reduced data coverage; ** Positions in regions with data coverage; Homopolymer tracts and GGT-like highlighted; CR – control region, HV1,HV2 – Hypervariable sequence blocks; Codon 3rd wobble position in protein coding genes marked with [w]*

In total we found 13 positions with 1000 heteroplasmic individuals reported across our reviewed studies. Of these positions, 6 of these positions are also reported by 10 or more studies. Positions with most heteroplasmic samples reported is position 16093, reported in 6656 individuals samples across 15 studies. All but one of the 13 positions are located within the D-loop region, with the only exception being position 7521G found within the Aspartic acid tRNA.

Of the 13 positions with ≥1000 heteroplasmic samples reported, only one site (16183) is found within a GGT-like sequence context of the type shown earlier to be potentially associated with increased sequencing error rates. This is the second most commonly found heteroplasmy, found in 5617 individuals, across 9 studies. A further 5 positions are found close to or immediately adjacent to other homopolymer tracts of at least 4nt, a sequence characteristic often reported as problematic for sequencing and alignment in literature. These are highlighted in orange, and include position 310, a well-known problem site. This as well as two other commonly reported positions fall within the sequence region where we have reduced available data coverage.

Comparatively more positions appear reported in at least 10 studies, with a total of 48 such sites found. The majority (31 sites, 65%) fall within the D-loop region, with further 13 (27%) found in protein coding genes, 3 in rRNA and 1 in tRNA regions. Among the 13 very frequently reported heteroplasmic sites found in protein coding genes, 8 are found at the codon 3rd 'wobble' positions, while the other 5 are at codon 1st positions, with none found in 2nd position. 5 of the 48 high heteroplasmy incidence positions are found within some form of 'GGT' type contexts (highlighted in red, 2 GGGT, 3 GGT), with three other positions found near other homopolymer tracts. Four reported frequently heteroplasmic positions (64, 146, 150 and 152) are within regions of reduced available data coverage.

Overall, the relative proportion of potentially problematic positions in terms of GGT type motifs, homopolymer tracts and low sequence coverage is lower among the 48 sites reported by 10 or more studies compared to the 13 positions reported in over 1000 samples, suggesting that using number of heteroplasmy reporting studies could be a slightly more robust indicator that the site is genuinely more commonly associated with heteroplasmy.

Base composition of heteroplasmies reported in literature

It is interesting to consider what the differences are between positions previously reported as heteroplasmic versus those without known heteroplasmies. The sites with known reported heteroplasmies can be further grouped based on the heteroplasmy report frequency, with the differences between subsets potentially informative about the sequence characteristics associated with heteroplasmy. The simplest property to look at is base type at the site of heteroplasmy. We know that the heteroplasmy sets identified in our study show clear base frequency deviations from MT genome baseline frequencies, so the obvious question is whether we see similar deviations in the set of sites reported in literature.

To address this, we compare the base composition of the position subsets with and without known heteroplasmies as well as between subsets of positions grouped by heteroplasmy report frequency. **Figure 5.3.2 F1** shows the overall base composition of the set of 11180 mtDNA positions reported in literature as heteroplasmic compared to the set 5388 positions for which no heteroplasmy reports were found (**panel A**) as well as illustrates the base composition across mtDNA positions grouped by the number of studies found reporting a site as heteroplasmic (**panel B**) or the total number of heteroplasmic samples reported for that across all reviewed studies (**panel C**). **Supplementary figure 5.3.2 S2** compares the base composition deviations from mtDNA baseline observed among mtDNA position subsets with different heteroplasmy report frequencies to those observed among our identified heteroplasmies above 0.1% and 1% MAF thresholds.

Figure 5.3.2 F2: Base composition of known heteroplasmic mtDNA positions reported in previous literature

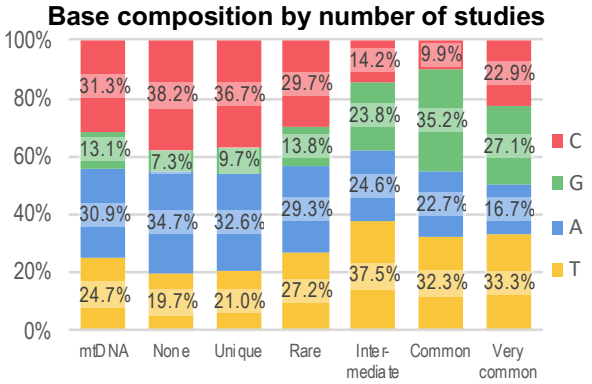
Base composition of known heteroplasmic mtDNA position groups reported in literature

(a) Overall base composition of sets of positions with and without reported heteroplasmies

Positions reported as	Positions	% mtDNA	Base type counts and relative frequencies A	C	G	T
Heteroplasmic	11180	67.5%	3252 *(29.1%)*	3122 *(27.9%)*	1775 *(15.9%)*	3031 *(27.1%)*
Not heteroplasmic	5388	32.5%	1872 *(34.7%)*	2059 *(38.2%)*	394 *(7.3%)*	1063 *(19.7%)*

(b) Known heteroplasmic positions grouped by number of studies heteroplasmy is reported in

Group	Positions *(% mtDNA)*	Base type frequencies A	C	G	T
Unique *1 study*	4011 *(24.2%)*	32.6%	36.7%	9.7%	21.0%
Rare *2 to 3 studies*	4346 *(26.2%)*	29.3%	29.7%	13.8%	27.2%
Intermediate *4 to 5 studies*	1783 *(10.8%)*	24.6%	14.2%	23.8%	37.5%
Common *6 to 9 studies*	992 *(6.0%)*	22.7%	9.9%	35.2%	32.3%
Very common *10+ studies*	48 *(0.3%)*	16.7%	22.9%	27.1%	33.3%



(c) Known heteroplasmic positions grouped by number of samples heteroplasmy is reported in

Group	Positions *(% mtDNA)*	Base type frequencies A	C	G	T
Unique *1 sample*	2751 *(16.6%)*	30.9%	34.7%	11.1%	23.4%
Very rare *2 to 9 samples*	6016 *(36.3%)*	29.7%	30.8%	14.2%	25.4%
Rate *10 to 99 samples*	2207 *(13.3%)*	25.6%	12.0%	25.7%	36.7%
Intermediate *100 to 999 samples*	193 *(1.2%)*	24.9%	27.5%	25.4%	22.3%
Very common *1000+ samples*	13 *(0.1%)*	15.4%	0%	15.4%	69.2%

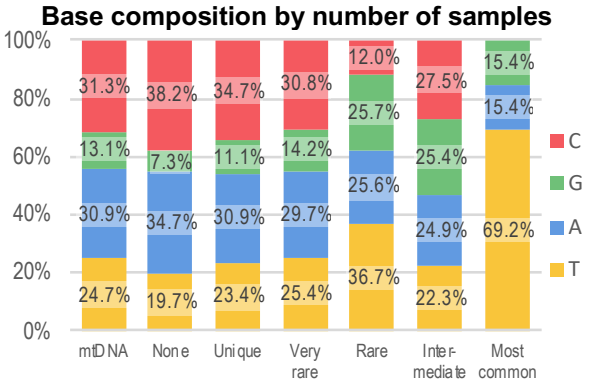


Figure 5.3.2 F2 demonstrates clear differences in the base composition between the sets of positions with and without reported heteroplasmies, with Gs and Ts relatively overrepresented among known heteroplasmic sites. When we break sites down by heteroplasmy report frequency, base composition of the sites with unique heteroplasmies is the most similar to that of sites without previous heteroplasmy reports, with clearly diverging base composition displayed by position sets with increasing heteroplasmy report frequency.

There are some unusual differences between the sets of positions with most heteroplasmy reports compared to the sets with more intermediate levels of heteroplasmy report frequency, however this could be a consequence of the most commonly reported heteroplasmic site subsets being relatively smaller and thus more vulnerable to noise. Despite some discrepancies among the smaller position subsets with most commonly reported heteroplasmies, the overall trend shows a clear link between bases G and T and increasing heteroplasmy report frequency.

This is also broadly consistent with our data, where the major allele base G and to a lesser extent T show overrepresentation among our detected heteroplasmies, especially when the 1% MAF threshold is used. There is however one notable difference, whereby major allele A is overrepresented among our reported heteroplasmies with MAF≥0.1%, yet in the literature heteroplasmies at positions where the base is A tend to be less common. This discrepancy may indicate false positives contaminating our data at this lower MAF threshold. At the same time however, none of the studies reviewed attempted to look at MAF thresholds this low, with the lowest reported thresholds being 0.2% and 0.5% each used with a single sample of a cell line by Duan *et al.*, 2019. All other studies reviewed used MAF thresholds of 1% or above. Therefore if higher prevalence at positions with base A was an intrinsic characteristic of very low level heteroplasmy, this would not necessarily appear in our reviewed literature. Considering the typically higher MAF thresholds used in the literature, comparison between previously reported heteroplasmies and our data at 1% MAF thresholds is likely more appropriate. Nevertheless the discrepancies with our data at 0.1% MAF threshold warrant caution.

It should also be noted that in the case of heteroplasmies reported in literature, rCRS sequence at the reported positions was used to determine subset base composition, and the number of heteroplasmies reported at each position were not taken into account at all. Whereas in the previous section investigating major allele base composition in our putative heteroplasmy sets, we are calculating base composition across the entire set of heteroplasmies identified, using information of each individual's major allele base type. Also sites where more individuals are flagged up as putative heteroplasmic are contributing proportionally more to the overall set base composition than sites with few heteroplasmic individuals found. Both these factors mean that any direct comparison between base composition of heteroplasmies reported in literature against those identified in this study requires some caution.

Another important point is that the thresholds used for grouping highest frequency heteroplasmies reported in literature were chosen arbitrarily, and the number of sites in the highest frequency groups both by number of samples and by number of studies are small and likely more affected by chance events. In contrast, the slightly lower frequency groups with more sites reported are probably a better reflection of the mechanisms underlying heteroplasmy.

5.3.3 Further comparison between heteroplasmies reported in this study with those reported in literature

Heteroplasmy discovery rate comparison

Literature shows a variation in heteroplasmy report frequency across different mtDNA positions. The percentage of mtDNA reported to be heteroplasmic ranges from 0.2% to over 50% between different studies, but this is very dependent on how large a sample size is used by an individual study. Generally in literature few highly heteroplasmic positions are reported in any one individual, but overall 67.5% of mtDNA positions having been reported as heteroplasmic at least once and at 43.5% of mtDNA positions heteroplasmies have been reported by 2 studies or more. Some positions clearly are more often found heteroplasmic than others, with the most extreme case being heteroplasmy at position 16093 which is reported by 75% of studies reviewed.

In comparison, in this study we report putative heteroplasmies with MAF ≥ 1% at 35.3% of mitochondrial genome positions and with MAF ≥ 0.1% at 65.2% of mtDNA (or 66.1% when using the less stringent standard method), which are not dissimilar to the overall number of mtDNA positions with some heteroplasmy reports found in literature.

The fraction of mtDNA positions reported as heteroplasmic in any one study ranges between 0.02% and 50.8% with an average of 10.6% but studies vary hugely in sample size and sequence length analysed, so we cannot compare this directly to our observed putative heteroplasmic sequence fraction. Instead we can either compare our observations with only studies of comparable size, or compare heteroplasmy discovery rates normalized for sequence length and sample size differences. **Figure 5.3.3 F1** shows both these comparisons.

Figure 5.3.3 F1: Heteroplasmic sequence fraction and discovery rate comparison between the present study and comparable previous studies

Comparison between putative heteroplasmic sequence fraction and discovery rates in this study with those found reported in comparable previous studies in literature

(a) Heteroplasmic mtDNA sequence fraction comparison

This study		Comparable* studies reported in literature				
MAF threshold	Putative heteroplasmic sequence fraction	MAF threshold	Sample size range	Heteroplasmic sequence fraction Mean	Median	Range
1% *Filter 7*	**35.3%** *of analysed positions*	**1%** *(n = 4)*	1085 - 3052 *1950 avg.*	**20.9%** *SD: 20.5%*	**14.3%** *IQR: 11.1 - 24.2%*	*min:* 4.15% *max:* 50.8%
		1% - 5% *(n = 6)*	1085 - 3052 *2155 avg.*	**16.4%** *SD: 17.9%*	**14.0%** *IQR: 6.5 - 15.1%*	*min:* 0.14% *max:* 50.8%
0.1% *Filter 5*	*Consensus method*: **65.2%** *Standard method*: **66.1%**	*No studies with comparable MAF and sample size between 1000 and 5000 individuals were reviewed*				

** Comparable studies required to have sample size between 1000 and 5000. Sample size in this study is 3202*

(b) Heteroplasmy discovery rate* comparison

This study		Comparable studies reported in literature			
MAF threshold	Putative heteroplasmy discovery rate	MAF threshold	Heteroplasmy discovery rates Mean	Median	Range
1% *Filter 7*	**0.26** *putative heteroplasmies per kb per sample*	**1%** *(n = 6)*	**0.19** *SD: 0.16*	**0.16** *IQR: 0.05 - 0.35*	*min:* 0.02 *max:* 0.39
		1% - 5% *(n = 13)*	**0.28** *SD: 0.45*	**0.11** *IQR: 0.04 - 0.31*	*min:* 0.0005 *max:* 1.69
0.1% *Filter 5*	*Consensus method:* **2.33** *Standard method:* **2.92**	**0.2%**, **0.5%** *(n = 1)*	**7.24** *heteroplasmies per kb per sample*		

** Discovery rate expressed as heteroplasmies identified per sample per kb of sequence analysed*

Figure 5.3.3 F1 panel A compares our observed heteroplasmic sequence fraction at MAF ≥ 1% to the reported average heteroplasmic sequence fraction across 2 sets of studies deemed comparable in size to this study, i.e. with sample populations between 1000 and 5000 samples. One set consists of 4 studies using MAF threshold of exactly 1% only, the other also includes 2 additional studies of similar size but with MAF thresholds up to 5%. Whilst our observed putative heteroplasmic sequence fraction of 35.3% is lower than the highest comparable reported value of 50.8%, our observed frequency of heteroplasmic positions is above the average and within the top quartile of values reported across either of the two comparable sets of studies.

Since we did not find any studies of similar size to ours that would use MAF threshold comparable to our chosen value of 0.1%, it is hard to comment on how well our observed heteroplasmic sequence fraction at this MAF threshold fits expectations. A different metric however, the heteroplasmy discovery rate per given sequence length, can be used for comparing studies regardless of the sample size. This has been done in **Figure 5.3.3 F1 panel B**, where our observed heteroplasmy discovery rate per 1kb of sequence studied is being compared all other studies which use comparable MAF thresholds to ours.

Comparing our putative heteroplasmy discovery rate of 0.26 at MAF threshold of 1% against 6 reported studies which use the same 1% threshold, our putative heteroplasmy discovery rate is within the interquartile range of the comparable reported rates, albeit slightly higher than both the mean and median reported rates. When additional studies with MAF thresholds up to 5% are included in the comparison, the median reported heteroplasmy discovery rate across 13 studies remains lower than our observed rate, but the mean across the set exceeds our observed putative heteroplasmy discovery rate. So even though theoretically increasing MAF threshold stringency should reduce the overall reported heteroplasmy discovery rate, there is very large variation between different studies and this does not always hold true.

Our observed putative heteroplasmy rate at the lower MAF threshold of 0.1% (filter 5) is 2.33 when using the consensus frequency based method and 2.92 when using the standard method for heteroplasmy identification. There is only one published study we reviewed that uses a somewhat comparably low MAF threshold (Duan *et al*., 2019) which reports heteroplasmies above slightly higher MAF thresholds of 0.2% and 0.5% from two cell line samples. Normalised heteroplasmy discovery rate reported by Duan *et al*. is 7.24, which is over nearly 2.5 times higher than our observed putative heteroplasmy discovery rate with the standard method and over 3.1 time higher than what we see using the consensus method despite our MAF threshold being lower. This is a very limited comparison however, with just a single study reporting data from just 2 samples.

It should also be noted that the data in Duan *et al*., 2019, comes from cell lines (K562 cancer cell line derived from a CML patient, and GM12878 lymphoblastoid cell line), rather than primary tissues. While immortalised cell lines are widely used in mitochondrial heteroplasmy research, it is important to consider whether the heteroplasmy patterns detected could be affected by adaptation to the in-vivo conditions, and thus perhaps not as reflective of the original source tissue or the human population in general (Kaur *et al*., 2012; Jimenez-Tellez *et al*., 2019). Nevertheless, it is an encouraging observation, which suggests our putative heteroplasmy discovery rates, at least as far as we can say from this very limited comparison with one published study, are of order of magnitude and no higher than those previously reported.

Overall, the rates at which we are detecting putative heteroplasmies at either of the two MAF thresholds appear to be generally within the same range as those reported in previous studies, and we have no reason to suspect that our most stringently filtered putative heteroplasmy sets identified at the two MAF thresholds we have used would be disproportionately more inflated with false positives compared to sets reported in literature.

Overlap between heteroplasmic positions reported in this study with those reported in literature

As our demonstrated heteroplasmy detection rate is comparable to heteroplasmy discovery rates reported in literature and we know that many of the heteroplasmic site sets reported by different studies show considerable overlaps, the obvious next question is to what extent the set of our identified heteroplasmies overlaps with the set of positions reported as heteroplasmic in literature previously.

In this study, we did not analyse all mtDNA positions for putative heteroplasmies due to 165 sites (most of these are within the D-loop) not meeting our minimum read depth criteria. Therefore here we shall specifically compare our results with previous literature reports across the 16,403 mtDNA sites that we did analyse. In the literature heteroplasmies are reported at 11,098 (67.7%) of these sites, whereas for 5,303 positions (32.3%) we found no heteroplasmy reports. In comparison, using our proposed consensus frequency based heteroplasmy measurement approach, we report heteroplasmies with MAF ≥ 0.1% at 10,804 positions (65.9% of all analysed sites), whereas our reported heteroplasmies with MAF≥1% are present at 5,842 positions (35.6% of all analysed sites).

To what extent these sets of identified heteroplasmies overlap with data from literature is summarised in **Figure 5.3.3 F2**. For an equivalent summary of the overlap between literature and the heteroplasmic positions sets identified using the standard method see **Supplementary figure 5.3.3 S1**.

We report heteroplasmies with MAF≥0.1% at 8,367 or 75.4% of positions with known previous heteroplasmy reports from literature. We also report novel heteroplasmies at 2,437 positions representing 45.9% of the MT genome sites without previous known heteroplasmy reports. At the higher MAF threshold ≥1%, we find heteroplasmies at 5,066 previously known location representing 45.6% of all previously reported heteroplasmic positions, with novel heteroplasmies found at 776 positions representing 14.6% of all sites without previous heteroplasmy reports in literature.

Generally we see that with more stringent allele filtration, the ratio between our identified novel heteroplasmies compared to heteroplasmies at previously known sites decreases. At MAF threshold ≥0.1% heteroplasmies found at novel positions represent 22.6% of all our identified heteroplasmies, whereas at MAF threshold ≥1% only 13.3% of our reported heteroplasmies are at novel sites.

When we compare how our identified heteroplasmies are distributed among sets of previously known heteroplasmic positions with different previous report frequencies, we see that the more commonly a site has been reported previously, the more likely we are to find the site heteroplasmic as well.

This is consistent with what we would expect and does not indicate any unusual divergence between our results and previous heteroplasmy literature in terms of heteroplasmy detection power or trends.

Overall, these observations are consistent with what we would generally expect from the existence of some level of a non-random heteroplasmy landscape on the population scale with heteroplasmies at some sites being more common than others. We do not observe any unusual divergence between our data and the previous heteroplasmy literature in terms of heteroplasmic position detection power or trends. This indicates that the signal we are interpreting as heteroplasmy in our data at the higher filtration stringency levels, whether it be true heteroplasmy or some combination of artefacts plus real signal, appears to largely follow the same principles as the signal reported as heteroplasmy in the published literature.

Figure 5.3.3 F2: Overlap between previously reported heteroplasmic mtDNA positions and the heteroplasmies identified in the present study

Overlap between putative heteroplasmies identified using consensus method and the sets of heteroplasmic mtDNA positions reported in literature previously

(a) mtDNA positions with and without putative heteroplasmies found using consensus method

Consensus method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Positions with one or more putative heteroplasmies	16,339 *(99.61%)*	16,210 *(98.82%)*	11,155 *(68.01%)*	10,804 *(65.87%)*	10,804 *(65.87%)*	5,842 *(35.62%)*	5,842 *(35.62%)*
Positions without any putative heteroplasmies	64 *(0.39%*	193 *(1.18%)*	5,248 *(31.99%)*	5,599 *(34.13%)*	5,599 *(34.13%)*	10,561 *(64.38%)*	10,561 *(64.38%)*

Percentages indicate fraction of the total 16,403 mtDNA positions analysed in this study

(b) Putative heteroplasmies found at known heteroplasmic positions reported in literature previously

Consensus method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Number of positions	11,063	11,018	8,569	8,367	8,367	5,066	5,066
as % of all putative heteroplasmies	67.71%	67.97%	76.82%	77.44%	77.44%	86.72%	86.72%
% of all previously reported sites	99.68%	99.28%	77.21%	75.39%	75.39%	45.65%	45.65%

(c) Putative heteroplasmies found at novel positions not previously reported in literature

Consensus method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Number of positions	5,276	5,192	2,586	2,437	2,437	776	776
as % of all putative heteroplasmies	32.29%	32.03%	23.18%	22.56%	22.56%	13.28%	13.28%
% of all previously unknown sites	99.45%	97.87%	48.75%	45.94%	45.94%	14.63%	14.63%

(d) Representation of sites with variable previous heteroplasmy report frequency in literature among positions identified as heteroplasmic using consensus method in this study

Group of sites from literature	Reported in studies	Percentage of each group found heteroplasmic in this study						
		Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Unique	1	99.55%	98.87%	60.83%	58.34%	58.34%	24.81%	24.81%
Rare	2 to 3	99.86%	99.47%	79.32%	77.42%	77.42%	43.75%	43.75%
Intermediate	4 to 5	99.61%	99.55%	96.11%	95.05%	95.05%	72.02%	72.02%
Common	6 to 9	99.59%	99.59%	98.98%	98.78%	98.78%	88.11%	88.11%
Very common	10 plus	100%	100%	97.83%	97.83%	97.83%	95.65%	95.65%

Correlation between putative heteroplasmy counts and reported heteroplasmy frequency in previous literature

The evidence presented earlier that we are seeing heteroplasmies at a higher proportion of sites where heteroplasmies have been more frequently reported in literature is encouraging. Here we test how well our observed heteroplasmy counts at any given position correlated to the number of heteroplasmies previously reported at that site. The previous heteroplasmy report frequency in literature serves as a measure for heteroplasmy-proneness of a given site that is independent from our data.

Assuming that heteroplasmy is not entirely random and thus more likely to occur in some mtDNA positions compared to others, one would expect such positions emerge as being more frequently reported across the existing heteroplasmy literature. With a sufficiently large volume of previous heteroplasmy research across a broad enough spectrum of source populations, the aggregate heteroplasmy report frequency at any given site could thus serve as a measure of heteroplasmy-proneness of that site. Importantly, such data is independent from our study, and allows us to test whether our heteroplasmy reports are replicating the same patterns. This is of course assuming we are studying a source population that is representative of the general population. If that was not the case, then we would not necessarily expect our observed heteroplasmy frequencies to match the amalgamated literature as well. For example, if we were studying a specific group of mitochondrial disease patients or a specific family. That is not the case however. The source population our data comes from (extended 1000 genomes phase 3 cohort) was

specifically designed to be representative of the general human variation. It is therefore reasonable to expect that our data should to some extent replicate the patterns of heteroplasmy incidence at different mtDNA sites reported in the wider literature. Major differences in heteroplasmy frequencies between our study and the broader literature would be alarming as this could indicate problems with our data.

Figure 5.3.3 F3 provides a summary of Pearson correlation between the heteroplasmy frequencies detected in our dataset with the frequency at which heteroplasmies are reported at the same site in literature. In addition to the correlation across all analysed mtDNA positions, **panel A** provides a breakdown by position subsets with different levels of previous report frequencies, whereas **panel B** shows the breakdown by position base type. **Figure 5.3.3 F1** focuses on heteroplasmies detected using our proposed consensus frequency based identification method only. An equivalent visualisation addressing the data produced using the standard method is shown in **Supplementary figure 5.3.3 S2**

Figure 5.3.3 F3: Correlation between heteroplasmic sample fraction found in the present study and the frequency of previous heteroplasmy reports in literature

Correlation between the number of heteroplasmies reported in previous literature and the heteroplasmic sample fraction identified in this study using consensus method

(a) Break-down by heteroplasmy report frequency in previous literature

Group of sites from literature	Reported in studies	Pearson correlation coefficient at different allele filtration stringencies						
		Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
All	**any**	0.038	0.044	0.080	0.080	0.080	0.199	0.199
Unique	1	0.050	0.019	-0.002	-0.002	-0.002	0.005	0.005
Rare	2 to 3	-0.015	0.003	0.003	0.003	0.003	-0.008	-0.008
Intermediate	4 to 5	-0.002	0.004	0.092	0.097	0.097	0.215	0.215
Common	6 to 9	0.046	0.056	0.061	0.057	0.057	0.241	0.241
Very common	10 plus	0.016	0.010	0.327	0.379	0.379	0.883	0.883

(b) Break-down by heteroplasmic position base type

Filter level	Correlation by base type				Overall correlation
	A	C	G	T	
Filter 1	0.02	0.10	0.16	0.03	0.04
Filter 2	0.02	0.15	0.25	0.04	0.04
Filter 3	0.02	0.39	0.31	0.10	0.08
Filter 4	0.02	0.44	0.33	0.10	0.08
Filter 5	0.02	0.44	0.33	0.10	0.08
Filter 6	0.31	0.67	0.29	0.22	0.20
Filter 7	0.31	0.67	0.29	0.22	0.20

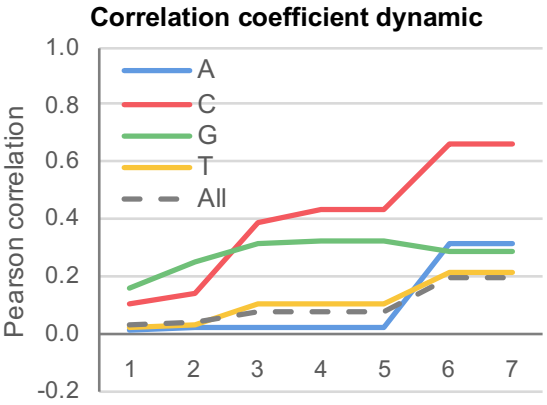


Figure 5.3.3 F3 shows relatively little correlation between our measured heteroplasmic sample fraction and the previous literature. Even though with more stringent allele filtration the correlation coefficients increase slightly, even at filter level 7 where the most stringent allele filters were applied we still see extremely low overall correlation between the heteroplasmy frequencies we observe in the population with those previously reported in literature.

Although when we split up the dataset into how commonly reported different positions are in literature, we do see a stronger correlation between our heteroplasmy counts at position reported in a larger number of papers, the only instance where the correlation is strong is across the 48 positions reported in at least 10 studies and only when looking at heteroplasmies with MAF≥1% only (r=0.88). Even among this set of most commonly reported positions, reducing MAF to 0.1% results in a significant drop in our heteroplasmy count correlation to literature (r=0.38).

Interestingly, when we break the analysed dataset down by the position base type, different bases vary considerably in how well our heteroplasmic fraction correlates to literature, with strongest correlation observed across positions with reference base C (MAF≥0.1%: r=0.44, MAF≥1%: r=0.67). At the same time, among the positions where the reference sequence base is A or T, the correlation is much lower at MAF≥1% and essentially non-existent at MAF≥0.1%. Curiously, our detected heteroplasmy fractions across positions with reference base G is marginally stronger when all heteroplasmies with MAF≥0.1% are counted compared to only the subset with MAF≥1%. Whether this difference is in any way meaningful is unclear.

In itself this data could be interpreted as suggesting that our putative heteroplasmy dataset contains a considerable level of incorrectly identified putative sites at positions of specific base types, especially A and T at the lower 0.1% MAF.

There are however two aspects that speak against this interpretation. Firstly, when we look at the putative heteroplasmy set identified by the standard method, which we expect to contain a larger proportion of artefacts, we actually see slightly higher base-type specific correlations between our heteroplasmy counts and the literature (**Supplementary figure 5.3.3 S2)**. The differences are small, but could potentially hint at the presence of artefacts in the published literature also contributing to the differences we observe.

The second line of evidence comes from pairwise comparisons between the reported heteroplasmy data by 6 largest studies with 2000+ samples and >100 heteroplasmic positions (Wei *et al.,* 2019, Bolze *et al.,* 2020, Ding *et al.,* 2015, Calabrese *et al.,* 2022, Laricchia *et al.,* 2022, Calabrese *et al.,* 2014). This was carried out to test whether our measured heteroplasmic fraction correlating better to literature for some bases than others is just an issue of our study, or whether this is true for the studies we reviewed as well, with the results, shown in **Supplementary figure 5.3.3 S3**, indicating that our study is not at all an outlier.

In fact, we see highly variable levels of correlation between the heteroplasmy counts reported by these studies, with the levels of correlation even between the same two studies being very different depending on the position base type. No one base type emerges to have a consistently stronger or weaker correlation in terms of heteroplasmic fraction reported and in no instances do we see similarly high correlation between two studies across all bases. Comparing how the heteroplasmy counts reported in each of these 6 largest studies correlate with our data, we see that our heteroplasmy counts at MAF >1% show strongest correlation with data from Calabrese *et al.,* 2014, at positions with base is A or T (r=0.68), whereas for positions where the base is C or G, our heteroplasmy counts closer match those reported by Laricchia *et al.,* 2022, (C: r=0.71, G: r=0.44). This generally fits what we see in terms of the overall lack of consistent correlation between individual studies.

We can speculate why the reported heteroplasmic fraction correlation differences between base types are present in all compared studies but do not follow the same patterns between studies. This could potentially be a result of some experimental pipeline step shared by some but not all studies biasing their resulting data in the same way? All analysed studies here use the Illumina platform and either a whole genome or a whole exome sequencing approach, but differ in what source tissue was used (e.g. saliva by Bolze *et al.,* 2020, vs blood/LCLs by others) as well as what sequence mapping and clean-up pipelines were employed and what heteroplasmy detection and filtering thresholds were set. Alternatively, could the observations be due to data from partially overlapping source populations clustering together? Source populations analysed in the compared studies are largely independent, with only minor overlap present between two studies (Calabrese *et al.,* 2014, Laricchia *et al.,* 2022). In contrast, the population used in our study does overlap substantially with Calabrese *et al.,* 2014, yet we do not see a consistent correlation with our results across all base types. This suggests different base type heteroplasmic sample count correlation patterns between these studies probably cannot be due to just non-independent sample populations clustering together.

While chance effects could play a role with smaller sample sizes, all six studies report data from large populations and large numbers of heteroplasmic positions. Thus most likely this base type correlation pattern variation is arising through some technical reason, and perhaps studies sharing some analysis pipeline step or procedure result in similar some bias towards including some signals more than others, which then appears as a stronger correlation between some parts of their results.

In summary, given the high levels of variability uncovered within the published literature in terms of the reported heteroplasmy counts at specific MT genome positions, the low positive correlation of our data with the aggregate heteroplasmy counts in literature is balanced by an otherwise good agreement between our data and published literature in terms of heteroplasmy discovery rates, base composition and general trend of observing more putative heteroplasmies at previously reported compared to novel sites.

Correlation between heteroplasmy counts above different MAF and reported heteroplasmy frequency in previous literature

In the previous section we addressed the overall correlation between our reported heteroplasmy counts with the heteroplasmy reports in existing literature. Here, we shall address how our data correlates to literature when specifically heteroplasmy counts above different MAF thresholds are considered. This is illustrated in **figure 5.3.3 F4,** focusing on the heteroplasmy dataset generated using 0.1% minimum MAF threshold.

Figure 5.3.3 F4 panel A shows how many previous studies reported heteroplasmies at different groups of positions, and **panel B** shows the average number of reported heteroplasmies per position for the different groups. Positions in these panels were grouped by number of samples (vertically) that we identified with MAF above different thresholds (horizontally). **Panel C** shows how many positions fall in each of the groups looked at. Lastly, **panel D** provides Pearson correlation coefficients describing the correlation between the number of samples found at a position above certain MAF thresholds and either the number of studies heteroplasmic position has been reported in (row labelled 'Papers') or the total number of putative heteroplasmies reported at a site across all 20 analysed studies (row labelled 'Individuals').

Figure 5.3.3 F4: Relationship between heteroplasmies found in the present study at different MAF and the frequency of previous heteroplasmy reports in literature

Relationship between number of samples identified at a position with different MAF levels and the frequency of previous heteroplasmy reports at that position in literature

(a) Average number of papers that putative heteroplasmic position has been reported in previously

*Putative heteroplasmies found by **consensus method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	2.31	2.77	3.03	3.76	4.14
3+	2.99	3.95	4.60	6.33	7.48
5+	3.45	4.89	5.75	8.60	10.2
10+	4.34	6.69	8.01	11.4	12.8
50+	5.28	9.46	14.5	15	n/a
100+	4.11	7.5	15	n/a	n/a
500+	3.29	n/a	n/a	n/a	n/a
1000+	2.81	n/a	n/a	n/a	n/a

*Putative heteroplasmies found by **standard method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	2.29	2.72	3.00	3.74	4.13
3+	2.97	3.86	4.51	6.26	7.38
5+	3.41	4.78	5.65	8.54	9.87
10+	4.30	6.30	7.39	10.4	12.5
50+	5.16	6.47	6.14	15	n/a
100+	3.93	5.25	5.25	n/a	n/a
500+	3.06	3.29	2	n/a	n/a
1000+	2.88	1.75	2	n/a	n/a

(b) Average number of individuals in which putative heteroplasmic position has been reported

*Putative heteroplasmies found by **consensus method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	13.1	17.2	19.8	28.0	36.0
3+	19.0	32.9	46.5	150.4	274.5
5+	24.7	56.3	95.5	514.5	821.6
10+	40.7	162.1	372.1	1211.3	2155.8
50+	141.0	981.5	3886.5	6656	n/a
100+	148.0	1946.3	6656	n/a	n/a
500+	23.5	n/a	n/a	n/a	n/a
1000+	10.1	n/a	n/a	n/a	n/a

*Putative heteroplasmies found by **standard method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	13.2	17.0	19.5	28.5	36.5
3+	19.6	32.8	45.4	143.0	331.5
5+	24.3	53.0	88.9	476.8	1016.2
10+	40.6	137.5	303.6	1194.3	1896.7
50+	134.6	465.8	1120.1	6656	n/a
100+	131.8	567.6	1667.8	n/a	n/a
500+	18.7	14.4	6.5	n/a	n/a
1000+	9.33	4	6.5	n/a	n/a

(c) Number of sites in each group with different sample counts above different MAF thresholds

*Putative heteroplasmies found by **consensus method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	10,804	7,380	5,843	2,824	1,808
3+	6,393	2,845	1,559	266	90
5+	4,481	1,274	543	57	20
10+	2,163	286	88	11	5
50+	203	13	2	1	0
100+	107	4	1	0	0
500+	34	0	0	0	0
1000+	16	0	0	0	0

*Putative heteroplasmies found by **standard method***

Samples with MAF above threshold	Minimum sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
1+	10,948	7,682	6,053	2,887	1,848
3+	6,547	3,130	1,692	282	95
5+	4,640	1,481	618	63	23
10+	2,289	382	116	13	6
50+	267	30	7	1	0
100+	142	16	4	0	0
500+	51	7	2	0	0
1000+	24	4	2	0	0

(d) Correlation between previous heteroplasmy report frequency in literature and the number of samples found in this study with MAF above different thresholds

*Putative heteroplasmies found by **consensus method***

Literature reports by number of	Our identified sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
Papers	0.10	0.41	0.44	0.40	0.37
Individuals	0.06	0.43	0.54	0.56	0.52

*Putative heteroplasmies found by **standard method***

Literature reports by number of	Our identified sample MAF threshold				
	≥ 0.1%	≥ 0.5%	≥ 1%	≥ 5%	≥ 10%
Papers	0.09	0.06	0.06	0.40	0.37
Individuals	0.05	0.06	0.08	0.56	0.54

When we look at all heteroplasmic positions identified at filter 5 (1+ samples, MAF ≥0.1%) average 2.3 studies report a position as heteroplasmic in on average 13.1 samples across all reporting studies. These averages increase when we look at subsets of positions with higher minimum number of heteroplasmic samples found, and the averages also go up when looking at position subsets where a certain number of samples must reach increasing MAF thresholds. Across the 1,808 putative heteroplasmic positions with at least one sample with MAF ≥ 10% we find on average 36 previous putative heteroplasmy reports across an average of 4.1 papers. For the set of 2163 positions where we found 10 or more putative heteroplasmies with MAF ≥ 0.1%, on average 40 heteroplasmic samples are reported per position between an average of 4.3 studies. But when we look at a subgroup of 88 positions where at least 10 samples with MAF ≥1% were found (consensus method), now on average 372 heteroplasmies are reported per position across an average of 8 studies.

How many positions there are in each group by minimum sample number and MAF threshold varies dramatically, with very few positions identified where large numbers of putative heteroplasmies with very high MAF are found, so the average literature report frequencies across these much smaller groups should be interpreted with caution. Nevertheless, across the larger groups of positions a clear trend emerges, with positions with more heteroplasmies with higher MAF found reported previously in literature more often.

Panel D shows Pearson correlation coefficients between number of heteroplasmy reports found in literature per site, either as a number of papers or as the total number of heteroplasmies reported, and the number of heteroplasmic samples with MAF above different thresholds identified at each position in this study. With heteroplasmies identified using the consensus method (table on the left), we see a positive correlation between literature reports and our identified putative heteroplasmies from MAF threshold of 0.5% onwards, and only at the lowest MAF threshold of 0.1% is there essentially no correlation between our identified putative heteroplasmy counts and literature report frequency. These results are true whether we look at how our sample counts above different MAF thresholds correlate to the number of heteroplasmy reporting studies, or to the number of total heteroplasmies reported across all studies, although correlation to the latter tends to be slightly higher.

The differences between standard and consensus methods results should be noted here. Unlike with the consensus method, when looking at the number of samples above different MAF thresholds identified by standard approach, we only see a reasonable level of positive correlation to literature report frequency from 5% MAF threshold onwards. With the standard method, the number of samples identified with MAF >0.1%, 0.5% and even 1% shows essentially no correlation to how often heteroplasmies at that site are reported in literature. This is interesting, and lends weight to our proposed sequencing data clean-up strategy. If literature report frequency is treated as a measure of true heteroplasmy probability, then these results support the notion that potentially erroneous excess read removal as employed by the consensus frequency based heteroplasmy identification method does indeed give a higher heteroplasmy detection accuracy at lower MAF.

5.4 The landscape of low level heteroplasmy prevalence in the population

The landscape of low level heteroplasmy prevalence in the population We originally speculated that, assuming some level of variable landscape background mutation rate landscape being intrinsic to any DNA sequence as a consequence of the equilibrium between replication errors plus the ongoing DNA damage and repair processes, we might be able to see the evidence of such a landscape in the mitochondrial genome through heteroplasmy more easily compared to the nuclear genome. This assumption is based on presumed relaxation of the selection pressures on any one mtDNA copy, as long as sufficient amount of wild-type mtDNA remains in the cell that can compensate for small proportion of copies carrying deleterious mutations that would not be able to persist in the nuclear genome with lower redundancy of only 2 copies.

Any mutation landscape of a DNA sequence represents the balance between existing mutation rates being countered by purifying selection. The exact balance would differ from position to position due to, on one hand, presumably different positions and sequence context displaying variable vulnerability to mutation, and on the other hand, variable fitness effects of mutations between different positions, with, for example, amino acid altering mutations at 1st and 2nd codon positions in protein coding genes being more strongly selected against compared to silent mutations at the codon 3rd position.

The mitochondrial genome should be no exception to these basic principles of mutation-selection balance, however considering its multi-copy nature, it is not unreasonable that the balance might be shifted, with deleterious (and therefore strongly selected against) variants being able to persist as very low level heteroplasmies longer provided that there is sufficient proportion of wild type MT genomes capable of compensating for it. If this is the case, one can envision a landscape of such very low level heteroplasmies throughout the mitochondrial genome, with their frequencies representing a snapshot of the mutation-selection balance at any given site, as postulated at the outset of this thesis.

As discussed previously, the detection of very low level heteroplasmies is bound to be hindered by sequencing artefacts, and thus heteroplasmies below MAF threshold of approximately 1% are rarely studied. In this study, using our proposed approach for mitigating the sequencing error impact through exclusion of excess read fractions unsupported by both sequencing directions, we attempt to describe and quality the mitochondrial heteroplasmy landscape of heteroplasmies as low as 0.1% MAF. The fundamental questions we address here are whether such a landscape of low-level mitochondrial heteroplasmies exists, and if so, what are its common patterns shared across the population.

Despite our best attempts at minimising the impact of sequencing errors, it is highly unlikely that our dataset is entirely devoid of such artefacts. With our chosen allele filtration criteria we are trying to balance the rate of false positives of sequencing errors against that of false negatives of missed genuine heteroplasmies. Nevertheless, it is possible that there may be no point where both the false positive and false negative rates are sufficiently low for truly reliable very low level heteroplasmy detection. This must be considered when interpreting the results, and any detected heteroplasmy landscapes must be assessed for known markers associated with errors.

5.4.1 The landscape of low level heteroplasmy frequency in the population

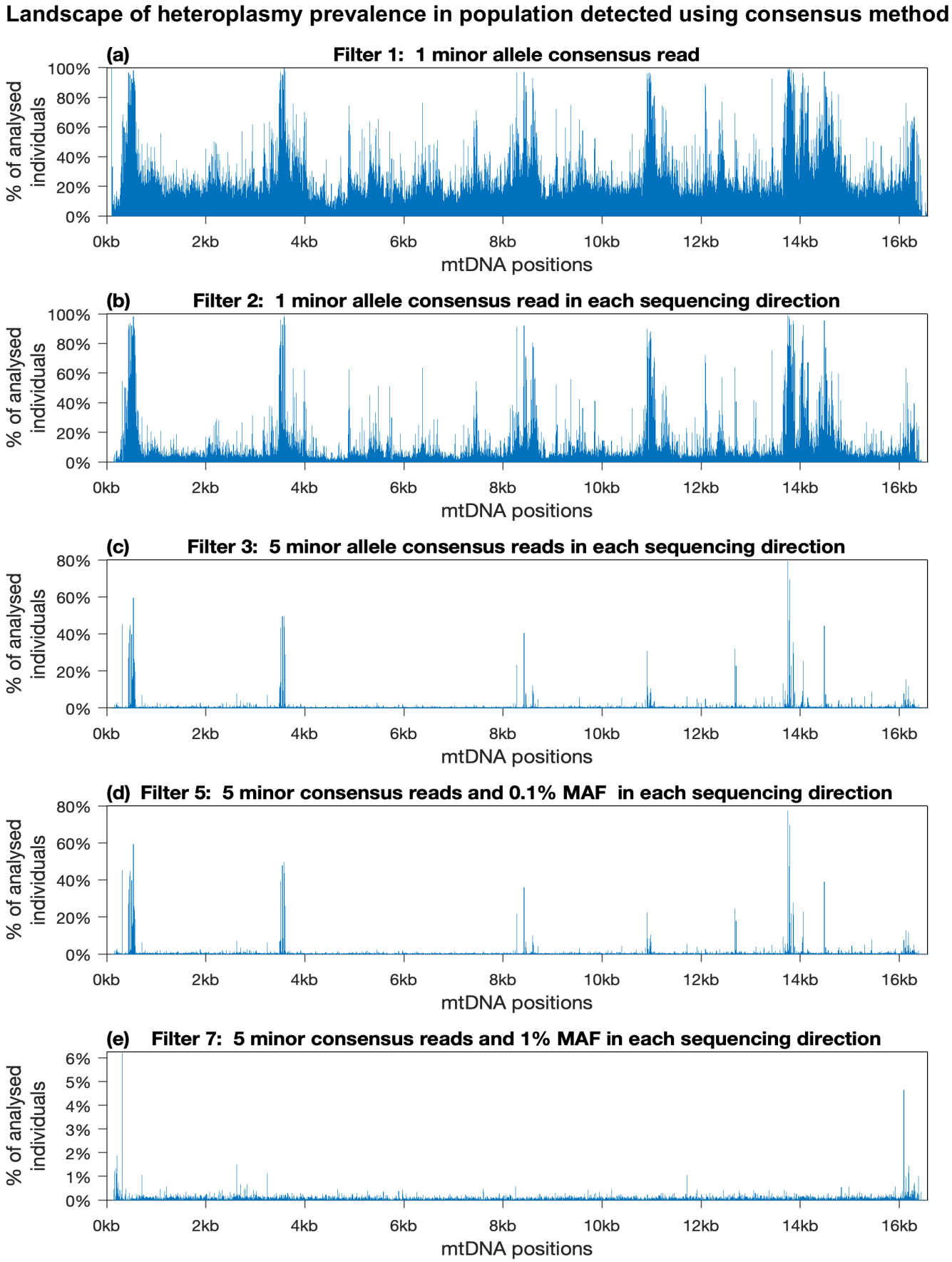
As we are interested in whether common patterns of low level mitochondrial heteroplasmy landscape exists across the population, assessing how our identified heteroplasmies are distributed along the mitochondrial genome informs us about how typical heteroplasmies are in different MT genome regions, showing any obvious heteroplasmy hotspot areas if such exist.

In this section we shall largely focus on heteroplasmies that pass minimum MAF threshold of 0.1%, as well as the smaller subset of these with MAF≥1%. We focus on heteroplasmies identified using our proposed method where MAF is estimated based on consensus reads only, with data relating to heteroplasmies identified using the more standard method provided in the Supplementary information.

Figure 5.4.1 F1 below presents the landscapes of our identified heteroplasmy prevalence landscape along the mitochondrial genome. In addition to the two heteroplasmy sets of interest for us with MAF ≥0.1% and MAF ≥1% that passed allele filtration steps 5 and 7 respectively, the figure also shows the distributions of the less stringently filtered putative heteroplasmy sets at filter levels 1, 2 and 3 to allow assessment of excluded artefact distributions.

As the number of analysed samples varies between positions, the data shown in **Figure 5.4.1 F1** is normalised, showing the relative proportion of heteroplasmic samples analysed at each MT genome position. For an equivalent figure showing the total identified heteroplasmy count at each position see **Supplementary figure 5.4.1 S1**.

Figure 5.4.1 F1: Detected heteroplasmy prevalence at each mitochondrial genome position



Heteroplasmy prevalence – proportion of analysed samples with detectable heteroplasmy at a given MT genome position

Looking at the heteroplasmy prevalence landscapes presented in **Figure 5.4.1 F1,** there are a few immediate observations. First, rather than being fairly uniform across the genome, even at filter 1 when no additional allele filtration beyond removal of excess reads has taken place, we see a clear landscape of heteroplasmy frequency peaks and dips. Comparing landscapes between different allele filtration stringency levels, we see that as stricter allele filtration takes place, the number of putative heteroplasmies detected decreases. This decrease however is not uniform across the mitochondrial genome, with some areas appearing much less sensitive to increased filtration stringency, ultimately resulting in the extreme pattern at filter 5 where the number of putative heteroplasmies being much reduced nearly everywhere except a few most extreme landscape peak areas. The patterns of the very obvious peaks of high putative heteroplasmy counts only disappear fully at filter level 7 as MAF threshold is raised from 0.1% to 1%.

If we compare this to the landscape of heteroplasmic sample fraction identified using the standard method (**Supplementary figure 5.4.1 S3**), the most obvious differences are at lower stringency filter levels, with the landscape resulting from standard method being much more saturated at filter 1. At higher filtration stringency levels, the standard method tends to result in a similar overall landscape pattern, although the absolute height of the peaks remains somewhat higher with the standard method compared to the consensus method.

The obvious question is, whether there is something special about the regions where we see these disproportionately high heteroplasmy counts, and whether these regions are somehow different from the rest of the mitochondrial genome. When we looked at the landscape of heteroplasmies reported in literature previously, no such obvious peak areas were found outside of the D-Loop region. But none of the studies reported heteroplasmies with low MAF thresholds as low as 0.1% at which we observed the peaks in question, and we know our observed peaks at MAF 0.1% disappear when MAF threshold is increased. So lack of such peaks in literature is not necessarily indicative that the signal we see is a false positives, it could be instead a feature of real low level heteroplasmy. Nevertheless, lack of similar observations from literature warrants caution.

To address the observed extremely high frequency <1% MAF heteroplasmy hotspot phenomenon we need to identify the positions at which such hotspots are found. At the minimum MAF threshold of 0.1%, across nearly 80% of the analysed MT genome sequence we detect less than 10 heteroplasmies per site, with between 10 and 100 heteroplasmies per site found across 19% of mtDNA. In total with the minimum MAF threshold of 0.1% there are only 107 sites with >100 heteroplasmies/site that constitute the extremely high frequency heteroplasmy hotspots. With the standard method there are slightly more such sites at 147.

With the minimum MAF threshold >1%, just over half of all such heteroplasmic sites are actually singletons, found heteroplasmic in only a single sample. At just over 20% of positions with heteroplasmies above 1% MAF, such heteroplasmies are found in 2 samples, with sites with more heteroplasmic samples with MAF<1% found at the same position being increasingly rare. We report only 88 positions where 10 or more heteroplasmies with MAF>1% were identified, constituting 1.5% of all positions with any heteroplasmies 1% MAF threshold detected with the consensus method.

For a more detailed break-down of the analysed sequence fractions with different identified heteroplasmy counts at 0.1% and 1% MAF thresholds using consensus or standard methods see **Supplementary figure 5.4.1 S5**.

5.4.2 Positions associated with highest heteroplasmy incidence

Positions with highest incidence of heteroplasmy with MAF≥0.1%

Using the consensus frequency based heteroplasmy detection method we identified 107 positions where heteroplasmies with MAF≥0.1% were present in 100 or more samples. These positions are listed in **Figure 5.4.2 F1**. For such positions identified using the standard method see **Supplementary figure 5.4.2 S1**.

Figure 5.4.2 F1: 107 positions with 100 or more identified heteroplasmies with MAF ≥0.1%

Positions with ≥100 putative heteroplasmies with MAF≥0.1% identified by consensus method

Position	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥0.1%	*n* with MAF ≥1%	Genomic context	
439	A	CCCCCC**A**	GTTAGT**T**	859	0	CR	
440	A	CCCCCA**A**	TGTTAG**T**	113	0	CR	
443	A	CCAACT**A**	ATGTGT**T**	422	0	CR	
446	A	ACTAAC**A**	ATAATG**T**	173	0	CR	
448	A	TAACAC**A**	AAATAA**T**	1115	0	CR	
464	A	CCTCCC**A**	TGGGAG**T**	1346	0	CR	
470	A	ACTCCC**A**	GTAGTA**T**	1430	0	CR	
471	T	CTCCCA**T**	AGTAGT**A**	163	0	CR	
503	A	CCGCCC**A**	GTAGGA**T**	1269	1	CR	
512	A	CTACCC**A**	GTGTGC**T**	486	0	CR	
517	A	CAGCAC**A**	TGTGTG**T**	238	0	CR	
519	A	GCACAC**A**	GGTGTG**T**	235	0	CR	
521	A	ACACAC**A**	GCGGTG**T**	230	0	CR	
523	A	ACACAC**A**	CAGCGG**T**	315	1	CR	
538	A	AACCCC**A**	GGGGTA**T**	1896	1	CR	
550	A	CGAACC**A**	TTTGGT**T**	313	0	CR	
554	A	CCAACC**A**	GGGGTT**T**	828	0	CR	
561	A	AACCCC**A**	TGTCTT**T**	761	0	CR	
567	A	AAAGAC**A**	GGGGGG**T**	135	0	CR	
574	A	CCCCCC**A**	AAACTG**T**	603	0	CR	
709	G	ATCCCC**G**	CTGGAA**C**	205	34	RNR1	
2623	A	ATAGGG**A**	TACAGG**T**	224	48	RNR2	
2702	G	AGGCGG**G**	GTTATG**C**	109	21	RNR2	
3244	G	GGCAGA**G**	ACCGGG**C**	199	36	TL1	
3489	A	GCCCCT**A**	GGGTTT**T**	220	0	ND1	[w]
3501	A	CGCCAC**A**	GGTAGA**T**	223	0	ND1	[w]
3508	A	TCTACC**A**	GGGTGA**T**	1267	0	ND1	
3518	A	CCCTCT**A**	GTGATG**T**	236	0	ND1	
3547	A	CTCACC**A**	GAGCGA**T**	1528	2	ND1	
3561	A	TCTACT**A**	GGTTCA**T**	272	0	ND1	[w]
3577	A	CTCCCC**A**	TGGGTA**T**	1596	0	ND1	
3578	T	TCCCCA**T**	TTGGGT**A**	1193	1	ND1	
3583	A	ATACCC**A**	GGGGGT**T**	1397	0	ND1	
3584	A	TACCCA**A**	AGGGGG**T**	321	1	ND1	
3601	A	AACCTC**A**	CTAGGT**T**	839	0	ND1	
8279	T	CCCCTC**T**	GGGGGT**A**	623	3	NC	
8426	T	ACACTA**T**	TGAGGA**A**	1152	1	ATP8	
8459	A	AACACA**A**	GGTAGT**T**	111	0	ATP8	
8463	A	CAAACT**A**	AGGTGG**T**	213	1	ATP8	
8597	T	TGATCA**T**	AATAGA**A**	320	1	ATP6	
8603	T	TTCTAT**T**	GGGGGA**A**	196	0	ATP6	
8613	A	CCCTCT**A**	GATCAA**T**	155	0	ATP6	[w]
8701	A	ATGATA**A**	GTATGG**T**	132	6	ATP6	
10398	A	GACTGA**A**	ATTCGG**T**	139	7	ND3	
10873	T	CATCCC**T**	TAGTAG**A**	108	3	ND4	[w]
10906	A	CAACCT**A**	GCTAAA**T**	723	1	ND4	[w]
10915	T	TAGCTG**T**	TGGGGA**A**	266	0	ND4	[w]
10921	A	TTCCCC**A**	AAAGGT**T**	200	0	ND4	[w]
10959	T	TCCTAA**T**	GTTAGT**A**	176	2	ND4	
10963	A	AATACT**A**	GGTAGT**T**	271	3	ND4	[w]
10967	A	CTAACT**A**	GTCAGG**T**	135	2	ND4	
10977	T	GACTCC**T**	AGGGGT**A**	334	0	ND4	
10978	A	ACTCCT**A**	GAGGGG**T**	131	2	ND4	[w]
10987	A	CCTCAC**A**	CATGAT**T**	134	0	ND4	[w]
10988	A	CTCACA**A**	CCATGA**T**	215	1	ND4	

(cont.) Position	rCRS base	Upstream context L strand	Upstream context H strand	n with MAF ≥0.1%	n with MAF ≥1%	Genomic context	
11711	G	ATAATC**G**	CGTGGG**C**	174	34	ND4	
11914	G	AACCAC**G**	GGAGAA**C**	125	14	ND4	[w]
12684	G	TAATCA**G**	GAAGAA**C**	789	0	ND5	[w]
12705	C	ACTCAT**C**	TAGGAA**G**	574	1	ND5	[w]
13105	A	GCAGGA**A**	AGAAGA**T**	122	8	ND5	
13276	A	CTCATA**A**	TAACTA**T**	121	1	ND5	
13433	T	AAACCA**T**	AGAGGT**A**	157	3	ND5	
13650	C	CTTCCC**C**	AAGGGT**G**	299	1	ND5	[w]
13681	A	AACCCC**A**	GTAGGG**T**	101	5	ND5	
13696	A	AACCCC**A**	GTTTAA**T**	179	0	ND5	
13744	A	ATTACT**A**	TGTTGT**T**	177	0	ND5	
13747	A	ACTAAC**A**	AAATGT**T**	582	0	ND5	
13750	A	AACAAC**A**	GGGAAA**T**	2355	0	ND5	
13751	T	ACAACA**T**	GGGGAA**A**	556	0	ND5	
13752	T	CAACAT**T**	GGGGGA**A**	2482	1	ND5	[w]
13753	T	AACATT**T**	CGGGGG**A**	103	0	ND5	
13776	A	CCAAAC**A**	GATTGT**T**	342	0	ND5	[w]
13777	A	CAAACA**A**	GGATTG**T**	299	0	ND5	
13779	A	AACAAC**A**	GGGGAT**T**	1515	0	ND5	[w]
13780	A	ACAACA**A**	GGGGGA**T**	841	2	ND5	
13781	T	CAACAA**T**	AGGGGG**A**	693	4	ND5	
13789	T	CCCCTC**T**	TTAGGT**A**	2227	2	ND5	
13790	A	CCCTCT**A**	TTTAGG**T**	534	3	ND5	
13794	A	CTACCT**A**	GAGTTT**T**	104	1	ND5	[w]
13803	A	ACTCAC**A**	GAGGGC**T**	140	0	ND5	[w]
13819	T	GTCACT**T**	CTAGGA**A**	693	0	ND5	
13853	A	TCAACT**A**	GTTAGG**T**	137	0	ND5	
13857	A	CTACCT**A**	GTTGGT**T**	886	0	ND5	[w]
13861	A	CTAACC**A**	GTTTGT**T**	670	0	ND5	
13864	A	ACCAAC**A**	TAAGTT**T**	122	0	ND5	
13868	T	ACAAAC**T**	ATTTTA**A**	453	0	ND5	
13887	A	CCCACT**A**	TGTGCA**T**	120	0	ND5	[w]
13894	T	TGCACA**T**	AATAAA**A**	133	2	ND5	
14046	A	GCACCA**A**	GGAGAT**T**	196	0	ND5	[w]
14047	A	CACCAA**A**	TGGAGA**T**	289	2	ND5	
14053	A	ATCTCC**A**	TGGAGG**T**	199	5	ND5	
14059	A	ACCTCC**A**	TGATGA**T**	731	0	ND5	
14487	T	CCATCA**T**	GGGGGA**A**	1246	2	ND6	
14513	A	AAAACT**A**	GTTTAA**T**	109	0	ND6	
14783	T	ATAAAA**T**	TAATTA**A**	116	1	CYB	
15043	G	CATCGG**G**	GCCTCG**C**	138	4	CYB	[w]
15301	G	CATCTT**G**	GAAGGG**C**	162	3	CYB	[w]
15443	C	GGCTTA**C**	AGAGAA**G**	250	0	CYB	
16093	T	ATGTAT**T**	GTACGA**A**	237	149	CR	
16111	C	TGCCAG**C**	ATGGTG**G**	106	6	CR	
16126	T	ATATTG**T**	TACCGT**A**	119	5	CR	
16129	G	TTGTAC**G**	TGGTAC**C**	413	31	CR	
16171	A	AACCCA**A**	TGTGGA**T**	147	0	CR	
16189	T	ACCCCC**T**	ATGGGG**A**	304	30	CR	
16192	C	CCCTCC**C**	AGCATG**G**	128	39	CR	
16223	C	TCAACC**C**	AGTTGA**G**	105	2	CR	
16293	A	CTACCC**A**	TAAGGG**T**	162	20	CR	

Nearby positions forming visually identifiable peaks shaded blue; Upstream sequence GGGT and GGGT-like motifs are highlighted;
Genomic context: control region, rRNA, tRNA, non-coding, protein; [w] – codon 3rd wobble position in protein coding genes

We can make several observations from **Figure 5.4.2 F1**. First, in terms of the genomic context, 29 of our identified 107 heteroplasmy hotspot positions fall within the D-loop region of the mitochondrial genome, with 3 found in rRNA, 1 in tRNA and 1 at a non-coding position. The majority, 73 positions are found within protein coding genes, of which 45.2% are at codon 1st position, with 20.5% found at codon 2nd position and 34.2% found at codon 3rd position. This lack of synonymous site overrepresentation among the highest heteroplasmy incidence positions makes synonymous flexibility an unlikely sole explanation for our observed heteroplasmy hotspots.

It is also interesting to note that across the 107 high heteroplasmy incidence sites the number of samples with MAF>0.1% does not appear to correlate with the number of the detected heteroplasmies with MAF>1% (Pearson's *r* = -0.14). This could suggests that how high MAF actually is may be a random variable. Alternatively it could indicate that there may be two different mechanism at play – one resulting in lower prevalence high MAF heteroplasmies and a different one causing the low MAF high prevalence peaks.

Only 12 of the heteroplasmy hotspot positions are found within some type of GGGT-like adjacent sequence contexts, suggesting that at least the systematic sequencing error rates linked to GGT-type motifs also cannot fully account for our observed heteroplasmy incidence.

The most striking observation however relates to the base composition of and around the heteroplasmy hotspot sites. This is summarised in **Figure 5.4.2 F2**.

Figure 5.4.2 F2: Base composition at and around the sites with ≥100 identified heteroplasmies with MAF ≥0.1%

Base composition of and around positions with 100 or more putative heteroplasmies identified by consensus method at filter level 5

Sequence subset	Total base pairs	Overall base composition by rCRS base type A	T	C	G	Base pair orientation distributions: AT pairs	AT orientation A_LT_H	A_HT_L	CG pairs	CG orientation C_LG_H	C_HG_L
Positions with 100+ putative heteroplasmies with MAF ≥ 0.1%	107	68 63.6%	24 22.4%	6 5.6%	9 8.4%	92 86.0%	73.9%	26.1%	15 14.0%	40.0%	60.0%
Sequence +/- 6bp from 107 positions with 100+ putative heteroplasmies	866	237 27.4%	204 23.6%	362 41.8%	63 7.3%	441 50.9%	53.7%	46.3%	425 49.1%	85.2%	14.8%
MT genome overall	all	*30.9%*	*24.7%*	*31.3%*	*13.1%*	*55.6%*	*55.6%*	*44.4%*	*44.4%*	*70.5%*	*29.5%*

Looking at the 107 sites themselves, the base composition of these positions is heavily skewed towards A and depleted in C and G. A/T base pairs account for 86% all sites with 100+ heteroplasmies with MAF≥0.1%, with strong bias towards an A_LT_H orientation. The adjacent sequence context around the heteroplasmy hotspot sites also shows unusual base composition. If we look at the 866 positions that fall within 6 base pairs from or more of the heteroplasmy hotspot sites, considerable overrepresentation of Cs is evident, with CG base pairs displaying much stronger skew towards C_LG_H orientation than found across mitochondrial genome overall.

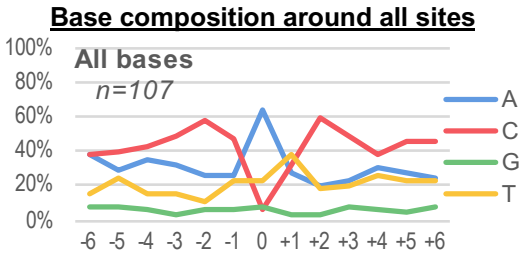
The sequence patterns around the low level heteroplasmy, addressed in **Figure 5.4.2 F3,** warrant exploration.

Figure 5.4.2 F3: Adjacent sequence contexts around 107 sites with ≥100 identified heteroplasmies with MAF ≥0.1%

Adjacent sequence context base composition around 107 positions where at least 100 putative heteroplasmies with MAF ≥ 0.1% were found by consensus method at filter 5

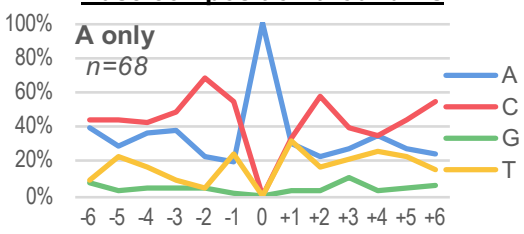
(a) Base composition around all 107 putative heteroplasmy hotspot sites regardless of base type

	L strand upstream						All	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	38%	29%	36%	33%	25%	25%	63.6%	28%	20%	23%	30%	27%	24%
C	38%	39%	43%	49%	58%	47%	5.6%	32%	59%	49%	38%	45%	46%
G	8%	7%	7%	3%	6%	6%	8.4%	3%	3%	8%	6%	5%	7%
T	15%	24%	15%	16%	11%	22%	22.4%	37%	19%	20%	26%	23%	22%



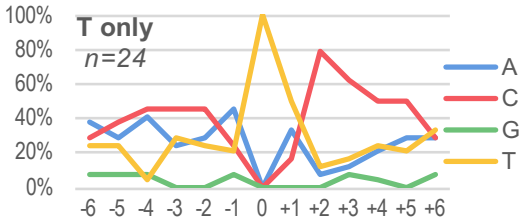
(b) As only: Base composition around 68 positions where rCRS base is A

	L strand upstream						A	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	40%	29%	37%	38%	24%	19%	100%	31%	24%	28%	35%	28%	25%
C	44%	44%	43%	49%	68%	54%	-	34%	57%	40%	35%	44%	54%
G	7%	3%	4%	4%	4%	1%	-	3%	3%	10%	3%	4%	6%
T	9%	24%	16%	9%	4%	25%	-	32%	16%	22%	26%	24%	15%



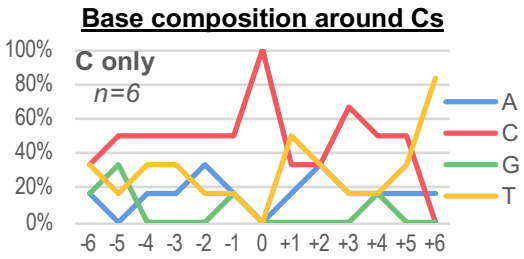
(c) Ts only: Base composition around 24 positions where rCRS base is T

	L strand upstream						T	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	38%	29%	42%	25%	29%	46%	-	33%	8%	13%	21%	29%	29%
C	29%	38%	46%	46%	46%	25%	-	17%	79%	63%	50%	50%	29%
G	8%	8%	8%	0%	0%	8%	-	0%	0%	8%	4%	0%	8%
T	25%	25%	4%	29%	25%	21%	100%	50%	13%	17%	25%	21%	33%



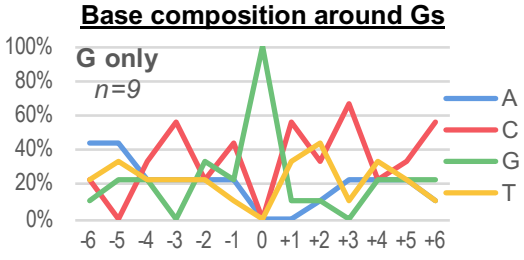
(d) Cs only: Base composition around 6 positions where rCRS base is C

	L strand upstream						C	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	17%	0%	17%	17%	33%	17%	-	17%	33%	17%	17%	17%	17%
C	33%	50%	50%	50%	50%	50%	100%	33%	33%	67%	50%	50%	0%
G	17%	33%	0%	0%	0%	17%	-	0%	0%	0%	17%	0%	0%
T	33%	17%	33%	33%	17%	17%	-	50%	33%	17%	17%	33%	83%



(e) Gs only: Base composition around 9 positions where rCRS base is G

	L strand upstream						G	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	44%	44%	22%	22%	22%	22%	-	0%	11%	22%	22%	22%	11%
C	22%	0%	33%	56%	22%	44%	-	56%	33%	67%	22%	33%	56%
G	11%	22%	22%	0%	33%	22%	100%	11%	11%	0%	22%	22%	22%
T	22%	33%	22%	22%	22%	11%	-	33%	44%	11%	33%	22%	11%



Looking at the 6bp regions upstream and downstream of each of the 107 hotspot sites, not only is the overall upstream and downstream region base composition very different from that of mtDNA L strand overall, but there is also a the lack of reverse complementarity between upstream and downstream regions, with very high frequency of Cs and low frequencies of Gs present on both sides of heteroplasmy hotspots sites. Considering our previous work on sequencing errors where upstream context on the specific strand is the main error determining factor, the striking lack of reverse complementarity between upstream and downstream patterns points strongly away from CSEs as the main explanation behind low level heteroplasmy hotspots. This observation generally is quite contrary to what one might normally expect if the mechanism behind the apparent heteroplasmy hotspots, whether that be error or genuine heteroplasmy, was directional.

Before speculating further on the source mechanisms, we should assess whether these sequence characteristics are specific to only the very low level heteroplasmies (MAF between 0.1% and 1%), or whether we see something similar when looking at positions where highest frequencies of higher level heteroplasmies (MAF>1%) are detected.

Positions with highest incidence of heteroplasmy with MAF ≥1%

As stated, using the consensus frequency based heteroplasmy identification method we find 88 positions where heteroplasmies with MAF >1% are detected in 10 or more samples, listed in **Figure 5.4.3 F4**. For sites detected using the standard method see **Supplementary figure 5.4.2 S2**

Figure 5.4.2 F4: 88 positions with 10 or more identified heteroplasmies with MAF ≥1%

Positions with ≥10 heteroplasmies with MAF ≥1% identified by consensus method

Position	rCRS base	Upstream context L strand	Upstream context H strand	n with MAF ≥0.1%	n with MAF ≥1%	Genomic (context.)
152	T	TCATCC**T**	AATAAT**A**	13	10	CR
189	A	AGGCGA**A**	AGTATG**T**	64	42	CR
195	T	ACATAC**T**	TTAGTA**A**	53	35	CR
200	A	CTTACT**A**	ACACTT**T**	19	12	CR
204	T	CTAAAG**T**	TAACAC**A**	98	60	CR
214	A	TTAATT**A**	ATTAAT**T**	17	11	CR
215	A	TAATTA**A**	CATTAA**T**	33	16	CR
234	A	GGACAT**A**	TATTAT**T**	41	24	CR
251	G	TTGAAT**G**	TGCAGA**C**	25	14	CR
279	T	CAGACA**T**	GTTATG**A**	26	13	CR
385	A	AGCCTA**A**	ATCTGG**T**	40	15	CR
709	G	ATCCCC**G**	CTGGAA**C**	205	34	RNR1
1082	A	TTAGAT**A**	GTGGGG**T**	67	15	RNR1
1169	G	CTCAAA**G**	CAGGTC**C**	43	12	RNR1
1201	A	TAGAGG**A**	ACAGGC**T**	53	18	RNR1
1435	A	CAGTAA**A**	TCTTAG**T**	19	10	RNR1
1598	G	TTGGAC**G**	CTGGTT**C**	46	11	RNR1
2075	T	TCTAAA**T**	AAGGGG**A**	19	11	RNR2
2174	G	TAGTAG**G**	TTTAGG**C**	23	10	RNR2
2275	T	GACCAA**T**	TGATAG**A**	28	12	RNR2
2333	G	TCCTCC**G**	CTTATG**C**	74	14	RNR2
2492	G	TACCCC**G**	AACAGG**C**	51	11	RNR2
2623	A	ATAGGG**A**	TACAGG**T**	224	48	RNR2
2702	G	AGGCGG**G**	GTTATG**C**	109	21	RNR2
2778	T	CACAGG**T**	TTTAGG**A**	42	15	RNR2
2805	A	TTAAAA**A**	CCGAAA**T**	42	15	RNR2
2833	A	AGCAGA**A**	GTTGGG**T**	84	21	RNR2
2960	T	CTAGAG**T**	ATATGG**A**	48	14	RNR2
3022	G	GCAGCC**G**	TAATAG**C**	62	12	RNR2
3242	G	ATGGCA**G**	CGGGCT**C**	48	11	TL1
3244	G	GGCAGA**G**	ACCGGG**C**	199	36	TL1
3421	G	CCCAAC**G**	CTACAA**C**	25	10	ND1
4665	G	GCAACC**G**	TGGATG**C**	47	15	ND2
5262	G	AAATGG**G**	TAATGG**C**	54	12	ND2
5585	G	GTAACA**G**	CCTTAG**C**	33	11	NC
5894	A	TACCTC**A**	TGGGGG**T**	48	12	NC
5970	G	TTATTC**G**	ATGCGC**C**	63	14	CO1
6260	G	AGTGGA**G**	TCCGGC**C**	41	11	CO1 [w]
7598	G	CATGCA**G**	CTTGCG**C**	35	15	CO2
7609	T	AGTAGG**T**	TTGTAG**A**	39	11	CO2 [w]
8251	G	AATAGG**G**	TACGGG**C**	67	18	CO2 [w]
9055	G	GGAAGC**G**	GGGTGG**C**	42	10	ATP6
9716	T	ACTGGG**T**	ATAGAG**A**	34	14	CO3 [w]
9868	G	TCATCC**G**	AGTTGG**C**	46	10	CO3

Position (cont.)	rCRS base	Upstream context L strand	Upstream context H strand	n with MAF ≥0.1%	n with MAF ≥1%	Genomic context
9947	G	AGATGT**G**	TCAAAC**C**	37	11	CO3 [w]
10197	G	TCCCCC**G**	CGCGGG**C**	58	15	ND3
11225	G	CTAGTA**G**	GGGAGC**C**	24	10	ND4
11456	G	CTTGCC**G**	GTACTG**C**	42	10	ND4
11711	G	ATAATC**G**	CGTGGG**C**	174	34	ND4
11914	G	AACCAC**G**	GGAGAA**C**	125	14	ND4 [w]
12172	A	TTGTGA**A**	GTCAGA**T**	40	10	TH
12457	G	ATTGTC**G**	TGGATG**C**	31	13	ND5
12771	G	CTGAGA**G**	TACGCC**C**	48	12	ND5 [w]
13063	G	ACCCCA**G**	CTGAGA**C**	54	13	ND5
13369	T	TCCGGG**T**	TGATGG**A**	57	13	ND5
13676	A	AAAATA**A**	GTGGGG**T**	74	14	ND5
13722	A	AAGCCT**A**	TGCGAA**T**	44	11	ND5 [w]
13879	T	ATAAAA**T**	GTGGGG**A**	46	10	ND5
14831	G	ATCTCC**G**	ATCATG**C**	58	17	CYB
14905	G	AGCCAT**G**	GTAGTG**C**	52	10	CYB [w]
14985	G	TCATCC**G**	AGGTAG**C**	70	18	CYB
15200	G	CTATCC**G**	GGATGG**C**	36	11	CYB
15884	G	AAATGG**G**	GACAGG**C**	55	12	CYB
15924	A	TTGTAA**A**	CTCCGG**T**	93	18	TT
16065	G	AGTATT**G**	GTGAGT**C**	42	12	CR
16086	T	AACCGC**T**	ATACAT**A**	56	15	CR
16092	T	TATGTA**T**	TACGAA**A**	53	15	CR
16093	T	ATGTAT**T**	GTACGA**A**	237	149	CR
16124	T	GAATAT**T**	CCGTAC**A**	64	11	CR
16129	G	TTGTAC**G**	TGGTAC**C**	413	31	CR
16172	T	ACCCAA**T**	ATGTGG**A**	57	11	CR
16183	A	ATCAAA**A**	AGGGGG**T**	24	14	CR
16187	C	AAACCC**C**	GGGGAG**G**	49	11	CR
16189	T	ACCCCC**T**	ATGGGG**A**	304	30	CR
16192	C	CCCTCC**C**	AGCATG**G**	128	39	CR
16256	C	CCAAAG**C**	GGGGTG**G**	51	13	CR
16261	C	GCCACC**C**	GGTGAG**G**	43	13	CR
16266	C	CCCTCA**C**	TAGTGG**G**	49	10	CR
16291	C	ACCTAC**C**	AGGGTG**G**	52	20	CR
16292	C	CCTACC**C**	AAGGGT**G**	54	12	CR
16293	A	CTACCC**A**	TAAGGG**T**	162	20	CR
16298	T	CACCCT**T**	ACTGTT**A**	33	12	CR
16301	C	CCTTAA**C**	TGTACT**G**	22	10	CR
16309	A	GTACAT**A**	ATGTAC**T**	46	16	CR
16311	T	ACATAG**T**	TTATGT**A**	77	23	CR
16327	C	CATTTA**C**	TGTACG**G**	31	12	CR
16362	T	TTCTCG**T**	ATGGGG**A**	29	10	CR
16390	G	GATAGG**G**	AGGGAC**C**	57	31	CR

*Positions with ≥100 heteroplasmies with MAF≥0.1% shown in **bold**; Upstream sequence GGGT-like motifs are highlighted;*
Genomic context: control region , rRNA , tRNA , non-coding , protein ; [w] – codon 3rd wobble position in protein coding genes

Of the 88 sites with highest number of heteroplasmies found >1% MAF, 35 are located in the D-loop region, with 29 found in protein coding genes, 18 found in rRNA, 4 found in tRNA and 2 found in non-coding sites. Interestingly, again the split of the 29 sites in the protein coding genes between the three codon positions (58% at position 1, 10% at position 2, 31% at position 3) does not seem to fit what we would expect if synonymous sites were more likely to display higher heteroplasmy levels. We address the implications of this in a later section.

There are several notable differences between the 107 positions of >0.1% MAF heteroplasmy hotspots and this site of 88 positions with highest numbers of heteroplasmies found with MAF>1%. Firstly, a small overlap exists between the two sets of positions, with only 11 sites found among the 107 positions with 100+ heteroplasmies with MAF ≥0.1% discussed earlier as well as the 88 positions with 10+ heteroplasmies with MAF>1%.

Secondly, unlike the 107 sites at the hotspots of heteroplasmies ≥0.1% MAF, this set of 88 sites with 10+ heteroplasmies >1% MAF shows much stronger correlation between the actual numbers of samples found with MAF ≥0.1% and with MAF ≥1% (Pearson's r = 0.56). Secondly, the number of samples with MAF≥0.1% in most cases is much lower compared to that observed among the 107 lower MAF heteroplasmy hotspot sites.

Thirdly, the base composition of the 88 sites where the highest frequency of heteroplasmies above 1% MAF is observed is quite different from that of positions at the lower level heteroplasmy hotspots. This is summarised in **Figure 5.4.2 F5**, with the base composition patterns of the adjacent sequence contexts addressed in **Figure 5.4.2 F6.**

Figure 5.4.2 F5: Base composition of and around the 88 sites with ≥10 identified heteroplasmies with MAF>1%

Base composition of and around positions with 10 or more putative heteroplasmies identified by consensus method at filter level 7

Sequence subset	Total base pairs	Overall base composition by rCRS base type A	T	C	G	Base pair orientation distributions AT pairs	AT orientation A_LT_H	A_HT_L	CG pairs	CG orientation C_LG_H	C_HG_L
Positions with 10+ putative heteroplasmies with MAF ≥ 1%	88	20 22.7%	21 23.9%	9 10.2%	38 43.2%	41 46.6%	48.8%	51.2%	47 53.4%	19.1%	80.9%
Sequence +/- 6bp from 88 positions with 10+ putative heteroplasmies	898	261 29.1%	204 22.7%	293 32.6%	140 15.6%	465 51.8%	56.1%	43.9%	433 48.2%	67.7%	32.3%
MT genome overall	all	*30.9%*	*24.7%*	*31.3%*	*13.1%*	*55.6%*	*55.6%*	*44.4%*	*44.4%*	*70.5%*	*29.5%*

Over 40% of the positions with 10+ heteroplasmies with MAF >1% are Gs, compared to just over 8% of sites with 100+ heteroplasmies with MAF ≥0.1%. And while over 60% of the lower MAF heteroplasmy hotspot sites were As, As makes up only slightly over 20% of positions with 10+ heteroplasmies with MAF>1%.

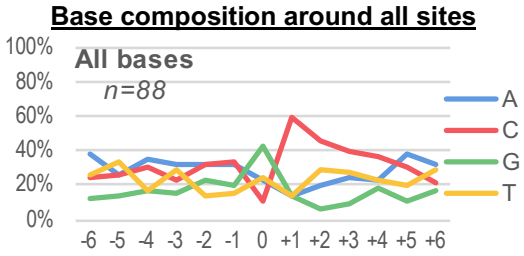
While the base composition of the sequence adjacent to the 107 very low level heteroplasmy hotspot sites displays major deviations from MT genome baseline, the overall base composition of the 898 base pairs worth of sequence found within 6bp of one or more of the 88 positions with 10+ heteroplasmies at 1% MAF threshold is actually very similar to the mitochondrial genome overall base composition.

Figure 5.4.2 F6: Adjacent sequence contexts around 88 sites with ≥10 identified heteroplasmies with MAF ≥1%

Base composition of the adjacent sequence context around 88 positions where at least 10 putative heteroplasmies with MAF ≥ 1% were found by consensus method at filter 7

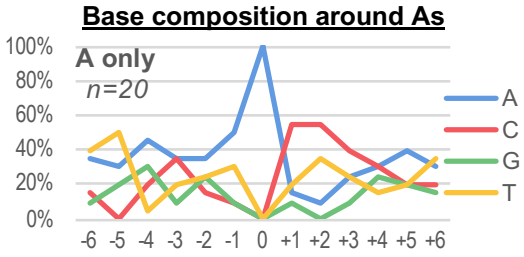
(a) Base composition around 88 most common putative heteroplasmy sites regardless of base type

	L strand upstream						All	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	38%	26%	35%	32%	32%	32%	22.7%	14%	19%	24%	23%	39%	32%
C	24%	26%	31%	23%	32%	34%	10.2%	59%	45%	40%	36%	31%	22%
G	13%	14%	17%	16%	23%	19%	43.2%	14%	6%	9%	18%	11%	17%
T	26%	34%	17%	30%	14%	15%	23.9%	14%	30%	27%	23%	19%	30%



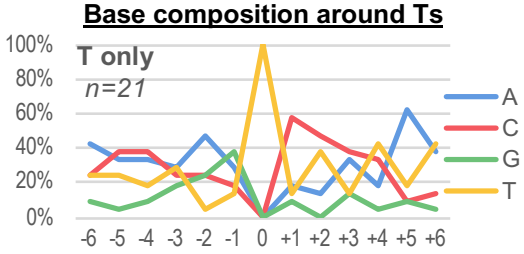
(b) As only: Base composition around 20 positions where rCRS base is A

	L strand upstream						A	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	35%	30%	45%	35%	35%	50%	100%	15%	10%	25%	30%	40%	30%
C	15%	0%	20%	35%	15%	10%	-	55%	55%	40%	30%	20%	20%
G	10%	20%	30%	10%	25%	10%	-	10%	0%	10%	25%	20%	15%
T	40%	50%	5%	20%	25%	30%	-	20%	35%	25%	15%	20%	35%



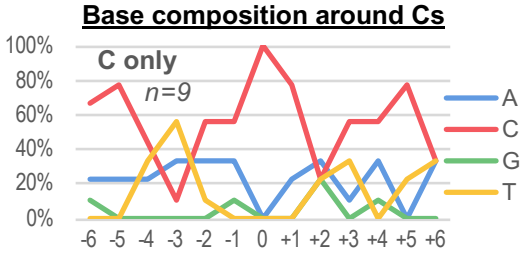
(c) Ts only: Base composition around 21 positions where rCRS base is T

	L strand upstream						T	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	43%	33%	33%	29%	48%	29%	-	19%	14%	33%	19%	62%	38%
C	24%	38%	38%	24%	24%	19%	-	57%	48%	38%	33%	10%	14%
G	10%	5%	10%	19%	24%	38%	-	10%	0%	14%	5%	10%	5%
T	24%	24%	19%	29%	5%	14%	100%	14%	38%	14%	43%	19%	43%



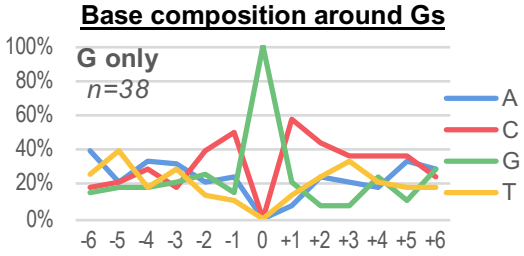
(d) Cs only: Base composition around 9 positions where rCRS base is C

	L strand upstream						C	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	22%	22%	22%	33%	33%	33%	-	22%	33%	11%	33%	0%	33%
C	67%	78%	44%	11%	56%	56%	100%	78%	22%	56%	56%	78%	33%
G	11%	0%	0%	0%	0%	11%	-	0%	22%	0%	11%	0%	0%
T	0%	0%	33%	56%	11%	0%	-	0%	22%	33%	0%	22%	33%



(e) Gs only: Base composition around 38 positions where rCRS base is G

	L strand upstream						G	L strand downstream					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	39%	21%	34%	32%	21%	24%	-	8%	24%	21%	18%	34%	29%
C	18%	21%	29%	18%	39%	50%	-	58%	45%	37%	37%	37%	24%
G	16%	18%	18%	21%	26%	16%	100%	21%	8%	8%	24%	11%	29%
T	26%	39%	18%	29%	13%	11%	-	13%	24%	34%	21%	18%	18%



In terms of the adjacent sequence context patterns around the 88 most common >1% MAF heteroplasmy positions, the most noticeable is a large increase in frequency of Cs immediately downstream of these positions. We also see lower frequency of As and Ts immediately downstream as well as considerable drop in the immediate upstream T frequency. Looking at the positions of each base type separately, we see that regardless of the base type, there is always a peak in C frequency immediately downstream of the position of interest. However only for positions where rCRS is C or G is there a similar increase in the upstream C frequency as well. For positions with As and Ts on the other hand, there are much fewer Cs found at positions -1 and -2 than would be expected based on the mitochondrial genome overall base composition.

The adjacent region base composition patterns appear to be overall somewhat more symmetric around Cs and Gs, with those around As and Ts showing a bigger difference between upstream and downstream regions. We are seeing some extent of reverse complement symmetry, such as peaks of Gs upstream of Ts and matching peaks of Cs downstream of As, or the increase in frequency of As upstream of A positions being matched by increased levels of Ts downstream of T positions, although this is less clear cut.

Overall, the data suggests that sites within C-rich sequences tend to be enriched in heteroplasmy, with some evidence of directional sequence context bias.

Overlap between sets of most common heteroplasmic position found at 0.1% and 1% MAF thresholds

As noted earlier, only 11 positions among the set of sites with 10+ heteroplasmies at >1% MAF are also found among the set of sites with 100+ heteroplasmies with MAF>0.1%. In terms of base composition, this subset of overlap sites is far more consistent with the 88 positions with 10+ heteroplasmies >1% MAF, showing much higher proportion of Gs compared to sites where low level heteroplasmy hotspots were observed (see **Supplementary figure 5.4.2 S3**). In fact, the even though the set of 11 overlap sites makes up only 9.2% of all positions with over 100 heteroplasmies above 0.1% MAF threshold, the 6 Gs among the overlap sites represent up 66.7% of all Gs found to have over 100 lower level heteroplasmies.

Comparison with sites most commonly reported as heteroplasmic in literature

Across the previous literature we have reviewed, we find 48 positions where heteroplasmy has been reported in 10 or more studies, and 13 sites where heteroplasmies are reported in 1000 or more individuals. Comparing these sets of most commonly reported sites in literature with the sets of 107 positions where we find 100 or more heteroplasmies with MAF>0.1% and the 88 positions where we find 10 or more heteroplasmies with MAF>1%, we see partial overlap with the data.

Of the 48 positions reported in 10 or more papers, 4 are present among our set of 107 MAF>0.1% hotspot positions only, 16 are present among our set of 88 MAF>1% highest frequency site set, and 3 are present among both of our highest heteroplasmy frequency site sets. We did not find either 100+ heteroplasmies above 0.1% MAF nor 10+ heteroplasmies above 1% MAF at 25 of the 48 positions with 10+ previous literature reports.

Of the 13 positions previously reported in 1000 or more samples, 2 are present among our datasets of both 107 sites found with 100+ heteroplasmies above 0.1% MAF as well as 88 sites with 10+ heteroplasmies above 1% MAF. Further 5 sites are present among our identified sites with 10+ heteroplasmies above 1% MAF. The remaining 6 positions with 1000+ reported heteroplasmies are not found among either of our identified position sets with highest heteroplasmy incidence. The overlaps between the different datasets are summarised in **Figure 5.4.2 F7** below.

Figure 5.4.2 F7: Overlap between the most frequently found heteroplasmic positions in the present study with the positions most often reported as heteroplasmic in the literature

Overlap between most commonly reported heteroplasmic positions in literature with the most common putative heteroplasmic sites identified at filters 5 and 7 in this study

(a) Overview of which known heteroplasmic sites are commonly found in this study

Position	rCRS base	In literature: 10+ studies	In literature: 1000+ samples	In this study: 100+ with MAF≥0.1%	In this study: 10+ with MAF≥1%
146	T	yes	yes	no	no
150	C	yes	no	no	no
152	T	yes	yes	no	yes
189	A	yes	no	no	yes
194	C	yes	no	no	no
195	T	yes	yes	no	yes
204	T	yes	yes	stnd. only	yes
214	A	yes	no	no	yes
215	A	yes	no	no	yes
228	G	yes	no	no	no
234	A	yes	no	no	yes
310	T	no	yes	no	stnd. only
318	T	no	yes	no	no
319	T	no	yes	no	no
323	G	no	yes	no	no
593	T	yes	no	no	no
988	G	yes	no	no	no
2623	A	no	no	yes	yes
2706	A	yes	no	no	no
5054	G	yes	no	no	no
5147	G	yes	no	no	no
5460	G	yes	no	no	no
7028	C	yes	no	no	no
7521	G	no	yes	no	no
8577	A	yes	no	no	no
10398	A	yes	no	yes	no
12775	G	yes	no	no	no

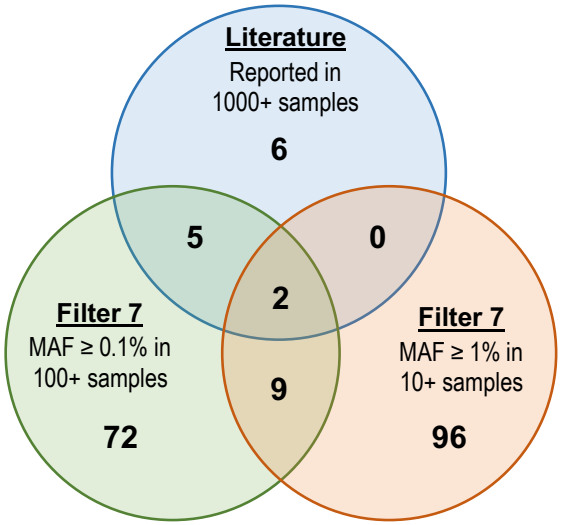
(cont.)

Position	rCRS base	In literature: 10+ studies	In literature: 1000+ samples	In this study: 100+ with MAF≥0.1%	In this study: 10+ with MAF≥1%
13708	G	yes	no	no	no
14040	G	yes	no	no	no
14569	G	yes	no	no	no
14798	T	yes	no	no	no
14869	G	yes	no	no	no
15043	G	yes	no	yes	no
16086	T	yes	no	no	yes
16092	T	yes	no	no	yes
16093	T	yes	yes	yes	yes
16126	T	yes	no	yes	no
16129	G	yes	no	yes	yes
16131	T	yes	no	no	no
16148	C	yes	no	no	no
16172	T	yes	no	no	yes
16183	A	no	yes	no	yes
16189	T	no	yes	yes	yes
16223	C	yes	no	yes	no
16234	C	yes	no	no	no
16256	C	yes	no	no	yes
16291	C	yes	no	no	yes
16292	C	yes	no	no	yes
16296	C	yes	no	no	no
16311	T	yes	no	stnd. only	yes
16324	T	yes	no	no	no
16362	T	yes	no	no	yes
16390	G	yes	no	no	yes
16519	T	yes	no	no	no

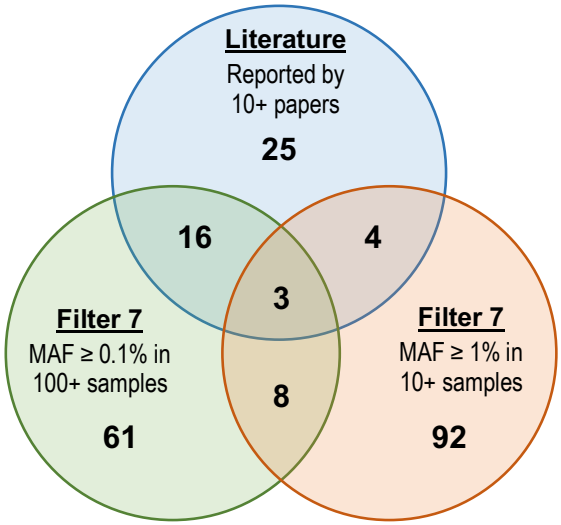
'stnd. only' – site passes most common putative heteroplasmy criteria with standard method only

(b) Venn diagrams illustrating overlaps between sets of most common heteroplasmic sites

Overlap with heteroplasmies reported in 1000+ individuals



Overlap with heteroplasmies reported in 10+ studies



There is a small overlap between positions with very high number of heteroplasmies with MAF > 0.1% and the positions with highest number of heteroplasmies with MAF > 1%. Plus, whatever overlap there is does not include positions with highest number of samples with MAF >0.1%. These two facts suggest that we are possibly looking at two different mechanisms: one which results in low level presumed heteroplasmies at the hotspot sites and another which results in the higher MAF heteroplasmies scattered throughout the mitochondrial genome. This is further supported by the clearly different base composition of each position set and their adjacent regions.

The very different base composition and distributions along the MT genome between heteroplasmies with MAF as low as 0.1% versus those with MAF >1% raises suspicion. Especially when considering that among the set of 107 lowest MAF heteroplasmy hotspot sites number of heteroplasmies with MAF>1% does not correlate with the total number of heteroplasmies above the lower 0.1% MAF threshold. Furthermore, previous heteroplasmy literature does not indicate a higher concentration of heteroplasmies reported within our identified hotspot regions, tilting the balance in favour of the artefact hypothesis.

On the other hand, there are individual positions within these regions which are reported in multiple of previous papers. Therefore if these low level heteroplasmy hotspots are in fact entirely a result of artefacts, these same artefacts must have also been repeatedly incorrectly reported as heteroplasmies in the previous literature.

The other possibility of course is that these could be genuine heteroplasmies that happen to behave differently from what we might have expected. It could perhaps point to existence of different mechanisms leading to different heteroplasmy classes. If, as our data indicates, it is indeed the case that hotspots of very common yet essentially always low-level heteroplasmy do exist, we would not necessarily expect to see large peaks of previous heteroplasmy reports in literature as studies rarely look at MAF below 1%. Thus literature not corroborating the patterns of heteroplasmy hotspot peaks is not necessarily a definitive argument for these peaks being artefacts. If these were indeed real heteroplasmies, this is a very interesting possibility with potentially profound implications raising important questions about what underlying processes might be causing such high population incidence yet low MAF heteroplasmy peaks, and what kinds of selection pressures must be present that could explain the nearly complete absence of emergence of at least some higher MAF sample proportion.

Equally, if these low level presumed heteroplasmy hotspots are actually an artefact, the underlying mechanism question is just as important. Given the base composition bias associated with apparent heteroplasmy hotspot sites shown earlier in **Figure 5.4.2 F2**, these hotspots do not look random visually, however it is unclear what might cause these types of artefacts. Some of the possibilities could include coincidence of more minor (and hence unidentified) error linked patterns affecting the same sites in both sequencing directions, or perhaps artefacts in read mapping around indels, or some other source of systematic errors such as variants arising from misidentified NUMTs. How one may begin to answer this, and what implications would presence of such artefact peaks mean for the next steps of the analysis is addressed in more detail in the discussion chapter.

As our ultimate goal is to analyse the lowest level heteroplasmy landscape that can be feasibly identified, this means looking at sites with MAF frequencies between 0.1% and 1%. Therefore we must make sure we handle our observed very low MAF heteroplasmy hotspot areas with care when making any subsequent conclusions about the heteroplasmy landscapes we are investigating

5.5 Heteroplasmy levels in the population

In the previous section we addressed the population scale patterns of the putative heteroplasmy incidence observed across the mitochondrial genome. While the heteroplasmy prevalence is one component of the population landscape, another component is the observed MAF itself and how this varies between positions and samples. The previous section already has provided some striking examples of differences between heteroplasmy frequency landscapes when all instances of MAF ≥0.1% are considered compared to landscapes of only the subsets of heteroplasmies with MAF ≥ 1%.

To understand the relationship between the MAF of a particular heteroplasmic locus and its prevalence among human samples better, in this section we investigate the observable heteroplasmy patterns in terms of the measured MAF. We assess the population average MAF across the mitochondrial genome and explore MAF distribution variability between loci and between samples.

5.5.1 Average measured MAF in the population

In order to assess the patterns of the heteroplasmy levels observed in the population, we can look at the landscapes of the average measured MAF.

One approach is to look at the landscapes of the MAF averaged across all analysed samples. The average MAF measured this way is affected by samples where no heteroplasmies were found. As expected, our MAF averages measured this way are very low, with no unexpected patterns in the landscapes and any peaks mirroring locations where higher numbers of heteroplasmic samples were identified (See **Supplementary figures 5.5.1 S1** and **5.5.1 S2**). With increasing allele filtration stringency the average MAF measured across the entire population decrease as a consequence of decreasing number of heteroplasmic samples identified at each position. Only very few positions persist where average MAF remain relatively high, with all such positions found within regions of analysed sample count. **Figure 5.5.1 F1** provides a summary of proportions of MT genome with different levels of the population average MAF.

Figure 5.5.1 F1: The average MAF observed across the mitochondrial genome

Overview of population mean MAF landscape found across the mitochondrial genome

(a) Number of positions found with different overall population mean MAF levels

Mean MAF across all samples	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
< 0.01%	14,170 *86.7%*	14,652 *90.4%*	10,101 *90.6%*	9,765 *90.4%*	4,985 *85.3%*	2,872 *17.5%*	13,252 *81.1%*	10,373 *90.0%*	9,831 *89.8%*	4,950 *84.7%*
0.01% - 0.02%	1,453 *8.89%*	1,038 *6.40%*	775 *6.95%*	764 *7.07%*	673 *11.5%*	8,480 *51.7%*	1,929 *11.81%*	816 *7.08%*	798 *7.29%*	698 *11.95%*
0.02% - 0.03%	384 *2.35%*	266 *1.64%*	164 *1.47%*	163 *1.51%*	121 *2.07%*	2,637 *16.1%*	532 *3.26%*	182 *1.58%*	168 *1.53%*	125 *2.14%*
0.03% - 0.04%	138 *0.84%*	99 *0.61%*	50 *0.45%*	48 *0.44%*	31 *0.53%*	960 *5.85%*	194 *1.19%*	57 *0.49%*	56 *0.51%*	34 *0.58%*
0.04% - 0.05%	66 *0.40%*	48 *0.30%*	16 *0.14%*	16 *0.15%*	8 *0.14%*	464 *2.83%*	107 *0.65%*	23 *0.20%*	24 *0.22%*	9 *0.15%*
0.05% - 0.1%	88 *0.54%*	71 *0.44%*	30 *0.27%*	29 *0.27%*	17 *0.29%*	661 *4.03%*	203 *1.24%*	45 *0.39%*	41 *0.37%*	18 *0.31%*
0.1% - 1%	39 *0.24%*	35 *0.22%*	18 *0.16%*	18 *0.17%*	6** *0.10%*	318 *1.94%*	120 *0.73%*	32 *0.28%*	29 *0.26%*	6** *0.10%*
> 1%	1 *0.01%*	1 *0.01%*	1 *0.01%*	1 *0.01%*	1* *0.02%*	8 *0.05%*	2 *0.01%*	2 *0.02%*	1 *0.01%*	1* *0.02%*
No heteroplasmies	64	193	5,248	5,599	10,561	3	64	4,873	5,455	10,562

Percentages indicate fraction of all MT genome positions with any putative heteroplasmies found at that filter level;
** position 310, mean MAF 3.66% (consensus) & 6.56% (standard); **positions 152, 195, 204, 16129, 16192, 16093*

Another aspect of the MAF landscape is how many individuals exhibit MAF exceeding certain threshold levels at each position, summarised in **Supplementary figure 5.5.1 S3.** We find that on average only a very small fraction of samples at any given site have high MAF, which when averaged across the entire MT genome results in only 0.15 samples per positions on average found with MAF of 10% or higher and 0.26 samples on average found per position with MAF of 5% or higher. The standard method results in a slightly higher number of high MAF samples per position, as well as higher variability between positions in terms of heteroplasmic sample fraction found above MAF thresholds of 0.5% or higher compared to the consensus method.

Considering that as the allele filtration stringency increases, only a small minority of individuals with putative heteroplasmies are found at most mtDNA positions, averaging MAF across the entire population is not useful for interrogating the actual behaviour of heteroplasmies that we are detecting. A more useful measure instead is MAF central tendency across only the heteroplasmic sample subsets.

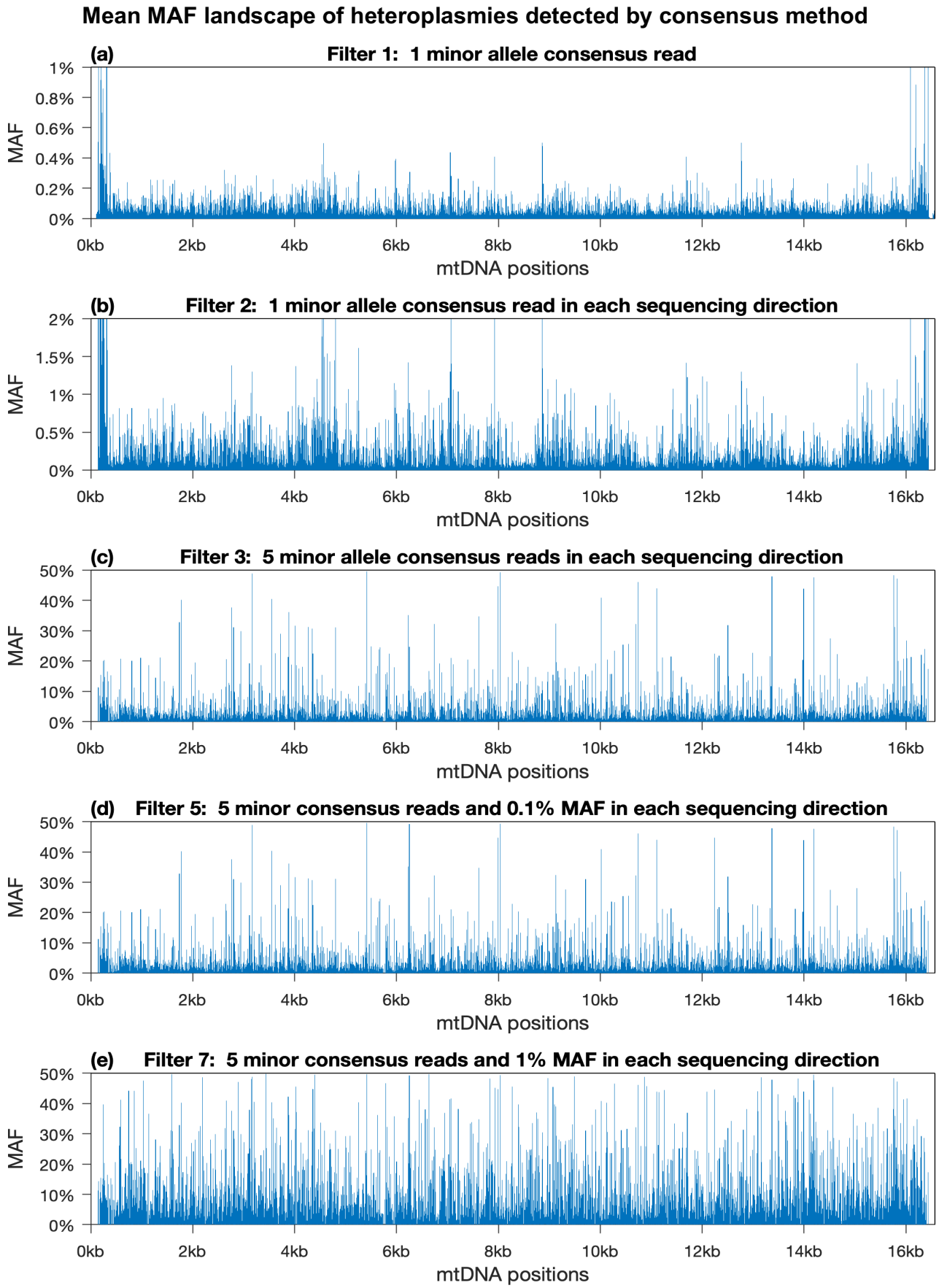
5.5.2 Average MAF across heteroplasmic individual subset

Two distinct parameters affect the shape of population average heteroplasmy level landscapes explored in the previous section. On one hand the landscape depends on the level of minor allele frequencies measured in heteroplasmic individuals. On the other hand, the landscape is also determined by the frequency of heteroplasmic individuals in the population itself. To tease these two parameters apart and to investigate specifically heteroplasmy levels independently of how common in population any particular heteroplasmy is, we need to look at the MAF averaged across the putative heteroplasmic individual subsets only. This is averaged for each locus across the subset of individuals that are heteroplasmic at that locus.

Landscape of mean MAF of the observed heteroplasmies

Figure 5.5.2 F1 visualises the mean observed MAF at each mtDNA position measured across our identified heteroplasmies using the consensus method at different allele filtration stringency levels. For the landscape of mean MAF landscapes across heteroplasmies identified by standard method see **Supplementary figure 5.5.2 S1.** An overview of the overall position-specific average heteroplasmy MAF level distribution is provided in **Figure 5.5.2 F2** showing how many positions we found with mean heteroplasmy levels in different MAF ranges.

Figure 5.5.2 F1: Mean MAF of heteroplasmies passing different filtration stringency levels at each mtDNA site



MAF – minor allele frequency; MAF mean calculated across the heteroplasmic sample subset at each MT genome position

Figure 5.5.2 F2: Overview of observed heteroplasmy mean MAF

Overview of the putative heteroplasmy mean MAF level landscape across the MT genome

(a) Number of positions found with different heteroplasmic sample subset mean MAF levels

Mean MAF across het. samples	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
< 0.01%	6 *0.04%*	0	0			6 *0.04%*	0	0		
0.01% - 0.05%	14,182 *86.8%*	11,013 *67.9%*	14 *0.13%*			15,209 *92.7%*	13,189 *80.7%*	19 *0.16%*		
0.05% - 0.1%	1,373 *8.40%*	2,255 *13.9%*	461 *4.13%*			835 *5.09%*	2,016 *12.3%*	407 *3.53%*		
0.1% - 0.5%	762 *4.66%*	2,522 *15.6%*	5,145 *46.1%*	4,977 *46.1%*		331 *2.02%*	1,100 *6.73%*	5,469 *47.4%*	4,845 *44.3%*	
0.5% - 1%	10 *0.06%*	335 *2.07%*	2,104 *18.9%*	2,156 *20.0%*		11 *0.07%*	21 *0.13%*	2,239 *19.4%*	2,308 *21.1%*	
1% - 5%	5 *0.03%*	79 *0.49%*	2,770 *24.8%*	2,955 *27.4%*	3,595 *61.5%*	7 *0.04%*	12 *0.07%*	2,806 *24.3%*	3,098 *28.3%*	3,520 *60.3%*
5% - 10%	1 *0.01%*	6 *0.04%*	427 *3.83%*	450 *4.17%*	1,141 *19.5%*	1 *0.01%*	1 *0.01%*	378 *3.28%*	442 *4.04%*	1,176 *20.1%*
> 10%	0	0	234 *2.10%*	266 *2.46%*	1,106 *18.9%*	0	0	212 *1.84%*	255 *2.33%*	1,145 *19.6%*

Percentages indicate fraction of all MT genome positions with any putative heteroplasmies found at that filter level

The average MAF across putative heteroplasmies increases with increasing allele filtration stringency as expected, and the higher average MAF peaks in the D-loop disappear as the average MAF across the rest of the genome catches up. While we do not see any unusually high average MAF peaks outside of the D-loop with the consensus method, several such peaks exist when MAF is calculated using the standard method. Interestingly, the high average MAF peaks with the standard method coincide with regions where we identified very low level high heteroplasmy incidence hotspots in the earlier section. The absence of the high average MAF peaks outside if the D-loop with the consensus method tells us that the difference in the measured MAF in the peak regions compared to adjacent areas comes entirely from excess minor allele reads in one sequencing direction only.

Comparing **Figure 5.5.2 F1** with **Supplementary figure 5.5.2 S1** shows that the average MAF landscape overall height outside of the aforementioned peak areas tends to be higher for the consensus method compared to the standard method. These differences between landscapes produced by the two methods decrease as allele filtration stringency increases. This is consistent with the presence of many more low confidence lower MAF heteroplasmies with the standard method that are not flagged as real heteroplasmies using the consensus method, especially at lower filtration stringency levels.

At filter levels 5 and 7 that are of most interest to us, the majority of heteroplasmic sites show average MAF between 0.1% and 0.5%, and between 1% and 5% respectively. At filter level 5 we see average MAF above 10% at approximately 2% of heteroplasmic sites, which rises to just under 20% when minimum MAF is increased to 1% at filter level 7.

Landscape of median MAF of the observed heteroplasmies

So far we have looked at the mean MAF landscapes across different mtDNA positions. But since the mean can be significantly affected by outlier values and we know that MAF distributions tend to be skewed, we should compare this to landscapes of the median MAF as well, which should be less affected by outliers. The landscapes of the identified heteroplasmic position median MAF are visualised in **Supplementary figures 5.5.2 S3 and 5.5.2 S4** with an overview of the mitochondrial genome overall position-specific median putative heteroplasmy MAF distribution shape presented in **Supplementary figure 5.5.2 S5.**

As predicted, assessing how the mean heteroplasmy MAF compares to the median MAF, we see generally lower median MAF, indicating likely skewing of the mean MAF by presence of a smaller number of high heteroplasmy samples at many positions.

One noteworthy difference from the mean MAF allele landscapes is the clear median MAF peaks above the surrounding baseline at filter levels 1 and 2 not just with the standard method, but also with the consensus method. These peaks are much larger with standard method, but they are still unmistakably present even when MAF is calculated using the consensus allele frequencies. Looking closely at MAF between the two sets of landscapes, we see that the peaks present in the consensus method median MAF landscape come from the surrounding areas showing bigger differences between mean and median MAF compared to these peak areas. This shifts the median MAF landscape downwards less in the aforementioned peak areas. This tells us that at this lowest allele filtration stringency, we are looking at unusually symmetrical putative heteroplasmy MAF distributions in these peak areas, while the rest of the mtDNA follows the more typical highly skewed MAF distribution with the majority of sample MAF values clustering at or closer to zero.

Frequencies of MAF values above different thresholds

In addition to the landscapes of heteroplasmy subset MAF central tendency averages explored above, we assess what proportion of our identified heteroplasmies per site on average reaches different MAF thresholds. This data is visualised in **Supplementary figure 5.5.2 S6**.

Comparing consensus and standard methods, we find that at filter 1 on average at each putative heteroplasmic position the standard method results in nearly 4.5 times more heteroplasmic samples identified compared to the consensus method. At filter 2 this ratio is approximately 3, and decreases to 1.4 by filter and 1.2 at filter 5. At filter level 7 both methods result in very similar heteroplasmic sample count on nearly 2.3 per putative heteroplasmic site.

In terms of the MAF distribution between heteroplasmic positions, at filter 5 when minimum MAF threshold is 0.1%, with consensus method on average 66.6% of putative heteroplasmies show MAF below 0.5% and 33.5% have MAF above 0.5%. MAF between 0.5% and 1% is found on average in 12.9% of cases, MAF between 1% and 5% in 13.9% of cases, MAF between 5% and 10% in 2.94% of cases and above 10% in the remaining 3.64% of cases. At filter 7 with minimum MAF of 1%, of all heteroplasmic positions 68.7% show average MAF between 1% and 5%, with a further 14% with MAF between 5% and 10%, and 17.3% with MAF above 10%.

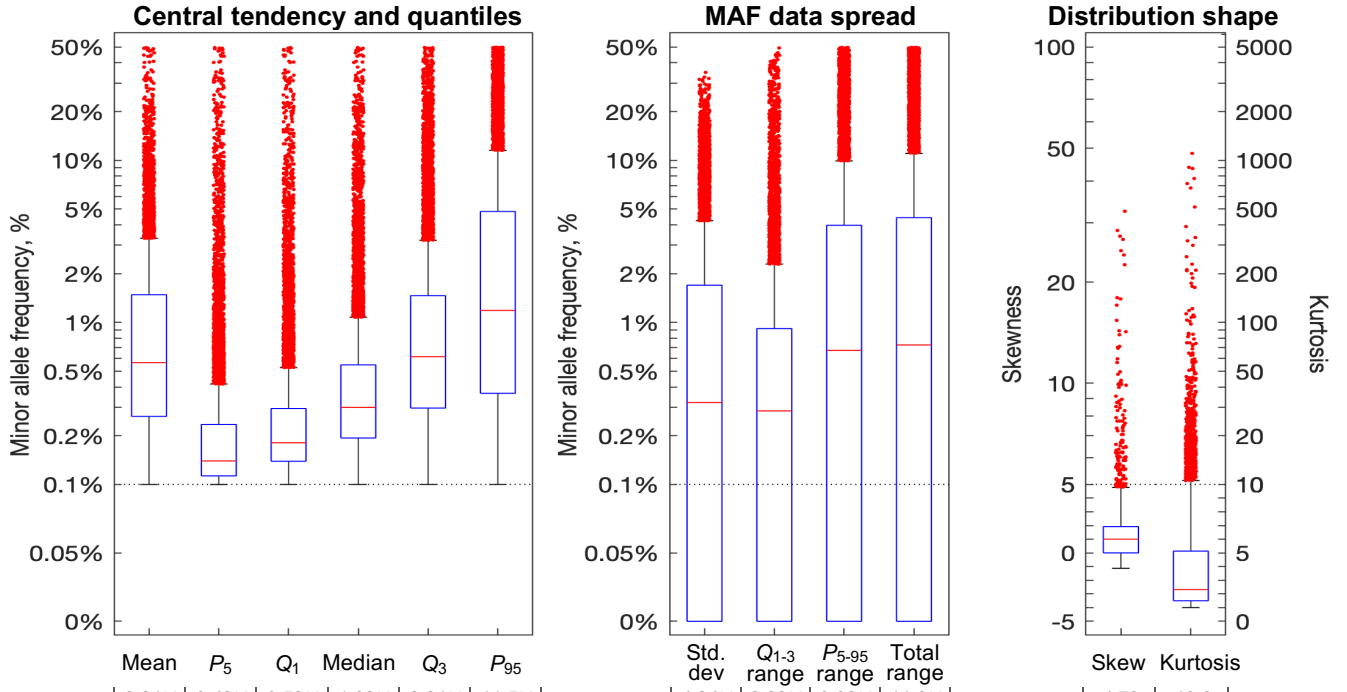
Analysing a multidimensional distribution

All MAF landscapes looked at so far focus on some form of an average MAF behaviour. Yet in reality we are looking at a complicated system where we have 16568 mitochondrial genome positions, at each of which variable numbers of heteroplasmic samples with variable minor allele frequency levels may be found. In essence we are looking at a distribution consisting of 16568 unique and slightly different position-specific MAF distributions, each made up of MAF values detected at that position in variable numbers of individuals. Thus in order to assess MAF overall behaviour across the mitochondrial genome, we necessarily have to reduce this complexity to some single variable. This is what the previous sections have attempted to do. But looking at any one landscape parameter in isolation misses a lot of information. Therefore the following figures attempt to provide a more global overview of the population-wide MAF behaviour across the mitochondrial genome by illustrating the distributions of various descriptive statistics that each describe some feature of the observed heteroplasmy MAF distributions found at each mitochondrial genome position.

Figure 5.5.2 F3: Observed heteroplasmy MAF population distribution shape variability between positions

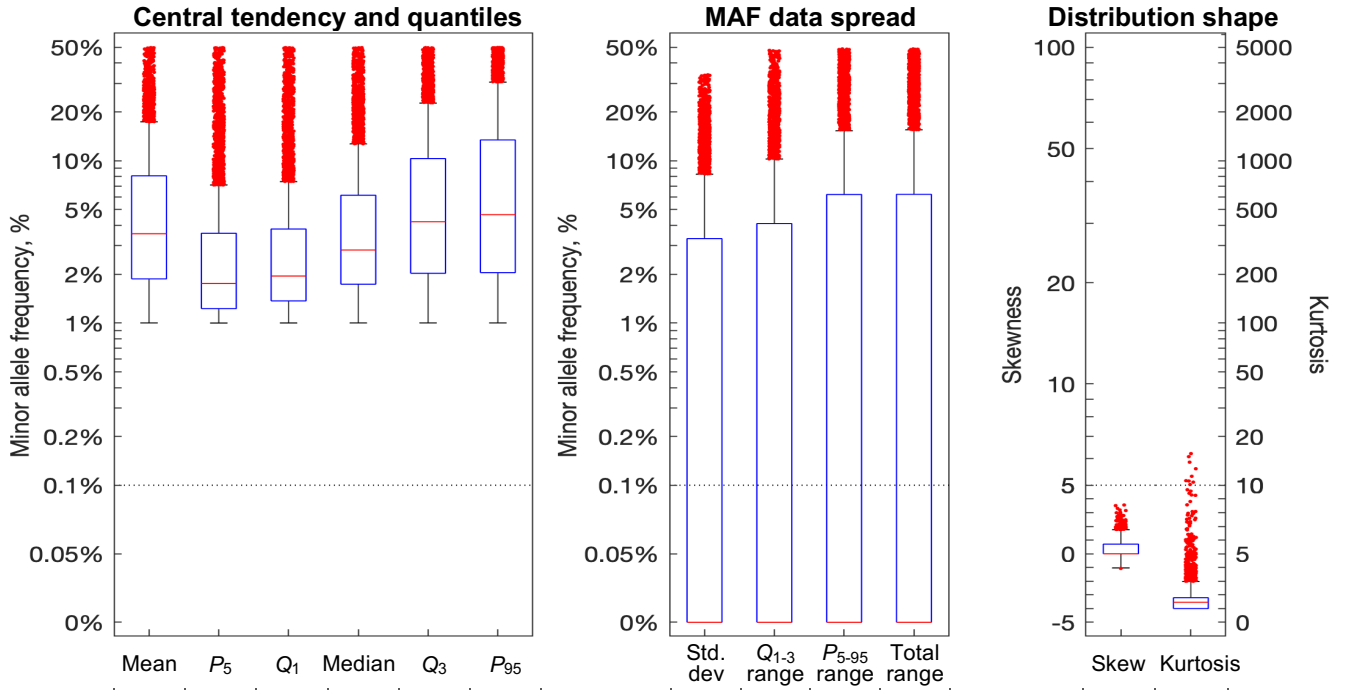
Overview of the distributions of position-specific MAF population distribution summary statistics and shape properties at filter levels 5 and 7 identified using consensus method

(a) MAF distribution features of 10,804 putative heteroplasmic positions at filter 5 (MAF ≥ 0.1%)



	Mean	P_5	Q_1	Median	Q_3	P_{95}	Std. dev	$Q_{1\text{-}3}$ range	$P_{5\text{-}95}$ range	Total range	Skew	Kurtosis	
Whisker Hi	3.31%	0.42%	0.53%	1.08%	3.21%	11.5%	4.24%	2.29%	9.92%	11.0%	4.79	10.6	*Whisker Hi*
Box Hi - Q_3	1.48%	0.23%	0.29%	0.55%	1.46%	4.84%	1.70%	0.92%	3.98%	4.43%	1.92	5.13	*Box Hi - Q_3*
Median - Q_2	0.56%	0.14%	0.18%	0.30%	0.61%	1.19%	0.32%	0.28%	0.67%	0.73%	1.00	2.32	*Median - Q_2*
Box Lo - Q_1	0.26%	0.11%	0.14%	0.19%	0.30%	0.37%	0%	0%	0%	0%	0	1.50	*Box Lo - Q_1*
Whisker Lo	0.10%	0.10%	0.10%	0.10%	0.10%	0.10%	0%	0%	0%	0%	-1.13	1	*Whisker Lo*
Outliers	1235	1528	1373	1383	1348	1406	1456	1283	1459	1547	143	721	*Outliers*

(b) MAF distribution features of 5,842 putative heteroplasmic positions at filter 7 (MAF ≥ 1%)



	Mean	P_5	Q_1	Median	Q_3	P_{95}	Std. dev	$Q_{1\text{-}3}$ range	$P_{5\text{-}95}$ range	Total range	Skew	Kurtosis	
Whisker Hi	17.4%	7.09%	7.44%	12.7%	22.7%	30.5%	8.27%	10.3%	15.3%	15.5%	1.76	2.98	*Whisker Hi*
Box Hi - Q_3	8.08%	3.58%	3.80%	6.13%	10.3%	13.4%	3.31%	4.10%	6.19%	6.21%	0.70	1.79	*Box Hi - Q_3*
Median - Q_2	3.55%	1.75%	1.95%	2.82%	4.21%	4.66%	0%	0%	0%	0%	0	1.46	*Median - Q_2*
Box Lo - Q_1	1.87%	1.23%	1.37%	1.73%	2.02%	2.04%	0%	0%	0%	0%	0	1	*Box Lo - Q_1*
Whisker Lo	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%	0%	0%	0%	0%	-1.03	1	*Whisker Lo*
Outliers	501	810	784	694	608	564	835	790	846	848	101	290	*Outliers*

Figure 5.5.2 F3 focuses on the identified heteroplasmy MAF distributions found using the consensus method using the minimum MAF thresholds of 0.1% (**panel A**) and 1% (**panel B**). For equivalent data relating to the standard method, see **Supplementary figure 5.5.2 S7**. In all panels each boxplot illustrates the distribution of values of one specific descriptive statistic or shape parameter of MAF distributions of all mtDNA sites where at least one putative heteroplasmic sample was found at that filter level. Each panel consists of three tiles, with boxplots shown in the left-most tile illustrating distributions of the position-specific MAF distribution central tendency (mean, median) and 5^{th}, 25^{th}, 75^{th} and 95^{th} percentile values, while the middle tile focuses on MAF value dispersion with boxplots showing positions specific distribution standard deviation as well as interquartile, 5^{th} to 95^{th} percentile and total range value distributions. Finally, boxplots in the right-most tile illustrate distributions of position-specific MAF distribution skewness and kurtosis parameters. In each boxplot the box itself represents the middle 50% of values (25^{th} to 75^{th} percentile), with the red line indicating that distribution median. Whiskers encompass all values that fall within 1.5 times interquartile range distance above or below 25^{th} and 75^{th} percentiles, with any outlier values outside this shown as red points. Each boxplot median and box upper and lower values, as well as whisker end point values and the number of outlier positions are listed directly underneath that boxplot.

The data shows that with the standard method, all quantile distribution medians are slightly higher than their consensus method counterparts. This is also true when we look at the 1^{st} and 3^{rd} quartiles of various quantile distribution values (i.e. boxplot box lower and upper edges), all of which are consistently slightly higher for the standard method compared to the consensus method, confirming a consistent tendency for the standard method to result in a slight increase in the measured MAF exactly as expected.

When we are looking at any kind of averages, as done in earlier sections, or when analysing value distribution parameter distributions as we are doing in this section, the presence of many positions with only a single or very few heteroplasmic samples identified must be taken into account. At the higher allele filtration stringency levels analysed here, many heteroplasmic positions are singletons with only one heteroplasmic sample identified in the whole population. At MAF >0.1%, with the consensus method 2821 sites or 26.1% of all heteroplasmic positions are singletons. At MAF>1% the singleton heteroplasmic site percentage increases to 2964 sites or 50.7% of all heteroplasmic positions. High prevalence of singleton sites means that the distribution of measures of position-specific MAF between different heteroplasmic individuals will be by definition skewed towards zero, which is what we see.

This section gives an overview of the overall position specific MAF distribution landscape. The obvious question now is what is the logical position-specific MAF measure and its threshold to choose for identifying candidate positions with highest MAF to investigate in more detail. The issue of singleton sites complicates this as we cannot judge whether the MAF observed is typical of the human population as a whole or not at such sites (i.e. assuming that there are more individuals outside our test population with heteroplasmies at these singleton sites). In this sense, excluding singleton positions, or looking at several groups of positions with different numbers of samples identified separately would make sense.

No obvious MAF threshold values (peaks, etc) appeared when looking at the landscapes or at the distributions. So based on this, perhaps an approach of ranking positions by heteroplasmy median or mean MAF (or perhaps IQR or coefficient of variation if we want to know how variability related to sequence features) would be most useful, as then we can select subsets of highest ranked positions based on how many samples were found at these sites, and then investigate these groups separately for sequence context or base composition patterns, and also test for overlap with literature reported sites, and with sites that had highest numbers of putative heteroplasmies detected from the precious major section.

5.5.3 MAF and the base type

We have observed in previous sections that a strong base composition bias exists in terms of the observed heteroplasmy incidence. The obvious next question therefore is to address how position base type relates to the measured heteroplasmy levels.

This is explored in **figure 5.5.3 F1**. **Panel A** shows the relative base composition of the set of positions at which any putative heteroplasmies were identified with the consensus method at different allele filtration stringency levels. Colour scale indicates base composition deviation from MT genome bassline (red – up, blue – down). **Panel B** provides an overview of at what proportion of all analysed positions heteroplasmies were found. For equivalent data relating to the standard method see **Supplementary figure 5.5.3 S1.**

Figure 5.5.3 F1: Heteroplasmic position base composition and distribution

Heteroplasmic position base composition

(a) Base composition of heteroplasmic position sets by consensus method

	Positions	A	C	G	T
Filter 1	16,339	31.0% *(+0.02%)*	31.3% *(-0.03%)*	12.9% *(-0.03%)*	24.7% *(+0.04%)*
Filter 2	16,210	31.1% *(+0.11%)*	31.2% *(-0.10%)*	12.9% *(-0.01%)*	24.7% *(+0.004%)*
Filter 3	11,155	26.4% *(-4.6%)*	21.9% *(-9.4%)*	18.2% *(+5.2%)*	33.5% *(+8.9%)*
Filter 5	10,804	25.8% *(-5.2%)*	21.3% *(-10.0%)*	18.7% *(+5.7%)*	34.2% *(+9.5%)*
Filter 7	5,842	20.6% *(-10.4%)*	11.8% *(-19.6%)*	27.9% *(+14.9%)*	39.7% *(+15.1%)*

Values in brackets indicate base type frequency change from base composition of 16,403 analysed mtDNA positions

(b) Heteroplasmic fraction of all analysed mtDNA positions

	Heteroplasmic fraction	Heteroplasmic fraction of analysed positions by base type			
		A	C	G	T
Filter 1	99.6%	99.7%	99.5%	99.4%	99.8%
Filter 2	98.8%	99.2%	98.5%	98.7%	98.8%
Filter 3	68.0%	57.8%	47.6%	95.3%	92.4%
Filter 5	65.9%	54.8%	44.8%	94.9%	91.3%
Filter 7	35.6%	23.6%	13.4%	76.7%	57.3%

Figure 5.5.3 F1 demonstrates that as the allele filtration stringency increases, we start seeing substantial and increasing base composition differences from MT genome baseline among our identified heteroplasmic positions: decrease in A and C frequencies and an increase in G and T frequencies. G is the base type showing strongest association with heteroplasmies, with T being the next base with highest filter-passing heteroplasmy likelihood. Positions where the base is C on the other hand have the lowest probability to have filter passing heteroplasmies.

How these findings relate to MAF is explored in the **Figure 5.5.3 F2**, which summarises average heteroplasmic position mean and median MAF observed across subsets of heteroplasmic positions grouped by the position base type, with the colour scale showing indicating the relative differences between base-specific average MAF and the overall heteroplasmic position set average MAF. For equivalent data based on the standard method, see **Supplementary figure 5.5.3 S1**.

Figure 5.5.3 F2: Relationship between position base type and the average heteroplasmy MAF

Relationship between position base type and the average heteroplasmy MAF level

(a) Average position-specific mean observed heteroplasmy MAF by consensus method

Consensus method	Overall average	Average median by base type							
		A		C		G		T	
Filter 1	0.034%	0.028%	*(-0.006%)*	0.023%	*(-0.011%)*	0.062%	*(+0.028%)*	0.040%	*(+0.006%)*
Filter 2	0.09%	0.06%	*(-0.02%)*	0.05%	*(-0.04%)*	0.19%	*(+0.10%)*	0.10%	*(+0.02%)*
Filter 3	1.45%	1.51%	*(+0.06%)*	1.45%	*(-0.004%)*	1.36%	*(-0.09%)*	1.46%	*(+0.004%)*
Filter 5	1.61%	1.70%	*(+0.09%)*	1.59%	*(-0.02%)*	1.49%	*(-0.12%)*	1.61%	*(+0.004%)*
Filter 7	6.57%	7.17%	*(+0.61%)*	7.01%	*(+0.45%)*	5.99%	*(-0.58%)*	6.52%	*(-0.04%)*

Values in brackets show base-type specific average mean MAF difference from the overall heteroplasmic position mean MAF average

(b) Average position-specific median observed heteroplasmy MAF by consensus method

Consensus method	Overall average	Average median by base type							
		A		C		G		T	
Filter 1	0.018%	0.018%	*(+0.001%)*	0.016%	*(-0.002%)*	0.017%	*(-0.001%)*	0.020%	*(+0.002%)*
Filter 2	0.025%	0.026%	*(+0.001%)*	0.022%	*(-0.003%)*	0.031%	*(+0.006%)*	0.025%	*(-0.0002%)*
Filter 3	0.84%	1.07%	*(+0.23%)*	1.25%	*(+0.40%)*	0.42%	*(-0.43%)*	0.63%	*(-0.21%)*
Filter 5	0.97%	1.23%	*(+0.26%)*	1.38%	*(+0.42%)*	0.50%	*(-0.47%)*	0.77%	*(-0.20%)*
Filter 7	5.90%	6.75%	*(+0.85%)*	6.86%	*(+0.96%)*	4.84%	*(-1.05%)*	5.91%	*(+0.01%)*

Values in brackets show base-type specific average median MAF difference from the overall heteroplasmic position median MAF average

Figure 5.5.3 F2 indicates differences in the average MAF at heteroplasmic positions of different base types. When we consider the heteroplasmies identified at the more stringent filtration levels when minimum MAF thresholds of 0.1% or 1% are applied, we see higher average heteroplasmy MAF at positions with reference base A or C compared to positions where the reference base is G or T. The differences between base types at these higher allele filtration stringencies are larger when considering median MAF compared to mean MAF, which is consistent with higher level of variability between the heteroplasmies that are outliers in terms of MAF, which therefore affect the median MAF more compared to the mean MAF.

Overall, the data presented here indicates that base types differ considerably in what proportion of the initially detected putative heteroplasmies 'survives' the steps of increasingly stringent allele filtration. G is the base type with the lower heteroplasmy sensitivity to filtration, with the largest proportion of positions still found reported as heteroplasmic at G positions even when the minimum MAF threshold is raised to 1% at filter level 7. At the same time though, G positions also show lowest average MAF across the found heteroplasmies, indicating that even though heteroplasmies at G sites are more common, they tend to be characterised by lower minor allele frequencies compared to other bases. On the other hand, the base type with highest heteroplasmy sensitivity to filtration is C, with any heteroplasmies with MAF>1% at all only found at 13.3% of all analysed C positions. Yet the average MAF across these sites that we do find appear to be higher than for other base types, suggesting that while heteroplasmies at C positions are rarer, when these happen, minor allele frequencies are higher.

Curiously, the average heteroplasmy MAF, especially the median MAF appear to follow a roughly inverse pattern to that of how commonly heteroplasmies are found at positions of each base type, with bases found at higher-than-baseline frequency among heteroplasmic position sets (Gs and Ts) showing lower-than-average MAF and vice versa. On one hand this could hint to a complex system of interlinked mechanisms of potentially different error types combined with different types of real heteroplasmy and so on. On the other hand, we are dealing with a system subject to a certain amount of noise, which will affect any exploratory analysis of central tendencies and averages. In this case, the differences between patterns associated with averages of MAF distribution means versus medians that we see could well be consistent with a noisy dataset, warranting extreme caution when interpreting the patterns observed here, but especially so for filter levels where MAF averages are extremely low and we do not have a way to assess the significance of these observations.

5.5.4 Positions associated with highest average MAF

We have shown in the previous sections that mitochondrial genome positions vary by average MAF of the putative heteroplasmies. Focusing on filter levels 5 and to some extent 7, this section aims to identify the sets of positions associated with the highest overall MAF across the detected heteroplasmies, and assess whether we can identify any characteristics distinguishing such positions from the rest of the MT genome. We also aim to assess how much overlap there is between highest MAF positions versus positions associated with highest heteroplasmy incidence either in our study, or in literature reports.

Due to the variable nature of the population level MAF data, different measures can be used for defining mtDNA positions with highest MAF. Positions could be ranked by the observed heteroplasmy MAF distribution central tendency or some other descriptive statistic. Alternatively, positions can also be sorted by how many samples exceeding certain minimum MAF we detect. Different sorting methods may not necessarily result in the same sets of sites being identified at the top of the list. Therefore in this section we assess how the different ways of looking at MAF affect what positions we identify as being linked to highest MAF.

Relationship between heteroplasmy frequency and MAF level

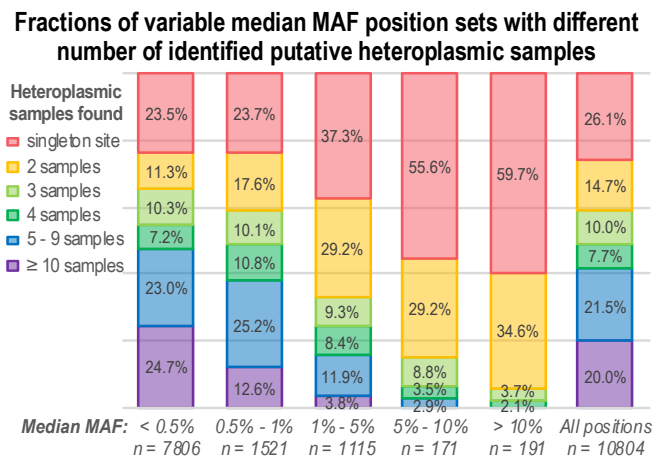
Central tendency of the MAF distribution across all putative heteroplasmies identified at a position can be used to identify which mtDNA positions are associated on average with the highest MAF. Here we focus on using the median as the central tendency measure due to its higher resistance to outlier effects, and assess how positions with different frequencies of detected heteroplasmic samples are distributed position subsets with different median MAF. This is explored in **Figure 5.5.4 F1** focusing on the MAF data measured using the consensus method. For equivalent data relating to the standard method see **Supplementary figure 5.5.4 S1.**

At the minimum MAF threshold of 0.1% using the consensus method we identified a total of 120,754 heteroplasmies distributed across 65.9% of the analysed mtDNA sequence, with the position-specific heteroplasmy median MAF ranging from 0.10% to 49.62%.

Figure 5.5.4 F1: Relationship between the number of heteroplasmies and the average heteroplasmy MAF level

Relationship between the number of heteroplasmies with MAF≥0.1% and the average MAF level

Position median MAF	Total sites	Number of samples with MAF ≥ 0.1% 1	2	3	4	5 - 9	10+
0.1% - 0.5%	7806	1836 *23.5%*	881 *11.3%*	802 *10.3%*	561 *7.2%*	1796 *23.0%*	1930 *24.7%*
0.5% - 1%	1521	360 *23.7%*	267 *17.6%*	154 *10.1%*	165 *10.8%*	384 *25.2%*	191 *12.6%*
1% - 5%	1115	416 *37.3%*	326 *29.2%*	104 *9.3%*	94 *8.4%*	133 *11.9%*	42 *3.8%*
5% - 10%	171	95 *55.6%*	50 *29.2%*	15 *8.8%*	6 *3.5%*	5 *2.9%*	0
≥ 10%	191	114 *59.7%*	66 *34.6%*	7 *3.7%*	4 *2.1%*	0	0
Total het. positions	10804	2821 *26.1%*	1590 *14.7%*	1082 *10.0%*	830 *7.7%*	2318 *21.5%*	2163 *20.0%*



Data shown for consensus method. Percentages indicate fraction of total heteroplasmic sites in each set grouped by position median MAF

Figure 5.5.4 F1 demonstrates that when we compare subsets of sites with different median MAF measured across the identified heteroplasmies, the ratios between how many sites are found with different number of heteroplasmic samples changes, with the relative proportion of sites with only 1 or 2 heteroplasmic samples found going up with increasing position median MAF. A clear inverse relationship exists between how common heteroplasmy is at a given site and the observed average MAF. Even though there are singleton positions with low minor allele frequencies,

these are relatively less common whereas when median MAF is high, the vast majority of sites are actually singletons. While this is at least partially due to us looking at the MAF median, which means high MAF at singleton sites will be more visible compared to sites with more than one heteroplasmy, it does not explain the full extent of the inverse relationship between the number of heteroplasmies per site found and their median MAF that we observe. When MAF distributions across all detected heteroplasmies is compared to that of the singleton heteroplasmy positions only, we see a mismatch in the distributions, with higher MAF enriched among singleton heteroplasmy positions.

To assess the characteristics of positions linked to high MAF, we will address two main groups of positions. Groups 1 and 2 represent sites where we are detecting high or very high average MAF across all analysed heteroplasmies at that site. We have chosen arbitrarily minimum median MAF cut-offs 10 and 100 times above the minimum heteroplasmy MAF threshold of 0.1% as the criteria for defining the high (group 1) and very high (group 2) average MAF positions respectively.

Group 1: High heteroplasmy positions with median MAF ≥1%

To assess the characteristics of positions linked to high MAF, we use the median MAF threshold of ≥1%. Using this threshold with the consensus method, we find 1477 sites meeting our 1% median MAF threshold for high heteroplasmy level. Of these, 191 also meet the 10% median MAF threshold for very high heteroplasmy level, representing 1.7% of all heteroplasmic positions.

The base composition and MAF distribution of the 1477 high heteroplasmy sites with median MAF >1% measured by the consensus method is summarised in **Figure 5.5.4 F2** below. For equivalent data relating to the standard method see **Supplementary figure 5.5.4 S2.**

Figure 5.5.4 F2: Features of 1,477 mtDNA positions with high observed heteroplasmy median MAF ≥1%

MAF distribution descriptive statistics and base composition of 1,477 mtDNA positions with putative heteroplasmy median MAF ≥ 1% found at filter 5 with consensus method

(a) MAF descriptive statistic averages for sets of sites with variable putative heteroplasmy counts

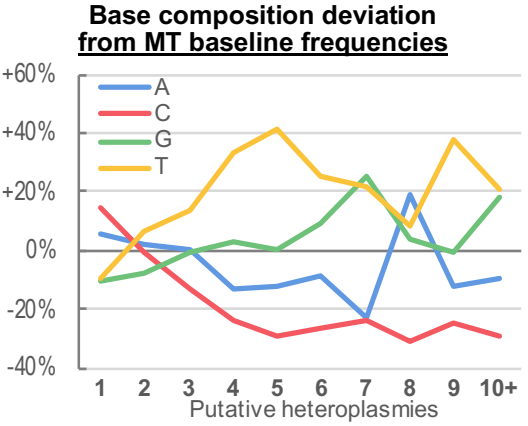
Het. count	High MAF positions	High MAF positions by rCRS base type A	C	G	T	Avg. values of MAF descriptive statistics Mean	SD	Median	IQR
Any	1477	473	458	109	437	5.7%	3.6%	5.0%	3.3% - 7.8%
1	625 (42.3%)	229 (48.4%)	287 (62.7%)	15 (13.8%)	94 (21.5%)	6.7%	n/a	6.7%	n/a
2	442 (29.9%)	147 (31.1%)	134 (29.3%)	22 (20.2%)	139 (31.8%)	4.9%	5.8%	4.9%	0.8% - 9.0%
3	126 (8.5%)	39 (8.2%)	23 (5.0%)	16 (14.7%)	48 (11.0%)	5.6%	6.7%	3.4%	1.2% - 10.6%
4	104 (7.0%)	19 (4.0%)	8 (1.7%)	17 (15.6%)	60 (13.7%)	4.9%	6.4%	2.8%	0.6% - 9.2%
5	53 (3.6%)	10 (2.1%)	1 (0.2%)	7 (6.4%)	35 (8.0%)	4.4%	5.9%	1.9%	0.6% - 7.1%
6	44 (3.0%)	10 (2.1%)	2 (0.4%)	10 (9.2%)	22 (5.0%)	4.3%	6.4%	1.8%	0.4% - 5.5%
7	13 (0.9%)	1 (0.2%)	1 (0.2%)	5 (4.6%)	6 (1.4%)	5.4%	8.2%	1.6%	0.4% - 7.3%
8	12 (0.8%)	6 (1.3%)	0	2 (1.8%)	4 (0.9%)	4.8%	7.8%	1.5%	0.5% - 6.1%
9	16 (1.1%)	3 (0.6%)	1 (0.2%)	2 (1.8%)	10 (2.3%)	4.7%	7.2%	1.9%	0.4% - 6.0%
10+	42 (2.8%)	9 (1.9%)	1 (0.2%)	13 (11.9%)	19 (4.3%)	4.6%	7.6%	1.5%	0.4% - 4.4%

Percentages in brackets indicate relative proportions of total putative heteroplasmic positions of that base type

(b) Base composition of high MAF position subsets with different putative heteroplasmy counts

Het. count	High MAF position subset base composition A	C	G	T
Any	32.0% (+1.1%)	31.0% (-0.3%)	7.4% (-5.7%)	29.6% (+4.9%)
1	36.6% (+5.7%)	45.9% (+14.6%)	2.4% (-10.7%)	15.0% (-9.7%)
2	33.3% (+2.3%)	30.3% (-1.0%)	5.0% (-8.1%)	31.4% (+6.7%)
3	31.0% (+0.0%)	18.3% (-13.0%)	12.7% (-0.4%)	38.1% (+13.4%)
4	18.3% (-12.7%)	7.7% (-23.6%)	16.3% (+3.3%)	57.7% (+33.0%)
5	18.9% (-12.1%)	1.9% (-29.4%)	13.2% (+0.1%)	66.0% (+41.3%)
6	22.7% (-8.2%)	4.5% (-26.7%)	22.7% (+9.6%)	50.0% (+25.3%)
7	7.7% (-23.2%)	7.7% (-23.6%)	38.5% (+25.4%)	46.2% (+21.4%)
8	50.0% (+19.1%)	0% (-31.3%)	16.7% (+3.6%)	33.3% (+8.6%)
9	18.8% (-12.2%)	6.3% (-25.0%)	12.5% (-0.6%)	62.5% (+37.8%)
10+	21.4% (-9.5%)	2.4% (-28.9%)	31.0% (+17.9%)	45.2% (+20.5%)

Percentages in brackets show deviation from mtDNA baseline frequencies



The average heteroplasmy mean MAF we observe across this set of sites is 5.7%, however when subsets of positions with different numbers of heteroplasmies are compared, singleton sites display on average higher MAF compared to non-singleton sites. Base composition of the total set of 1477 high heteroplasmy positions shows a slight excess of Ts and depletion of Gs relative to the MT genome overall, however when positions are split into singleton versus non-singleton sites, we actually observe an inverse pattern of base composition between the sets, with significant C and A overrepresentation among singleton sites and T and G overrepresentation among non-singleton sites.

The visible differences in singleton sites compared to non-singleton positions suggests that compared to heteroplasmy frequency, the actual heteroplasmy levels may not be as strongly linked to MT genome sequence context. This conclusion is also further supported by a lack of obvious adjacent sequence context patterns shared

between the singleton and non-singleton high median MAF position subsets, visualised in **Supplementary figure 5.5.4 S3,** indicating that perhaps singleton heteroplasmies might more likely be random events compared to the more common heteroplasmy positions, where some sequence contexts might be more common than others.

Among the 1477 high heteroplasmy positions with median MAF>1% we found 48 sites where 10 or more heteroplasmies were detected. This subset of the most commonly represented high average MAF sites, listed in **Supplementary figure 5.5.4 S4**, is worth addressing. While these 42 positions represent sites with the highest number of reported heteroplasmies among the high median MAF positions subset, these are by no means the positions with highest median MAF. These are however positions that have all been reported in previous literature by at least 3 separate studies and include several of the most frequently reported heteroplasmy sites, with the previous literature report frequencies following a relatively similar pattern to our reported heteroplasmy counts.

Interestingly, while 15 of the 42 sites are located within protein coding genes, only a single site is at a codon 'wobble' position, suggesting persistence of heteroplasmy despite presumably active purifying selection against mutations in these positions.

Group 2: Very high heteroplasmy positions with median MAF ≥ 10%

So far we have focused on positions with high average heteroplasmy levels (median MAF >1%). Here we look at positions with very high average heteroplasmy levels (median MAF >10%). In total, among the heteroplasmic sites identified using consensus method with the minimum MAF threshold 0.1%, we find 191 such very high average MAF positions, summarised in **Figure 5.5.4 F3**.

Compared to the wider set of positions with median MAF >1%, heteroplasmies at the sites with average MAF >10% are found in much fewer samples. 59.7% of these positions are singleton sites, with 2, 3 and 4 heteroplasmies found at 34.6%, 3.6% and 2.1% of the sites respectively. None of the 191 very high MAF positions showed more than 4 heteroplasmic samples, nor do we see any positions where MAF>10% is detected in more than 2 of the samples.

Figure 5.5.4 F3: Features of 191 mtDNA positions with very high observed heteroplasmy median MAF ≥10%

MAF distribution descriptive statistics and base composition of 191 mtDNA positions with putative heteroplasmy median MAF ≥ 10% found at filter 5 with consensus method

(a) MAF descriptive statistic averages for sets of sites with variable putative heteroplasmy counts

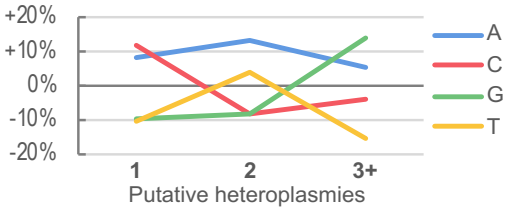
Het. count	High MAF positions	Very high MAF positions by rCRS base type				Avg. values of MAF descriptive statistics			
		A	C	G	T	Mean	SD	Median	IQR
Any	191	78	67	10	36	17.9%	20.3%	20.5%	14.8% - 26.5%
1	114 *(59.7%)*	45 *(57.7%)*	49 *(73.1%)*	4 *(40%)*	16 *(44.4%)*	23.0%	n/a	23.0%	0%
2	66 *(34.6%)*	29 *(37.2%)*	15 *(22.4%)*	3 *(30%)*	19 *(52.8%)*	16.9%	20.6%	16.9%	2.3% - 31.4%
≥ 3	11 *(5.8%)*	4 *(5.1%)*	3 *(4.5%)*	3 *(30%)*	1 *(2.8%)*	17.9%	18.4%	16.7%	3.7% - 32.2%

Percentages in brackets indicate relative proportions of total putative heteroplasmic positions of that base type

(b) Base composition of very high MAF site subsets with different putative heteroplasmy counts

Het. count	Very high MAF position subset base composition			
	A	C	G	T
Any	40.8% *(+9.9%)*	35.1% *(+3.8%)*	5.2% *(-7.9%)*	18.8% *(-5.9%)*
1	39.5% *(+8.5%)*	43.0% *(+11.7%)*	3.5% *(-9.6%)*	14.0% *(-10.7%)*
2	43.9% *(+13.0%)*	22.7% *(-8.5%)*	4.5% *(-8.5%)*	28.8% *(+4.1%)*
≥ 3	36.4% *(+5.4%)*	27.3% *(-4.0%)*	27.3% *(+14.2%)*	9.1% *(-15.6%)*

Percentages in brackets show deviation from mtDNA baseline frequencies



The set of 191 very high heteroplasmy positions follow broadly similar patterns to those observed in the much larger high heteroplasmy position set in terms of base composition, with bases C and A found at the absolute majority of singleton very high heteroplasmy positions, whereas a larger proportion of Gs and Ts are found at the very high median MAF positions with 2 to 4 heteroplasmic samples. In terms of the adjacent sequence context around the 191 very high heteroplasmy positions, visualised in **Supplementary figure 5.5.4 S5,** the most notable pattern is an enrichment of As in the immediate upstream region, and the fact that this is equally present whether looking at singleton or non-singleton sites. Interestingly, this is different from the larger set of heteroplasmic positions with median MAF above 1%, where such upstream A enrichment was not observed in association with singleton positions. However, given the low number of samples found at any of the very high average MAF sites, it is hard to speculate about the significance of the singleton versus non-singleton distinction in this dataset.

5.5.5 Positions associated with largest number of high MAF heteroplasmies

Data presented in the previous section shows that no MT genome sites exist where large numbers of heteroplasmies are identified that all have very high MAF. Instead, at sites where large number of heteroplasmies are detected, only a small proportion show high MAF, with high median MAF only really observed at positions where heteroplasmies are detected in a small number of samples.

Because of this, another way to look at the heteroplasmy levels and identify positions associated with higher minor allele frequencies would be to identify the sites where a large number of samples with high MAF were found regardless of the overall median MAF across all heteroplasmies detected at that site. We shall again focus on heteroplasmies passing minimum 0.1% MAF threshold identified using the consensus method. **Supplementary figure 5.5.5 S1 panel A** provides a summary of the number of MT genome positions where we find different numbers of heteroplasmies above 0.1%, 0.5%, 1%, 5% and 10% MAF thresholds.

The approach we have used here was to identify the set of the most extreme positions in terms of the incidence of high MAF heteroplasmies, and assess how these sets of positions compare to mitochondrial genome as a whole. Specifically, we looked at sets of positions where at least 5 samples exhibited heteroplasmies with MAF exceeding 1%, 5% and 10%. This approach and the specific thresholds were chosen for ease of assessing base composition deviations of the resulting subsets of the most extreme sites compared to the MT genome overall, with the thresholds chosen yielding a reasonable number of sites in each of the three groups.

It is important to note that these thresholds are arbitrary, and what one may consider as high MAF or high heteroplasmy incidence will depend on the specifics of the study. In our case, the aim is to gain a broader insight into whether there are some obvious distinguishing characteristics of the positions where high MAF heteroplasmies occur most frequently. To address the criticism that such thresholding is not necessary, that is indeed correct. and an alternative stronger approach would have been statistical modelling of the heteroplasmy behaviours.

In total we find 543 MT genome positions with 5 heteroplasmies above 1% MAF. Of these, at 57 we find 5 heteroplasmies ≥ 5% MAF, and 20 with 5 heteroplasmies ≥10%. We shall be studying these three sets of positions in further detail.

The 20 positions with at least 5 samples with MAF>10% are listed in **Supplementary figure 5.5.5 S1 panel B**. All of these positions have also been reported in heteroplasmy literature, with many having been reported in over 10 studies. In terms of the genomic context, 14 of the sites are within the D-loop region, with 3 found in protein coding genes (none in 'wobble' positions), two in rRNA and 1 in tRNA.

Figure 5.5.5 F1 below addresses the base compositions of all three position sets of interest, with the adjacent sequence context patterns addressed in **Supplementary figure 5.5.5 S2.**

Figure 5.5.5 F1: Base composition mtDNA position sets where high to very high heteroplasmy levels (MAF≥1%, 5% or 10%) are found in ≥5 samples

Base composition of high heteroplasmy (5+ samples with MAF ≥1%, 5% or 10%) position sets

High heteroplasmy position set	Total positions	Base composition			
		A	C	G	T
Set 1: *5+ samples with MAF ≥1%*	543	13.6% *(-17.3%)*	4.60% *(-26.7%)*	58.0% *(+44.9%)*	23.8% *(-0.95%)*
Set 2: *5+ samples with MAF ≥5%*	57	17.5% *(-13.4%)*	12.3% *(-19.0%)*	40.4% *(+27.3%)*	29.8% *(+5.11%)*
Set 3: *5+ samples with MAF ≥10%*	20	20.0% *(-10.9%)*	15.0% *(-16.3%)*	30.0% *(+16.9%)*	35.0% *(+10.3%)*

Percentages in brackets indicate deviation from MT genome overall base frequencies of 30.9% A, 31.3% C, 13.1% G & 24.7% T

It is evident that strong sequence composition bias exists among positions where the largest number of high MAF heteroplasmies are observed. Consistent to our previous observations, such heteroplasmic positions are enriched for Gs and at higher MAF thresholds TS, and contain significantly fewer As or Cs compared to the rest of MT genome.

It is interesting to note how increasing the required MAF threshold results in change in position set base composition, with overrepresentation of base T only emerging when we are considering positions with 5 heteroplasmies above at least 5% MAF. This suggests that base G is strongly linked to slightly lower level heteroplasmies whereas T appears to be associated more with higher MAF heteroplasmies.

In terms of the adjacent sequence context, several adjacent sequence context patterns emerge. First, we see an excess of Gs upstream at positions -1 and -2, and an excess of Cs downstream at positions +1 and +2. Secondly, there is a depletion of As in the immediate downstream area between positions +1 and +4 is depleted in As. It is hard to say whether there is a clear change in the strength of this pattern or not between sets. One feature that does clearly change between sets 1, 2 and 3 is the frequency of Cs immediately upstream at position -1 and G immediately downstream at position +1, both of which increase as MAF threshold goes up from 1% in set 1 to 10% in set 3. At the same time, frequency of As at position -1 drops from 38.7% in set 1 which is 7.7% above A baseline frequency to 25% (5.9% below baseline) in set 3.

A caveat with interpreting these observations is the substantial overlap between the three sets of positions investigated. But it is interesting that even though we see some fairly clear shared patterns between the sets as should be expected from the overlap, there are also some changes taking place in the adjacent base composition patterns as MAF threshold changes between sets.

Overlap between measures of site-specific MAF and previous heteroplasmy reports in literature

In the previous sections we discuss different ways of defining high heteroplasmy level positions, which yield different sets of sites. Here we address which measures of heteroplasmy levels show the best correlation to heteroplasmy report frequency in previous literature. We are comparing our results against the combined reports of heteroplasmy across studies reviewed and analysed in **section 5.3**.

This is addressed in **Supplementary figures 5.5.6 S1 and 5.5.6 S2** below, focusing on heteroplasmies identified using the consensus method and minimum MAF threshold 0.1%. **Supplementary figure 5.5.6 S1** looks at the overlaps between positions reported as heteroplasmic in literature at different frequencies and sets of positions we identify as being associated with high MAF based on different criteria. The criteria we use for selecting highest MAF sites include the average observed heteroplasmy MAF of at least 1% or 10% at the site (either on its own or in combination with some minimum number of heteroplasmies that must be detected at that site), or the number of high MAF samples (MAF>1%, >5% or >10%) detected in some minimum number of samples at the site.

Generally, when we look at sets of positions that differ by the minimum number of individuals in which MAF has to reach a specific threshold, both higher number of required individuals as well as higher MAF threshold leads to better overlap with sets of sites previously reported in literature more frequently. In contrast, positions where median MAF are higher do show as good an overlap with more frequently reported sites in literature. This is further supported by correlation data between various measures we use for defining high MAF sites and the literature heteroplasmy report frequencies, shown in **Supplementary figure 5.5.6 S2**, measured both across all our detected heteroplasmic sites as well as specifically for different subsets of high MAF positions.

Overall, based on this data we can reasonably say that there is some link between how often heteroplasmy has been reported at a particular site in literature previously and some of our position MAF measures, however it is the number of high MAF samples we find per position rather than the average MAF across all detected heteroplasmies

at a site that shows better correlation with literature heteroplasmy report frequency. This is very likely a consequence of most previous studies using higher MAF thresholds, which would logically predispose positions where higher MAF heteroplasmies are more common to being reported more frequently, whereas any positions where low levels of heteroplasmy are more likely, if any such exist, would also be less likely to have been reported previously.

5.6 Relationship between the observed mitochondrial heteroplasmy and its genomic context

Thus far we have not addressed in detail the genomic context of our observed heteroplasmy landscapes. While it is probable that the presence of numerous mitochondrial genome copies within a cell could lead to any single mtDNA copy experiencing lower selection constraint compared to the that experienced by the nuclear DNA, the critical importance of the mitochondrial function to the cell energy metabolism means that a sufficient proportion of wild-type mtDNA copies must remain within the cell to ensure mitochondrial function. Thus mitochondrial mutations show threshold effects (Rossignol *et al*., 2003; Craven *et al*., 2017), with variable minimum disease-causing heteroplasmy thresholds reported for different deleterious mtDNA mutations.

Not all mitochondrial genomic contexts are under the same level of selective pressure. For example, certain areas in the mitochondrial genome D-loop region, non-coding intergenic nucleotides and the protein coding gene codon 3rd 'wobble' positions are thought to be more permissive to mutations compared to tRNA, rRNA or protein coding gene 1st and 2nd codon positions (Voets *et al*., 2011; Lake *et al*., 2024). This is evidenced by both higher evolutionary sequence divergence in these regions between vertebrates (Kumar, 1996) as well as higher prevalence of homoplasmic mtDNA variants in the human population (Voets *et al*., 2011).

Genomic context of the identified heteroplasmies has already been touched upon briefly in some of the earlier sections. In this section we shall focus on it in more detail, addressing whether our observed heteroplasmy incidence and levels differ between different functional regions of the mitochondrial genome. We also assess how this compares to the existing heteroplasmy literature.

Considering that heteroplasmic mutations are known to exhibit threshold effects, and in this study we explore heteroplasmy landscapes of MAF as low as 0.1%, the relationships between our observed heteroplasmy landscapes of different genomic contexts and MAF can inform about the selection pressures mitochondrial genome is subject to below the more typically studied higher MAF heteroplasmies. However, this must also be considered in the context of potential artefact presence in our data, with deviation of our observed genomic context specific heteroplasmy landscapes from those emerging in the previous literature a potential starting point for further clarifying sources of artefacts.

In the sections below we explore how our observed heteroplasmy differences between functional elements correspond to expectations based on the expected selection pressure, and assess how this compared to the existing heteroplasmy literature. We also explore the patterns of nonsynonymous to synonymous heteroplasmy ratios both in our data as well as in the previous heteroplasmy literature, and address the differences between heteroplasmy.

5.6.1 Heteroplasmy differences between mitochondrial genome functional elements

It is reasonable to expect that that different mtDNA positions may exhibit different levels of tolerance to heteroplasmy, depending on how functionally important that position is and how significant the effect of even few mutant mtDNA copies at the particular site could be on the overall mitochondrial function (Ruiz-Pesini *et al.*, 2006), as well as whether the mutation represents gain or loss of function. Pathogenic gain-of-function mtDNA point-mutations however appear to be extremely rare, with only one such variant reported in the recent literature (Elliott *et al*., 2008) affecting a tRNA gene, and the vast majority of reported mitochondrial point mutations represent loss-of-function instead.

Broadly, in the protein coding genes the degeneracy of the genetic code implies a higher expected selection constraint at codon 1st and 2nd positions where base changes will result in amino acid change, with higher mutation rates typically observed at the less constrained codon 3rd 'wobble' positions (Bofkin *et al*., 2007). Thus codon 3rd positions as a group generally should be more tolerant to higher heteroplasmy levels.

Selection pressures at protein coding gene 1st and 2nd codon positions may also vary from site to site depending on both how critical the specific encoded amino acid is to the protein function, as well as how chemically different the amino acid encoded by the mutant codon is compared to its wild type counterpart. While we would expect some level of tolerance to exist to low levels of nonsynonymous heteroplasmic mutations as long as a sufficient proportion of wild type mtDNA encoded proteins exists that is able to compensate, once the mutation load exceeds the phenotypic threshold, we would expect the phenotypic effect and consequently selection pressure to be proportional to the mutation load (Nunnari *et al*., 2012).

Although the selection constraints on RNA genes does not follow the same pattern as for protein coding genes, the relationship between how functionally important a given position is and its sensitivity to mutation should still hold true [REF]. Not all positions within the comparatively large rRNA genes will experience the same level of constraint, with strongest purifying selection experienced by specific areas most crucial to ribosome function, while other areas show evidence of adaptive selection (McFarland *et al*., 2004; Ruiz-Pesini *et al*., 2006). Similarly, because the specific loop structures of tRNA molecules are critical to their function, many residues within the mitochondrial tRNA genes are under tight constraint (Vilmi *et al*., 2005). mtDNA variants disrupting tRNA structure are often linked to mitochondrial diseases. A prime example of this is the m.3243 A>G variant – a mutation affecting a highly conserved residue in the Leu(UUR) gene that results in mitochondrial dysfunction and is the most common mtDNA variant associated with MELAS (Mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes) in humans (Goto *et al*., 1990; Chinnery *et al*., 1997).

Unlike the rest of the mitochondrial genome, the mtDNA control region (CR) is non-coding and mostly shows a somewhat higher overall mutation rate compared to the coding regions (Suissa *et al*., 2009). Multiple functional elements involved in the regulation of mtDNA replication and transcription are located within the control region (Suissa *et al*., 2009). Sequence constraint across the CR has been shown to be variable, with some regions, such as Conserved Sequence Block 3, found under tight constraint, while others are highly polymorphic (Lake *et al*., 2024). Hypervariable sequence fragments in particular appear to be mutational hotspots with high rates of heteroplasmy (Stoneking, 2000; Irwin *et al*., 2009). Evidence is emerging linking the CR regulatory functions to its ability to form DNA secondary structures, providing further explanation of the highly variable patterns of CR sequence constraint (Pereira *et al*., 2008).

Finally, intergenic non-coding nucleotides, the smallest functional element group, are scattered throughout the mitochondrial genome outside of the control region and do not have a known function besides being a buffer

between various coding elements. As such, we should expect the non-coding nucleotides as a group to be subject to least selection constraint of all functional element groups.

Based on all this, we should generally expect lower heteroplasmy prevalence to exist across the mtDNA coding regions. The most predictable pattern should be present in protein coding genes, with higher heteroplasmy incidence expected at codon 3rd positions due to the lower selective pressure. We would also expect to see slightly higher find higher heteroplasmy across the control region, but especially in the hypervariable sequence segments. This should be true both in our data as well as across the previous heteroplasmy literature. In this section we assess how well our observed heteroplasmy landscapes fit these predictions.

Figure 5.6.1 F1 summarises the distribution of our observed heteroplasmies across each of these different mitochondrial genome functional element groups, while **Figure 5.6.1 F2** summarises the reported heteroplasmy incidence across 20 previous studies (reviewed earlier in **section 5.3**) in the different mitochondrial genome regions.

Figure 5.6.1 F1: Mitochondrial heteroplasmy incidence observed in different genomic contexts

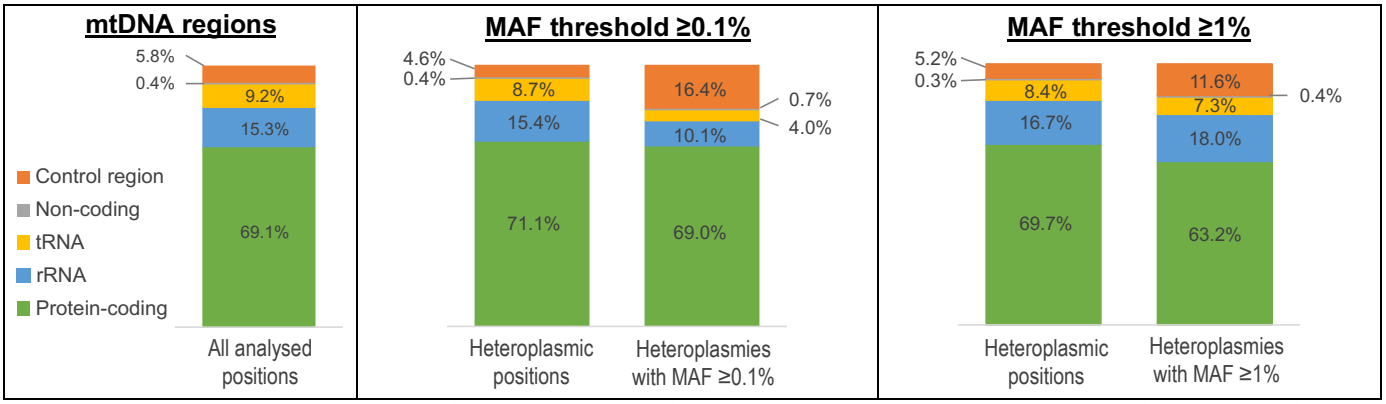
Observed mitochondrial heteroplasmy incidence in different genomic contexts

(a) Heteroplasmies found across different mitochondrial genome functional element groups

MT genome functional element group	Analysed positions[1]	MAF threshold ≥0.1%				MAF threshold ≥1%			
		Heteroplasmic positions[2]	Heteroplasmies Total	Heteroplasmies Per site[3]	HDR[4]	Heteroplasmic positions[2]	Heteroplasmies Total	Heteroplasmies Per site[3]	HDR[4]
Control region	959 (-163)	495 *(50.1%)*	19,848	20.1 *(118)*	7.58	302 *(30.6%)*	1,556	1.57 *(6.44)*	0.59
Non-coding	58	44 *(75.9%)*	875	15.1 *(81.7)*	4.88	19 *(32.8%)*	53	0.91 *(2.21)*	0.30
tRNA	1,504 (-2)	942 *(62.6%)*	4,876	3.2 *(5.5)*	1.02	488 *(32.4%)*	976	0.65 *(1.25)*	0.20
rRNA	2,510	1,662 *(66.2%)*	12,224	4.9 *(10.1)*	1.52	974 *(38.8%)*	2,420	0.96 *(2.15)*	0.30
Protein coding	11,341	7,682 *(67.7%)*	83,334	7.3 *(59.8)*	2.30	4,070 *(35.9%)*	8,479	0.75 *(1.47)*	0.23
codon 1st position	3802	2,582 *(67.9%)*	37,469	9.9 *(82.0)*	3.08	1,474 *(38.8%)*	3,441	0.91 *(1.75)*	0.28
codon 2nd position	3797	2,686 *(70.7%)*	20,899	5.5 *(29.7)*	1.72	1,529 *(40.3%)*	3,068	0.81 *(1.36)*	0.25
codon 3rd position	3796	2,454 *(64.6%)*	25,222	6.6 *(55.4)*	2.08	1,088 *(28.7%)*	2,006	0.53 *(1.21)*	0.17
Overall	16,403	10,804 *(65.9%)*	120,754	7.4 *(58.0)*	2.33	5,842 *(35.6%)*	13,408	0.82 *(2.21)*	0.26

(1) Sites not analysed in red; (2) Heteroplasmic fraction of all analysed sites in brackets; (3) Values averaged across all analysed sites, standard deviation in brackets; (4) Heteroplasmy discovery rate (HDR) expressed as heteroplasmies found per sample per kb analysed

(b) Identified heteroplasmy relative distribution between different mitochondrial genome regions



Column segments represent the proportion of analysed sites, identified heteroplasmic sites or total heteroplasmies found in each of the following mitochondrial genome sequence regions: control region, non-coding nucleotides, tRNA, rRNA and protein coding genes

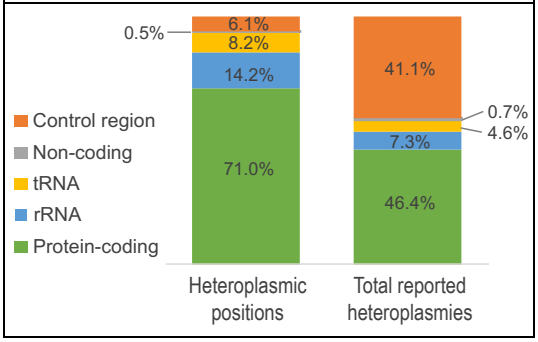
Figure 5.6.1 F1 shows lowest proportion of analysed heteroplasmic sites in the control region. This is however the result of lower number of samples having been analysed for some proportion of CR sites, and when we account for lower sample count, our heteroplasmy discovery rate per kb of analysed sequence is in fact the highest in the control region.

Within the coding regions, we observe lowest heteroplasmy incidence across tRNA genes, and highest incidence across protein coding genes at 0.1% MAF threshold and across rRNA at 1% MAF threshold. Overall we see 6.5 heteroplasmies per analysed site across all coding regions, which is 3 times lower than in the D-loop region. Contrary to expectation, when protein coding genes are broken down by codon positions, higher heteroplasmy incidence is observed at codon 1st position, with lowest incidence at 0.1% MAF threshold found in position 2, and at 1% MAF threshold found in position 3.

Figure 5.6.1 F2: Previously reported mitochondrial heteroplasmy incidence in different genomic contexts

Previously reported mitochondrial heteroplasmy incidence in different genomic contexts

MT genome elements	Heteroplasmic positions[1]	Heteroplasmies[2] Total	Per site	HDR[3] Aggregate	Mean
Control region	682 (69.0%)	65,799	66.6 (348)	0.0114	2.12 (5.35)
Non-coding	51 (87.9%)	1,096	18.9 (48.0)	0.0038	0.82 (2.01)
tRNA	915 (60.8%)	7,355	4.8 (34.5)	0.0010	0.40 (1.19)
rRNA	1,581 (63.0%)	11,770	4.7 (16.4)	0.0009	0.75 (2.40)
Protein coding	7,880 (69.5%)	74,318	6.6 (21.0)	0.0013	0.48 (1.18)
codon 1st position	2,410 (63.4%)	22,382	5.9 (18.9)	0.0012	0.47 (1.05)
codon 2nd position	2,094 (55.1%)	12,611	3.3 (11.2)	0.0007	0.18 (0.38)
codon 3rd position	3,415 (90.0%)	39,673	10.5 (28.5)	0.0021	0.78 (2.30)
Overall	11,098 (67.7%)	160,158	9.8 (89.3)	0.0020	0.61 (1.62)



(1) Sites not analysed in this study excluded, heteroplasmic site fraction given in brackets; (2) Aggregate heteroplasmies reported across 20 reviewed studies. Per site values averaged across analysed sites of each MT genome region, with standard deviation given in brackets; (3) Heteroplasmy discovery rate (HDR) expressed as heteroplasmies reported per sample per kb analysed: Aggregate – HDR determined pooling heteroplasmies reported across all reviewed studies; Mean – Average study-specific HDR, standard deviation between studies given in brackets; (4) Column segments represent proportion of reported heteroplasmic sites or total heteroplasmies in each mtDNA region.

Comparing this to **Figure 5.6.1 F2,** where the heteroplasmy prevalence reported in different contexts in 20 previous studies is summarised, we see that similarly to our data the highest incidence of previously reported heteroplasmies is found in the control region followed by non-coding nucleotides, with the lowest heteroplasmy incidence reported in the coding regions. The differences between regions however are more extreme compared to our data. Across literature the average reported number of heteroplasmies per site in D-loop is 10 times higher than in the coding regions (6.1 per site), while in our data the difference is only 3 times at 0.1% MAF and 2 times at 1% MAF. Heteroplasmies in the CR represent 41.1% of all reported heteroplasmies in literature but only 16.4% and 11.5% of the heteroplasmies we identified at 0.1% and 1% MAF thresholds respectively.

The biggest difference between literature and our data, which could suggest a real concern about the accuracy of our approach, is the observed heteroplasmy distribution between the three codon positions in protein coding genes. Codon 3rd position shows the highest frequency of previously reported heteroplasmies at the average of 10.5 heteroplasmies per site, compared to 5.9 and 3.3 heteroplasmies per site on average at positions 1 and 2 respectively. In contrast, our data at both 0.1% and 1% MAF thresholds show highest heteroplasmy frequency in codon position 1, which is contrary to what we would expect based on presumably different selection pressures at synonymous and non-synonymous sites. This shall be addressed in more depth in the next section.

There are several possible explanations why we might observe different patterns of heteroplasmy incidence at different genomic contexts from those reported in literature. First, in this study we are looking at heteroplasmies at

a much lower minimum MAF threshold compared to most the previous studies. Even tightly constrained positions could be more permissive to such low level heteroplasmy, which could result in our observed differences between our data and previous literature. If this were the case, we might expect to see a higher proportion of very low level heteroplasmies and a lower average heteroplasmy MAF at sites with a higher expected selection pressure (e.g. codon 1st and 2nd positions compared to the 3rd position in protein coding genes). This is however not the case, with little difference found between heteroplasmic 1st, 2nd and 3rd codon position average MAF, and a lack of higher MAF heteroplasmy excess at the codon 3rd positions (see **Supplementary figure 5.6.1 S1**), indicating MAF differences between codon positions as unlikely to be the main explanation for our observed discrepancy between our data and the previous literature.

The second way how the discrepancy between previously reported and our observed heteroplasmy incidence at different codon positions could emerge without contradicting higher selection constraint at 1st and 2nd positions could be via differences in heteroplasmy frequencies observed per heteroplasmic site. High overall heteroplasmy incidence across specific genomic context can emerge when most heteroplasmic positions contain an average number of heteroplasmies. Alternatively, it can also be a result of many sites with just one or two heteroplasmic samples in combination with an extreme heteroplasmy incidence at some small proportion of sites. The latter scenario would not be inconsistent with purifying selection acting across the majority of sites of a particular genomic context despite relatively high average heteroplasmy level across that context. Nevertheless, our data indicates that this is also an unlikely explanation, as we do not see an extreme excess of singleton heteroplasmies among codon 1st and second positions (**Supplementary figure 5.6.1 S2**)

Another possible explanation for the discrepancies between heteroplasmy incidence between previous literature and our data could be misinterpretation of artefacts. Considering that we are looking at MAF as low as 0.1%, this is not an unreasonable possibility. However, if the artefacts were being introduced due to the low MAF threshold, the majority of differences between our data and the literature should be found among the lowest MAF heteroplasmies. However this is not the case, and we see in fact an even more unexpected heteroplasmy incidence pattern across codon positions at the higher 1% MAF threshold than what we see at 0.1% MAF (see **Figure 5.6.1 F1**), suggesting that if the codon position heteroplasmy incidence differences were due to artefacts, these artefacts must be having an even more significant impact on our observed heteroplasmy incidence at the higher 1% MAF threshold.

Possibility that our reported results are a consequence of errors in how the data was handled should also be considered. Precautions were taken throughout the work to test new code, trial data manipulation steps on smaller test datasets, and manually spot-check results throughout the pipelines. While this does not fully eliminate the possibility that our results are due to some fundamental bug in the code, however if such a bug exists, we have not been able to identify it.

Thus at present the cause behind the differences between our results and literature remains an open question, that raises significant concerns about the data presented here.

Previously reported heteroplasmy discovery rate variation between studies

Something not yet addressed is the extent of variability between different reviewed studies, and how this compares with the extent of the differences between our observed data and the results reported in the wider literature. Specifically, focusing on the 20 previous studies reviewed earlier in **Section 5.3**, we can compare the reported heteroplasmy discovery rates (HDR) associated with specific genomic contexts between the different papers. This is addressed in **Supplementary figure 5.6.1 S3**. As the absolute heteroplasmy discovery rates vary between our reviewed studies by several orders of magnitude, it is hard to directly compare the different datasets. Scaling

heteroplasmy discovery rates each study reports in different genomic contexts relative to the study overall heteroplasmy discovery rate however allows for a clearer comparison between the datasets, revealing some clearly shared common patterns between the studies.

Highest heteroplasmy discovery rates are consistently found across the control region (18/20 studies), with the reported control region specific heteroplasmy discovery rates found to be between 1.2 to over 8 times above the study-specific overall heteroplasmy discovery rate, highlighting a considerable variation between heteroplasmy landscapes reported in different research papers.

Of all genomic contexts, HDR at non-coding (NC) nucleotides outside the D-loop varies the most between studies. In some studies, heteroplasmy discovery rate across non-coding nucleotides emerges as the second highest of all regions, while in other studies we see a lower than average reported heteroplasmy counts across these non-coding positions (NC-specific HDR to study overall HDR ratio range: 0 - 5.6). It is worth noting however that the intergenic non-coding nucleotides represent the least abundant distinct genomic context at just 58 positions, and thus is likely to be subject to more noise compared to other more abundant contexts, and especially for studies with lower overall heteroplasmy discovery rates.

Across our reviewed studies the heteroplasmy discovery rates are generally lower across coding regions, with lowest HDR usually observed in either tRNA or rRNA regions depending on the study. In the protein coding genes HDR at codon 3rd position tends to be highest, and while the HDR at codon 2nd position tends to be lowest. However there is considerable variation between studies, with only half of the reviewed studies fully fitting the aforementioned pattern. We see a considerable variation between studies in terms of how much different codon position specific heteroplasmy discovery rates differ from the protein coding gene overall HDR, with for example codon 1st position HDR found to exceed protein overall HDR in 6 studies, and codon 2nd position HDR found above the protein overall HDR in 4 of the reviewed studies.

The differences between studies in terms of HDR variation between genomic contexts does not appear to correlate with the different MAF thresholds employed by the various reviewed papers. Nor is there a clear association with the study size. High levels of variability in the reported heteroplasmy incidence between different genomic contexts is not restricted to the smaller studies which could, presumably, be more vulnerable to noise. In fact, we see considerable differences in the reported HDR patterns across 1st, 2nd and 3rd codon positions between even the largest studies. As noted earlier, our reported HDR pattern across codon positions are quite different from those of the aggregate data across all of our reviewed previous studies. Yet when we compare our data to the spread of the patterns seen across the individual reviewed papers, our results no longer are as obvious an outlier. In a way this is even more concerning.

If the patterns of heteroplasmy incidence across different MT genome functional regions were highly consistent across the reviewed previous literature, our results being different would be an obvious red flag. But at least it would be a clear indication that it is specifically the present study that is problematic. The fact that with all its flaws the present study falls within the range of variability seen across the published heteroplasmy research spanning the past 20 years should, in a way, be even more concerning. It does not make the present study any more valid. Instead it highlights worrying level of inconsistencies across the literature, which also makes it harder to assess the extent of the problems with the present work.

5.6.2 Heteroplasmy patterns in protein coding genes

Figure 5.6.2 F1 summarises the distribution of our observed mitochondrial protein coding gene heteroplasmies of different MAF between different codon positions.

Figure 5.6.2 F1: Observed heteroplasmy distribution between protein coding gene codon positions

Protein coding gene heteroplasmy distribution between codon positions

(a) Observed heteroplasmy relative distribution between codon 1st, 2nd and 3rd positions

Codon position	All heteroplasmies with MAF ≥0.1%	Observed heteroplasmies with MAF level within a specific range					Reported in literature
		0.1 - 0.5% MAF	0.5 - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF	
	N = 83,334	*N* = 68,869	*N* = 5,986	*N* = 5,811	*N* = 1,169	*N* = 1,499	*N* = 74,318
1st	45.0%	45.8%	41.0%	40.7%	39.6%	41.0%	30.1%
2nd	25.1%	22.9%	34.4%	36.6%	37.2%	33.9%	17.0%
3rd	30.3%	31.6%	24.7%	23.2%	23.8%	25.5%	53.4%
dN/dS ratio	*2.33*	*2.19*	*3.06*	*3.38*	*3.28*	*2.97*	*0.89*

(b) Observed heteroplasmic site relative distribution between codon 1st, 2nd and 3rd positions

Codon position	All heteroplasmic positions	Positions with highest observed heteroplasmy MAF level within a specific range					Reported in literature
		0.1 - 0.5% MAF	0.5 - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF	
	N = 7,682	*N* = 2,526	*N* = 1,085	*N* = 2,146	*N* = 703	*N* = 1,222	*N* = 7,880
1st	33.6%	30.2%	31.7%	35.0%	34.4%	39.4%	30.6%
2nd	35.0%	30.8%	34.9%	38.1%	41.0%	34.6%	26.6%
3rd	31.9%	39.5%	33.7%	27.4%	25.3%	26.4%	43.3%
N/S ratio	*2.17*	*1.55*	*1.98*	*2.71*	*3.04*	*2.83*	*1.33*

*dN/dS ratio - nonsynonymous-to-synonymous ratio, calculated as **(a)** the number of heteroplasmies found at codon 1st and 2nd positions divided by all heteroplasmies found across codon 3rd 'wobble' positions or **(b)** the number of codon 1st and 2nd positions with observed heteroplasmies divided by all codon 3rd 'wobble' positions with observed heteroplasmies. Literature data derived from aggregate heteroplasmy reports across our reviewed 20 previous published studies.*

Figure 5.6.2 F1 demonstrates that regardless of MAF threshold used, we observe highest heteroplasmy frequency at codon position 1. This contrasts the aggregate data of previously reported heteroplasmies across 20 earlier studies, and is not consistent with what we would expect based on the assumption of stronger selection constraint at non-synonymous positions.

Our observed heteroplasmy frequency at codon 2nd position is lower than both 1st and 3rd position only when we look at the lowest MAF heteroplasmies (between 0.1% and 0.5%). At MAF thresholds above 0.5%, more than 1/3 of our identified protein coding gene heteroplasmies are found at codon position 2. This shows that differences between our observed heteroplasmy distribution between codon positions compared to that reported in previous literature is unlikely to be due to our dataset constraining a higher proportion of lower MAF heteroplasmies.

Interestingly, when we consider what proportion of our identified heteroplasmic sites represent each codon position, if we split our identified heteroplasmic sites in subsets based on the highest observed MAF at each site, the codon position proportions are different depending on MAF. For heteroplasmic positions where we see slightly higher proportion of codon 3rd positions compared to 1st or 2nd positions are found among heteroplasmic sites where only very low level heteroplasmies are found (MAF between 0.1% and 0.5%). In contrast, among the heteroplasmic sites where maximum observed MAF is highest (>10%), codon 1st positions are most common while 3rd positions are least common. What is the significance of this is unclear.

In contrast to our observed heteroplasmy counts, the aggregate heteroplasmy counts reported in literature across the 20 previous studies we reviewed in **section 5.3** show a much better fit with the expected differences in selection constraint between different codon positions. The aggregate previously reported heteroplasmy dataset shows both a much higher heteroplasmy frequency across protein coding gene 3rd codon position (consistent with higher selective pressure at non-synonymous positions) as well as slightly higher heteroplasmy counts at codon position 1

compared to 2 (consistent with strongest constraint at 2nd position due to any mutation there resulting in an amino acid change).

The data however is not as clear cut when we look at the heteroplasmy counts at different codon positions reported by specific studies (**Supplementary figure 5.6.2 S1**). When individual studies are considered separately, we see a considerable variation in the nonsynonymous-to-synonymous heteroplasmy ratios (dN/dS), with dN/dS >1 found in 10 of the 19 reviewed studies, 3 of which show dN/dS >2. Similarly, it is not always the case that codon 2nd position shows lower heteroplasmy frequency compared to codon 1st position, with codon 2nd to 1st position heteroplasmy count ratios >1 emerging in 3 reviewed studies.

This means that multiple reviewed studies report considerably higher dN/dS ratios individually compared to the dN/dS ratio we calculated from the aggregate heteroplasmy dataset across all 20 studies. The discrepancy stems from large variation in study size and reported overall heteroplasmy discovery rates. This leads to some of the papers contributing a disproportionate number of heteroplasmies to the aggregate dataset and therefore having a bigger impact on the aggregate dN/dS ratio.

Figure 5.6.2 F2 panel A summarises the variation between heteroplasmy dN/dS ratios in 20 previous studies and how this compares to the dN/dS ratios of our identified heteroplasmies above 0.1% and 1% MAF thresholds. **Panel B** illustrates the dN/dS ratios across protein coding genes grouped by their respective OXPHOS complexes. In addition to data from the 20 studies review in detail in **section 5.3** we also compare our data to a further, more recent 2023 study (Hong *et al*. 2003).

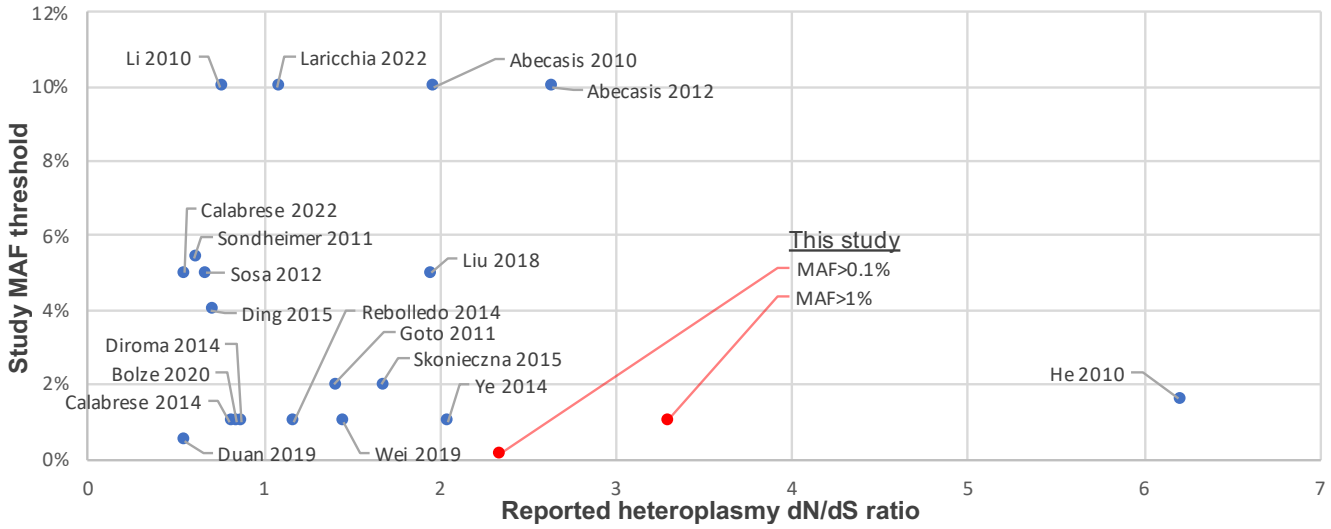
While higher, our identified heteroplasmy dN/dS ratio at 0.1% MAF threshold follows a similar pattern to dN/dS ratios based on the aggregate literature heteroplasmies, with lower dN/dS ratio values observed for complexes I and IV. This is consistent with previous literature, where highest selection constraint is reported for complexes I and IV (Stewart *at al.*, 2007, Lake *et al.*, 2023). When we compare this to individual study dN/dS ratios, the dN/dS values from our data fall within the range of values from the previously published studies, albeit in the upper quartile.

OXPHOS complexes V and III, thought to be under weaker selection pressure (Stewart *at al.*, 2007, Lake *et al.*, 2023), do indeed show higher dN/dS ratios compared to complexes I and IV both in our data as well as across the previous literature, albeit all complex-specific dN/dS ratio values are higher in our data when compared to the literature aggregate.

Figure 5.6.2 F2: Nonsynonymous to synonymous heteroplasmy ratio

Nonsynonymous-to-synonymous heteroplasmy ratio

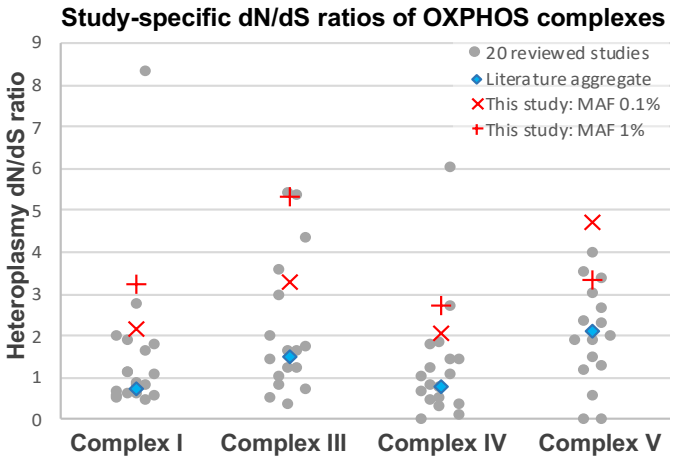
(a) Nonsynonymous-to-synonymous heteroplasmy ratios found in the reviewed literature



(b) Nonsynonymous-to-synonymous heteroplasmy ratios of different OXPHOS complexes

Reported in LITERATURE	20 reviewed studies Study average[1]	Aggregate[2]	Reported by Hong *et al.*, 2023[3]
Complex I	1.5 *SD 1.8*	0.7	0.87
Complex III	2.1 *SD 1.6*	1.5	1.32
Complex IV	1.2 *SD 1.4*	0.8	1.20
Complex V	2.0 *SD 1.2*	2.1	1.92

Observed in THIS STUDY	Heteroplasmies at MAF thresholds ≥0.1%	≥0.5%	≥1%	≥5%	≥10%
Complex I	2.2	3.1	3.2	3.1	3.1
Complex III	3.3	4.8	5.3	4.8	4.8
Complex IV	2.1	2.7	2.7	2.5	2.2
Complex V	4.7	3.5	3.4	3.3	3.0



dN/dS – nonsynonymous-to-synonymous heteroplasmy ratio, calculated as total heteroplasmies found across the protein coding gene codon 1st and 2nd positions divided by total heteroplasmies found in codon 3rd 'wobble' positions. Protein coding genes grouped by oxidative phosphorylation (OXPHOS) complex: Complex I – ND1, ND2, ND3, ND4L, ND4, ND5, ND6; Complex III – CYB; Complex IV – CO1, CO2, CO3; Complex V – ATP8, ATP6. Heteroplasmy dN/dS ratios from literature: (1) the mean and standard deviation of dN/dS ratios of heteroplasmies reported in each of the 20 previous studies reviewed earlier in section 5.3; (2) aggregate dN/dS ratio based on pooled heteroplasmies reported across the 20 reviewed studies; (3) dN/dS ratios reported by Hong et al., 2023, a more recent study not among the 20 originally reviewed papers, reporting mitochondrial heteroplasmies with MAF≥5% in a WGS dataset of 194,871 UK Biobank samples

5.6.3 Relationships between heteroplasmy and the base type in different genomic contexts

The differences between rCRS bases in terms of the number of heteroplasmies found per analysed position in different genomic contexts are addressed in **Figure 5.6.3 F1** below, which demonstrates that largest differences between our reported heteroplasmy counts per site and those reported in the previous literature are found in the MT genome control region and the non-coding nucleotides outside of D-loop, mainly for bases A and T. In contrast, in the coding regions our identified base type specific average heteroplasmy counts per site follow patterns of the previously reported heteroplasmies more closely.

Figure 5.6.3 F1: Heteroplasmies per site found in different genomic contexts at positions with different reference bases

Average number of heteroplasmies found per site found in different genomic contexts

(a) Heteroplasmies with MAF ≥0.1% identified per site in this study at different rCRS base types

Position rCRS base	Overall	MT genome functional elements					Protein codon positions		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	9.5	47.1	7.0	1.8	2.7	9.1	14.3	4.2	7.3
C	1.5	5.2	1.1	0.7	0.7	1.4	1.0	0.7	2.0
G	12.5	11.9	14.0	8.8	12.9	13.3	12.4	9.8	24.2
T	9.4	9.2	92.7	4.3	6.7	10.4	13.0	8.4	12.2

(b) Heteroplasmies reported per site in previous literature at different rCRS base types

Position rCRS base	Overall	MT genome functional elements					Protein codon positions		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	8.2	54.2	26.8	3.7	3.4	5.8	5.0	3.0	8.0
C	5.2	36.9	8.4	2.0	2.2	3.2	1.8	1.6	5.2
G	16.6	63.3	31.3	11.1	8.5	16.2	13.0	5.6	52.4
T	13.9	131.2	29.6	5.6	6.4	7.4	6.4	4.1	17.3

*(**a**) Heteroplasmies with MAF ≥0.1% identified using consensus read frequency based method; (**b**) Heteroplasmies reported across 20 previous studies reviewed earlier in section 5.3; Values represent total identified/reported heteroplasmies divided by number positions within each genomic context. Literature heteroplasmy data at 163 positions not analysed here is excluded. MAF – minor allele frequency; CR – control region, NC – non-coding nucleotides*

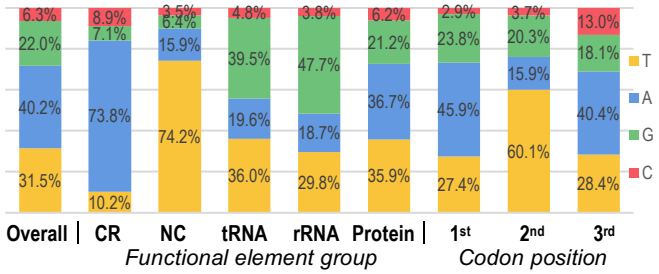
As the base composition of various mitochondrial genome functional elements differs, **Figure 5.6.3 F2** compares MT genome functional elements in terms of base composition of our identified heteroplasmic positions as well as the major allele base composition of all of our identified heteroplasmies. In this case, both the base composition of the heteroplasmies themselves as well as the heteroplasmic positions found in each MT genome region are compared against the MT genome baseline.

Figure 5.6.3 F2: Observed heteroplasmy major allele base composition differences between gnomic contexts

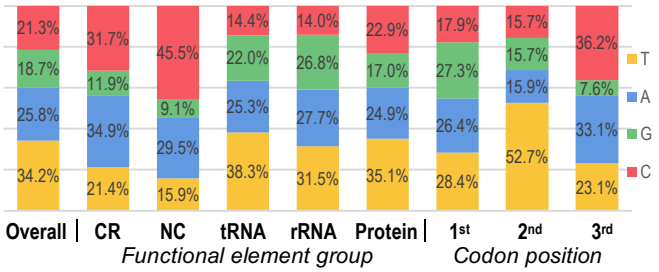
Identified heteroplasmy base composition differences between different genomic contexts

(a) Base composition of heteroplasmies with MAF ≥0.1% found in different genomic contexts

Observed heteroplasmy major allele base composition



Observed heteroplasmic position base composition



Major allele base composition deviation from MT baseline

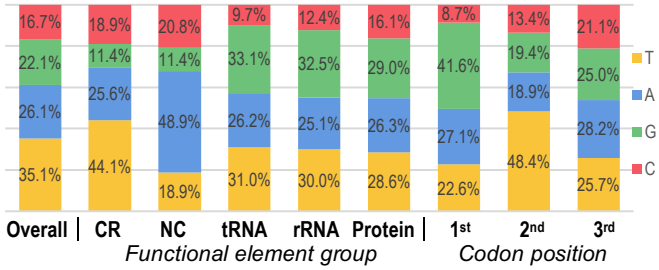
Major allele	Overall	MT genome functional elements					Codon position		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
C	-25.0%	-24.5%	-42.8%	-18.6%	-22.0%	-27.1%	-25.9%	-24.5%	-29.7%
G	+9.1%	-4.3%	-0.7%	+25.0%	+29.7%	+9.5%	+4.9%	+8.9%	+13.1%
A	+9.1%	+40.9%	-18.5%	-15.4%	-15.7%	+7.0%	14.3%	-4.9%	+3.7%
T	+6.8%	-12.1%	+62.1%	+9.1%	+8.0%	+10.6%	+6.6%	+20.5%	+12.9%

Position base composition deviation from MT baseline

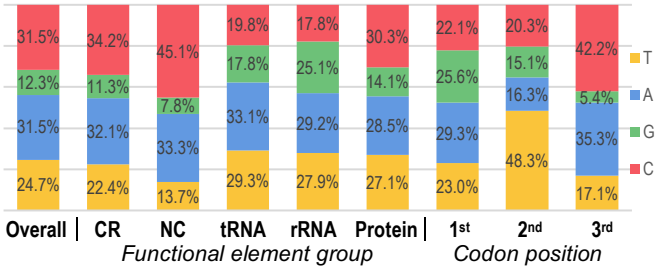
rCRS base	Overall	MT genome functional elements					Codon position		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
C	-10.0%	-2.5%	-1.1%	-9.0%	-11.9%	-10.3%	-10.9%	-12.4%	-6.5%
G	+5.7%	0.0%	+2.2%	+7.4%	+8.8%	+5.3%	+8.5%	+4.3%	+2.6%
A	-5.2%	+3.5%	-4.9%	-9.8%	-6.7%	-4.8%	-5.2%	-4.9%	-3.7%
T	+9.5%	-1.0%	+3.8%	+11.4%	+9.7%	+9.8%	+7.6%	+13.1%	+7.5%

(b) Base composition of heteroplasmies reported in previous literature

Reported heteroplasmy major allele base composition



Reported heteroplasmic position base composition



Major allele base composition deviation from MT baseline

Major allele	Overall	MT genome functional elements					Codon position		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
C	-14.5%	-14.5%	-25.5%	-13.8%	-13.5%	-17.1%	-20.1%	-14.7%	-21.6%
G	+9.1%	-0.1%	+4.3%	+18.6%	+14.5%	+17.2%	+22.8%	+8.0%	+20.0%
A	-5.0%	-7.2%	+14.5%	-8.9%	-9.2%	-3.4%	-4.5%	-2.0%	-8.5%
T	+10.4%	+21.8%	+6.8%	+4.0%	+8.2%	+3.3%	1.8%	+8.7%	+10.2%

Position base composition deviation from MT baseline

rCRS base	Overall	MT genome functional elements					Codon position		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
C	+0.2%	0.0%	-1.5%	-3.7%	-8.1%	-3.0%	-6.6%	-7.9%	-0.6%
G	-0.7%	-0.7%	+0.9%	+3.3%	+7.1%	+2.4%	+6.7%	+3.7%	+0.5%
A	+0.5%	+0.6%	-1.1%	-1.9%	-5.1%	-1.2%	-2.2%	-4.5%	-1.4%
T	+0.1%	+0.1%	+1.7%	+2.4%	+6.1%	+1.8%	+2.2%	+8.7%	1.6%

*(**a**) Heteroplasmies with MAF ≥0.1% identified using consensus read frequency based method; (**b**) Heteroplasmies reported across 20 previous studies reviewed earlier in section 5.3; Graph columns represent Graph columns visualise heteroplasmy major allele (left) or heteroplasmic position (right) relative base composition within a specific genomic context. Table values represent the observed genome heteroplasmy (left) or heteroplasmic position (right) base composition deviation from the overall MT genome base composition*

Figure 5.6.3 F2 demonstrates relative to MT genome baseline a depletion of heteroplasmic positions with reference bases A and C and an excess of heteroplasmic positions where reference base is G or T across the coding regions. In contrast, the only base type with an excess of heteroplasmic positions in the control region is A. Generally base composition of heteroplasmic positions reported in previous literature broadly follow the same patterns as our identified heteroplasmic sites, but differs in its extent. Our reported heteroplasmic site composition deviation from MT baseline tends to be more extreme across all genomic contexts, while in the literature we see bigger differences between different contexts, but lower overall deviation from the MT baseline compared to our data. Nevertheless the directions in which the heteroplasmic position base composition deviates from the baseline match nearly perfectly between our data and the previous literature.

One of the most interesting observations is that both in the literature as well as in our data stronger base composition deviation from MT genome baseline is found in regions under expected higher selective pressure (rRNA/tRNA and protein coding gene codon 1st and 2nd positions).

In contrast to heteroplasmic position base composition, when we look at the aggregate heteroplasmy major allele base composition in each genomic context, the direction of different region heteroplasmy composition deviation from the baseline match less well between previously reported and our identified heteroplasmies.

The only regions where we see a full match between our data and the literature in terms of the direction in which heteroplasmy major allele composition deviates from the baseline are tRNA, rRNA and protein coding region codon 2nd positions. In all other regions some bases deviate from the baseline in opposite directions between our data and the literature. In the control region this is A (up in our data, down in literature) and T (down in our data, up in literature). In non-coding nucleotides the only major allele base in excess among our identified heteroplasmies is T, while in literature we also see an excess of heteroplasmies with major alleles G and A. In protein coding genes the base that differs most between literature and our data is A, which we see in excess among codon 1st and 3rd position heteroplasmy major alleles in, while in literature such heteroplasmies are depleted across all codon positions.

The only heteroplasmy major allele base that deviates from MT genome baseline consistently in the same direction both in our data and in literature across all genomic contexts is C, which incidentally is also the base with strongest correlation between number of heteroplasmies per positions found in our study and reported in previous literature.

To address how much of the heteroplasmy major allele base composition variation between genomic contexts can be explained by the combination of different MT genome functional element base composition and different heteroplasmy detection rates at positions with different reference base, we can estimate the expected heteroplasmy base composition assuming base-type specific heteroplasmy detection rate independence from genomic context, and compare this to our actual data. Given our overall heteroplasmy detection rates at 0.1% MAF threshold of 3.01, 0.47, 3.96 and 2.97 per 1000 analysed base pairs for A, C, G and T respectively, we can estimate how many heteroplasmies of each major allele base type we should be observing in each genome functional element region, based on the base composition of each region. Comparison between this and our actual observed heteroplasmy base composition is visualised in **Supplementary figure 5.6.3 S1**, which also provides major allele base composition breakdown of our identified heteroplasmy subsets of different MAF.

Supplementary figure 5.6.3 S1 shows that the biggest base composition differences between subsets of heteroplasmies emerge when comparing heteroplasmies with MAF between 0.1%-0.5% to those with MAF between 0.5% and 1%. Genomic contexts with the biggest major allele base composition differences between these two subsets of heteroplasmies are control region (A 81.6% to 24.2%), non-coding nucleotides (T: 79.5% > 22.7%, A: 12.9% > 59.1%) and protein coding genes (A: 41.1% > 16.1%, G: 17.7% > 36.3%). In protein coding genes the most extreme A and G frequency change takes place in codon 1st position (A: 51.7% > 15.9%, G: 18.3% > 50.7%). Heteroplasmy major allele base composition differences between any of the higher MAF heteroplasmy subsets are much less dramatic.

Comparing the observed base-specific heteroplasmy rates between genomic contexts, we see a clear indication of variable heteroplasmy probability for the same base types, which also changes depending on the heteroplasmy MAF, suggesting interaction of multiple different mechanisms determining the resulting observed heteroplasmy landscape. Interestingly, while at the higher MAF we do see an increase in heteroplasmy detection rates for all base types in some of the regions where we would expect the lowest constraint (control region non-coding nucleotides and protein coding gene codon 3rd positions), no single base type displays a consistently elevated heteroplasmy rate across all three expected lower constraint genomic contexts. Perhaps the most striking in this regard is G, which is associated with higher heteroplasmy rates in protein codon 3rd positions while at the same time exhibiting the lowest heteroplasmy probability of all genomic contexts in the control region.

Major allele base type and the nonsynonymous-to-synonymous heteroplasmy ratio

We have already addressed this in the nonsynonymous and synonymous mutation ratio, one of the most commonly employed measures of selection constraint, in the context of our overall identified heteroplasmy sets. Here we discuss the differences between specific heteroplasmy major allele base types and the heteroplasmy dN/dS ratios.

Figure 5.6.3 F3 below summarises our observed nonsynonymous-to-synonymous heteroplasmy ratios for different major allele base types and at different MAF.

Heteroplasmies at Gs and Ts contribute strongly to high dN/dS ratio by virtue of being far more common in codon 1st and 2nd positions compared to 3rd position. This means that if we have a high probability of heteroplasmy affecting these two bases, we are going to have higher dN/dS ratio.

Expected dN/dS ratio is calculated assuming the same heteroplasmy probability across all instances of a base in the protein coding genes. Below 0.5% MAF level, nonsynonymous heteroplasmy proportion is higher than expected purely based on mtDNA base composition only for heteroplasmies with major allele A, while dN/dS ratio of heteroplasmies with major allele G shows a considerable reduction below the expected level.

Above 0.5% MAF the dN/dS ratios of major allele A heteroplasmies are very close to those expected in absence of any selection. For C and G major allele heteroplasmies we do see a slight reduction in dN/dS ratios. And, interestingly when the major allele is T, between MAF of 0.5% and 5% we actually see a higher than expected proportion of non-synonymous heteroplasmies.

Bases do not exist in isolation, and it is reasonable that the adjacent sequence context could be biasing the heteroplasmy probability sufficiently to overpower existing purifying selection acting on the non-synonymous sites. However this does not answer why evidence for comparable levels of heteroplasmy at non-synonymous sites is absent from previously published literature. As discussed earlier however, we do observe a considerable variation between different previous studies in terms of the dN/dS ratios, with the level of inter-study variability being relatively comparable to our data.

Figure 5.6.3 F3: Nonsynonymous-to-synonymous heteroplasmy ratios depending on position reference base type

Observed nonsynonymous-to-synonymous heteroplasmy ratios

(a) Observed heteroplasmy dN/dS ratios

Heteroplasmy major allele	Reported in literature	All observed with MAF ≥0.1%	Heteroplasmy subsets with specific MAF levels				
			0.1% - 0.5% MAF	0.5% - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF
A	0.8	2.0	2.1	1.5	1.4	1.5	1.4
C	0.4	0.6	0.5	0.7	1.1	1.0	1.2
G	1.2	2.9	2.4	4.7	5.5	6.3	4.8
T	1.1	3.2	3.0	4.4	4.3	3.7	3.5
Overall	0.9	2.3	2.2	3.1	3.4	3.3	3.0

(b) Observed heteroplasmic position N/S ratios

Heteroplasmic position rCRS base	Reported in literature	All observed with MAF ≥0.1%	Heteroplasmic site subsets with specific maximum observed MAF levels				
			0.1% - 0.5% MAF	0.5% - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF
A	0.9	1.4	1.3	1.7	1.4	1.5	1.3
C	0.7	1.0	1.0	0.9	1.1	1.1	1.2
G	5.2	6.2	5.8	4.6	7.9	7.3	5.1
T	2.7	3.8	4.3	3.7	3.9	3.9	3.3
Overall	1.3	2.2	1.5	2.0	2.7	3.0	2.8

(c) Observed heteroplasmy dN/dS ratio deviation from expected

Heteroplasmy major allele	Expected dN/dS	All observed heteroplasmies	Heteroplasmy subsets with specific MAF levels				
			0.1% - 0.5% MAF	0.5% - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF
A	1.4	+0.59	+0.65	+0.03	-0.02	+0.09	-0.04
C	1.3	-0.77	-0.83	-0.63	-0.27	-0.34	-0.17
G	6.3	-3.41	-3.95	-1.60	-0.83	-0.05	-1.47
T	3.9	-0.72	-0.90	+0.45	+0.33	-0.20	-0.47
Overall	2.0	+0.31	+0.18	+1.04	+1.36	+1.26	+0.95

(d) Observed heteroplasmic position N/S ratio deviation from expected

Heteroplasmic position rCRS base	Expected N/S	All observed with MAF ≥0.1%	Heteroplasmic site subsets with specific maximum observed MAF levels				
			0.1% - 0.5% MAF	0.5% - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF
A	1.4	-0.06	-0.16	+0.23	-0.04	+0.07	-0.12
C	1.3	-0.34	-0.35	-0.48	-0.25	-0.26	-0.13
G	6.3	-0.08	-0.49	-1.71	+1.55	+1.02	-1.18
T	3.9	-0.10	+0.40	-0.20	-0.06	-0.02	-0.61
Overall	2.0	+0.15	-0.47	-0.04	+0.69	+1.02	+0.81

*(**a**) Nonsynonymous-to-synonymous heteroplasmy ratios (dN/dS), expressed as total heteroplasmies found at codon 1st and 2nd positions divided by heteroplasmies found at codon 3rd 'wobble' positions; (**b**) Nonsynonymous-to-synonymous heteroplasmic position ratios (N/S), expressed as all heteroplasmic codon 1st and 2nd positions divided by all heteroplasmic codon 3rd positions; (**c**,**d**) Deviation between the observed dN/dS or N/S ratios from those expected based protein coding gene codon position overall base composition differences*

5.7 Prevalence of mtDNA variants linked mitochondrial disease

Although we have thoroughly described and analysed the patterns of our observed mitochondrial heteroplasmy landscapes, one aspect not touched upon so far is the prevalence of known disease-associated mtDNA mutations. The mtDNA mutations that result in mitochondrial disease are well studied, with extensive research available on their prevalence patterns. The purpose of this section is to compare our observed prevalence of known mtDNA mutations with literature, and explore any differences between our results and wider body of the existing scientific research. For this purpose, we shall focus on 42 known disease-associated variants spanning 41 mtDNA positions (Ryzhkova *et al.*, 2018).

These variants are listed in **Figure 5.7 F1** below, which compares our observed variant prevalence of these variants to that reported in three other sources:

- Mitomap database, which collates variant frequencies across 61,134 GenBank mtDNA sequences (Lott *et al*., 2013)
- GnomAD database, where mtDNA variants are reported from 56,434 individual WGS sequences (Laricchia *et al*., 2022)
- A 2008 study reporting prevalence of specific mtDNA variants across 3168 neonatal cord-blood samples (Elliott *et al*., 2008)

Mitomap variant prevalence data does not provide information about heteroplasmic variants, whereas GnomAD data does, distinguishing between homoplasmies and heteroplasmies with MAF ≥10%. The third dataset by Elliott *et al.*, 2008, also distinguishes between homoplasmies and heteroplasmies, although here the heteroplasmy detection threshold is variable, with authors estimating it between ~1.5% to 10% MAF depending on the sample. This dataset is also more restricted, with data only available for 9 of the 42 disease-associated mtDNA variants that we are looking at.

Figure 5.7 F1: Prevalence of mtDNA variants associated with mitochondrial disease

Population prevalence of mtDNA variants associated with mitochondrial diseases

mtDNA variant	Locus	Disease association	Reported variant prevalence							Observed variant prevalence					
			GnomAD			Mitomap	Elliott *et al.*			MAF ≥0.1%			MAF ≥1%		
			Total	Hom.	Het.	Total	Total	Hom.	Het.	Total	Hom.	Het.	Total	Hom.	Het.
1095 T>C [1]	MT-RNR1	AIHD	0.04%	0.04%	0.002%	0.11%		-		0.25%	0.06%	0.19%	0.06%	0.06%	0%
1494 C>T	MT-RNR1	AIHD	0.02%	0.02%	0.002%	0.01%		-		0.06%	0%	0.06%	0.03%	0%	0.03%
1555 A>G	MT-RNR1	AIHD	0.13%	0.11%	0.02%	0.14%	0.07%	0%	0.07%	0.09%	0%	0.09%	0.03%	0%	0.03%
*1658 T>C	MT-TV	CPEO/PEO	0.01%	0.01%	0%	0.01%		-		0.16%	0%	0.16%	0%	0%	0%
3200 T>C [1]	MT-RNR2	NIDDM	0.09%	0.09%	0.002%	0.13%		-		0.28%	0.09%	0.19%	0.16%	0.12%	0.03%
*3242 G>A [2]	MT-TL1	NIDDM	0%	0%	0%	0%		-		1.50%	0%	1.50%	0.34%	0%	0.34%
3243 A>G	MT-TL1	MELAS | CMEM | CPEO/PEO	0.01%	0%	0.01%	0.02%	0.14%	0%	0.14%	0.78%	0%	0.78%	0.25%	0%	0.25%
3252 A>G [1]	MT-TL1	NIDDM	0%	0%	0%	0%		-		0.06%	0%	0.06%	0%	0%	0%
3256 C>T	MT-TL1	MELAS	0%	0%	0%	0%		-		0.03%	0%	0.03%	0.03%	0%	0.03%
3260 A>G	MT-TL1	CMEM	0%	0%	0%	0%		-		0%	0%	0%	0%	0%	0%
3264 T>C [1]	MT-TL1	NIDDM	0.002%	0%	0.002%	0%		-		0.47%	0%	0.47%	0.12%	0%	0.12%
3271 T>C	MT-TL1	MELAS	0%	0%	0%	0%		-		0.09%	0%	0.09%	0%	0%	0%
3316 G>A [1]	MT-ND1	NIDDM	0.48%	0.46%	0.03%	0.94%		-		2.00%	0.94%	1.06%	1.33%	1.19%	0.13%
*3394 T>C	MT-ND1	NIDDM	0.93%	0.91%	0.02%	1.31%		-		0.56%	0.41%	0.16%	0.44%	0.44%	0%
*3421 G>A [1]	MT-ND1	CPEO/PEO	0.23%	0.22%	0.01%	0.13%		-		0.91%	0.16%	0.75%	0.44%	0.16%	0.28%
*3460 G>A [1]	MT-ND1	LHON	0.002%	0%	0.002%	0.05%	0.11%	0%	0.11%	0.34%	0%	0.34%	0.03%	0%	0.03%
*4269 A>G	MT-TI	CMEM	0%	0%	0%	0.002%		-		0.06%	0%	0.06%	0.03%	0%	0.03%
4833 A>G	MT-ND2	NIDDM	0.15%	0.15%	0%	1.04%		-		0.69%	0.62%	0.06%	0.63%	0.62%	0%
7587 T>C	MT-CO2	CMEM	0%	0%	0%	0%		-		0.12%	0%	0.12%	0.03%	0%	0.03%
8296 A>G	MT-TK	CMEM	0.05%	0.04%	0.002%	0.06%		-		0.03%	0%	0.03%	0.03%	0%	0.03%
8344 A>G	MT-TK	MERRF	0.01%	0%	0.01%	0.01%	0%	0%	0%	0.25%	0%	0.25%	0.12%	0%	0.12%
8348 A>G [2]	MT-TK	CMEM	0.11%	0.09%	0.02%	0.22%		-		0.19%	0%	0.19%	0.03%	0%	0.03%
8356 T>C	MT-TK	MERRF	0.004%	0%	0.004%	0%		-		0.03%	0%	0.03%	0.03%	0%	0.03%
8363 G>A	MT-TK	MERRF | CMEM	0.002%	0%	0.002%	0%		-		0.09%	0%	0.09%	0.06%	0%	0.06%
8993 T>C	MT-ATP6	LS | NARP	0.004%	0.002%	0.002%	0.003%		-		0.09%	0%	0.09%	0.06%	0%	0.06%
8993 T>G	MT-ATP6	LS | NARP	0%	0%	0%	0.01%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9804 G>A [2]	MT-CO3	LHON	0.39%	0.36%	0.03%	0.30%		-		0.47%	0.12%	0.34%	0.16%	0.12%	0.03%
9957 T>C	MT-CO3	CMEM	0.03%	0.03%	0%	0.08%		-		0.12%	0.09%	0.03%	0.09%	0.09%	0%
9997 T>C	MT-TG	CMEM	0%	0%	0%	0.002%		-		0.06%	0%	0.06%	0%	0%	0%
10197 G>A [1]	MT-ND3	MELAS | LS	0%	0%	0%	0.01%		-		1.81%	0%	1.81%	0.47%	0%	0.47%
*11778 G>A	MT-ND4	LHON	0.04%	0.02%	0.02%	0.31%	0.11%	0.04%	0.07%	0.28%	0%	0.28%	0.06%	0%	0.06%
*12192 G>A [1]	MT-TH	CMEM	0.40%	0.40%	0.005%	0.21%		-		0.53%	0.09%	0.44%	0.13%	0.09%	0.03%
12297 T>C [1]	MT-TL2	CMEM	0.16%	0.16%	0.002%	0.08%		-		0.41%	0.03%	0.37%	0.09%	0.06%	0.03%
12315 G>A [1]	MT-TL2	CPEO/PEO	0%	0%	0%	0%		-		0.47%	0%	0.47%	0.16%	0%	0.16%
13042 G>A [1]	MT-ND5	MELAS | MERRF	0%	0%	0%	0.003%		-		0.47%	0%	0.47%	0.19%	0%	0.19%
13513 G>A [2]	MT-ND5	LS	0%	0%	0%	0.002%	0%	0%	0%	0.16%	0%	0.16%	0.03%	0%	0.03%
*14459 G>A [1]	MT-ND6	LHON	0.002%	0%	0.002%	0.005%	0%	0%	0%	0.31%	0%	0.31%	0.09%	0%	0.09%
14484 T>C	MT-ND6	LHON | CMEM	0.07%	0.05%	0.01%	0.11%	0.11%	0.07%	0.04%	0.19%	0%	0.19%	0.03%	0%	0.03%
14577 T>C	MT-ND6	NIDDM	0.09%	0.09%	0.004%	0.70%		-		0.06%	0%	0.06%	0%	0%	0%
*15059 G>A [1]	MT-CYTB	CMEM	0%	0%	0%	0%		-		0.53%	0%	0.53%	0.13%	0%	0.13%
15257 G>A	MT-CYTB	LHON	1.28%	1.27%	0.002%	1.53%		-		0.56%	0.31%	0.25%	0.41%	0.41%	0%
15533 A>G	MT-CYTB	MELAS	0.01%	0.01%	0%	0.01%		-		0.03%	0%	0.03%	0%	0%	0%

Mitochondrial disease-associated mtDNA variant prevalence *expressed as a percentage of population individuals carrying the variant allele. Variant set adapted from Ryzhkova et al., 2018. Genomic context: rRNA, tRNA, protein. Colour scale: lowest (white) to highest (red) prevalence.* ***Reported variant prevalence****: Population frequency data from GnomAD database (Laricchia et al., 2022), Mitomap database (Lott et al., 2013), and study by Elliott et al., 2008.* ***Observed variant prevalence****: Proportion of samples analysed in the present study found carrying disease-associated allele at minimum frequency ≥0.1% MAF and ≥1% MAF. Allele frequencies estimated using consensus method.* ***Total*** *– variant overall prevalence,* ***Hom.*** *– homoplasmic variant prevalence,* ***Het.*** *– heteroplasmic variant prevalence. * Positions with less than 3202 analysed samples: 1658 (n=3201), 3421 (n=3198), 3460 (n=3197), 4269 (n=3149), 10197 (n=3198), 11778 (n=3197), 12192 (n=3199), 15059 (n=3196).* ***1,2*** *– Positions potentially affected by NUMTs, where (**1**) disease-associated allele matches a known NUMTs-derived variant or (**2**) an alternative allele at the same position matches a known NUMTs-derived variant (3242 G>T, 8348 A>C, 9804 G>C, 13513 G>T). NUMTs-affected variant data obtained from Li et al., 2012.* ***Mitochondrial diseases****: LHON – Leber hereditary optic neuropathy; LS – Leigh syndrome; MIDD – maternally inherited diabetes and deafness; MELAS – mitochondrial myopathy, encephalopathy, lactic acidosis and stroke-like episodes; MERRF – myoclonic epilepsy associated with ragged red fibres; NIDDM – noninsulin-dependent diabetes mellitus; NARP – neuropathy, ataxia and pigmentary retinopathy; CPEO/PEO – chronic progressive external ophthalmoplegia syndrome/ progressive external ophthalmoplegia; AIHD – Aminoglycoside induced hearing disorders; CMEM – Cardiomyopathy and encephalomyopathy*

Figure 5.7 F1 reveals considerable differences between all compared studies in terms of how many of the 42 disease-associated variants are found in any instances in the source data. Of the 42 disease associated mtDNA variants investigated, for 11 there are no reported instances in Mitomap dataset, and 14 variants are entirely absent from GnomAD dataset. Of the 9 variants studied by Elliott *et al.*, no instances were found for 4. In contrast to this, at the lowest 0.1% MAF threshold, we found instances of all but one of the 42 disease-associated variants within our data set. Some of these variants are however only present as very low-level heteroplasmies. Increasing the minimum MAF threshold to 1% excludes all instances of 7 of the disease-associated variants, while raising the minimum MAF to 10% filters out all instances of further 15 variants. Thus only 23 of the 42 disease-associated variants are represented in our dataset as either homoplasmies or heteroplasmies with MAF ≥10%.

There are also some notable differences in terms of the actual variant prevalence patterns reported by different sources. Focusing on the previously published datasets first, GnomAD data is particularly useful as it provides variant prevalence information separately for homoplasmies and heteroplasmies, which across the disease-associated variant set show a positive correlation, although it is relatively weak (Pearson's r = 0.37). Mitomap variant prevalence, which does not distinguish between homoplasmies and heteroplasmies, correlates better with the prevalence pattern of GnomAD homoplasmies (Pearson's r = 0.86) than it does to the GnomAD heteroplasmies (Pearson's r = 0.40), which would be consistent with the majority of Mitomap variant data coming from homoplasmies. Variant prevalence reported by Elliott *et al.* also distinguish between heteroplasmic and homoplasmic variants. This much smaller study however is restricted to only 9 of the disease associated variants we are interested in, which makes it harder to compare it to the larger databases. On one hand, there are some variants observed by Elliott *et al*, such as 1555 A>G or 14484 T>C, where the prevalence estimates match fairly well with those of GnomAD and Mitomap. Similarly, some variants that Elliott *et al*. did not detect in their study at all, such as 8993 T>G, 13513 G>A and 14459 G>A, are also very rare or fully absent from one or both of the larger databases. On the other hand however, we see some major prevalence estimate disagreements too. The most striking example of this is the MELAS-associated variant 3243 A>G, where the 0.14% population prevalence observed by Elliott *et al*. is an order of magnitude above the prevalence estimates based on the Mitomap and GnomAD data. Another example is 3460 G>A, where the 0.11% prevalence estimate by Elliott *et al*. is twice as high as the Mitomap estimate and over 50 times above the GnomAD estimate.

Overall, the variant prevalence reported by Elliott *et al*. correlates somewhat better to Mitomap data (r = 0.51) than it does to GnomAD data overall (r = 0.33). Looking at GnomAD homoplasmies and heteroplasmies separately, Elliott *et al*. results correlate better with the heteroplasmy prevalence data (r=0.53) than the homoplasmy data (r=0.26). This is be consistent with the fact that 80% of the variants identified by Elliott *et al*. are heteroplasmies.

This leads us to the next set of observations – the differences between the ratios of homoplasmies versus heteroplasmies in the different datasets. One of the immediate differences between our data and GnomAD is the much lower proportion of heteroplasmies among GnomAD variants. Across the 42 disease associated variants, only 4.3% of all variant instances found in GnomAD are heteroplasmic. This is even lower across the wider GnomAD dataset, where the proportion of heteroplasmies drops to just 1.2% of all reported variants. In contrast, looking at our data, 81% of all our identified disease-associated variant instances are heteroplasmies. To some extent this is due to the different MAF thresholds used. It is however important to note the differences in MAF thresholds between the datasets. Restricting our data to variants with MAF≥1%, the proportion of heteroplasmies drops to 46.5%, and raising minimum MAF to 10%, decreases our overall observed heteroplasmy proportion to 12.2%. Although more comparable, this is still three times higher than the heteroplasmy proportion in GnomAD.

The extreme differences in heteroplasmy proportion between GnomAD and Elliott *et al*. data are only partially explained by the different MAF thresholds, as at least a third of the heteroplasmies reported by Elliott *et al*. exceed

the 10% MAF threshold. A more important factor is the fact that Elliott *et al.* focuses on a smaller subset of specific variants, which may be potentially more deleterious than some of the others among the wider set of 42 disease-associated variants. Consistent with this hypothesis, looking across just the 5 variants Elliott *et al*. found in their dataset, 26.8% of all instances in GnomAD are heteroplasmies. This is even more extreme in our dataset, where we found no homoplasmic instances of these 5 variants. The link between how deleterious variants are and the relative proportion of heteroplasmies is also supported by the fact that the GnomAD heteroplasmy proportion across the 42 disease-associated variants is more than 3 times higher than it is across the full database.

Despite some of the differences between heteroplasmic variant prevalence in GnomAD and the present study being related to the different MAF thresholds, it is not the full explanation. Firstly, as already noted, even at the same MAF threshold as the one used in GnomAD, we still see a higher proportion of heteroplasmies among the disease-associated variants. Secondly, there are also differences between the databases and our data in terms of the actual patterns of how frequent various disease-associated variants are.

This is illustrated in **Figure 5.7 F2** below, which gives Pearson correlation coefficient estimates for pairwise comparisons between our observed mtDNA disease associated variant prevalence to the prevalence patterns in GnomAD and Mitomap. The correlation coefficients shown are calculated based on both the overall variant prevalence as well as that of heteroplasmies and homoplasmies separately where available. Elliott *et al*. dataset was not included due to its more limited number of investigated variants.

The correlation between the observed and reported variant prevalence across the full set of 42 disease-associated mtDNA variants is addressed in **Panel A.** which reveals only a weak positive correlation between the overall variant prevalence we observe at the lower 0.1% MAF threshold to the frequencies in Mitomap (r = 0.31) or GnomAD (r = 0.28) when looking at all variants with MAF≥0.1%. Although this increase slightly , although this does increase for

At the lowest 0.1% MAF threshold our observed disease-associated variant prevalence is only weakly correlated to Mitomap (r = 0.31) or GnomAD (r = 0.28) overall variant frequencies data. And although the correlation increases for the higher MAF≥1% threshold (Mitomap: r=0.60, GnomAD: r=48), separating heteroplasmies from homoplasmies shows that this is largely due to our observed homoplasmies, where our observed prevalence is a relatively good match with the databases. Our observed heteroplasmic variant prevalence however is not a good match at all, with only a very weak positive correlation to GnomAD heteroplasmies found at MAF ≥0.1% (r = 0.11). This also disappears entirely at the higher MAF ≥1% threshold (r = -0.04).

These are concerning findings that require explanation. The set of 42 variants that we focus on here are obviously deleterious and associated with a range of mitochondrial diseases (Ryzhkova *et al*., 2018). We would therefore expect such variants to be strongly selected against. Given the threshold effects characteristic to mitochondrial diseases (Rossignol *et al*., 2003), it is not unreasonable for the selection against homoplasmies and heteroplasmies with higher MAF levels to be stronger compared to the very low MAF heteroplasmies. When combined with disease-associated variants being found in different genomic contexts, with different modes of action and different phenotypic thresholds, this could potentially lead to heteroplasmies of different MAF levels following different prevalence patterns. If so, then indeed we might at 0.1% MAF see a different pattern of heteroplasmies compared to GnomAD which only looks at variants above 10% MAF. Logically however, applying a higher MAF threshold to our data should yield a set of variants with more similar prevalence pattern to the GnomAD homoplasmies. But this does not happen. Instead, at the 1% MAF threshold our observed heteroplasmy prevalence pattern correlates even less to GnomAD heteroplasmies. This suggests that the prevalence differences we see are not a result of high- and low-MAF heteroplasmies being under different levels of selection, and it is more likely that there is actually a bleed-through of some other signal in our dataset besides true heteroplasmy.

The primary candidates for such a signal are NUMTs – nuclear DNA sequences that are paralogous to the mitochondrial genome fragments that may be mistaken for heteroplasmies (Calvignac *et al.*, 2011). Comparing our investigated set of 42 disease-associated mtDNA variants with a database of known NUMTs-affected mtDNA variants (Li *et al.*, 2008), 14 variants directly match a known NUMTs-affected variant. Further 4 are found at mtDNA positions where another alternative allele is a known NUMTs-affected variant, leaving only 24 of the 42 disease associated variants without a potential NUMTs match.

Figure 5.7 F2: Correlation between observed and reported disease associated-variant prevalence patterns

Correlation between the observed and previously reported disease-associated mtDNA variant population prevalence: a potential impact from NUMTs-derived artefacts

(a) Correlation between observed vs reported prevalence of all 42 disease-associated variants

Pearson's *r*		Observed variants with MAF ≥0.1%			Observed variants with MAF ≥1%		
		All variants	Homoplasmies	Heteroplasmies	All variants	Homoplasmies	Heteroplasmies
GnomAD	All variants	0.28	0.59	0.05	0.48	0.60	-0.15
	Homoplasmies	0.27	0.58	0.04	0.48	0.59	-0.15
	Heteroplasmies	0.27	0.42	0.11	0.39	0.43	-0.02
Mitomap	All variants	0.31	0.76	0.00	0.60	0.75	-0.20

(b) Correlation between observed vs reported prevalence of 18 NUMTs-affected variants

Pearson's *r*		Observed variants with MAF ≥0.1%			Observed variants with MAF ≥1%		
		All variants	Homoplasmies	Heteroplasmies	All variants	Homoplasmies	Heteroplasmies
GnomAD	All variants	0.33	0.72	0.06	0.56	0.70	-0.18
	Homoplasmies	0.33	0.71	0.06	0.55	0.69	-0.17
	Heteroplasmies	0.28	0.69	0.02	0.53	0.67	-0.20
Mitomap	All variants	0.49	0.96	0.15	0.81	0.95	-0.14

(c) Correlation between observed vs reported prevalence of 24 variants unaffected by NUMTs

Pearson's *r*		Observed variants with MAF ≥0.1%			Observed variants with MAF ≥1%		
		All variants	Homoplasmies	Heteroplasmies	All variants	Homoplasmies	Heteroplasmies
GnomAD	All variants	0.54	0.62	0.14	0.65	0.70	-0.22
	Homoplasmies	0.54	0.63	0.14	0.64	0.70	-0.23
	Heteroplasmies	0.37	0.09	0.43	0.19	0.09	0.29
Mitomap	All variants	0.65	0.83	0.09	0.80	0.87	-0.28

(d) Internal correlation between the prevalence of heteroplasmies versus homoplasmies

Pearson's *r*		All disease-associated variants	18 NUMTs-affected variants	24 unaffected variants
GnomAD variants		0.37	0.78	0.24
Observed variants	MAF ≥0.1%	0.18	0.25	-0.01
	MAF ≥1%	-0.04	-0.01	-0.27

***Correlation between disease-associated mtDNA variant prevalence observed in the present study and the frequencies reported in GnomAD (Laricchia et al., 2022) and Mitomap (Laricchia et al., 2022) databases**. Pearson correlation coefficients (Pearson's r) shown. Positive correlation in green, negative correlation in red. Observed variant prevalence measured as the proportion of 3202 analysed samples found carrying the disease-associated allele at a minimum frequency ≥0.1% MAF or ≥1% MAF. Allele frequencies estimated using consensus method. Disease-associated mtDNA variant set based on Ryzhkova et al., 2018. **(a)** Correlation across all 42 disease-associated variants; **(b)** Correlation across 18 variants potentially affected by NUMTs; **(c)** Correlation across 24 variants unaffected by NUMTs; **(d)** Correlation between prevalence of heteroplasmic versus homoplasmic variants: comparison between NUMTs-affected and unaffected variant subsets in GnomAD and the present study. Variant NUMTs-affected vs unaffected status based on Li et al., 2012. NUMTs-affected disease-associated variants: 1095T>C, 3200T>C, 3242G>A, 3252A>G, 3264T>C, 3316G>A, 3421G>A, 3460G>A, 8348A>G, 9804G>A, 10197G>A, 12192G>A, 12297T>C, 12315G>A, 13042G>A, 13513G>A, 14459G>A, 15059G>A. NUMTs – nuclear sequences of mitochondrial origin.*

These sets of potentially NUMTS-affected versus unaffected variants are labelled in **Figure 5.7 F1,** and we explore the differences between these variant sets in terms of our observed prevalence correlation with Mitomap and GnomAD data in **Figure 5.7 F2 panels B and C**. Pearson correlation coefficients are again shown for pairwise comparisons of the variant prevalence patterns in our data to GnomAD or Mitomap. **Panel B** shows this specifically for the subset of 18 potentially NUMTs affected variants, while **panel C** focuses on the 24 unaffected variants.

Comparing the behaviour of the two variant subsets, there are notable differences, especially when looking at the correlation between our observed heteroplasmy prevalence with that of heteroplasmies in GnomAD. With MAF ≥0.1% threshold, correlation to GnomAD heteroplasmies is essentially non-existent for NUMTs affected variants (r=0.02), and at the higher MAF ≥1% threshold even a weak inverse correlation appears (r= -0.20). Whereas for the variant set not affected by NUMTs we do actually see a positive correlation between GnomAD and our observed heteroplasmy prevalence patterns at both MAF≥0.1% (r=0.43) and MAF≥1% MAF (r=0.29) thresholds, although the latter is weaker.

This alone strongly points to misinterpreted NUMTs-derived variants as a likely source of some of the differences between heteroplasmy prevalence patterns between our data and GnomAD database. But there are some important details. First, which of the datasets is the one affected by NUMTs? The fact that any correlation between heteroplasmy prevalence in the two studies is restricted to the variant subset less likely to be NUMTs-derived supports NUMTs as the source of the differences between studies. It does not in itself however identify which of the two compared datasets the NUMTs signal is coming from. It could be one, it could be the other, or it could be both. Although for the latter to be true, NUMTs artefacts would have to be distributed differently through both datasets, as otherwise we should be seeing correlation between variant prevalence in GnomAD and our data due to sharing of NUMTs signals.

Second, even though the prevalence NUMTs-affected variant heteroplasmies in GnomAD does not correlate to prevalence of heteroplasmies we observe, it does correlate relatively strongly to our observed homoplasmy prevalence. This is the case however only when looking at the NUMTs-affected variant subset (r=0.69), but not when we look at NUMTs-unaffected variants (r=0.09). This observation prompts a comparison between the prevalence of heteroplasmies and the prevalence of homoplasmies within each of the studies, with the results shown in **panel D** of **Figure 5.7 F2.**

Interestingly, GnomAD data shows a fairly strong positive correlation between heteroplasmy and homoplasmy prevalence specifically across the NUMTs-affected variants (r=0.79), whereas this is much weaker for the unaffected variants (r=0.24). Looking at our results, at the 0.1% MAF threshold there is only a weaker weak correlation between heteroplasmy and homoplasmy prevalence across the NUMTs-affected variants (r=0.25) and no correlation at all across the variants unaffected by NUMTs (r=-0.01). With the higher MAF 1≥% threshold, any correlation between prevalence of heteroplasmies and homoplasmies disappears across the NUMTs-affected variants, while a weak negative correlation between these variable emerges across variants unaffected by NUMTs. These observations have some key implications.

It is evident that there are marked differences between the potentially NUMTs-affected versus unaffected mtDNA variants in terms of their prevalence patterns, and this is true both in our data as well as in GnomAD. That strongly suggests some presence of NUMTs-derived artefact signals is both these data sets. That however allows us to explore differences between the variants that are more versus less likely to be impacted by NUMTs. One such difference is the trend that both our data and GnomAD data follows, where the link between prevalence patterns of heteroplasmies and homoplasmies appears to be weaker for the variants unaffected by NUMTs. Assuming that the variant classification as unaffected by NUMTs by Li *et al*. is correct, and these are indeed less likely to be erroneous

signals, a larger proportion of these variants should logically be true heteroplasmies. Which would then imply that the lower observed correlation between heteroplasmic and homoplasmic variant prevalence is a characteristic feature of real variants. Although initially counterintuitive, it does actually make sense that, at least for highly deleterious variants we have been focusing on, there would indeed be a weaker relationship between heteroplasmy and homoplasmy prevalence among true variants than there would necessarily be among NUMTs-derived artefacts. This may of course not be true for more neutral variation.

While the data we have addressed here is limited to just a small number of mtDNA mutations, it illustrates a potential use case for further exploration of variant prevalence patterns across potentially NUMTs-affected versus unaffected sites. Quantifiable characteristics that distinguish NUMTs signals from genuine variants could provide an extremely useful indicator of data contamination level. Should a more thorough investigation confirm that our suggested relaxed relationship between homoplasmy and true heteroplasmy prevalence of deleterious variants is indeed a feature linked to genuine signals rather than NUMTs-derived ones, this could potentially serve as a measurable parameter for assessing NUMTs contamination. Without further verification and in depth study however, this remains only a speculation.

5.8 Chapter discussion

In this chapter we have presented the population level landscapes of the low level mitochondrial heteroplasmy with minor allele frequencies as low as 0.1%. At this MAF threshold we identify 120,754 putative heteroplasmies across mtDNA sequences of 3002 individuals, which an average of 37.5 heteroplasmic sites SD: 19.0 identified per sample at median MAF of 0.21% (IQR: 0.14%-0.37%). Use of our proposed approach for context dependent sequencing error component reduction via bi-directionally unsupported excess read fraction exclusion results in on average 9.6 fewer heteroplasmies per sample with median MAF by 0.05% lower compared to the more standard method.

Validity of the proposed CSE component minimisation strategy

One of the key aims of this chapter is the assessment of our proposed strategy for minimising the impact of context dependent sequencing errors in low level heteroplasmy analyses. Our data presented here clearly demonstrates that our employed excess read removal results in less extreme observations, whether that be heteroplasmy base composition, heteroplasmy levels or discovery rate. Furthermore, removing excess read fraction not only results in reduced MAF, it also leads to lower MAF variability between samples, which is encouraging in the context of unexpectedly high level of context-dependent sequencing error rate variability between samples demonstrated earlier in chapter 4. Overall, the various facets of analysis discussed through the chapter paint a picture of excess read removal as a valid and a potentially useful strategy for handling CESs that cannot be filtered out otherwise.

It is important to note however that our proposed CSE-reduction strategy via excess read removal does not fully eliminate all of CSE impact, and might remove true signal. Instead it is a heuristic relying on error rate effects in most instances being different between the sequencing directions, that allows us to exclude context dependent error components that are different between sequencing directions. Because of this, there are some potential problem cases that need to be kept in mind:

Scenario 1: A sequencing error affecting the major allele base at a true heteroplasmic site produces an incorrect base call identical to the minor allele, resulting in erroneous inflation of the true minor allele frequency. In this case however, as long as the CSE profile in the opposite sequencing direction is different, consensus frequency based heteroplasmy measurement method will correctly remove the erroneous heteroplasmy estimate inflation.

Scenario 2: A sequencing error affecting the minor allele base at a true heteroplasmic site produces an incorrect base call identical to the major allele, resulting in an erroneous back-correction essentially reducing the full extent of the true minor allele frequency and the true heteroplasmy level underestimation.

Scenario 3: Sequencing errors affecting a homoplasmic site in both sequencing directions by chance introducing incorrect calls of a complementary base pair would result in erroneous identification of a non-existent heteroplasmy. In this case the estimated level of such incorrect apparent heteroplasmy would be that of the lower of the two direction specific error rates.

Scenarios 2 and 3 do indeed present problem cases where our proposed consensus frequency based heteroplasmy detection approach fails to eliminate the erroneous reads. In both these cases however, a more standard approach where MAF is estimated based on total read frequencies would not result in a more accurate true heteroplasmy estimate either. Therefore existence of such possible problem cases does not negate the potential benefit of a consensus frequency based approach in reducing error impact on the heteroplasmy level measurements.

Lastly, there is a further problem scenario related to the technology rather than biology. Software bugs resulting in incorrect calculations and mishandling of data may be introducing false signals or removing real signals. To prevent this, precautions were taken in testing new code, running data manipulation step on dummy test datasets with as

many edge cases as one could think of, incorporating intermediate output steps and checksums, and manually spot-checking results. All data manipulation and analyses pipelines were re-run until no further code errors could be identified. We have not found further result-altering bugs after the final datasets were produced. Nevertheless, the possibility of undiscovered software errors remaining cannot be fully excluded, which would invalidate this study.

Characteristics of the observed heteroplasmy landscapes

We demonstrate a link between heteroplasmy and both the heteroplasmic site reference base type as well as its adjacent sequence context base composition. There are however differences depending on whether we look at all heteroplasmies with minimum 0.1% MAF or only at the subset where MAF is at least 1%.

Using minimum MAF threshold of 0.1% the base most frequently affected by heteroplasmy is A, whereas heteroplasmies affecting G and T become more common when minimum MAF threshold is raised to 1%. When we look at the previous literature, we also see a base composition bias of the reported heteroplasmies, heteroplasmies at bases G and T are much more common compared to A and C. Interestingly though, there are significant differences in the previously reported heteroplasmy base composition between different studies, with some similarly to our data reporting excess of heteroplasmies at base A, while others do not.

The correlation between our observed and previously reported heteroplasmy incidence depends on the base type. Our observed heteroplasmy incidence correlates best with the existing literature for bases C and G, whereas for base A we see no correlation to literature at all when using minimum MAF threshold of 1%. For this base some correlation to literature only appears at minimum MAF threshold of 1%..

In fact, at the minimum MAF threshold of 0.1% our reported heteroplasmy incidence landscape shows multiple heteroplasmy hotspots, characterised by low-level heteroplasmies (MAF between 0.1% and 1%) observed in a large proportion of analysed samples. These hotspot sites show strong base composition bias towards A and to a lesser extent T, while the surrounding sequences are enriched in C and depleted in Gs. Interestingly, there is no evidence of the adjacent sequence context directionality, with both upstream and downstream areas around the hotspots showing equal bias towards C.

At the higher 1% MAF threshold these heteroplasmy hotspots disappear. We also observe base composition change among the positions with highest heteroplasmy incidence, with reference base G becoming the most frequent base at such sites. The adjacent sequence contexts however still remain enriched in C and depleted in G, albeit this is less extreme, and some asymmetry between upstream and downstream regions begins to emerge.

Not only is the heteroplasmy incidence linked to the DNA base type, so is the MAF level, with higher probability of observing multiple high MAF heteroplasmies at positions where the reference sequence base is G or T.

Heteroplasmy in different genomic contexts

In this chapter we show that heteroplasmy discovery rates differ between different genomic contexts both in our study as well as in the previous literature. Broadly, our data matches the previous literature in terms of highest heteroplasmy discovery rates being observed in the mtDNA D-loop region, while lower overall heteroplasmy frequencies are found across the coding regions. There are however considerable differences between our reported heteroplasmy landscapes and existing previous literature in terms of the nonsynonymous versus synonymous heteroplasmy ratios in the protein coding genes, with much lower apparent selection against nonsynonymous heteroplasmic mutations emerging from our analyses compared to previous literature.

Lower minimum MAF thresholds employed in this study do not explain this, as we actually observe higher heteroplasmy dN/dS ratios among heteroplasmy subsets with higher MAF. Generally, the biggest differences in heteroplasmy discovery rate patterns across genome functional elements emerge when comparing heteroplasmies below and above 0.5% MAF threshold, hinting to this as the likely point where it changes what the dominant mechanisms contributing to the apparent heteroplasmy landscape are.

Overall, the complexity of our presented heteroplasmy landscapes strongly suggest that our observed heteroplasmy patterns are most likely arising through interaction of multiple mechanisms. The fact that our data shows good correlation to previous literature in some aspects supports genuine heteroplasmy as being one of the mechanisms. At the same time, other aspects of our data where our results diverge from the literature, such as the very low level heteroplasmy hotspot sites, or the aforementioned differences in nonsynonymous to synonymous heteroplasmy ratios highlight the necessity for caution when interpreting our data, and points to a not-insignificant presence of errors or artefacts in our data

It is however important to consider the existing variability between heteroplasmy landscapes reported by different studies. Whenever we look at previous studies individually, we see a considerable variability, whether in terms of heteroplasmy base composition, discovery rates of dN/dS ratios, with the differences between our data and the previous literature being comparable to the differences between individual studies. This extent of variability between the existing studies and the fact that our study despite its many faults falls within that range, should be a significant cause for alarm.

6 Discussion and conclusions

At the outset of this work, the principal aim was to investigate the detectable human mitochondrial heteroplasmy patterns in the context of genome maintenance. We speculated that, compared to the nuclear genome, variable DNA sequence vulnerability to damage on one hand and differences in damage repair efficiency on the other hand could more readily leave observable signatures in the mitochondrial heteroplasmy landscapes. Considering that off-target data generated in unrelated sequencing experiments may contain large amounts of mitochondrial sequences as a by-product that can be extracted and studied (Picardi *et al*., 2012), this could potentially make the mitochondrial genome a cost-effective readout system for detecting changes in genome maintenance processes. This original rationale was borne from preliminary unpublished work in our lab having hinted to some link between the BRCA1 protein, which is involved in nDNA repair and is known to also localise to mitochondria (Coene *et al*., 2005), and what appeared to be changes in mtDNA heteroplasmy. Through the course of the project and upon discussion with reviewers it has however emerged that this rationale was erroneous, and mtDNA is unlikely to be useful as a model for processes affecting nDNA. Nevertheless, the area of mitochondrial heteroplasmy research is fascinating and rapidly developing, with relevance for mitochondria diseases, cancer, evolutionary biology, ecology, and forensics (Salas *et al*., 2001; Ju *et al*., 2014; Elorza *et al*., 2021; Calabrese *et al*., 2022; Wang *et al*., 2023).

When this study was originally proposed, we did not intend to dedicate a major part of the work to systematic analysis of context dependent sequencing errors. Rather, the questions relating to error landscapes evolved naturally from the emergent obvious necessity to understand how pervasive the context dependent sequencing error patterns were, and to find out whether it is even possible to separate heteroplasmy from errors in datasets such as these produced in experiments not specifically targeting the very low level mitochondrial variation. We are now able to look back to the original intended strategies and questions with the hindsight of having thoroughly analysed context dependent sequencing error pattern and having applied the gained insights to detection, measurement and analyses of the human mitochondrial genome heteroplasmy landscape in a high coverage whole genome sequence dataset.

Implication for applying insights about CSEs gained on one dataset to heteroplasmy analysis in a different dataset

The two major parts of this study were carried out on different mtDNA sequence datasets, with the 1000 genomes project (G1K) phase 3 dataset being used in sequence context dependent error profile analyses, and a higher coverage extended G1K phase 3 cohort dataset produced by New York Genome Center (NYGC) used for the subsequent identification and analyses of low-level mitochondrial heteroplasmy landscapes.

Prior to pivoting to the NYGC higher depth sequences we did briefly attempt to apply our proposed sequence data clean-up strategy to our original, lower coverage G1K phase 3 dataset. Not too surprisingly however, it became clear quite early on that the older dataset, which forms the basis of work presented in chapter 4, lacked sufficient depth of high quality mitochondrial data necessary in order to resolve low level variants and draw meaningful conclusions about heteroplasmy landscapes. While the much higher depth NYGC dataset ultimately used for the analyses presented in chapter 5 offers obvious advantages in terms of sequence depth and quality, we do need to address some of the limitations this brings in terms of how applicable our findings may be between the two major aspects of this work presented in chapters 4 and 5.

The NYGC dataset was produced on a different Illumina system (NovaSeq 4000) compared to the original G1K dataset (HiSeq 2500). While NovaSeq and HiSeq systems are fairly similar, there are some differences, the main being 2 dye chemistry used in NovaSeq versus 4 dyes used in the HiSeq platform. This has been cited as a potential

source for overconfidence of G calls, although other studies indicate that the quality differences between the two platform outputs are generally small (Zhou *et al*., 2019). A recent study systematically comparing context specific error profiles across various Illumina sequencing platforms reports comparable overall error rates between NovaSeq and HiSeq platforms, but a considerably higher overrepresentations of errors associated specifically with the 'GGT' trimer in HiSeq system compared to any other Illumina platform (Stoler *et al*., 2021). This certainly could be consistent with the lower than perhaps expected overrepresentation of 'GGT' patterns among the putative heteroplasmies flagged as erroneous when using our proposed heteroplasmy detection approach which excludes excess read fractions lacking proportional support in both sequencing directions. If 'GGT'-type patterns are not as CSE prone in NovaSeq than it is in HiSeq, then a smaller proportion of excess reads should be found at GGT sites than might have been expected based on our assessment of most error prone motifs with HiSeq platform.

These differences between platforms on which datasets used for sequencing error rate and heteroplasmy landscape analyses do not invalidate our approach of presumed CSE-derived MAF component subtraction from the observed putative heteroplasmy landscapes, that is the core principle of the data analysis presented in the final results chapter. At no point do we directly apply our estimated motif-specific sequencing error rate data from chapter 4 to our heteroplasmy analysis presented in chapter 5. Instead, we simply apply the same principles to likely error-derived putative heteroplasmy identification and removal as we did in the earlier chapter for identifying the likely CSE-derived signal component, except in chapter 5 we invert this to extract the signal component least likely to be CSE-derived. Thus, although the exact error profiles may be different between the two sequencing platforms, the strategy for minimising CSE impact in the heteroplasmy level estimation itself is equally applicable regardless of the platform that the variant the data was generated on.

That being said, as a next step, repeating the work of CSE pattern and variation analysis on the high coverage data could be very illuminating about the levels of error patterns variability. Our results demonstrated a high level of CSE proneness variation between samples, but we do not know whether such level of variability is the norm across sequencing experiments. Another interesting question that the higher coverage dataset might allow one to answer would be to test how the context dependent error rate measurements change depending on what initial read and mapping quality control thresholds are set in the first place.

Feasibility of studying low level mitochondrial heteroplasmy in datasets not specifically designed for it

We successfully demonstrate that even though the dataset employed in this study was not specifically designed for studying mitochondrial heteroplasmy, the population-level heteroplasmy landscapes we estimate follow broadly similar incidence, MAF distribution and base composition patterns to those reported across the previous literature. In specifically designed experiments with additional steps, such as for example duplex sequencing allowing to trace reads back to the source molecules, this error component could be reduced further. The use of our proposed excess read removal approach results in considerable reduction of the presumed sequence context-dependent error component from the raw sequence data, as evidenced by reduction in both the observed MAF and inter-sample variability. The latter is of particular note. One issue we originally observed in chapter 3 was the large variation between samples in terms of the raw non-reference read frequencies at any given site. In chapter 4 we also showed remarkable variability in context dependent error rates observed between samples for the same motifs, suggesting that a significant proportion of the non-reference read frequency variability seen in chapter 3 may likely be a result of highly variable sample proneness to sequencing errors. Unlike chapters 3 and 4, the heteroplasmy landscapes presented in chapter 5 are based on the analyses of a much higher quality and read depth dataset where far more stringent read and mapping quality control was possible. Nevertheless, we again see that exclusion of excess reads presumed to derive through CSEs results in a considerable reduction of variability in the observed MAF between

samples. This indicates that even with the more stringent read and mapping quality control used, relatively high levels of variability remain between samples in terms of the context-specific sequencing error rates in this higher coverage dataset, thus highlighting the important appropriate CSE-derived data component filtration approach in heteroplasmy analyses regardless of the dataset used.

Extreme low-level heteroplasmy frequency peaks observed in this study absent from previous literature

While the heteroplasmy landscapes we report broadly follow similar patterns to those derived observed across the previous literature, there are some differences that must be addressed. The first is our observed hotspot areas of extremely high frequencies of very low MAF (0.1% - 0.5%) heteroplasmies hotspot areas, that are not corroborated by the existing literature. There is essentially no correlation between literature heteroplasmy reports and the incidence of our observed heteroplasmies ≥0.1% MAF (r=0.06), whereas once variants <0.5% MAF are excluded, positive correlation to literature (MAF≥0.5%: r=0.43, MAF ≥1%: r=0.54) emerges.

The very lowest level heteroplasmies are far more likely to be at positions where the reference sequence base is A. Only 8.3% of all heteroplasmic variants affecting A have MAF≥0.5%, compared to 24.6%, 31.6% and 22.3% of variants affecting bases C,G and T respectively (**Figure 6 F1**). This is also very obvious when comparing the base composition of sites where the highest number of heteroplasmies above 0.1% MAF versus above 1% MAF are observed, with A being by far the most common base type among the former while G is the most common base among the latter.

Figure 6 F1: Major allele base type relationship to observed heteroplasmic variant MAF

Proportion of all observed heteroplasmies represented by variants within specific MAF range

Major allele base type	Heteroplasmy MAF level				
	0.1% - 0.5% MAF	0.5% - 1% MAF	1% - 5% MAF	5% - 10% MAF	≥10% MAF
A	91.7%	3.4%	3.2%	0.7%	0.9%
C	75.4%	10.9%	8.7%	2.2%	2.8%
G	68.4%	12.6%	13.4%	2.5%	3.1%
T	77.7%	9.2%	8.9%	1.8%	2.4%
Overall	81.2%	7.7%	7.6%	1.5%	2.0%

In the literature mtDNA mutations are generally reported to exhibit strong strand bias, likely as a consequence of the asymmetric nature of mtDNA replication, with T>C and G>A variants being more common on L strand while C>T and A>G are more prevalent on the heavy strand (Matkarimov *et al*., 2020). Based on this, across the population higher rates of heteroplasmy should be expected at positions where the reference (L strand) sequence base is T or G. Our observations at MAF above 0.5% are generally consistent with this, but the very lowest MAF heteroplasmy hotspots are a clear outlier, that require an explanation.

Even when including the lowest MAF heteroplasmies, our observed heteroplasmy transition-to-transversion (Ti/Tv) ratios at positions with reference bases C (10.1) and G (31.5) fall within or at least close to the range expected of mitochondrial mutations (Wang *et al*., 2015), whereas the Ti/Tv ratios at positions with reference bases A (0.30) and T (1.54) are much lower than would be expected of genuine mitochondrial mutations and could be an indicator or sequencing errors. In fact, sequencing patterns specific to Illumina platforms are reported to include higher frequencies of transversions, with the most common miscalls A>C and T>G (Abnizova *et al*., 2012) being transversions affecting A and T bases, strongly warranting caution when interpreting the putative heteroplasmic A and T variants below 0.5% MAF as true heteroplasmies.

If indeed the observed very low MAF variants affecting As and Ts represent artefacts, the question remains as to what mechanism can leads to the observed pattern of such extreme low MAF variant hotspots as observed in our data, where only a relatively small proportion of all mtDNA positions with reference bases A and T harbour these extreme low MAF hotspots. Observed sequence context bias around these hotspot sites suggests some type of

context dependent mechanism. Whatever the mechanism is, it must be something that evades removal through clean-up strategy specifically designed to exclude context-dependent errors, with a some proportion of erroneous yet reads present in both sequencing directions that must be complementary in order to evade the excess read removal step.

One possibility is that these positions happen to have error prone motifs upstream and downstream that by chance produce complementary errors, which then gets misinterpreted as a real signal. The extent of the error size would still be limited by whichever of the two directional error rates is lowest. Thus it would be reasonable that MAF observed due to such coincidence of two matching error prone sequences would rarely be high, which is consistent with our data. Another possibility not necessarily accounted for could be sequence misalignments around genuine deletion or insertion sites, especially considering that indel variation is not analysed in this study. However this is unlikely given the frequency of these low level heteroplasmies found at the hotspot sites, as it would require certain indel variants to be present in nearly 100% of the population.

It is thus more likely we are looking at sequencing errors associated with repetitive sequences or other harder-to-sequence sequence features. Why we are not finding many of our known most error prone motifs at the low MAF hotspot sites is likely because the actual error rate of a motif in this instance may not be not as important as the probability of there being a coincidence of complementary error rates on both strands. With the MAF threshold of 0.1%, at any given position with 5000 consensus reads in each direction, only 5 erroneous reads would need to successfully pass excess read removal for such a site to pass the MAF threshold. So this means two motifs, each with just 1 in 1000 to 1 in 200 probability of producing complementary mis-calls could conceivably generate erroneous MAF of within the 0.1%-0.5% near ubiquitously across individuals sharing those two sequence motifs. Neither of the motifs needs to be associated with an error rate much higher than that for the signal to still pass filters. Such an explanation is consistent with sequences rich in repetitive motifs and homopolymer runs found both upstream and downstream of our identified low level heteroplasmy hotspot sites.

Although it is not impossible that some proportion of the very high incidence 0.1%-0.5% MAF heteroplasmies at the hotspot A and T sites could represent genuine biological signal, the balance of probability favours artefact as the more likely explanation of these heteroplasmy hotspots at A and T positions, with the disappearance of pattern above 0.5% MAF pointing to this as the likely practicable heteroplasmy detection limit.

DNA damage can lead to inflated heteroplasmy estimates and lower Ti/Tv ratio

Errors during the sequencing process are not the only possible source of artefacts. Actual damage to the DNA samples, acquired during handling or storage, can also have a significant impact on the detected apparent heteroplasmy. For example, Rathbun *et al.*, 2017, demonstrated that hydrolytic damage accumulation in DNA samples stored in water results in an increased rate of transversions, especially G>T/C>A, leading to significant decrease in the observed heteroplasmy transition-transversion (Ti/Tv) ratio at 3.6:1 compared to the expected 15:1 to 30:1 ratio previously reported for mitochondrial heteroplasmy (Irwin *et al.*, 2009). Rathbun *et al.*, 2017, suggest that mtDNA sequence vulnerability to damage and consequently erroneously identified heteroplasmies is not random, with 27 damage hotspot sites found within the control region (see **Figure 6 F2**).

Reviewing previous heteroplasmy literature, we found that 7 of the damage hotspot positions reported by Rathbun *et al.*, 2017, (16390G, 16519T, 152T, 189A, 214A, 215A, and 234A) are also among the most commonly reported heteroplasmic sites across literature, with each reported in 10 or more studies, which is quite interesting.

While we did not analyse position 16519 due to limited data coverage, the other 6 sites plus position 385A are also amongst sites identified by our study as having the highest incidence of heteroplasmy ≥1% MAF (10+ samples ≥1%

MAF, **figure 5.4.2 F1**). Yet when the lower MAF threshold of 0.1% is used, none of the damage hotspot sites reported by Rathbun *et al.*, 2017, are among our identified lowest level heteroplasmy hotspots (100+ samples ≥0.1% MAF, **figure 5.4.2 F4**), with no heteroplasmies at all found at 5 positions, and most samples above 0.1% MAF being 64 at position 189. Unfortunately Rathbun *et al.* did not extend their analyses outside the control region where some of our observed unusually high incidence very low MAF heteroplasmy hotspot areas, however even just looking at the control region, there are multiple sites where we report over 1000 heteroplasmies with MAF ≥0.1%, yet none of these overlap with the reported damage hotspot positions. This suggests that while clearly DNA damage during sample handling and storage can be misinterpreted as heteroplasmy, it is unlikely to be the main explanation behind our observed unusual distribution of the lowest level heteroplasmies. This could however be an important factor potentially contributing to our observed high levels of variation in sequencing error rates between samples addressed in chapter 4.

The overlap between positions with reported propensity to damage and previous heteroplasmy report incidence as well as our observed heteroplasmy frequencies at higher MAF is also interesting. This could be consistent with certain sequence positions or contexts being more vulnerable to specific types of DNA damage *in vivo,* leading to higher mutation rates and thus genuine increased heteroplasmy. Increase in DNA damage and decreased repair activity has been shown to correlate with increased heteroplasmy levels and mitochondrial dysfunction in disease and aging (Canter *et al*., 2005; Lin *et al*., 2011), and it is not unreasonable that sequence context may influence DNA damage and maintenance processes, leading to some sites being more damage and therefore heteroplasmy prone.

Figure 6 F2: Heteroplasmy incidence at suspected DNA damage hotspot sites

Heteroplasmies observed at 27 mtDNA positions reported as damage hotspots (*Rathbun et al., 2017*)

Position	rCRS base	*Present study*						*Heteroplasmies in literature*	
		Samples analysed	Heteroplasmies found at each MAF threshold ≥0.1%	≥0.5%	≥1%	≥5%	≥10%	Papers reported in	Reported het. samples
16055	C	3202	2	1	0	0	0	2	2
16232	C	3201	3	0	0	0	0	4	8
16265	A	3198	46	4	2	1	1	8	147
16356	T	3200	7	5	2	0	0	6	113
16390	**G**	3079	57	53	31	7	6	**11**	395
16406	T	2424	0	0	0	0	0	0	0
16424	T	1045	0	0	0	0	0	1	2
16430	A	610	0	0	0	0	0	1	1
16471	G	0	-	-	-	-	-	1	1
16472	G	0	-	-	-	-	-	2	2
16473	G	0	-	-	-	-	-	2	2
16516	G	0	-	-	-	-	-	1	1
16519	**T**	0	-	-	-	-	-	**11**	699
37	A	0	-	-	-	-	-	0	0
67	G	0	-	-	-	-	-	2	2
102	A	1	-	-	-	-	-	1	1
131	T	63	-	-	-	-	-	4	14
152	**T**	825	13	13	10	5	3	**11**	1915
189	**A**	3118	64	61	42	15	9	**12**	633
199	T	3183	16	14	9	2	2	9	273
214	**A**	3198	17	15	11	3	3	**10**	844
215	**A**	3198	33	26	16	5	3	**10**	1170
234	**A**	3201	41	33	24	13	8	**11**	601
385	A	3202	40	23	15	5	2	8	96
496	C	3149	0	0	0	0	0	2	8
497	C	3149	2	0	0	0	0	5	25
169	A	2329	0	0	0	0	0	1	1

Implication of very low-level heteroplasmy detection

Unlike bases A and T, heteroplasmies we detect at positions with reference sequence bases C or G show better correlation to previous literature reports even at MAF as low as 0.1%, indicating that these heteroplasmies may be more likely to be genuine. This raises questions about biological implications of heteroplasmies detected in MAF this low in sequence data derived from blood samples.

Leukocytes tend to typically contain only between 150 to 600 mtDNA copies per cell (although this may vary between cells by up to an order of magnitude) (Picard, 2021). MAF between 0.1% and 0.5% means detection of 1 variant in every 200 to 1000 mtDNA molecules. While blood heteroplasmy levels this low are not unprecedented, with a previous study reporting blood heteroplasmies at MAF threshold as low as 0.2% (Payne *et al*., 2013), the lowest 0.1% MAF threshold does likely imply less than 1 variant mitochondrial genome per cell, necessitating there to exist tissue heterogeneity, whereby mtDNA variants are present only in some proportion of the analysed cells. This is not unlikely, with previous evidence demonstrating high levels of mtDNA heterogeneity in bone marrow and clonal expansion of heteroplasmic mtDNA variants in hematopoietic progeny (Shin *et al*., 2004).

Clonal expansion of variant mtDNA molecules that result in tissue mosaicism mean that even though the overall heteroplasmy level within the tissue may be low, a not insignificant proportion of cells may exist with much higher heteroplasmy load that can impact the overall function of the tissue (Trifunovic, 2006). Secondly, the well-established phenomenon of heteroplasmy level variability between different tissues (Kirches *et al*., 2001; Krjutškov *et al*., 2014; Li *et al*., 2015) means that even though deleterious heteroplasmies may be found well below the expected phenotypic thresholds in the tissues, such as blood or buccal epithelium typically analysed in large-scale studies, much higher heteroplasmy levels may be present in other tissues, highlighting the potential importance of exploring the very low-level heteroplasmy landscape patterns.

Technical error and artefact mitigation

Identification and mitigation of potential sources of errors and artefacts is crucial in low level mitochondrial heteroplasmy analyses, as even minor error components in the data could obscure the signal from genuine heteroplasmy. Systematic context dependent sequencing errors are addressed in depth throughout this study and to some extent mitigated by the approach we employ where excess read fractions not supported by both sequencing directions are excluded. However CSEs are not the only source of potential artefacts. Other potential sources of artefacts that can affect heteroplasmy detection include PCR errors, alignment artefacts and inadvertent capture of nuclear-encoded mitochondrial sequences (NUMTs).

PCR errors during both the mtDNA amplification as well as during the sequencing itself can be a major source of false heteroplasmy detection. Possible PCR error mitigation strategies include heteroplasmy verification through sequencing of multiple, independent replicates (Robasky *et al*., 2014; Korolija *et al*., 2024) and appropriate minimum heteroplasmy threshold choice based on the PCR error rate estimated in control libraries of known PCR products (He *et al*., 2010). However replicate or control library data may not always be available, especially when mitochondrial heteroplasmy is analysed using already existing datasets.

Errors during the sequencing process itself as well as those of the alignment are another major contributor to false heteroplasmy detection. This can be to some extent mitigated by setting stringent read and alignment quality score thresholds (Ye *et al*., 2014), as well as assessing the quality of the detected variants. One suggested criterion for mitochondrial variant quality control is the assessment of Ti/Tv ratio (Ye *et al*., 2014) which for mitochondrial variants is reported to be from 15:1 to over 30:1 (Irwin *et al*., 2009; Guo *et al*., 2012; Ye *et al*., 2014), with higher Ti/Tv ratios reported for heteroplasmic variants compared to homoplasmic variants (Guo *et al*., 2012).

rCRS sequence contains an artificially inserted 'N' at position 3107, which, unless removed prior to alignment, results in alignment artefacts with false indels and base mismatches that must be accounted for when analysing mtDNA heteroplasmy (Ye *et al*., 2014). Similarly, use of a linear reference sequence when in fact MT genome is a circular molecule tends to result in reduced sequence coverage near the artificial break-point at the position 1/16569. This can be mitigated by using sequence alignment algorithms specifically designed for circular genomes, or by performing the alignment against reference sequence extended beyond the artificial 1/16569 so that reads spanning it can be appropriately mapped (Tao *et al*., 2023).

The possible impact of nuclear mtDNA-derived sequences (NUMTs)

MT genome data can be obtained either by extracting mtDNA sequences from non-targeted experiments such as whole genome or whole exome sequencing, where between 1 and 5% of reads tend to align to mtDNA (Ye *et al*., 2014), or directly in targeted experiments where mtDNA is isolated and enriched prior to sequencing. Methods for directly targeting MT genome include isolation of mitochondria from cells, or mtDNA enrichment and capture through CsCL density gradient ultracentrifugation, microarray hybridization, and PCR amplification (He *et al*., 2010; Ye *et al*., 2014; Xue *et al*., 2023). Although capture-based and PCR amplification based enrichment provide higher throughput compared to older methods such as CsCL density gradient ultracentrifugation, these techniques, especially capture-enrichment may still result in contamination by NUMT-derived signals (Li *et al*., 2012; Albayrak *et al*., 2016) due to possible probe/primer alignment to the homologous nuclear sequences rather than the intended mtDNA target sequences. Indirect approaches where MT genome sequence data is extracted from a larger scale experiment are also subject to NUMT contamination, although measures can also be taken in the experiment design to mitigate NUMT impact. Detected mitochondrial variant confirmation via a second method can help reduce NUMT impact, however unless the method used specifically isolates mtDNA from nDNA prior to sequencing and heteroplasmy detection, false NUMT derived heteroplasmies may even pass such verification (Ye *et al*., 2014).

NUMTs may result in sequencing reads that may align to both nDNA and mtDNA. How such reads are treated may have a vastly different impact on heteroplasmy detection. In one extreme, if the nuclear genome is ignored entirely and all raw reads mapped to mtDNA included, the resulting alignment will contain a significant proportion of NUMT-derived variants reads resulting in false positive heteroplasmy detection (Li *et al*., 2012; Ye *et al*., 2014). In the other extreme, a 2-step approach can be used where first any reads aligning to nuclear DNA are excluded, followed by mtDNA alignment generated from only the remaining definitive non-nDNA. This however will exclude a considerable proportion of genuine mtDNA-derived reads with sufficient similarity to NUMTs, resulting in significantly reduced MT genome coverage and thus decreased chance of true heteroplasmic variant detection (Ye *et al*., 2014). Even using a single step alignment approach where the raw reads are aligned using the entire human genome (nDNA + mtDNA) as the reference (where any read mapping to both nuclear and MT references is randomly assigned to either nDNA or mtDNA) results in significant reduction of coverage in mtDNA areas homologous to known NUMTs (Li *et al*., 2012; Maude *et al*., 2019; Laricchia *et al*., 2022).

A possible approach for NUMT-derived artefact reduction that does not lead to as severe a reduction in homologous region sequence coverage involves generation of a de-novo MT reference sequence specific to the individual, followed by read re-alignment variant calling based on the new reference. Compared to rCRS, using an individual-specific MT reference sequence can improve the alignment quality, sequence coverage and the confidence of detected variants, enabling elimination of a significant proportion of false NUMT-derived variants (Ring *et al*., 2018).

MT genome positions where potential variants may be introduced by NUMTs can be identified relatively easily, as the locations and sequences of many NUMTs are known, especially for longer sequences with high similarity to mtDNA (Lascaro *et al*., 2008; Calabrese *et al*., 2012). Such an approach for exclusion of NUMTs-derived variants is in

previous studios (Li *et al*., 2012). It does however only mitigate against known, common NUMTs, but does not account for the ongoing identification and reporting of new NUMT sequences, nor does it account for NUMTs that are rare and may be unique to specific individuals (Wei *et al*., 2022). Furthermore, exclusions of sites of possible NUMT-derived variants would also exclude genuine heteroplasmies at these sites.

One of the main existing strategies, including in the present study, for increasing confidence in the sequencing data being analysed being genuinely mtDNA-derived is to use paired end sequencing data and only analyse reads where both paired reads are mapped to the mtDNA at the correct expected distance and orientation from one other, with read pairs where only one read maps to mtDNA considered to be NUMT derived and excluded (Albayrak *et al*., 2016). A shortcoming of this approach is its reliance on the NUMT-derived read extending past the ends of the NUMT, which makes it ineffective for exclusion of variants arising from within some of the largest known, several kilobases long NUMTs (Calabrese *et al*., 2012).

Thus it is likely that some proportion of our reported heteroplasmies may be NUMT-derived false positives. In particular, this could be relevant in the context of the lowest level heteroplasmies with MAF between 0.1% and 0.5% that form the high incidence peaks in our reported heteroplasmy landscapes discussed at length throughout chapter 5. Considering the much lower abundance of nDNA compared to mtDNA, we should expect NUMT-derived false heteroplasmies to be characterised by low MAF but potentially high incidence in the population, making this a potential explanation for our observed unusual hotspots of very high incidence of low level heteroplasmy.

The 2016 study by Albayrak *et al.* reported predicted NUMT interference maps of the mitochondrial genome based on the number and lengths of the highest fidelity NUMTS found corresponding to each mtDNA positions, with highest NUMT-derived read contamination expected in mtDNA regions for which more and longer corresponding identical NUMT regions are found in the nDNA. These identical NUMT sequences are referred to by authors as "Longest Common Sub-sequences" or LCS. Comparing these NUMT interference maps with our observed heteroplasmy frequency landscapes, there are no obvious overlaps visually between the predicted highest interference regions and our identified heteroplasmy incidence hotspots. In fact, our identified highest heteroplasmy incidence appears to be at sites with lower than average expected nuclear interference reported by Albayrak *et al.*, 2016, with both fewer (median LCS count 34, IQR 32-40) and shorter (median average LCS size 17nt, IQR 16-22nt) average identical NUMT regions found overlapping our identified heteroplasmy hotspot sites with 100+ heteroplasmies above 0.1% MAF compared to MT genome overall (median LCS count 53, IQR 35-105, with median average length 30nt, IQR 18-63nt).

It is however important to note that Albayrak *et al*., 2016, define LCS as the longest continuous NUMT sub-regions found that are fully identical to the mtDNA sequence around a specific site. For false positive heteroplasmies to arise at specific positions at as high a frequency in the population as we observe however, these positions should correspond to sites where the NUMT sequence diverges from the mtDNA. Therefore if our observed highest frequency heteroplasmies were in fact NUMT-derived, we would expect these positions to more often represent the ends of identical NUMT fragments rather than being found within the identical regions, in which case seeing fewer fully identical overlapping NUMTs at such positions would not be unexpected. Considering this, our observations hint to NUMT-derived signals being at least one of the contributors behind our observed unusual low level heteroplasmy hotspot patterns.

There are however some counterpoints. Not only do we see fewer identical overlapping NUMTs at hotspot sites for heteroplasmies ≥0.1% MAF but also across positions with highest incidence of heteroplasmies above 1% and even 5% MAF thresholds (sites with 10+ heteroplasmies with MAF ≥1%: median 34, IQR 32-62; sites with 3+ heteroplasmies with MAF ≥5%: median 34, IQR 32-51). Furthermore, positions with highest heteroplasmy report

frequency in previous literature also show even fewer overlapping identical NUMT regions (median LCS count 33, IQR 31-39, across positions with 100+ heteroplasmy reports across 20 reviewed studies). If we interpret these observations as potential support for presence of NUMT-derived false positives within our reported lowest level heteroplasmy landscapes, not only would this suggest similar NUMT-derived signal presence among higher MAF heteroplasmies we report, it would also imply comparable NUMT-derived artefact presence among heteroplasmies reported most frequently in the previous literature. This is potentially a very interesting point, considering that we have been generally treating heteroplasmy report frequency in the previous literature as the gold standard for testing the validity of our heteroplasmy identification and measurement approaches.

While interesting, without further in-depth analyses, the relationship between nDNA interference and our observed heteroplasmy landscape patterns remains a speculation. It is also important to highlight that the very study reporting nDNA interference maps also shows that based on the lengths of identical NUMT sub-sequences found, even with single end reads, 150bp read length should be sufficient to eliminate approximately 70% of nDNA interference across the MT genome (Albayrak *et al*., 2016). The sequencing data used for heteroplasmy landscape analyses in the present study was generated using 150bp paired-end reads, which should in theory be even more effective at eliminating NUMTs signals further.

Nevertheless, without quantitative evidence we cannot eliminate the possibility that all of our detected signals could still be entirely due to NUMTs. Such possibility would rise some interesting consequences in the context of the broader heteroplasmy research. As discussed earlier, multiple facets of our reported results align well with the broader heteroplasmy literature. If the entire signal reported in this study was in fact NUMTs artefacts, it would suggest similar level of contamination in significant parts of the existing heteroplasmy literature.

An interesting database worth addressing is the set of potentially 'NUMTs-affected positions' reported by Li *et al*., 2012, which represent all MT genome sites where any identified NUMTs sequences differ from the homologous mtDNA sequence (rCRS) and thus may be misinterpreted as heteroplasmic variants. Considering that a variant position may be present in multiple NUMTs corresponding to a specific mtDNA region, if a significant proportion of detected heteroplasmies were indeed NUMT-derived, we should expect the observed heteroplasmy frequencies at the NUMTs-affected positions to correlate with the number of NUMTs found carrying the variant bases. When we look at our data however, we do not find any correlation between the variant NUMT counts reported by Li *et al*., 2012, and our observed heteroplasmy frequencies, whether looking at all heteroplasmies ≥0.1% MAF or only at the subsets ≥0.5%, 1%, 5% or 10% MAF thresholds (Pearson's r <<0.1 with p>0.05 in all cases). The only variable with a significant albeit very low positive correlation to the variant NUMTs counts reported by Li *et al*., 2012, is the number of previously published studies in which the position has been reported as heteroplasmic (r =0.15, p<<0.05), although the overall number of heteroplasmies reported at a site across the reviewed existing literature does not correlate with the number of variant NUMTs (r = 0.02, p =0.04).

Even though no correlation exists between heteroplasmy incidence and the number of variant NUMTs potentially affecting any given site, there is a statistically significant relationship between a site being potentially affected by NUMTs and being reported as heteroplasmic both in the present study (X^2 (df=1, N=16568)= 46.23, p = $1x10^{-11}$) or across the broader literature (X^2 (df=1, N=16568) =33.43, p = $7x10^{-9}$), with a slightly higher proportion of NUMTs-affected positions found among heteroplasmic sites compared to non-heteroplasmic sites (present study: 50.3% vs 44.7%, Literature: 53.2% vs 38.2%). While not definitive evidence, this does speak to the presence of some NUMT contamination. It should however be noted that even very small effect sizes may achieve statistical significance for sample sizes this large (Kelen *et al*., 1988; Lantz, 2013), warranting careful consideration of the practical significance of these findings.

Among the smaller subsets of positions where we report highest incidence of heteroplasmies above 1% and 5% MAF thresholds the proportion of NUMTs-affected sites is even higher (60% of the sites with 10+ heteroplasmies ≥1% MAF and 56% of the sites with 3+ heteroplasmies ≥5% MAF). This is a worryingly large proportion of high heteroplasmy incidence sites, indicating a significant potential impact of NUMTs even with higher MAF thresholds.

Unexpectedly though, in the subset of the 107 extremely low MAF heteroplasmy hotspot sites (100+ heteroplasmies ≥0.1% MAF), the proportion of potentially NUMTs-affected positions is considerably lower at 34.6% (37 of 107 hotspot positions). Yet if NUMTs was the sole mechanism causing the low level signals we are detecting, one should expect most if not all of this subset of extremely high incidence low MAF heteroplasmy hotspot sites to be at known NUMTs-affected positions. Instead, the proportion of NUMTs affected positions is actually lower among this particular subset than even among non-heteroplasmic positions (34.6% vs 44.7%).

NUMTs represent an attractive explanation for our discovered hotspots of extremely common very low MAF heteroplasmies. The depletion of known NUMTs-affected positions among these heteroplasmy hotspots, even compared to non-heteroplasmic sites, does however speak somewhat against this particular mechanism being the sole cause behind specifically the low-level heteroplasmy hotspot phenomenon. Unless of course NUMTs are the source of the entirely of signal at every heteroplasmic mtDNA position we report, which of course is not beyond the realm of possibility. Furthermore, if NUMTs were the primary cause of specifically the very low MAF artefacts, one would expect the correlation between NUMTs-affected positions and heteroplasmic sites in our study (which uses much lower MAF thresholds than nearly every other published study) to be stronger than that between NUMTs-affected positions and heteroplasmic sites reported in the literature. This is not true though, suggesting that NUMTs-derived artefacts can manifest as heteroplasmies with a wide range of MAF values.

All together this point to a potentially not-insignificant presence of NUMT-derived false positives both within our data as well as within the existing literature. Furthermore, such NUMT-derived contamination is unlikely to be restricted to only the very lowest MAF heteroplasmies, potentially making minimum MAF threshold increase less effective as a filtering tool against specifically this type of artefacts.

One of the major shortcomings of the present study is the fact that we did not attempt to quantify the level of NUMTs-derived artefact contamination in our data. Based on the multiple lines of evidence discussed here however, this is likely not insignificant. It is interesting though that the existing previous heteroplasmy literature appear to also suffer from comparable levels of NUMTs contamination. Furthermore, such NUMT-derived contamination is unlikely to be restricted to only the very lowest MAF heteroplasmies, potentially making minimum MAF threshold increase less effective as a filtering tool against specifically this type of artefacts.

The fact that NUMTs contamination in our data is comparable and might even be slightly lower than across our reviewed literature may seem counterintuitive, especially considering that we did not employ any further NUMTs mitigations beyond using relatively long paired-end reads and requiring proper read mapping distance and orientation. It does however indicate that just the use of sufficient read length and paired ends likely result in comparable data quality to many existing studies, all of which may be similarly contaminated. This highlights the importance of giving special consideration to how NUMTs may impact heteroplasmy detection, and necessity to devise strategies to identify and filter such artefacts beyond just setting higher MAF thresholds.

Finally, one should address the possibility that the relationship between potentially NUMT-affected sites and our identified heteroplasmic positions does not actually indicate NUMT-derived variants misinterpreted as heteroplasmies, but rather results from genuine heteroplasmy replicating the mutation patterns seen in the NUMT sequences. NUMTs sequence divergence from the ancestral MT genome is the result of mutation accumulation since the nuclear insertion event, where similarly to other pseudogenes NUMTs accumulate mutations under relaxed

selection (Calvignac *et al*., 2011). The heteroplasmy landscape of the actual MT genome also represents a snapshot of the ongoing mutation-selection balance, but on a much shorter time scale. The presence of multiple mtDNA copies within the cell enables deleterious variants to persist at low frequencies provided a sufficient proportion of WT mtDNA remains that can compensate for it. Provided that mutation rates are not entirely independent of the sequence context, it is not inconceivable that the same sequence contexts in the mtDNA and nDNA could be associated with higher mutation rates. While the former would manifest as heteroplasmy, the latter should appear as NUMT sequence divergence, with both potentially being more common in the same sequence contexts.

The analyses presented in this study does not provide a definitive answer whether this is indeed the case or not. However this is an intriguing possibility that highlights NUMT consideration in heteroplasmy study not only in the context of artefacts, but also potentially in the context of the underlying mutation and selection processes affecting both the nuclear and the mitochondrial genome.

Importance of DNA secondary structure in both heteroplasmy and sequencing fidelity

Our results from **Chapter 4** point strongly to the possibility of DNA secondary structure, especially G quadruplexes being involved in the context dependent sequencing error rate patterns that we observe. At the same time, DNA secondary structures are also known to be prone to mutation *in vivo*, leading to the problem where routinely filtering out sites with error signatures from the data, creates a bias where positions genuinely prone to heteroplasmy are biased against and become invisible. To add another layer of complication to this, G4 structures for example that are thought to have a major involvement in regulatory functions, could thus have an obvious mechanism for how even a low level variant could be able to influence the fitness of a particular mitochondrion or the particular mtDNA molecule. This multi-layered problem highlights why continuing the investigation of the systematic errors landscapes and behaviours could be quite useful.

Limitations and future perspectives

Among others, one of the obvious limitations of the data presented here is that this is very much an exploratory study, with very open-ended starting questions. As a result, much of the data presented in this thesis is quite descriptive, with different possible explanations that with the available data are hard to prove one way or the other. Therefore now that we have developed much more concrete hypotheses, next logical steps would be to address them with more experimental work.

One particularly fascinating area is the context dependent sequencing error rate variation we describe, both within and between samples, and how that relates to genuine heteroplasmy levels. If the same sequences happen to be both error prone in vitro and in vivo, and induce heteroplasmies that are nevertheless obscured by errors, being able to measure how large the heteroplasmy versus error components are would be fascinating. One way to approach this, if we suspect that polymerase stalling is the cause for the particular error/mutation, could be to run side by side experiments on a sequencing by synthesis machine and Nanopore sequencer, which uses a very different sequencing strategy. Such experiments have been done to measure specifically sequencing error rates, however to our knowledge not in the heteroplasmy context, and could be a way to actually measure the ratios between heteroplasmy and errors.

The overarching limitations of the chosen investigative approach

As already stated, this study is highly exploratory and descriptive. In our attempt to explore the highly complex system of mitochondrial heteroplasmy, the decisions taken for how we slice through the data have often been fairly arbitrary. Form what minimum MAF threshold to use in the initial heteroplasmy filtration, to which characteristics qualify a position as a 'heteroplasmy hotspot', to what parameters of heteroplasmy landscape to focus on – all of these and many others are fundamentally arbitrary choices, with many of the results and observations presented without a sufficient backing of precise statistical statements about given hypotheses.

A fundamentally better way to address many of the quantitative questions tackled here, from variability between samples or loci, to different observed MAF levels or heteroplasmy counts, would have been via the lens of a mechanistic model. Ideally a statistical model appropriately describing the processes leading to the observations being made should have been built, from which we could then infer parameters corresponding to quantities we are interested in, such as the true heteroplasmy levels. Such an approach could have been more powerful for addressing the scientific questions posed throughout this work. In the absence of a more mechanistic model, the insights that do come from our relatively descriptive treatment of the data represented here necessarily remain limited, especially so where the estimates of uncertainty are unclear.

7 References

1000 Genomes Project Consortium (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), 68–74. https://doi.org/10.1038/nature15393.

Abecasis, G.R. *et al*. (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), 1061–1073. https://doi.org/10.1038/nature09534.

Abecasis, G.R. *et al*. (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), 56–65. https://doi.org/10.1038/nature11632.

Abnizova, I. *et al*. (2012) 'Analysis of context-dependent errors for Illumina sequencing', *Journal of Bioinformatics and Computational Biology*, 10(2), 1241005. https://doi.org/10.1142/S0219720012410053.

Adebayo, M. *et al*. (2021) 'Mitochondrial Fusion and Fission: The fine-tune balance for cellular homeostasis', *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 35(6), e21620. https://doi.org/10.1096/fj.202100067R.

Albayrak, L. *et al*. (2016) 'The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome', *BMC Genomics*, 17(1), 1017. https://doi.org/10.1186/s12864-016-3375-x.

Alexeyev, M. *et al*. (2013) 'The maintenance of mitochondrial DNA integrity--critical analysis and update', *Cold Spring Harbor perspectives in biology*, 5(5), a012641. https://doi.org/10.1101/cshperspect.a012641.

Allhoff, M. *et al*. (2013) 'Discovering motifs that induce sequencing errors', *BMC Bioinformatics*, 14(S1), 1–10. https://doi.org/10.1038/nbt.1754.

Anderson, S. *et al*. (1981) 'Sequence and organization of the human mitochondrial genome', *Nature*, 290(5806), 457–465. https://doi.org/10.1038/290457a0.

Antes, A. *et al*. (2010) 'Differential regulation of full-length genome and a single-stranded 7S DNA along the cell cycle in human mitochondria', *Nucleic Acids Research*, 38(19), 6466–6476. https://doi.org/10.1093/nar/gkq493.

Archibald, J.M. (2015) 'Endosymbiosis and Eukaryotic Cell Evolution', *Current Biology*, 25(19), R911–R921. https://doi.org/10.1016/j.cub.2015.07.055.

Ashton, T.M. *et al*. (2018) 'Oxidative Phosphorylation as an Emerging Target in Cancer Therapy', *Clinical Cancer Research*, 24(11), 2482–2490. https://doi.org/10.1158/1078-0432.CCR-17-3070.

Becker, E.A. *et al*. (2012) 'Experimental Analysis of Sources of Error in Evolutionary Studies Based on Roche/454 Pyrosequencing of Viral Genomes', *Genome Biology and Evolution*, 4(4), 457–465. https://doi.org/10.1093/gbe/evs029.

Bedrat, A. *et al*. (2016) 'Re-evaluation of G-quadruplex propensity with G4Hunter', *Nucleic Acids Research*, 44(4), 1746–1759. https://doi.org/10.1093/NAR/GKW006.

Bezwada, D. *et al*. (2024) 'Mitochondrial complex I promotes kidney cancer metastasis', *Nature*, 633(8031), 923–931. https://doi.org/10.1038/s41586-024-07812-3.

Birkenmeier, K. *et al*. (2016) 'Hodgkin and Reed–Sternberg cells of classical Hodgkin lymphoma are highly dependent on oxidative phosphorylation', *International Journal of Cancer*, 138(9), 2231–2246. https://doi.org/10.1002/ijc.29934.

Bofkin, L. *et al*. (2007) 'Variation in Evolutionary Processes at Different Codon Positions', *Molecular Biology and Evolution*, 24(2), 513–521. https://doi.org/10.1093/molbev/msl178.

Bolze, A. *et al*. (2020) 'A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans', *bioRxiv*, 798264. https://doi.org/10.1101/798264.

Butler, T.J. *et al*. (2020) 'Mitochondrial genetic variation is enriched in G-quadruplex regions that stall DNA synthesis in vitro', *Human Molecular Genetics*, 29(8), 1292–1309. https://doi.org/10.1093/HMG/DDAA043.

Byrska-Bishop, M. *et al*. (2022) 'High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios', *Cell*, 185(18), 3426-3440.e19. https://doi.org/10.1016/J.CELL.2022.08.004.

Cai, S. *et al*. (2005) 'Mitochondrial Targeting of Human O6-Methylguanine DNA Methyltransferase Protects against Cell Killing by Chemotherapeutic Alkylating Agents', *Cancer Research*, 65(8), 3319–3327. https://doi.org/10.1158/0008-5472.CAN-04-3335.

Calabrese, C. *et al*. (2014) 'MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing', *Bioinformatics*, 30(21), 3115–3117. https://doi.org/10.1093/BIOINFORMATICS/BTU483.

Calabrese, C. *et al*. (2022) 'Heteroplasmic mitochondrial DNA variants in cardiovascular diseases', *PLOS Genetics*, 18(4), e1010068. https://doi.org/10.1371/JOURNAL.PGEN.1010068.

Calabrese, F.M. *et al*. (2012) 'Primates and mouse NumtS in the UCSC Genome Browser', *BMC Bioinformatics*, 13(4), S15. https://doi.org/10.1186/1471-2105-13-S4-S15.

Calvignac, S. *et al*. (2011) 'Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts)', *Mitochondrion*, 11(2), 246–254. https://doi.org/10.1016/j.mito.2010.10.004.

Canter, J.A. *et al*. (2005) 'Degree of heteroplasmy reflects oxidant damage in a large family with the mitochondrial DNA A8344G mutation', *Free Radical Biology & Medicine*, 38(5), 678–683. https://doi.org/10.1016/j.freeradbiomed.2004.11.031.

Cao, Y. *et al*. (2010) 'Detection of eight frequently encountered point mutations in mitochondria in Chinese patients suggestive of mitochondrial encephalomyopathies', *Mitochondrion*, 10(4), 330–334. https://doi.org/10.1016/j.mito.2010.01.008.

Caro, P. *et al*. (2012) 'Metabolic Signatures Uncover Distinct Targets in Molecular Subsets of Diffuse Large B Cell Lymphoma', *Cancer Cell*, 22(4), 547–560. https://doi.org/10.1016/j.ccr.2012.08.014.

Castellana, S. *et al*. (2011) 'Evolutionary Patterns of the Mitochondrial Genome in Metazoa: Exploring the Role of Mutation and Selection in Mitochondrial Protein–Coding Genes', *Genome Biology and Evolution*, 3, 1067–1079. https://doi.org/10.1093/gbe/evr040.

Chambers, A.F. *et al*. (2002) 'Dissemination and growth of cancer cells in metastatic sites', *Nature Reviews Cancer*, 2(8), 563–572. https://doi.org/10.1038/nrc865.

Chatre, L. *et al*. (2013) 'Prevalent coordination of mitochondrial DNA transcription and initiation of replication with the cell cycle', *Nucleic Acids Research*, 41(5), 3068–3078. https://doi.org/10.1093/nar/gkt015.

Chinnery, P.F. *et al*. (1997) 'Molecular pathology of MELAS and MERRF. The relationship between mutation load and clinical phenotypes', *Brain*, 120(10), 1713–1721. https://doi.org/10.1093/BRAIN/120.10.1713.

Chinnery, P.F. *et al*. (2002) 'Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: Is there a common mechanism?', *Lancet*, 360, 1323–25. https://doi.org/10.1016/S0140-6736(02)11310-9.

Clayton, D.A. (2000) 'Transcription and replication of mitochondrial DNA', *Human Reproduction*, 15(suppl 2), 11–17. https://doi.org/10.1093/humrep/15.suppl_2.11.

Clayton, D.A. *et al*. (1974) 'The Absence of a Pyrimidine Dimer Repair Mechanism in Mammalian Mitochondria', *Proceedings of the National Academy of Sciences*, 71(7), 2777–2781. https://doi.org/10.1073/pnas.71.7.2777.

Cline, S.D. *et al*. (2010) 'Arrest of human mitochondrial RNA polymerase transcription by the biological aldehyde adduct of DNA, M 1 dG', *Nucleic Acids Research*, 38(21), 7546–7557. https://doi.org/10.1093/nar/gkq656.

Cline, S.D. (2012) 'Mitochondrial DNA damage and its consequences for mitochondrial gene expression', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(9), 979–991. https://doi.org/10.1016/j.bbagrm.2012.06.002.

Coene, E.D. *et al*. (2005) 'Phosphorylated BRCA1 Is Predominantly Located in the Nucleus and Mitochondria', *Molecular Biology of the Cell*, 16(2), 997–1010. https://doi.org/10.1091/mbc.e04-10-0895.

Cotterill, M. *et al*. (2013) 'The activity and copy number of mitochondrial DNA in ovine oocytes throughout oogenesis in vivo and during oocyte maturation in vitro', *Molecular Human Reproduction*, 19(7), 444–450. https://doi.org/10.1093/molehr/gat013.

Cox, R.T. *et al*. (2021) 'The role of mitophagy during oocyte aging in human, mouse, and Drosophila: implications for oocyte quality and mitochondrial disease', *Reproduction and Fertility*, 2(4), R113–R129. https://doi.org/10.1530/RAF-21-0060.

Craven, L. *et al*. (2017) 'Recent Advances in Mitochondrial Disease', *Annual Review of Genomics and Human Genetics*, 18, 257–275. https://doi.org/10.1146/annurev-genom-091416-035426.

Dahal, S. *et al*. (2018) 'Homologous recombination-mediated repair of DNA double-strand breaks operates in mammalian mitochondria', *Cellular and Molecular Life Sciences*, 75(9), 1641–1655. https://doi.org/10.1007/s00018-017-2702-y.

Damas, J. *et al*. (2012) 'Mitochondrial DNA deletions are associated with non-B DNA conformations', *Nucleic Acids Research*, 40(16), 7606–7621. https://doi.org/10.1093/nar/gks500.

Das, B.B. *et al*. (2010) 'Role of tyrosyl-DNA phosphodiesterase (TDP1) in mitochondria', *Proceedings of the National Academy of Sciences*, 107(46), 19790–19795. https://doi.org/10.1073/pnas.1009814107.

De Grey, A.D.N.J. (1997) 'A proposed refinement of the mitochondrial free radical theory of aging', *BioEssays*, 19(2), 161–166. https://doi.org/10.1002/bies.950190211.

Diaz, F. *et al*. (2002) 'Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control', *Nucleic Acids Research*, 30(21), 4626–4633. https://doi.org/10.1093/nar/gkf602.

Ding, J. *et al*. (2015) 'Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools', *PLOS Genetics*, 11(7), e1005306. https://doi.org/10.1371/JOURNAL.PGEN.1005306.

Diroma, M.A. *et al*. (2014) 'Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data', *BMC genomics*, 15(Suppl. 3), S2. https://doi.org/10.1186/1471-2164-15-S3-S2.

Druzhyna, N.M. *et al*. (2008) 'Mitochondrial DNA repair in aging and disease', *Mechanisms of Ageing and Development*, 129(7), 383–390. https://doi.org/10.1016/j.mad.2008.03.002.

Duan, M. *et al*. (2019) 'Evaluating heteroplasmic variations of the mitochondrial genome from whole genome sequencing data', *Gene*, 699, 145–154. https://doi.org/10.1016/J.GENE.2019.03.016.

Dubie, J.J. *et al*. (2019) 'The conflict within: origin, proliferation and persistence of a spontaneously arising selfish mitochondrial genome', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1790), 20190174. https://doi.org/10.1098/rstb.2019.0174.

Elliott, H.R. *et al*. (2008) 'Pathogenic mitochondrial DNA mutations are common in the general population', *American Journal of Human Genetics*, 83(2), 254–260. https://doi.org/10.1016/j.ajhg.2008.07.004.

Elorza, A.A. *et al*. (2021) 'mtDNA Heteroplasmy at the Core of Aging-Associated Heart Failure. An Integrative View of OXPHOS and Mitochondrial Life Cycle in Cardiac Mitochondrial Physiology', *Frontiers in Cell and Developmental Biology*, 9. https://doi.org/10.3389/fcell.2021.625020.

Falabella, M. *et al*. (2021) 'Potential roles for G-quadruplexes in mitochondria', *Current Medicinal Chemistry*, 26(16), 2918-2932. https://doi.org/10.2174/0929867325666180228165527.

Falkenberg, M. (2018) 'Mitochondrial DNA replication in mammalian cells: overview of the pathway', *Essays in Biochemistry*, 62(3), 287. https://doi.org/10.1042/EBC20170100.

Fuller, C.W. *et al*. (2009) 'The challenges of sequencing by synthesis', *Nature Biotechnology,* 27, 1013–1023. https://doi.org/10.1038/nbt.1585.

González, M.M. *et al*. (2020) 'Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing', *Mitochondrion*, 50, 88–93. https://doi.org/10.1016/J.MITO.2019.10.006.

Gorman, G.S. *et al*. (2016) 'Mitochondrial diseases', *Nature Reviews Disease Primers*, 2(1), 1–22. https://doi.org/10.1038/nrdp.2016.80.

Goto, H. *et al*. (2011) 'Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study', *Genome Biology*, 12(6), 1–16. https://doi.org/10.1186/GB-2011-12-6-R59/FIGURES/5.

Goto, Y. *et al*. (1990) 'A mutation in the tRNA(Leu)(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies', *Nature*, 348(6302), 651–3. https://doi.org/10.1038/348651a0.

Guiblet, W.M. *et al*. (2018) 'Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate', *Genome Research*, 28(12), 1767–1778. https://doi.org/10.1101/GR.241257.118/-/DC1.

Guo, Y. *et al*. (2012) 'The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation', *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 744(2), 154–160. https://doi.org/10.1016/j.mrgentox.2012.02.006.

Harman, D. (1956) 'Aging: a theory based on free radical and radiation chemistry', *Journal of Gerontology*, 11(3), 298–300. https://doi.org/10.1093/geronj/11.3.298.

Hazkani-Covo, E. *et al*. (2010) 'Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes', *PLOS Genetics*, 6(2), e1000834. https://doi.org/10.1371/journal.pgen.1000834.

He, Y. *et al*. (2010) 'Heteroplasmic mitochondrial DNA mutations in normal and tumour cells', *Nature*, 464, 610–614. https://doi.org/10.1038/nature08802.

Holt, I.J. *et al*. (2012) 'Human Mitochondrial DNA Replication', *Cold Spring Harbor Perspectives in Biology*, 4(12), a012971. https://doi.org/10.1101/cshperspect.a012971.

Hopper, R.K. *et al*. (2006) 'Mitochondrial Matrix Phosphoproteome: Effect of Extra Mitochondrial Calcium', *Biochemistry*, 45(8), 2524–2536. https://doi.org/10.1021/bi052475e.

Hübner, A. *et al*. (2019) 'Sharing of heteroplasmies between human liver lobes varies across the mtDNA genome', *Scientific Reports*, 9(1), 11219. https://doi.org/10.1038/s41598-019-47570-1.

Irwin, J.A. *et al*. (2009) 'Investigation of heteroplasmy in the human mitochondrial DNA control region: A synthesis of observations from more than 5000 global population samples', *Journal of Molecular Evolution*, 68(5), 516–527. https://doi.org/10.1007/S00239-009-9227-4.

Itsara, L.S. *et al*. (2014) 'Oxidative Stress Is Not a Major Contributor to Somatic Mitochondrial DNA Mutations', *PLOS Genetics*, 10(2), e1003974. https://doi.org/10.1371/journal.pgen.1003974.

Jeong, Y.-C. *et al*. (2005) 'Pyrimido[1,2-a]-purin-10(3H)-one, M1G, is less prone to artifact than base oxidation', *Nucleic Acids Research*, 33(19), 6426–6434. https://doi.org/10.1093/nar/gki944.

Jeppesen, D.K. *et al*. (2011) 'DNA repair deficiency in neurodegeneration', *Progress in Neurobiology*, 94(2), 166–200. https://doi.org/10.1016/j.pneurobio.2011.04.013.

Ježek, P. *et al.* (2009) 'Mitochondrial reticulum network dynamics in relation to oxidative stress, redox regulation, and hypoxia', *The International Journal of Biochemistry & Cell Biology*, 41(10), 1790–1804. https://doi.org/10.1016/J.BIOCEL.2009.02.014.

Jiao, J. *et al*. (2008) 'Manganese-induced single strand breaks of mitochondrial DNA in vitro and in vivo', *Environmental Toxicology and Pharmacology*, 26(2), 123–127. https://doi.org/10.1016/j.etap.2007.12.009.

Jimenez-Tellez, N. *et al*. (2019) 'Cellular models for human cardiomyopathy: What is the best option?', *World Journal of Cardiology*, 11(10), 221–235. https://doi.org/10.4330/wjc.v11.i10.221.

Johns, D.R. *et al*. (1993) 'Leber's Hereditary Optic Neuropathy: Clinical Manifestations of the 14484 Mutation', *Archives of Ophthalmology*, 111(4), 495–498. https://doi.org/10.1001/archopht.1993.01090040087038.

Ju, Y.S. *et al*. (2014) 'Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer', *eLife*, 3, e02935. https://doi.org/10.7554/eLife.02935.

Kalsbeek, A.M.F. *et al*. (2018) 'Altered mitochondrial genome content signals worse pathology and prognosis in prostate cancer', *The Prostate*, 78(1), 25–31. https://doi.org/10.1002/pros.23440.

Kauppila, J.H.K. *et al*. (2015) 'Mitochondrial DNA: Radically free of free-radical driven mutations', *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1847(11), 1354–1361. https://doi.org/10.1016/j.bbabio.2015.06.001.

Kaur, G. *et al*. (2012) 'Cell lines: Valuable tools or useless artifacts', *Spermatogenesis*, 2(1), 1–5. https://doi.org/10.4161/spmg.19885.

Kelen, G.D. *et al*. (1988) 'Statistical reasoning in clinical trials: Hypothesis testing', *The American Journal of Emergency Medicine*, 6(1), 52–61. https://doi.org/10.1016/0735-6757(88)90207-0.

Kennedy, S.R. *et al*. (2013) 'Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage', *PLoS genetics*, 9(9), e1003794. https://doi.org/10.1371/journal.pgen.1003794.

Kirches, E. *et al*. (2001) 'Heterogeneous tissue distribution of a mitochondrial DNA polymorphism in heteroplasmic subjects without mitochondrial disorders', *Journal of Medical Genetics*, 38(5), 312–317. https://doi.org/10.1136/jmg.38.5.312.

Korolija, M. *et al*. (2024) 'Mitochondrial point heteroplasmy: insights from deep-sequencing of human replicate samples', *BMC Genomics*, 25(1), 48. https://doi.org/10.1186/s12864-024-09963-z.

Krjutškov, K. *et al*. (2014) 'Tissue-specific mitochondrial heteroplasmy at position 16,093 within the same individual', *Current Genetics*, 60(1), 11–16. https://doi.org/10.1007/s00294-013-0398-6.

Kukat, C. *et al*. (2011) 'Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA', *Proceedings of the National Academy of Sciences*, 108(33), 13534–13539. https://doi.org/10.1073/PNAS.1109263108.

Kumar, S. (1996) 'Patterns of Nucleotide Substitution in Mitochondrial Protein Coding Genes of Vertebrates', *Genetics*, 143(1), 537–548. https://doi.org/10.1093/genetics/143.1.537.

Kuntz, E.M. *et al*. (2017) 'Targeting mitochondrial oxidative phosphorylation eradicates therapy-resistant chronic myeloid leukemia stem cells', *Nature Medicine*, 23(10), 1234–1240. https://doi.org/10.1038/nm.4399.

Lagadinou, E.D. *et al*. (2013) 'BCL-2 Inhibition Targets Oxidative Phosphorylation and Selectively Eradicates Quiescent Human Leukemia Stem Cells', *Cell Stem Cell*, 12(3), 329–341. https://doi.org/10.1016/j.stem.2012.12.013.

Lake, N.J. *et al*. (2024) 'Quantifying constraint in the human mitochondrial genome', *Nature*, 635(8038), 390–397. https://doi.org/10.1038/s41586-024-08048-x.

Lantz, B. (2013) 'The large sample size fallacy', *Scandinavian Journal of Caring Sciences*, 27(2), 487–492. https://doi.org/10.1111/j.1471-6712.2012.01052.x.

Laricchia, K.M. *et al*. (2020) gnomAD v3.1 Mitochondrial DNA Variants. Available at: https://gnomad.broadinstitute.org/news/2020-11-gnomad-v3-1-mitochondrial-dna-variants/ (Accessed: 11 January 2023).

Laricchia, K.M. *et al*. (2022) 'Mitochondrial DNA variation across 56,434 individuals in gnomAD', *Genome Research*, 32(3), 569–582. https://doi.org/10.1101/GR.276013.121.

Larman, T.C. *et al*. (2012) 'Spectrum of somatic mitochondrial mutations in five cancers', *Proceedings of the National Academy of Sciences*, 109(35), 14087–14091. https://doi.org/10.1073/pnas.1211502109.

Lascaro, D. *et al*. (2008) 'The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify Human NumtS', *BMC Genomics*, 9(1), 267. https://doi.org/10.1186/1471-2164-9-267.

Lawless, C. *et al*. (2020) 'The rise and rise of mitochondrial DNA mutations', *Open Biology*, 10(5), 200061. https://doi.org/10.1098/rsob.200061.

LeDoux, S.P. *et al*. (1999) 'Repair of alkylation and oxidative damage in mitochondrial DNA', *Mutation Research/DNA Repair*, 434(3), 149–159. https://doi.org/10.1016/S0921-8777(99)00026-9.

Leng, Y. *et al*. (2015) 'The mitochondrial DNA 10197 G > A mutation causes MELAS/Leigh overlap syndrome presenting with acute auditory agnosia', *Mitochondrial DNA*, 26(2), 208–212. https://doi.org/10.3109/19401736.2014.905860.

Lewis, S.C. *et al*. (2016) 'ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells', *Science*, 353(6296), aaf5549. https://doi.org/10.1126/science.aaf5549.

Li, D. *et al*. (2022) 'Pathogenic mitochondrial DNA 3243A>G mutation: From genetics to phenotype', *Frontiers in Genetics*, 13, 951185. https://doi.org/10.3389/fgene.2022.951185.

Li, H. *et al*. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', *arXiv preprint arXiv*, 1303.3997. https://doi.org/10.48550/arXiv.1303.3997.

Li, M. *et al*. (2010) 'Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes', *The American Journal of Human Genetics*, 87(2), 237–249. https://doi.org/10.1016/J.AJHG.2010.07.014.

Li, M. *et al*. (2012) 'Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs', *Nucleic Acids Research*, 40(18), e137. https://doi.org/10.1093/nar/gks499.

Li, M. *et al*. (2015) 'Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations', *Proceedings of the National Academy of Sciences*, 112(8), 2491–2496. https://doi.org/10.1073/pnas.1419651112.

Lin, H. *et al*. (2011) 'Mitochondrial DNA Damage and Repair in RPE Associated with Aging and Age-Related Macular Degeneration', *Investigative Ophthalmology & Visual Science*, 52(6), 3521–3529. https://doi.org/10.1167/iovs.10-6163.

Liu, C. *et al*. (2018) 'Deep sequencing of the mitochondrial genome reveals common heteroplasmic sites in NADH dehydrogenase genes', *Human Genetics*, 137(3), 203–213. https://doi.org/10.1007/S00439-018-1873-4/FIGURES/5.

Liu, C. *et al*. (2022) 'Comparison of mitochondrial DNA sequences from whole blood and lymphoblastoid cell lines', *Scientific Reports*, 12(1), 1801. https://doi.org/10.1038/s41598-022-05814-7.

Liu, P. *et al*. (2010) 'DNA repair in mammalian mitochondria: Much more than we thought?', *Environmental and Molecular Mutagenesis*, 51(5), 417–426. https://doi.org/10.1002/em.20576.

Loman, N.J. *et al*. (2012) 'Performance comparison of benchtop high-throughput sequencing platforms', *Nature Biotechnology* 2012 30:5, 30(5), 434–439. https://doi.org/10.1038/nbt.2198.

Lorenzoni, P.J. *et al*. (2014) 'When should MERRF (myoclonus epilepsy associated with ragged-red fibers) be the diagnosis?', *Arquivos de Neuro-Psiquiatria*, 72, 803–811. https://doi.org/10.1590/0004-282X20140124.

Lott, M.T. *et al*. (2013) 'mtDNA Variation and Analysis Using Mitomap and Mitomaster', *Current Protocols in Bioinformatics*, 44, 1.23.1-1.23.26. https://doi.org/10.1002/0471250953.bi0123s44.

Lynch, M. (2010) 'Evolution of the mutation rate', *Trends in Genetics*, 26(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003.

Matkarimov, B.T. *et al*. (2020) 'DNA Repair and Mutagenesis in Vertebrate Mitochondria: Evidence for Asymmetric DNA Strand Inheritance', in D.O. Zharkov (ed.) *Mechanisms of Genome Protection and Repair*, Springer, Cham., 77–100. https://doi.org/10.1007/978-3-030-41283-8_6.

Maude, H. *et al*. (2019) 'NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele', *Frontiers in Cell and Developmental Biology*, 7, 201. https://doi.org/10.3389/fcell.2019.00201.

McFarland, R. *et al*. (2004) 'Assigning pathogenicity to mitochondrial tRNA mutations: when "definitely maybe" is not good enough', *Trends in Genetics*, 20(12), 591–596. https://doi.org/10.1016/j.tig.2004.09.014.

Meacham, F. *et al*. (2011) 'Identification and correction of systematic error in high-throughput sequence data', *BMC Bioinformatics*, 12(1), 1–11. https://doi.org/10.1186/1471-2105-12-451.

Mehlmann, L.M. (2005) 'Stops and starts in mammalian oocytes: recent advances in understanding the regulation of meiotic arrest and oocyte maturation', *Reproduction*, 130(6), 791–799. https://doi.org/10.1530/rep.1.00793.

Monnot, S. *et al*. (2011) 'Segregation of mtDNA throughout human embryofetal development: m.3243A>G as a model system', *Human Mutation*, 32(1), 116–125. https://doi.org/10.1002/humu.21417.

Muller, H.J. (1964) 'The relation of recombination to mutational advance', *Mutation Research*, 106, 2–9. https://doi.org/10.1016/0027-5107(64)90047-8.

Nakamura, K. *et al*. (2011) 'Sequence-specific error profile of Illumina sequencers', *Nucleic Acids Research*, 39(13). https://doi.org/10.1093/NAR/GKR344.

Nandakumar, P. *et al*. (2021) 'Nuclear genome-wide associations with mitochondrial heteroplasmy', *Science Advances*, 7(12), eabe7520. https://doi.org/10.1126/sciadv.abe7520.

Narendra, D.P. *et al*. (2010) 'PINK1 Is Selectively Stabilized on Impaired Mitochondria to Activate Parkin', *PLOS Biology*, 8(1), e1000298. https://doi.org/10.1371/journal.pbio.1000298.

Nicholls, T.J. et el. (2018) 'Separating and Segregating the Human Mitochondrial Genome', *Trends in Biochemical Sciences*, 43(11), 869–881. https://doi.org/10.1016/j.tibs.2018.08.007.

Nikolaou, C. *et al*. (2006) 'Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes', *Gene*, 381, 34–41. https://doi.org/10.1016/j.gene.2006.06.010.

Nunnari, J. *et al*. (2012) 'Mitochondria: In Sickness and in Health', *Cell*, 148(6), 1145–1159. https://doi.org/10.1016/j.cell.2012.02.035.

Osellame, L.D. *et al*. (2012) 'Cellular and molecular mechanisms of mitochondrial function', *Best Practice & Research Clinical Endocrinology & Metabolism*, 26(6), 711–723. https://doi.org/10.1016/J.BEEM.2012.05.003.

Palozzi, J.M. *et al*. (2022) 'Mitochondrial DNA quality control in the female germline requires a unique programmed mitophagy', *Cell Metabolism*, 34(11), 1809-1823.e6. https://doi.org/10.1016/j.cmet.2022.10.005.

Park, J.S. *et al*. (2009) 'A heteroplasmic, not homoplasmic, mitochondrial DNA mutation promotes tumorigenesis via alteration in reactive oxygen species generation and apoptosis', *Human Molecular Genetics*, 18(9), 1578–1589. https://doi.org/10.1093/hmg/ddp069.

Parr, R.L. *et al*. (2006) 'The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation', *BMC Genomics*, 7(1), 185. https://doi.org/10.1186/1471-2164-7-185.

Payne, B.A.I. *et al*. (2013) 'Universal heteroplasmy of human mitochondrial DNA', *Human Molecular Genetics*, 22(2), 384–390. https://doi.org/10.1093/hmg/dds435.

Pereira, F. *et al*. (2008) 'Evidence for Variable Selective Pressures at a Large Secondary Structure of the Human Mitochondrial DNA Control Region', *Molecular Biology and Evolution*, 25(12), 2759–2770. https://doi.org/10.1093/molbev/msn225.

Petros, J.A. *et al*. (2005) 'mtDNA mutations increase tumorigenicity in prostate cancer', *Proceedings of the National Academy of Sciences*, 102(3), 719–724. https://doi.org/10.1073/pnas.0408894102.

Phipson, B. *et al*. (2010) 'Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn', *Statistical Applications in Genetics and Molecular Biology*, 9(1). https://doi.org/10.2202/1544-6115.1585.

Picard, M. (2021) 'Blood Mitochondrial DNA Copy Number: What Are We Counting?', *Mitochondrion*, 60, 1–11. https://doi.org/10.1016/j.mito.2021.06.010.

Picardi, E. *et al*. (2012) 'Mitochondrial genomes gleaned from human whole-exome sequencing', *Nature Methods*, 9(6), 523–524. https://doi.org/10.1038/nmeth.2029.

Putignani, L. *et al*. (2012) 'Preliminary evidences on mitochondrial injury and impaired oxidative metabolism in breast cancer', *Mitochondrion*, 12(3), 363–369. https://doi.org/10.1016/j.mito.2012.02.003.

Quail, M.A. *et al*. (2012) 'A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13(1), 341. https://doi.org/10.1186/1471-2164-13-341.

Radzvilavicius, A.L. *et al*. (2022) 'Organelle bottlenecks facilitate evolvability by traversing heteroplasmic fitness valleys', *Frontiers in Genetics*, 13. https://doi.org/10.3389/fgene.2022.974472.

Rand, D.M. (2001) 'The Units of Selection of Mitochondrial DNA', *Annual Review of Ecology, Evolution, and Systematics*, 32, 415–448. https://doi.org/10.1146/annurev.ecolsys.32.081501.114109.

Rathbun, M.M. *et al*. (2017) 'Considering DNA damage when interpreting mtDNA heteroplasmy in deep sequencing data', *Forensic Science International: Genetics*, 26, 1–11. https://doi.org/10.1016/j.fsigen.2016.09.008.

Reardon, J.T. *et al*. (2005) 'Nucleotide Excision Repair', *Progress in Nucleic Acid Research and Molecular Biology*, 79, 183–235. https://doi.org/10.1016/S0079-6603(04)79004-2.

Rebolledo-Jaramillo, B. *et al*. (2014) 'Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA', *Proceedings of the National Academy of Sciences*, 111(43), 15474–15479. https://doi.org/10.1073/PNAS.1409328111.

Ricci, C. *et al*. (2008) 'Mitochondrial DNA damage triggers mitochondrial-superoxide generation and apoptosis', *American Journal of Physiology-Cell Physiology*, 294(2), C413–C422. https://doi.org/10.1152/ajpcell.00362.2007.

Ring, J.D. *et al*. (2018) 'Bioinformatic removal of NUMT-associated variants in mitotiling next-generation sequencing data from whole blood samples', *Electrophoresis*, 39(21), 2785–2797. https://doi.org/10.1002/elps.201800135.

Robasky, K. *et al*. (2014) 'The role of replicates for error mitigation in next-generation sequencing', *Nature Reviews Genetics*, 15(1), 56–62. https://doi.org/10.1038/nrg3655.

Rong, Z. *et al*. (2021) 'The Mitochondrial Response to DNA Damage', *Frontiers in Cell and Developmental Biology*, 9, 1212. https://doi.org/10.3389/FCELL.2021.669379/BIBTEX.

Rossignol, R. *et al*. (2003) 'Mitochondrial threshold effects', *The Biochemical Journal*, 370(Pt 3), 751–762. https://doi.org/10.1042/BJ20021594.

Røyrvik, E.C. *et al*. (2020) 'MtDNA sequence features associated with "selfish genomes" predict tissue-specific segregation and reversion', *Nucleic Acids Research*, 48(15), 8290–8301. https://doi.org/10.1093/NAR/GKAA622.

Ruiz-Pesini, E. *et al*. (2006) 'Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA', *Human Mutation*, 27(11), 1072–1081. https://doi.org/10.1002/humu.20378.

Ryzhkova, A.I. *et al*. (2018) 'Mitochondrial diseases caused by mtDNA mutations: a mini-review', *Therapeutics and Clinical Risk Management*, 14, 1933–1942. https://doi.org/10.2147/TCRM.S154863.

Sabharwal, S.S. *et al*. (2014) 'Mitochondrial ROS in cancer: initiators, amplifiers or an Achilles' heel?', *Nature Reviews Cancer*, 14(11), 709–721. https://doi.org/10.1038/nrc3803.

Sacconi, S. *et al*. (2008) 'A functionally dominant mitochondrial DNA mutation', *Human Molecular Genetics*, 17(12), 1814–1820. https://doi.org/10.1093/hmg/ddn073.

Sage, J.M. *et al*. (2010) 'Discovery of a Novel Function for Human Rad51', *Journal of Biological Chemistry*, 285(25), 18984–18990. https://doi.org/10.1074/jbc.M109.099846.

Salas, A. *et al*. (2001) 'Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report', *International Journal of Legal Medicine*, 114(3), 186–190. https://doi.org/10.1007/s004140000164.

Santos, T.A. *et al*. (2006) 'Mitochondrial content reflects oocyte variability and fertilization outcome', *Fertility and Sterility*, 85(3), 584–591. https://doi.org/10.1016/j.fertnstert.2005.09.017.

Sasaki, T. *et al*. (2017) 'Live imaging reveals the dynamics and regulation of mitochondrial nucleoids during the cell cycle in Fucci2-HeLa cells', *Scientific Reports*, 7(1), 11257. https://doi.org/10.1038/s41598-017-10843-8.

Schaefer, A.M. *et al*. (2008) 'Prevalence of mitochondrial DNA disease in adults', *Annals of Neurology*, 63(1), 35–39. https://doi.org/10.1002/ana.21217.

Schinwelski, M. *et al*. (2015) 'Changing phenotypic expression in a patient with a mitochondrial encephalopathy due to 13042G>A de novo mutation – a 5 year follow up', *Metabolic Brain Disease*, 30(4), 1083–1085. https://doi.org/10.1007/s11011-014-9645-x.

Schirmer, M. *et al*. (2016) 'Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data', *BMC Bioinformatics*, 17(125), 1–15. https://doi.org/10.1186/s12859-016-0976-y.

Schneider, T.D. *et al*. (1990) 'Sequence logos: a new way to display consensus sequences', *Nucleic Acids Research*, 18(20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schon, E.A. *et al*. (2003) 'Neuronal degeneration and mitochondrial dysfunction', *The Journal of Clinical Investigation*, 111(3), 303–312. https://doi.org/10.1172/JCI17741.

Schöpf, B. *et al*. (2020) 'OXPHOS remodeling in high-grade prostate cancer involves mtDNA mutations and increased succinate oxidation', *Nature Communications*, 11(1), 1487. https://doi.org/10.1038/s41467-020-15237-5.

Seneca, S. *et al*. (2015) 'Analysis of the whole mitochondrial genome: translation of the Ion Torrent Personal Genome Machine system to the diagnostic bench?', *European Journal of Human Genetics*, 23(1), 41–48. https://doi.org/10.1038/ejhg.2014.49.

Shigenaga, M.K. *et al*. (1994) 'Oxidative damage and mitochondrial decay in aging', *Proceedings of the National Academy of Sciences*, 91(23), 10771–10778. https://doi.org/10.1073/pnas.91.23.10771.

Shin, M.G. *et al*. (2004) 'Mitochondrial DNA sequence heterogeneity in circulating normal human CD34 cells and granulocytes', *Blood*, 103(12), 4466–4477. https://doi.org/10.1182/blood-2003-11-3949.

Shoubridge, E. (2000) 'Mitochondrial DNA segregation in the developing embryo', *Human Reproduction*, 15(Suppl. 2), 229–234. https://doi.org/10.1093/humrep/15.suppl_2.229.

Shtolz, N. *et al*. (2019) 'The Mitochondrial Genome–on Selective Constraints and Signatures at the Organism, Cell, and Single Mitochondrion Levels', *Frontiers in Ecology and Evolution*, 7. https://doi.org/10.3389/fevo.2019.00342.

Skonieczna, K. *et al*. (2015) 'Heteroplasmic substitutions in the entire mitochondrial genomes of human colon cells detected by ultra-deep 454 sequencing', *Forensic Science International: Genetics*, 15, 16–20. https://doi.org/10.1016/J.FSIGEN.2014.10.021.

Smith, A.L.M. *et al*. (2020) 'Age-associated mitochondrial DNA mutations cause metabolic remodeling that contributes to accelerated intestinal tumorigenesis', *Nature Cancer*, 1(10), 976–989. https://doi.org/10.1038/s43018-020-00112-5.

Sondheimer, N. *et al*. (2011) 'Neutral mitochondrial heteroplasmy and the influence of aging', *Human Molecular Genetics*, 20(8), 1653–1659. https://doi.org/10.1093/HMG/DDR043.

Sosa, M.X. *et al*. (2012) 'Next-Generation Sequencing of Human Mitochondrial Reference Genomes Uncovers High Heteroplasmy Frequency', *PLOS Computational Biology*, 8(10), e1002737. https://doi.org/10.1371/JOURNAL.PCBI.1002737.

de Souza-Pinto, N.C. *et al*. (2008) 'Mitochondrial DNA, base excision repair and neurodegeneration', *DNA repair*, 7(7), 1098–109. https://doi.org/10.1016/j.dnarep.2008.03.011.

de Souza-Pinto, N.C. *et al*. (2009) 'Novel DNA mismatch-repair activity involving YB-1 in human mitochondria', *DNA Repair*, 8(6), 704–719. https://doi.org/10.1016/j.dnarep.2009.01.021.

Spikings, E.C. *et al*. (2007) 'Regulated Mitochondrial DNA Replication During Oocyte Maturation Is Essential for Successful Porcine Embryonic Development', *Biology of Reproduction*, 76(2), 327–335. https://doi.org/10.1095/biolreprod.106.054536.

Sriskanthadevan, S. *et al*. (2015) 'AML cells have low spare reserve capacity in their respiratory chain that renders them susceptible to oxidative metabolic stress', *Blood*, 125(13), 2120–2130. https://doi.org/10.1182/blood-2014-08-594408.

Stoler, N. *et al*. (2021) 'Sequencing error profiles of Illumina sequencing instruments', *NAR Genomics and Bioinformatics*, 3(1), lqab019. https://doi.org/10.1093/NARGAB/LQAB019.

Stoneking, M. (2000) 'Hypervariable Sites in the mtDNA Control Region Are Mutational Hotspots', *The American Journal of Human Genetics*, 67(4), 1029–1032. https://doi.org/10.1086/303092.

Subramanian, A. *et al*. (2005) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102.

Suissa, S. *et al*. (2009) 'Ancient mtDNA Genetic Variants Modulate mtDNA Transcription and Replication', *PLOS Genetics*, 5(5), e1000474. https://doi.org/10.1371/journal.pgen.1000474.

Sykora, P. *et al*. (2011) 'Aprataxin localizes to mitochondria and preserves mitochondrial function', *Proceedings of the National Academy of Sciences*, 108(18), 7437–7442. https://doi.org/10.1073/pnas.1100084108.

Tadi, S.K. *et al*. (2016) 'Microhomology-mediated end joining is the principal mediator of double-strand break repair during mitochondrial DNA lesions', *Molecular biology of the cell*, 27(2), 223–35. https://doi.org/10.1091/mbc.E15-05-0260.

Tang, Z. *et al*. (2022) 'A Genetic Bottleneck of Mitochondrial DNA During Human Lymphocyte Development', *Molecular Biology and Evolution*, 39(5), msac090. https://doi.org/10.1093/molbev/msac090.

Tao, Y. *et al*. (2023) 'Comprehensive Identification of Mitochondrial Pseudogenes (NUMTs) in the Human Telomere-to-Telomere Reference Genome', *Genes*, 14(11), 2092. https://doi.org/10.3390/genes14112092.

Trifunovic, A. (2006) 'Mitochondrial DNA and ageing', *Biochimica et biophysica acta*, 1757, 611–7. https://doi.org/10.1016/j.bbabio.2006.03.003.

Trounce, I. *et al*. (1994) 'Cytoplasmic transfer of the mtDNA nt8993 T-->G (ATP6) point mutation associated with Leigh syndrome into mtDNA-less cells demonstrates cosegregation with a decrease in state III respiration and ADP/O ratio', *Proceedings of the National Academy of Sciences*, 91(18), 8334–8338. https://doi.org/10.1073/pnas.91.18.8334.

Twig, G. *et al*. (2008) 'Fission and selective fusion govern mitochondrial segregation and elimination by autophagy', *The EMBO Journal*, 27(2), 433–446. https://doi.org/10.1038/sj.emboj.7601963.

Viale, A. *et al*. (2014) 'Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function', *Nature*, 514(7524), 628–632. https://doi.org/10.1038/nature13611.

Vilmi, T. *et al*. (2005) 'Sequence Variation in the tRNA Genes of Human Mitochondrial DNA', *Journal of Molecular Evolution*, 60(5), 587–597. https://doi.org/10.1007/s00239-003-0202-1.

Voets, A.M. *et al*. (2011) 'Large scale mtDNA sequencing reveals sequence and functional conservation as major determinants of homoplasmic mtDNA variant distribution', *Mitochondrion*, 11(6), 964–972. https://doi.org/10.1016/j.mito.2011.09.003.

Vydt, T.C.G. *et al*. (2007) 'Cardiac involvement in adults with m.3243A>G MELAS gene mutation', *The American Journal of Cardiology*, 99(2), 264–269. https://doi.org/10.1016/j.amjcard.2006.07.089.

Wang, C. *et al*. (2009) 'The Role of Mitochondria in Apoptosis', *Annual Review of Genetics*, 43, 95–118. https://doi.org/10.1146/annurev-genet-102108-134850.

Wang, J. *et al*. (2015) 'Genome measures used for quality control are dependent on gene function and ancestry', *Bioinformatics*, 31(3), 318–323. https://doi.org/10.1093/BIOINFORMATICS/BTU668.

Wang, W. *et al*. (2023) 'Heteroplasmy and Individual Mitogene Pools: Characteristics and Potential Roles in Ecological Studies', *Biology*, 12(11), 1452. https://doi.org/10.3390/biology12111452.

Wanrooij, P.H. *et al*. (2010) 'G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation', *Proceedings of the National Academy of Sciences*, 107(37), 16072–16077. https://doi.org/10.1073/pnas.1006026107.

Wauchope, O.R. *et al*. (2018) 'Oxidative stress increases M1dG, a major peroxidation-derived DNA adduct, in mitochondrial DNA', *Nucleic Acids Research*, 46(7), 3458–3467. https://doi.org/10.1093/nar/gky089.

Wei, W. *et al*. (2012) 'A role for small RNAs in DNA double-strand break repair', *Cell*, 149(1), 101–12. https://doi.org/10.1016/j.cell.2012.03.002.

Wei, W. *et al*. (2019) 'Germline selection shapes human mitochondrial DNA diversity', *Science*, 364(6442), eaau6520. https://doi.org/10.1126/science.aau6520.

Wei, W. *et al*. (2022) 'Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes', *Nature*, 611(7934), 105–114. https://doi.org/10.1038/s41586-022-05288-7.

Weissensteiner, M.H. *et al*. (2022) 'Distinct sequencing success at non-B-DNA motifs', *bioRxiv*, 2022.06.13.495922. https://doi.org/10.1101/2022.06.13.495922.

Westermann, B. (2010) 'Mitochondrial fusion and fission in cell life and death', *Nature Reviews Molecular Cell Biology*, 11(12), 872–884. https://doi.org/10.1038/nrm3013.

Whitaker-Menezes, D. *et al*. (2011) 'Hyperactivation of oxidative mitochondrial metabolism in epithelial cancer cells in situ', *Cell Cycle*, 10(23), 4047–4064. https://doi.org/10.4161/cc.10.23.18151.

Wu, Y. *et al*. (2022) 'Targeting oxidative phosphorylation as an approach for the treatment of ovarian cancer', *Frontiers in Oncology*, 12. https://doi.org/10.3389/fonc.2022.971479.

Xue, L. *et al*. (2023) 'The Mighty NUMT: Mitochondrial DNA Flexing Its Code in the Nuclear Genome', *Biomolecules*, 13(5), 753. https://doi.org/10.3390/biom13050753.

Ye, F. *et al*. (2014) 'High-throughput sequencing in mitochondrial DNA research', *Mitochondrion*, 17, 157–163. https://doi.org/10.1016/j.mito.2014.05.004.

Ye, K. *et al*. (2014) 'Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals', *Proceedings of the National Academy of Sciences*, 111(29), 10654–10659. https://doi.org/10.1073/PNAS.1403521111.

Yuan, Y. *et al*. (2020) 'Comprehensive molecular characterization of mitochondrial genomes in human cancers', *Nature Genetics*, 52(3), 342–352. https://doi.org/10.1038/s41588-019-0557-x.

Zhang, H. *et al*. (2018) 'The mitochondrial DNA genetic bottleneck: inheritance and beyond', *Essays in Biochemistry*, 62(3), 225–234. https://doi.org/10.1042/EBC20170096.

Zheng, W. *et al*. (2006) 'Origins of human mitochondrial point mutations as DNA polymerase γ-mediated errors', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 599(1), 11–20. https://doi.org/10.1016/j.mrfmmm.2005.12.012.

Zhou, L. *et al*. (2019) 'Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing', *Scientific Reports*, 9(1), 10383. https://doi.org/10.1038/s41598-019-46875-5.

Acknowledgements

First of all, I would like to express my sincere gratitude to Oxford Interdisciplinary DTP and to Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) for supporting this research project. This work was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/M011224/1].

I would like to thank all the amazing scientists and academics who have supported me through this journey, and without whom I would not have been able to complete this research. My heartfelt thanks especially go to:

- My supervisor, Prof. David Vaux, for his enthusiasm for the project, and for his unending support, encouragement, and patience;
- My examiners, Prof. Joanna Poulton and Prof. Iain Johnston, for providing the much needed constructive criticism and guidance that helped shape this thesis into its final form;
- DTC programme director, Prof. Esther Becher, for keeping an eye on my progress over the years and providing invaluable advice and support for navigating the practicalities of my degree.

I am also deeply grateful to my friends and family, without whom I would not have made it through to the end of my doctoral degree. My deepest thanks to:

- The many amazing people I met through Wolfson College Boat Club, and especially Tom, James, Gareth, and Jess, for friendship and support, and for providing a much needed distraction from the academics and keeping me sane;
- My dear friend Elīze, for always being there for me, and for picking up the conversation wherever we left off even when years have passed in between;
- My parents Ilze and Kārlis, for their unconditional love and support, and for instilling in me from a young age a curiosity and love for learning. *Paldies maniem mīļajiem vecākiem, par Jūsu mīlestību un atbalstu!*
- And finally, the biggest thanks of all to my husband Scott, for being my rock and support when the going was tough, for your patience and encouragement, and for standing by me and believing in me even when I did not believe in myself. I could not have done this without you.

Supplementary Information

Table of Contents for Supplementary Information

Appendix 1 Supplementary information for Chapter 3 210

A1.1 Mitochondrial genome %NR landscapes 211
A1.1.1 Mitochondrial %NR landscape shape 211
A1.1.2 Site-specific %NR behaviour across the entire analysed population 215
A1.2 mtDNA sequence context and the overall %NR behaviour 216
A1.2.1 Comparison between different measures of %NR distribution behaviour 216
A1.2.2 Quantification of the association between %NR behaviour and DNA sequence 219
A1.2.3 Relative enrichment of different sequence contexts among positions ranked by %NR behaviours 222
A1.3 mtDNA sequence context and %NR differences between sequencing directions 227
A1.3.1 Comparison between three measures of sequencing direction specific %NR behaviour differences 227
A1.3.2 Link between sequencing direction %NR behaviour differences and the base type 228

Appendix 2 Supplementary information for Chapter 4 230

A2.1 Sequence contexts associated with highest error rates 231
A2.1.1 Is the pattern 'GGT' a good predictor for high sequencing error rate? 231
A2.1.2 Sequencing error rate comparison between closely related motifs 232
A2.1.3 Predictive power of motif patterns for identification of error sites 234
A2.2 Sequencing error rate variability between samples 235
A2.2.1 Motif error rate variation between individuals and relationship to population ΣMER level 235
A2.2.2 MER distribution variability between motifs 238
A2.2.3 Sample-level MER distribution variability between individuals 239
A2.3 Measuring sample 'error proneness' 240
A2.3.1 Absolute 'error proneness' of a sample 241
A2.3.2 Describing individual's relative 'error proneness' 247
A2.3.3 Motif likelihood of being a stratification outlier in terms of MER rank 251
A2.3.4 Sequence patterns commonly found among outlier motifs 256

Appendix 3 Supplementary information for Chapter 5 259

A3.1 Heteroplasmy measurement method comparison 261
A3.1.1 Putative heteroplasmies detected at different allele filtration stringencies 261
A3.1.2 Putative heteroplasmy versus base type 275
A3.1.3 Modelling CSE impact on detected heteroplasmy major allele base composition 277
A3.2 Comparison with literature 283
A3.2.1 Heteroplasmy report prevalence in literature 283
A3.2.2 Comparison between heteroplasmies reported in this study with those reported in literature 285
A3.3 Heteroplasmy prevalence landscapes 288
A3.3.1 Observed heteroplasmy prevalence landscapes across the mitochondria genome 288
A3.3.2 Heteroplasmy prevalence comparison between standard and consensus methods 293
A3.3.3 Positions with highest heteroplasmy prevalence detected by standard method 294
A3.4 Heteroplasmy levels 299
A3.4.1 Average MAF in population 299
A3.4.2 Average MAF of the identified heteroplasmies 303
A3.4.3 Observed heteroplasmy MAF variability 310
A3.4.4 Relationship between MAF and the sequence context 312
A3.4.5 Overlap between MAF levels and previous heteroplasmy reports in literature 320
A3.4.6 MAF at the positions with highest heteroplasmy prevalence in population 322
A3.5 Relationship between mitochondrial heteroplasmy and the genomic context 325
A3.5.1 Heteroplasmy differences between mitochondrial genome functional elements 325
A3.5.2 Heteroplasmy patterns in protein coding genes 327
A3.5.3 Relationships between heteroplasmy and the base type in different genomic contexts 328
A3.6 Comparison between novel versus previously known heteroplasmies 329
A3.6.1 Distributions of the known versus novel heteroplasmies across the mitochondrial genome 329
A3.6.2 Base composition comparison 333
A3.6.3 Heteroplasmy level comparison 336
A3.6.4 Heteroplasmic variants 338

Appendix 1 Supplementary information for Chapter 3

List of supplementary figures associated with Chapter 3

A1.1 Supplementary figures associated with section 3.2 of the main text

3.2.2 S1: High incidence %NR peak positions within the mtDNA region m.1000 - 1500 212
3.2.2 S2: Population fraction with %NR peaks above sample %NR 95th percentile level 213
3.2.2 S3: Full mtDNA non-reference read frequency landscapes in samples 413 and 579 214
3.2.3 S1: Position-specific %NR distributions at 20 randomly selected mtDNA positions 215

A1.2 Supplementary figures associated with section 3.3 of the main text

3.3.1 S1: mtDNA position rank order by three %NR distribution behaviour measures 217
3.3.1 S2: %NR median, 95th percentile and distribution shape atypicality comparison 218
3.3.2 S1: Base type enrichment among positions ranked by three %NR behaviour measures 220
3.3.2 S2: Quantification of the association between base type and different %NR behaviours 221
3.3.2 S3: Sequence context and %NR behaviour: Positively enriched downstream motifs 223
3.3.2 S4: Sequence context and %NR behaviour: Positively enriched upstream motifs 224
3.3.2 S5: Sequence context and %NR behaviour: Negatively enriched downstream motifs 225
3.3.2 S6: Sequence context and %NR behaviour: Negatively enriched upstream motifs 226

A1.3 Supplementary figures associated with section 3.4 of the main text

3.4 S1: mtDNA positions ranked by three measures of %NR behaviour differences between sequencing directions 227
3.4.1 S1: Different base type enrichment among mtDNA positions ranked by three measures of %NR behaviour differences between sequencing directions 228
3.4.1 S2: Quantification of the association between the base type and %NR behaviour differences between sequencing directions 229

A1.1 Mitochondrial genome %NR landscapes

A1.1.1 Mitochondrial %NR landscape shape

Supplementary figure 3.2.2 S1 shows 'peak' positions within the mtDNA region from position 1000 to 1500, where %NR in at least 50% of all analysed samples exceeds the sample %NR 95th percentile level. The exact peak heights of the 6 sample %NR landscapes visualised in the main text **figure 3.2.2 F1** are listed on the left, with the number of total analysed samples in which that position represents an %NR peak shown on the right. **Supplementary figure 3.2.2 S1** demonstrates that many %NR levels at many of the 'peak' positions are exceed the sample %NR 95th percentile multiple times. Most %NR peaks are single base wide. Individual %NR peak heights do not appear to track with sample %NR baseline level, or other nearby peak heights.

Supplementary figure 3.2.2 S2 visualises in what proportion of the analysed samples any given mtDNA position constitutes an %NR landscape 'peak' with %NR level within top 5% for that sample. The entire length of MT genome is shown with '+' marking positions where the observed %NR peak count significantly exceeds the expected if %NR levels were random and independent of MT genome location (permutation test, n=500,000, significance threshold $p<0.05$ after correction for multiple comparisons).

Supplementary figure 3.2.2 S3 visualises the two examples of the actual observed %NR landscapes. Unlike **3.2.2 S2** where Y axis represented proportion of analysed samples with %NR peaks above their P_{95} %NR level, in this figure Y axis represents the actual %NR level observed at every mtDNA position in samples 413 (above X axis, %NR range: 0% -10%) and 579 (below X axis, Y axis shown reversed, %NR range: 0%-10%).

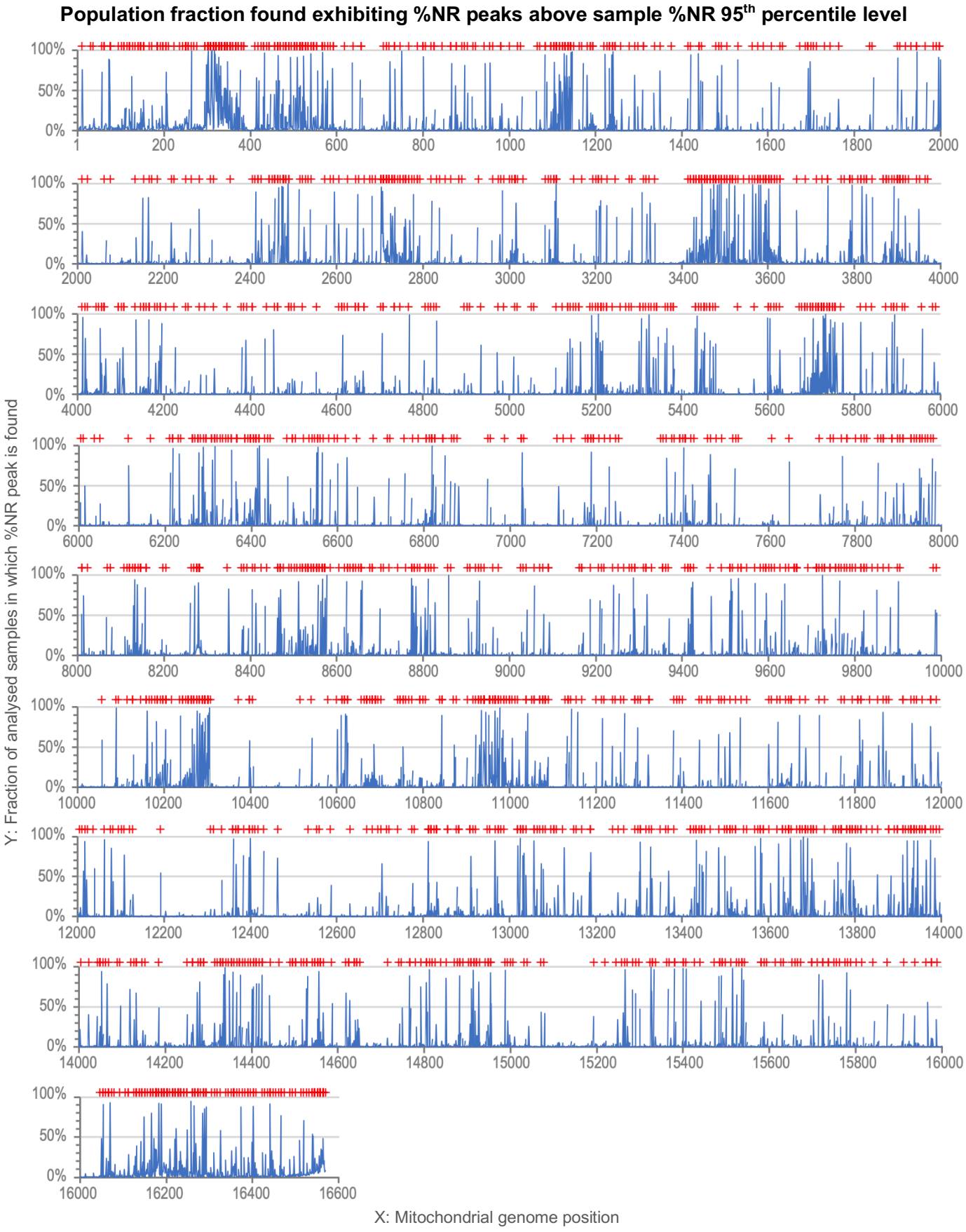
Supplementary figure 3.2.2 S1: High incidence %NR peak positions within the mtDNA region m.1000 - 1500

Positions with high incidence of %NR peaks within the mtDNA region m.1000 - 1500

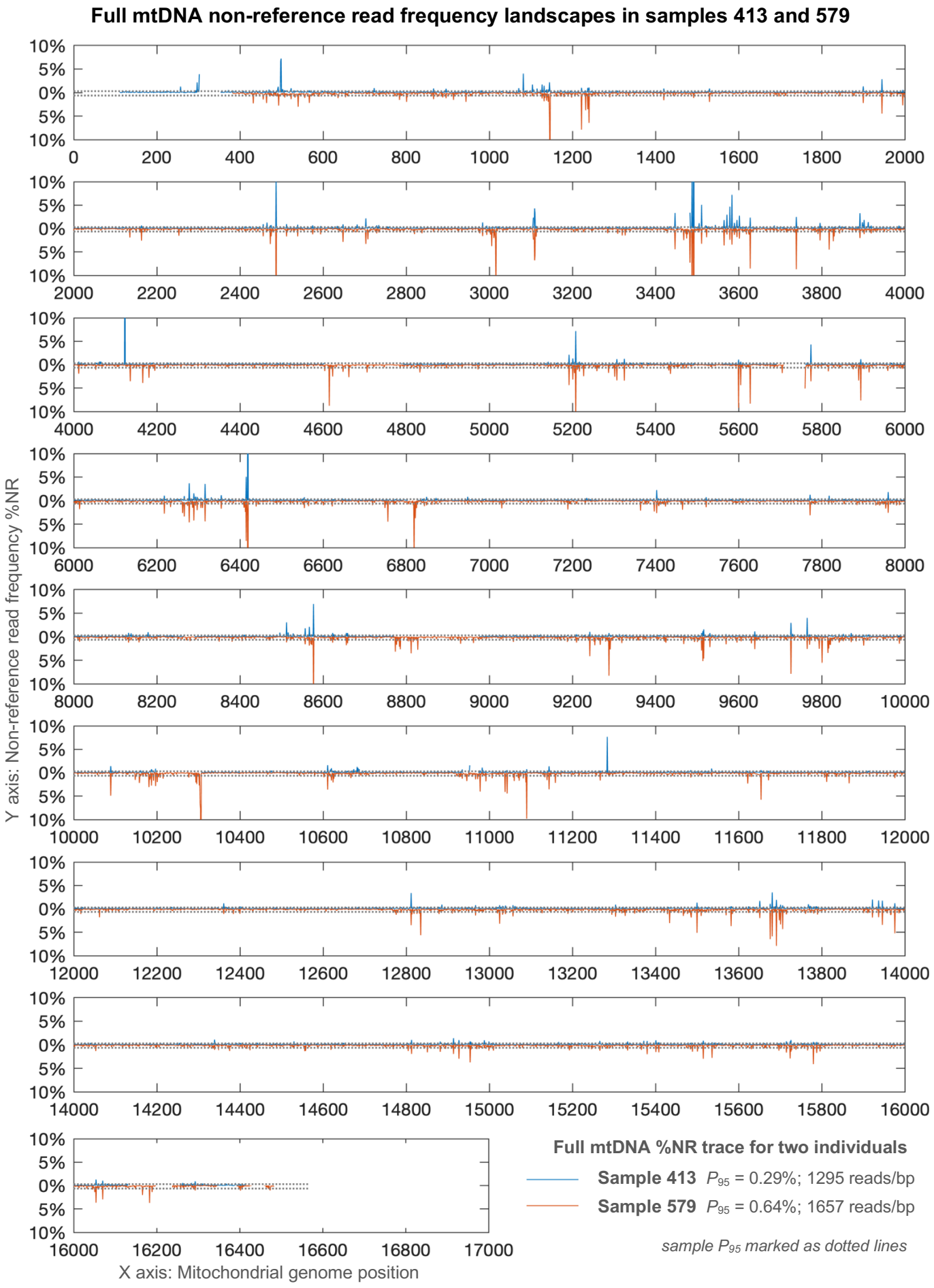
mtDNA position	rCRS base	Sample 114 P_{95} = 1.31%	Sample 413 P_{95} = 0.29%	Sample 579 P_{95} = 0.64%	Sample 1306 P_{95} = 0.29%	Sample 1960 P_{95} = 0.64%	Sample 1980 P_{95} = 3.00%	%NR peaks in the 6 visualised samples		%NR peaks among the full population	
1082	A		4.03%	0.99%	4.01%	4.22%	4.00%	5	83.3%	2397	94.6%
1087	A		0.39%				4.12%	2	33.3%	1531	60.4%
1104	A		1.70%	0.75%	0.82%	1.27%	3.86%	5	83.3%	2150	84.8%
1115	T		0.85%		0.31%	0.70%		3	50.0%	1761	69.5%
1119	T				0.31%			1	16.7%	1547	61.0%
1127	A	3.95%	1.72%	1.25%	1.34%	1.75%	4.60%	6	100%	2435	96.1%
1129	T	2.64%	0.47%		0.32%		3.29%	4	66.7%	1719	67.8%
1132	T	5.61%	1.44%	1.85%	0.99%	1.08%	5.33%	6	100%	2369	93.5%
1134	G	2.60%	0.39%	1.14%		1.09%	4.80%	5	83.3%	1697	66.9%
1138	G	3.45%	0.68%	1.73%		0.71%	4.61%	5	83.3%	1717	67.7%
1140	A			1.09%			3.60%	2	33.3%	1346	53.1%
1142	A	6.47%	0.65%	1.96%	0.94%	1.98%	7.51%	6	100%	2466	97.3%
1144	T	2.84%	0.38%	1.02%	0.33%		4.04%	5	83.3%	1730	68.2%
1145	A	8.28%	2.17%	6.09%	0.84%	1.78%	8.24%	6	100%	2487	98.1%
1146	C	4.49%	0.32%	17.86%	0.82%	4.19%		5	83.3%	2228	87.9%
1171	A	3.20%	0.55%	0.84%	0.42%	0.72%	4.43%	6	100%	2120	83.6%
1180	T	2.09%		0.97%	0.42%	0.84%	7.57%	5	83.3%	1751	69.1%
1222	A	3.42%	1.01%	7.88%	0.80%	1.08%	7.90%	6	100%	2409	95.0%
1232	A	1.67%	0.39%	3.76%	0.34%		8.33%	5	83.3%	1962	77.4%
1237	A	3.04%	1.01%	3.55%	1.03%	1.60%	13.54%	6	100%	2393	94.4%
1240	A	5.36%	1.07%	6.44%	2.12%	3.03%	14.21%	6	100%	2485	98.0%
1290	C			1.00%	1.04%	1.43%		3	50.0%	1753	69.2%
1309	A	1.32%			0.60%	0.90%	3.42%	4	66.7%	1284	50.7%
1349	T	1.53%			0.41%	1.21%	3.96%	4	66.7%	1881	74.2%
1420	T	3.19%	0.85%	1.64%	1.00%	1.95%	4.94%	6	100%	2389	94.2%
1443	T			0.70%	0.66%	1.34%		3	50.0%	1565	61.7%
1448	T				0.55%	0.84%		2	33.3%	1526	60.2%
1482	A				0.43%	1.29%	3.49%	3	50.0%	1726	68.1%
1492	A	1.82%		1.15%	1.22%	2.33%	3.20%	5	83.3%	2065	81.5%

%NP peak defined as the position-specific %NR level above sample MT genome %NR 95th percentile (P_{95}) level. Only positions with %NR peaks in ≥50% of all analysed individuals are included; %NR values below sample P_{95} level not shown. Columns on the right in blue show comparison to the total number of %NR peaks observed across the entire analysed population

Supplementary figure 3.2.2 S2: Population fraction with %NR peaks above sample %NR 95th percentile level



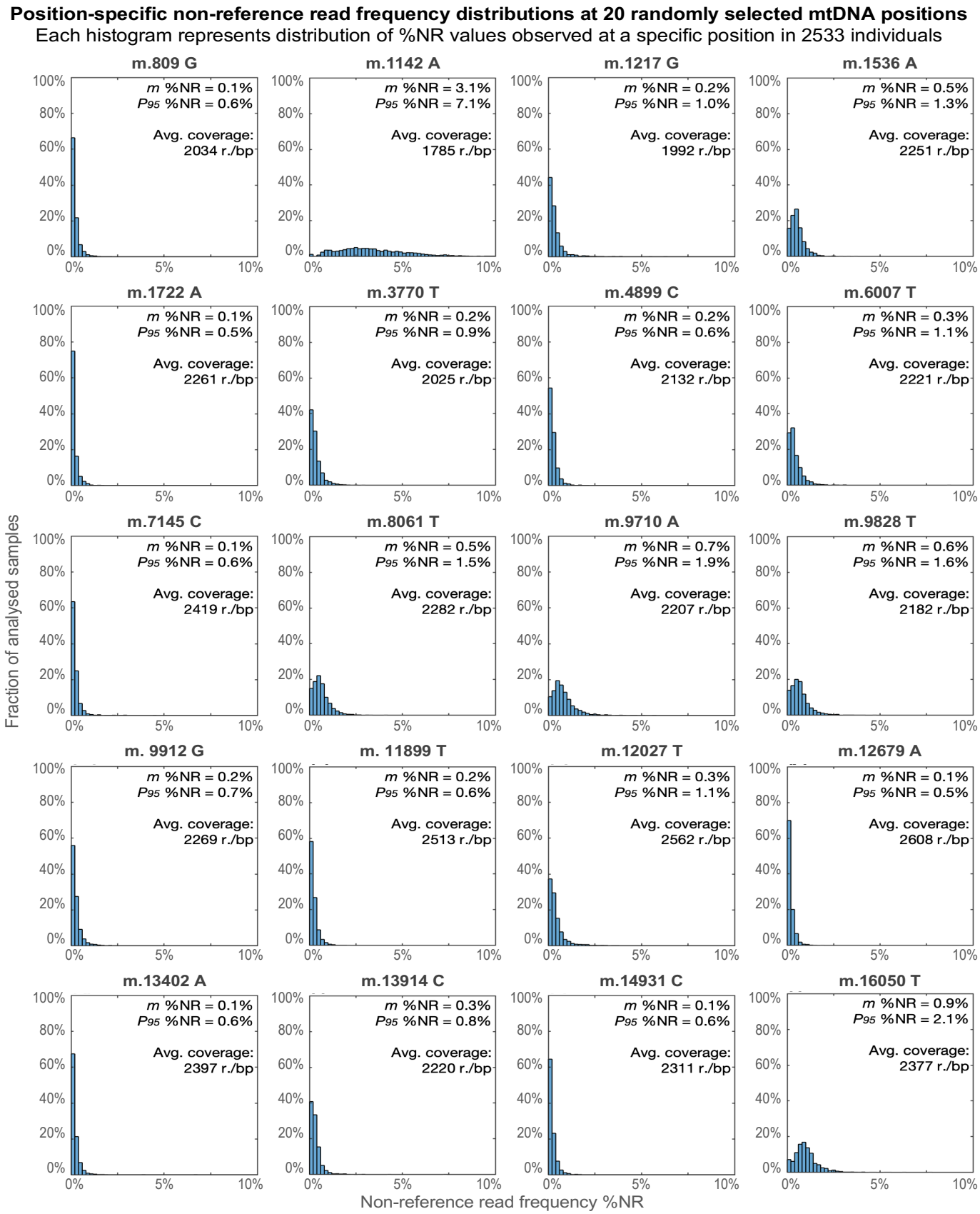
Supplementary figure 3.2.2 S3: Full mtDNA non-reference read frequency landscapes in samples 413 and 579



A1.1.2 Site-specific %NR behaviour across the entire analysed population

Supplementary figure 3.2.3 S1 visualises %NR population distributions observed at 20 randomly selected MT genome positions, demonstrating a range of distribution shapes comparable to that observed across 20 consecutive MT genome positions explored in **figure 3.2.3 F1** in the main text. For position-specific %NR distribution shape variability on the MT genome scale, see **figure 3.2.3 F2** in the main text.

Supplementary figure 3.2.3 S1: Position-specific %NR distributions at 20 randomly selected mtDNA positions



A1.2 mtDNA sequence context and the overall %NR behaviour

A1.2.1 Comparison between different measures of %NR distribution behaviour

Supplementary figures 3.3.1 S1 and **3.3.1 S2** compare three different mtDNA position-specific %NR distribution behaviour measures: median, 95th percentile and the overall distribution shape atypicality, expressed as KS2 test statistic comparing site-specific %NR distribution to the overall %NR distribution across all sites. **Supplementary figure 3.3.1 S1** visualises comparison between the outcomes of mtDNA position ranking by each of these %NR behaviour measures, while **Supplementary figure 3.3.1 S2** compates the actual mtDNA position %NR median, 95th percentile and distribution shape atypicality measure values. In both figures a three way comparison between all three measures is shown in **panel A**, whereas pairwise comparisons are presented in **panel B**.

Observations:

The oblong data point scatter observed when comparing mtDNA position rank order by %NR median and 95th percentile indicates that even though both measures behave in a largely coordinate fashion, this is a considerable variation in position ranking with some data points found at least 2-3 deciles away from the diagonal expected of highly similar rank orders. The inverse V scatter patterns when comparing %NR distribution atypicality with either %NR median or 95th percentile are a direct consequence of what these different measures represent. Position ranking from highest to lowest by %NR quantile values separates one extreme of a %NR behaviour range from the other, whereas ranking position by %NR distribution atypicality separates positions at both extremes of the possible %NR behaviour range from those in the middle of the range showing the most average %NR behaviours. Therefore we should expect positions with both highest and lowest %NR median and 95th percentile levels to be found among those ranked highest by %NR behaviour atypicality. This is exactly what we see.

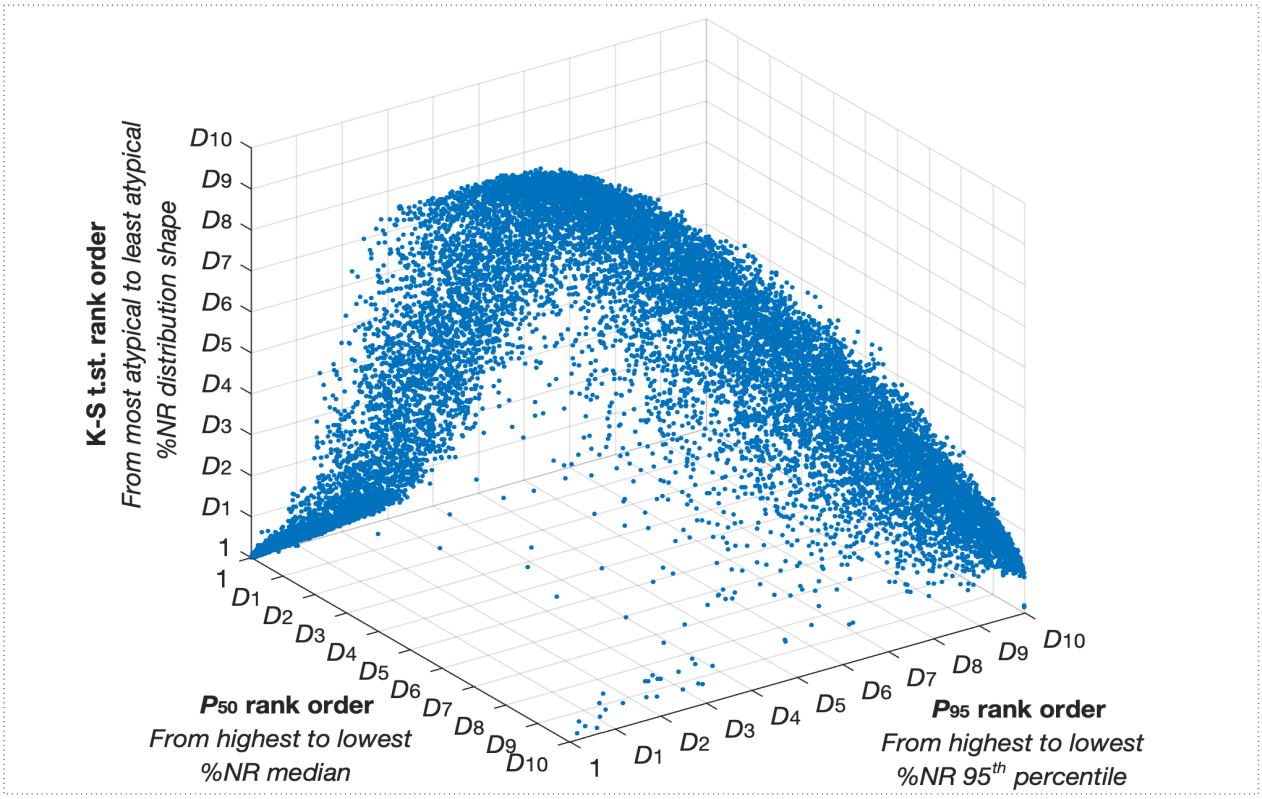
It is interesting that the inverse V pattern, while still present, in much less clear for the relationship between %NR distribution shape atypicality and the %NR 95th percentile compared to %NR median, with more variability between position atypicality ranking observed at high %NR 95th percentile values compared to low %NR 95th percentile values. This ties in with earlier assessment that a 'typical' mtDNA position has a single relatively narrow %NR distribution peak followed by a long, low positive tail, with more variation between positions observed in the %NR distribution tails. At higher %NR 95th percentile values the range of possible %NR distribution shapes with that specific %NR 95th percentile is higher, leading to more possible variation in the shape atypicality.

It is worth addressing two interesting points in the relationship between %NR levels and the distribution shape atypicality, that inform us about the shapes of an average. One point relates to the differences in composition of top 10% of sites with highest %NR atypicality compared to the rest of the dataset, with the top decile of the most atypical positions consist in almost entirely of sites with high median %NR, whereas those below the top 10% threshold represent a mixture of both high and low median %NR. Median %NR of position within this top decile are above $\sim$1.2%, representing a set of distinctly 'above average' positions in terms of %NR levels with KS2 test statistic values above ~ 0.4. The second point of interest is the location where the relationship between %NR atypicality and %NR median changes the direction from a direct relationship to an inverse one. This is around the border of deciles 4 and 5 in terms of position rank order by %NR atypicality, with around 60% of all analysed positions with 'below average' %NR levels characterised by %NR median below $\sim$0.2% and %NR 95th percentile below $\sim$0.8%-1%.

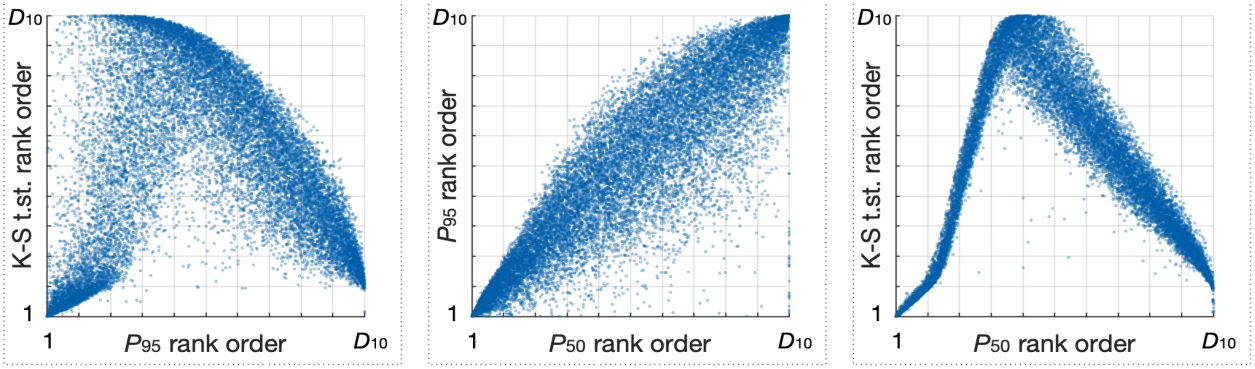
Supplementary figure 3.3.1 S1: mtDNA position rank order by three %NR distribution behaviour measures

Comparison of mtDNA position rank order by population %NR distribution median versus 95th percentile values versus the overall %NR distribution shape atypicality

(a) mtDNA position rank order by decreasing KS test statistic vs P_{50} vs P_{95} values



(b) Pairwise comparison of mtDNA position rank order by KS test statistic vs P_{50} vs P_{95} values



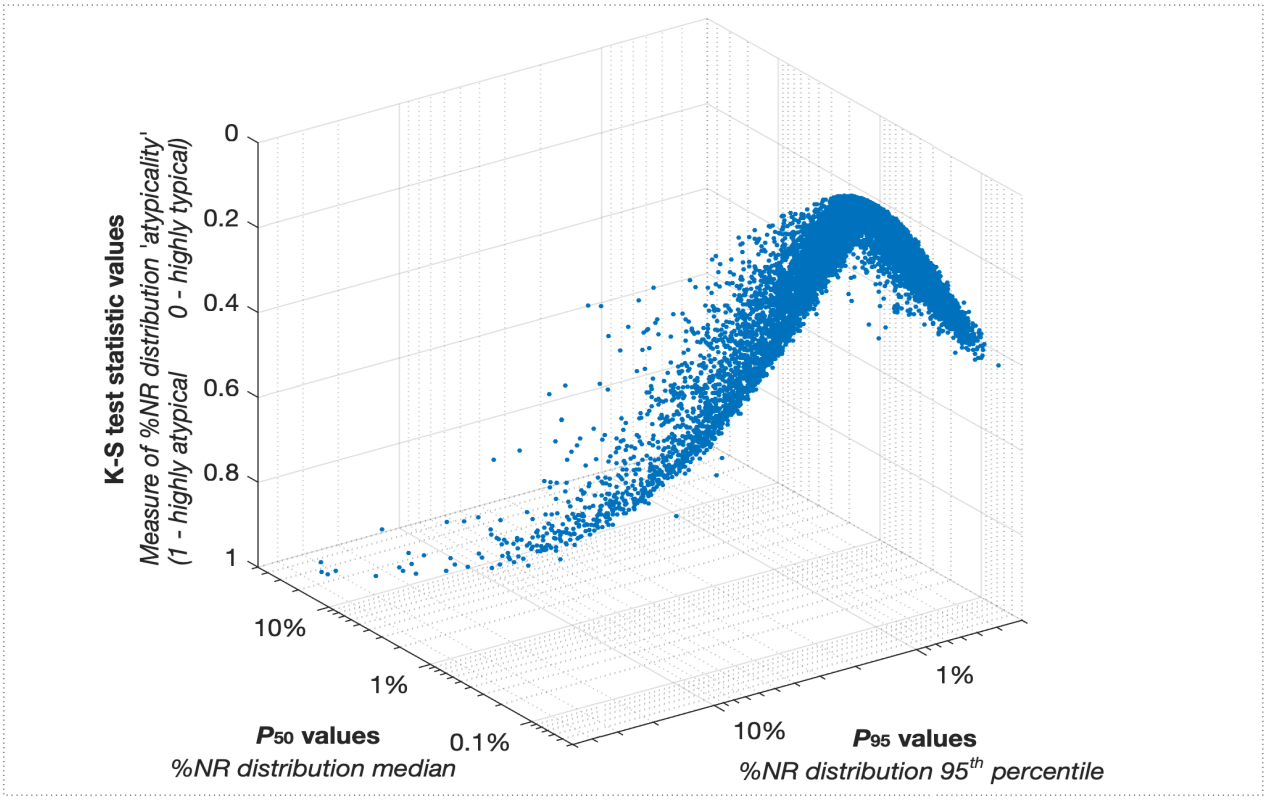
KS t.st. vs P_{95} **P_{95} vs P_{50}** **KS t.st. vs P_{50}**

Visual comparison between outcomes of mtDNA positions ranking based on different %NR distribution behaviour measures
A *– Three-way comparison of position rank order by %NR median (P_{50}, x-axis), 95th percentile (P_{95}, y-axis) and distribution shape atypicality (KS t.st., z-axis).* ***B*** *– Pairwise position ranking outcome comparisons. Positions are ranked from highest to lowest, with rank = 1 representing highest observed %NR distribution atypicality/median/95th percentile. %NR – non-reference read frequency. P_{50} – median, P_{95} – 95th percentile, KS t.st. – Kolmogorov-Smirnov test statistic. D_1-D_{10} – position rank order deciles*

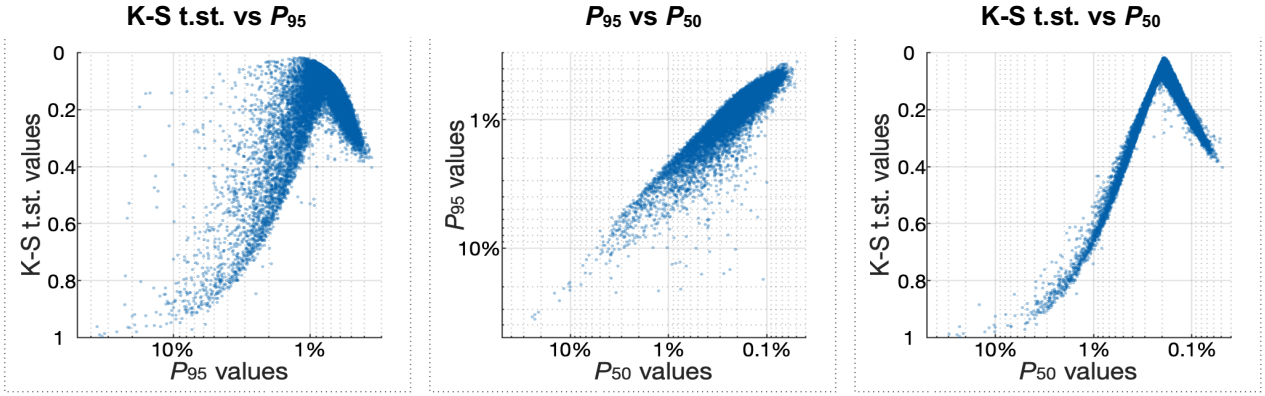
Supplementary figure 3.3.1 S2: %NR median, 95th percentile and distribution shape atypicality comparison

Comparison of mtDNA position population %NR median, 95th percentile and the overall distribution shape atypicality at each mtDNA position across 2533 samples

(a) mtDNA position %NR distribution P_{50} vs P_{95} values vs atypicality expressed as K-S test statistic



(b) Pairwise comparison of mtDNA position %NR distribution K-S test statistic vs P_{50} vs P_{95} values



KS t.st. vs P_{95} **P_{95} vs P_{50}** **KS t.st. vs P_{50}**

Visual comparison between three different measures of mtDNA position %NR distribution behaviour
A *– Three-way comparison of position %NR median (P_{50}, x-axis, log scale), 95th percentile (P_{95}, y-axis, log scale) and distribution shape atypicality measure (KS t.st., z-axis) values.* ***B*** *– Pairwise position %NR distribution behaviour measure comparisons. All axes inversed to maintain same scatter orientation as used in rank order visualisations. %NR – non-reference read frequency. P_{50} – median, P_{95} – 95th percentile, KS t.st. – Kolmogorov-Smirnov test statistic*

A1.2.2 Quantification of the association between %NR behaviour and DNA sequence

Supplementary figures 3.3.2 S1 and **S2** demonstrate the principles of how the strength of association between %NR behaviour and certain mtDNA sequence features, such as the position base type, can be quantified.

If a certain base type is linked to specific %NR behaviour, we would expect higher proportion this base type to be found among positions more strongly exhibiting the %NR behaviour in question. This is demonstrated in **Supplementary figures 3.3.2 S1,** where enrichment of different bases among mtDNA positions ranked by %NR distribution median, 95th percentile and the overall shape atypicality is visualised as cumulative frequency curves. The more overrepresented the base type at the top of the ranked position list, the more shifted above the diagonal baseline the corresponding cumulative frequency curve should be.

The strength of the association between base types and the different %NR behaviours can be quantified via enrichment scores, shown in **Supplementary figures 3.3.2 S2,** describing the average difference between our observed base-of-interest cumulative frequency curve and the neutral baseline expected if no association existed between that base and the %NR behaviour being used for position ranking. For details of enrichment score calculation see **Materials and Methods 2.3**

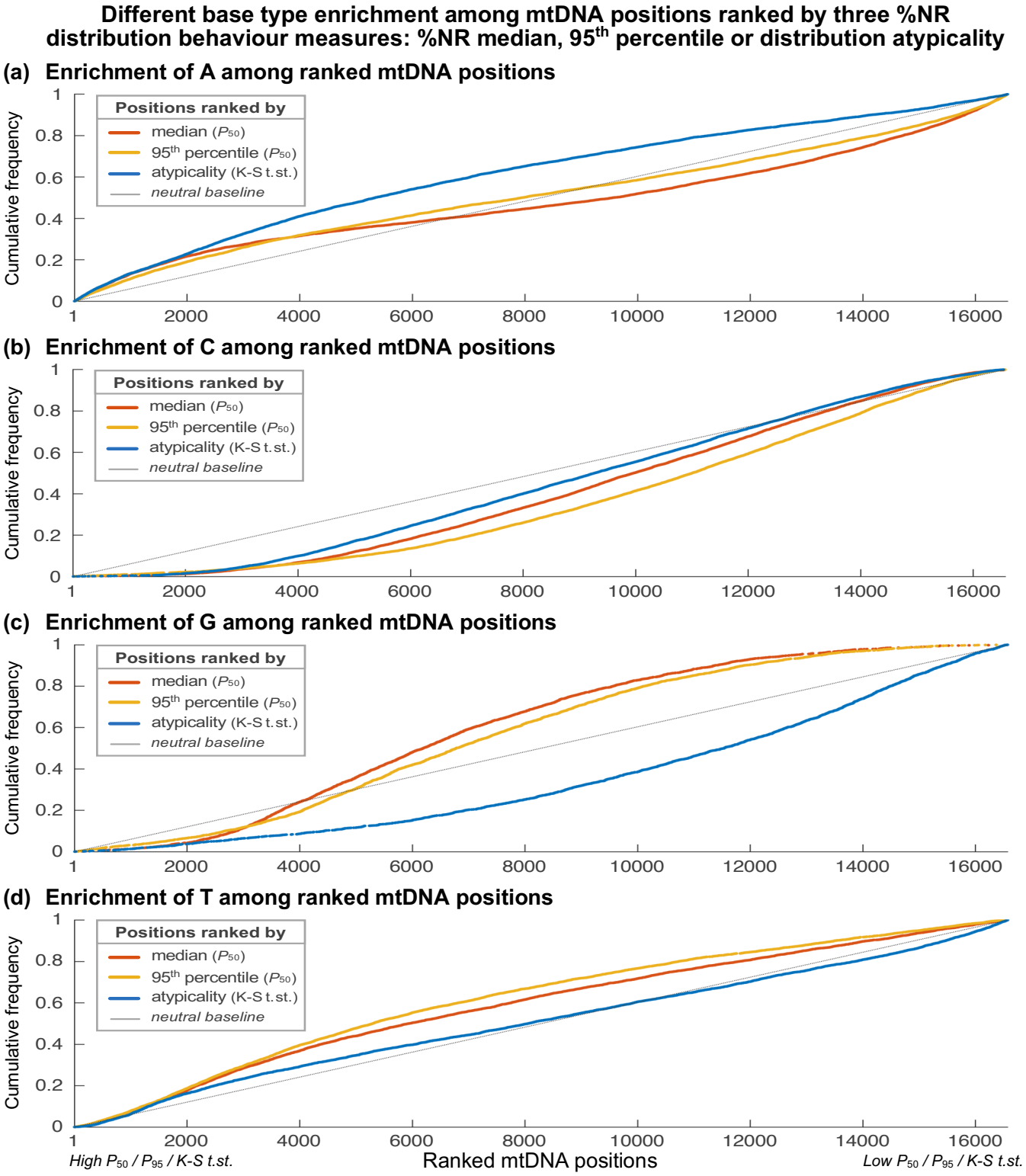
Enrichment score theoretical range is between -0.5 to +0.5. Values < 0 describe cumulative frequency curves below the neutral baseline indicating a negative association between the property of interest and the variable used for position ranking, whereas Enrichment scores > 0 describe cumulative frequency curves above neutral baseline indicating a positive association between the property of interest and the variable positions are ranked by. Enrichment index values further away from 0 indicate stronger association between the property of interest and the variable used to rank positions. For details of enrichment score calculation see **Materials and Methods 2.3**

Data presented in **Supplementary figures 3.3.2 S1** and **S2** indicate that the base types strongly associated with elevated %NR median and 95th percentile level are G and T. G shows slight stronger association with the %NR median levels while T is more strongly linked to elevated %NR 95th percentile. Meanwhile the base type showing strongest association with high overall %NR distribution shape atypicality is A.

It should be noted that different base types are not independent, with positive enrichment of some bases among the positions linked to specific %NR behaviours necessitating there to be negative enrichment of the other base types.

Base type enrichment quantification presented here confirms the qualitative observations from **figure 3.3.1 F1** in the main text.

Supplementary figure 3.3.2 S1: Base type enrichment among positions ranked by three %NR behaviour measures



Cumulative frequency of different DNA base types among mtDNA positions ranked by %NR distribution median (red), 95th percentile (yellow), or shape atypicality (blue). X-axis: position rank from highest (left) to lowest (right) by each %NR behaviour or interest. Y-axis: base of interest cumulative frequency. %NR – non-reference read frequency. P_{50} – median, P_{95} – 95th percentile, KS t.st. – Kolmogorov-Smirnov test statistic

Supplementary figure 3.3.2 S2: Quantification of the association between base type and different %NR behaviours

Comparison of base type enrichment between different mtDNA position ranking methods: by population %NR distribution median vs 95th percentile vs overall shape atypicality

Base type	Instances found	$E_{(P50)}$	$E_{(P95)}$	$E_{(KS2)}$
A	5124	-0.019 ** ↓	0.011 * ↑	0.114 ** ↑
C	5181	-0.091 ** ↓	-0.133 ** ↓	-0.057 ** ↓
G	2169	0.098 ** ↑	0.074 ** ↑	-0.145 ** ↓
T	4094	0.087 ** ↑	0.117 ** ↑	0.008 ‡

$E_{(P50)}$ – enrichment index for position rank order by %NR distribution median (P_{50}) level;
$E_{(P95)}$ – enrichment index for position rank order by %NR distribution 95th percentile (P_{95}) level;
$E_{(KS2)}$ – enrichment index for position rank order by %NR distribution atypicality, measured as K-S test statistic (KS2);
*↑ – positive enrichment, ↓ – negative enrichment; significance level: ** $p < 0.00002$, * $p = 0.0007$, ‡ $p = 0.025$*

A1.2.3 Relative enrichment of different sequence contexts among positions ranked by %NR behaviours

Enrichment scores can be used to quantify the %NR distribution behaviour relationship with not only the base at the position of interest itself but also to the adjacent sequence context. This is addressed in the main text, with **Figures 3.3.2 F1-F3** presenting the absolute enrichment scores associated with short, simple adjacent sequence motifs around the position of interest, for positions ranked by %NR distribution median, 95th percentile and shape atypicality measures. Motif enrichment addressed in the main text is the absolute enrichment, calculated against a neutral baseline. However as we have shown that the base identity at the position of interest can itself be significantly linked to %NR behaviour differences. To understand how the adjacent context is linked to %NR behaviours in itself, instead of absolute enrichment we can calculate motif relative enrichment, whereby the enrichment of the base type at position of interest is taken as the neutral baseline. This is presented in **Supplementary figures 3.3.2 S3**-**S6** below.

The upstream and downstream motifs with statistically significant positive relative enrichment among positions ranked by %NR median, 95th percentile and distribution shape atypicality are listed in **Supplementary figures 3.3.2 S3** and **S4**. The upstream and downstream motifs with statistically significant negative relative enrichment among positions ranked by %NR median, 95th percentile and distribution shape atypicality are listed in **Supplementary figures 3.3.2 S5** and **S6**.

Supplementary figure 3.3.2 S3: Sequence context and %NR behaviour: Positively enriched downstream motifs

Sequence motif relative enrichment: comparison between mtDNA position ranking methods

Significantly positively enriched downstream motifs ($E_N > 0$) with undefined variable position base type

(a) mtDNA position ranking: %NR P_{50}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
8	.... A CCCC	79	0.469
7	.... A CCC.	199	0.450
264	.... A CC.C	190	0.435
6	.... A CC..	515	0.414
265	.... A C.CC	184	0.410
5	.... A C...	1495	0.337
266	.... A .CCC	220	0.161
32	.... C TTTT	20	0.387
27	.... C GGG.	17	0.282*b
289	.... C GG.G	21	0.273*a
296	.... C .TTT	67	0.241
26	.... C GG..	80	0.234
20	.... C AAAA	39	0.221
279	.... C AA.A	150	0.202
18	.... C AA..	465	0.200
31	.... C TTT.	67	0.190
19	.... C AAA.	145	0.178
24	.... C CCCC	69	0.167
25	.... C G...	435	0.154
17	.... C A...	1534	0.108
23	.... C CCC.	224	0.099
280	.... C A.AA	145	0.083*c
40	.... G CCCC	26	0.414
39	.... G CCC.	93	0.313
304	.... G CC.C	79	0.274
305	.... G C.CC	53	0.255
38	.... G CC..	271	0.249
306	.... G .CCC	68	0.239
299	.... G AA.A	66	0.176
37	.... G C...	711	0.136
34	.... G AA..	201	0.108
56	.... T CCCC	50	0.414
55	.... T CCC.	108	0.391
324	.... T CC.C	125	0.349
325	.... T C.CC	158	0.316
54	.... T CC..	361	0.311
59	.... T GGG.	18	0.255*c
53	.... T C...	1204	0.186
52	.... T AAAA	46	0.174*a
326	.... T .CCC	135	0.162
58	.... T GG..	100	0.113*a

(b) mtDNA position ranking: %NR P_{95}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P95)}$ index
8	.... A CCCC	79	0.459
7	.... A CCC.	199	0.442
264	.... A CC.C	190	0.434
6	.... A CC..	515	0.409
265	.... A C.CC	184	0.395
5	.... A C...	1495	0.317
266	.... A .CCC	220	0.175
32	.... C TTTT	20	0.328
20	.... C AAAA	39	0.305
27	.... C GGG.	17	0.283*b
19	.... C AAA.	145	0.241
279	.... C AA.A	150	0.223
18	.... C AA..	465	0.204
296	.... C .TTT	67	0.180
31	.... C TTT.	67	0.166
24	.... C CCCC	69	0.166
26	.... C GG..	80	0.161
281	.... C .AAA	147	0.125
23	.... C CCC.	224	0.102
280	.... C A.AA	145	0.101
25	.... C G...	435	0.083
284	.... C CC.C	205	0.078*a
17	.... C A...	1534	0.056
22	.... C CC..	624	0.050*a
40	.... G CCCC	26	0.379
39	.... G CCC.	93	0.3
304	.... G CC.C	79	0.269
305	.... G C.CC	53	0.254
38	.... G CC..	271	0.239
306	.... G .CCC	68	0.213
299	.... G AA.A	66	0.127*c
37	.... G C...	711	0.124
34	.... G AA..	201	0.068*e
56	.... T CCCC	50	0.391
55	.... T CCC.	108	0.361
324	.... T CC.C	125	0.324
59	.... T GGG.	18	0.306
325	.... T C.CC	158	0.294
54	.... T CC..	361	0.283
52	.... T AAAA	46	0.202
326	.... T .CCC	135	0.161
53	.... T C...	1204	0.161
58	.... T GG..	100	0.131
51	.... T AAA.	129	0.112
320	.... T A.AA	110	0.104*c
319	.... T AA.A	149	0.083*b

(c) position ranking: %NR atypicality

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(KS2)}$ index
8	.... A CCCC	79	0.469
7	.... A CCC.	199	0.452
264	.... A CC.C	190	0.435
6	.... A CC..	515	0.405
265	.... A C.CC	184	0.387
5	.... A C...	1495	0.262
3	.... A AAA.	182	0.094
2	.... A AA..	524	0.041*e
29	.... C T...	1440	0.098
40	.... G CCCC	26	0.376
39	.... G CCC.	93	0.228
304	.... G CC.C	79	0.189
306	.... G .CCC	68	0.153
38	.... G CC..	271	0.14
37	.... G C...	711	0.038
56	.... T CCCC	50	0.392
55	.... T CCC.	108	0.363
324	.... T CC.C	125	0.308
54	.... T CC..	361	0.255
325	.... T C.CC	158	0.254
326	.... T .CCC	135	0.123
53	.... T C...	1204	0.119

$E_{N(P50)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution median (P_{50}) level;
$E_{N(P95)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution 95th percentile (P_{95}) level;
$E_{N(KS2)}$ – relative (base-specific) enrichment index for position rank order by %NR atypicality, measured as K-S test statistic;
A – adenine, C – cytosine, G – guanine, T – thymine, · – any base type;
Motifs ranked by enrichment relative to position of interest base type, only statistically significant E_N index values shown;
*Significance level: $p < 0.00002$, except: *a $p < 0.0001$, *b $p < 0.0002$, *c $p < 0.0003$, *d $p < 0.0004$, *e $p < 0.0005$*

Supplementary figure 3.3.2 S4: Sequence context and %NR behaviour: Positively enriched upstream motifs

Sequence motif relative enrichment: comparison between mtDNA position ranking methods

Significantly positively enriched upstream motifs ($E_N > 0$) with undefined variable position base type

(a) mtDNA position ranking: %NR P_{50}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
430	GG.G **A**	20	0.251*b
136	CCCC **A**	68	0.191
138	..GG **A**	123	0.168
424	C.CC **A**	162	0.139
135	.CCC **A**	163	0.135
425	CC.C **A**	191	0.113
426	CCC. **A**	198	0.100
134	..CC **A**	464	0.090
137	...G **A**	613	0.090
133	...C **A**	1534	0.038
160	TTTT **C**	16	0.308*a
159	.TTT **C**	70	0.252
449	G.GG **C**	26	0.237*a
154	..GG **C**	151	0.211
454	T.TT **C**	91	0.194
158	..TT **C**	308	0.153
441	AAA. **C**	168	0.151
153	...G **C**	711	0.092
157	...T **C**	1204	0.044
149	...C **C**	1771	0.021*a
474	T.TT **G**	29	0.213*b
170	..GG **G**	72	0.205
461	AAA. **G**	59	0.153*a
167	.CCC **G**	47	0.152*e
464	C.CC **G**	57	0.144*b
166	..CC **G**	141	0.114*a
174	..TT **G**	116	0.110*a
165	...C **G**	435	0.092
169	...G **G**	426	0.061
173	...T **G**	513	0.040*c
489	G.GG **T**	15	0.464
187	.GGG **T**	18	0.459
186	..GG **T**	80	0.434
490	GG.G **T**	17	0.432
185	...G **T**	419	0.336
184	CCCC **T**	63	0.255
180	AAAA **T**	41	0.238
183	.CCC **T**	190	0.195
484	C.CC **T**	176	0.168
485	CC.C **T**	154	0.160
481	AAA. **T**	125	0.160
486	CCC. **T**	160	0.156
179	.AAA **T**	110	0.154
182	..CC **T**	542	0.128
479	A.AA **T**	125	0.095*b
178	..AA **T**	376	0.059*a
181	...C **T**	1440	0.050

(b) mtDNA position ranking: %NR P_{95}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
136	CCCC **A**	68	0.205
424	C.CC **A**	162	0.139
138	..GG **A**	123	0.139
135	.CCC **A**	163	0.138
426	CCC. **A**	198	0.124
425	CC.C **A**	191	0.106
134	..CC **A**	464	0.076
137	...G **A**	613	0.067
449	G.GG **C**	26	0.221*a
441	AAA. **C**	168	0.200
159	.TTT **C**	70	0.193
148	AAAA **C**	52	0.156*a
154	..GG **C**	151	0.152
454	T.TT **C**	91	0.127
158	..TT **C**	308	0.095
153	...G **C**	711	0.075
149	...C **C**	1771	0.049
168	CCCC **G**	24	0.224*c
474	T.TT **G**	29	0.207*b
170	..GG **G**	72	0.180
167	.CCC **G**	47	0.156*b
464	C.CC **G**	57	0.151*b
166	..CC **G**	141	0.091*a
169	...G **G**	426	0.064
165	...C **G**	435	0.046*a
489	G.GG **T**	15	0.433
187	.GGG **T**	18	0.433
186	..GG **T**	80	0.397
490	GG.G **T**	17	0.375
184	CCCC **T**	63	0.282
185	...G **T**	419	0.262
180	AAAA **T**	41	0.212
183	.CCC **T**	190	0.209
486	CCC. **T**	160	0.185
484	C.CC **T**	176	0.177
485	CC.C **T**	154	0.171
481	AAA. **T**	125	0.162
179	.AAA **T**	110	0.143
182	..CC **T**	542	0.129
181	...C **T**	1440	0.042

(c) position ranking: %NR atypicality

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
130	..AA **A**	524	0.042*c
141	...T **A**	1373	0.036
446	CCC. **C**	205	0.213
152	CCCC **C**	69	0.175
444	C.CC **C**	201	0.107
445	CC.C **C**	201	0.102
151	.CCC **C**	224	0.101
145	...A **C**	1495	0.082
171	.GGG **G**	15	0.283*c
170	..GG **G**	72	0.129
489	G.GG **T**	15	0.435
186	..GG **T**	80	0.433
187	.GGG **T**	18	0.426
490	GG.G **T**	17	0.401
185	...G **T**	419	0.304
184	CCCC **T**	63	0.156*a

$E_{N(P50)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution median (P_{50}) level;
$E_{N(P95)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution 95th percentile (P_{95}) level;
$E_{N(KS2)}$ – relative (base-specific) enrichment index for position rank order by %NR atypicality, measured as K-S test statistic;
A – adenine, C – cytosine, G – guanine, T – thymine, · – any base type;
Motifs ranked by enrichment relative to position of interest base type, only statistically significant E_N index values shown;
*Significance level: $p < 0.00002$, except: *a $p < 0.0001$, *b $p < 0.0002$, *c $p < 0.0003$, *d $p < 0.0004$, *e $p < 0.0005$*

Supplementary figure 3.3.2 S5: Sequence context and %NR behaviour: Negatively enriched downstream motifs

Sequence motif relative enrichment: comparison between mtDNA position ranking methods

Significantly negatively enriched downstream motifs ($E_N < 0$) with undefined variable position base type

(a) mtDNA position ranking: %NR P_{50}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
3	. . . . **A** AAA.	182	-0.337
4	. . . . **A** AAAA	70	-0.323
2	. . . . **A** AA..	524	-0.299
259	. . . . **A** AA.A	172	-0.291
260	. . . . **A** A.AA	187	-0.285
1	. . . . **A** A...	1604	-0.237
261	. . . . **A** .AAA	175	-0.155
13	. . . . **A** T...	1230	-0.065
9	. . . . **A** G...	795	-0.054
29	. . . . **C** T...	1440	-0.131
21	. . . . **C** C...	1771	-0.024
310	. . . . **G** G.GG	26	-0.276
43	. . . . **G** GGG.	15	-0.262
45	. . . . **G** T...	419	-0.168
42	. . . . **G** GG..	72	-0.144*[a]
46	. . . . **G** TT..	104	-0.130
41	. . . . **G** G...	426	-0.097
63	. . . . **T** TTT.	63	-0.240
335	. . . . **T** T.TT	78	-0.217
334	. . . . **T** TT.T	65	-0.210
64	. . . . **T** TTTT	16	-0.204*[d]
62	. . . . **T** TT..	251	-0.199
61	. . . . **T** T...	1004	-0.169
49	. . . . **T** A...	1373	-0.046

(b) mtDNA position ranking: %NR P_{95}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P95)}$ index
3	. . . . **A** AAA.	182	-0.277
4	. . . . **A** AAAA	70	-0.253
2	. . . . **A** AA..	524	-0.244
259	. . . . **A** AA.A	172	-0.231
260	. . . . **A** A.AA	187	-0.218
1	. . . . **A** A...	1604	-0.183
261	. . . . **A** .AAA	175	-0.150
13	. . . . **A** T...	1230	-0.091
9	. . . . **A** G...	795	-0.084
29	. . . . **C** T...	1440	-0.107
310	. . . . **G** G.GG	26	-0.242
42	. . . . **G** GG..	72	-0.120*[a]
45	. . . . **G** T...	419	-0.094
41	. . . . **G** G...	426	-0.075
63	. . . . **T** TTT.	63	-0.211
62	. . . . **T** TT..	251	-0.180
334	. . . . **T** TT.T	65	-0.175
335	. . . . **T** T.TT	78	-0.169
61	. . . . **T** T...	1004	-0.129
49	. . . . **T** A...	1373	-0.048

(c) position ranking: %NR atypicality

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(KS2)}$ index
14	. . . . **A** TT..	330	-0.207
10	. . . . **A** GG..	174	-0.203
15	. . . . **A** TTT.	92	-0.202
274	. . . . **A** TT.T	104	-0.199
270	. . . . **A** G.GG	22	-0.190*[b]
9	. . . . **A** G...	795	-0.175
13	. . . . **A** T...	1230	-0.171
269	. . . . **A** GG.G	37	-0.149*[c]
275	. . . . **A** T.TT	77	-0.147
1	. . . . **A** A...	1604	-0.025
20	. . . . **C** AAAA	39	-0.145*[d]
19	. . . . **C** AAA.	145	-0.131
279	. . . . **C** AA.A	150	-0.107
18	. . . . **C** AA..	465	-0.104
25	. . . . **C** G...	435	-0.067
17	. . . . **C** A...	1534	-0.062
-	- - - -	-	-
61	. . . . **T** T...	1004	-0.067
49	. . . . **T** A...	1373	-0.041

$E_{N(P50)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution median (P_{50}) level;
$E_{N(P95)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution 95th percentile (P_{95}) level;
$E_{N(KS2)}$ – relative (base-specific) enrichment index for position rank order by %NR atypicality, measured as K-S test statistic;
A – adenine, C – cytosine, G – guanine, T – thymine, · – any base type;
Motifs ranked by enrichment relative to position of interest base type, only statistically significant E_N index values shown;
*Significance level: $p < 0.00002$, except: *[a] $p < 0.0001$, *[b] $p < 0.0002$, *[c] $p < 0.0003$, *[d] $p < 0.0004$, *[e] $p < 0.0005$*

Supplementary figure 3.3.2 S6: Sequence context and %NR behaviour: Negatively enriched upstream motifs

Sequence motif relative enrichment: comparison between mtDNA position ranking methods

Significantly negatively enriched upstream motifs ($E_N < 0$) with undefined variable position base type

(a) mtDNA position ranking: %NR P_{50}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P50)}$ index
132	AAAA **A**	70	-0.108*c
421	AAA. **A**	172	-0.089
419	A.AA **A**	175	-0.087*a
131	.AAA **A**	182	-0.078*b
420	AA.A **A**	187	-0.071*b
130	..AA **A**	524	-0.059
129	...A **A**	1604	-0.042
141	...T **A**	1373	-0.033
446	CCC. **C**	205	-0.222
152	CCCC **C**	69	-0.173
444	C.CC **C**	201	-0.142
151	.CCC **C**	224	-0.136
445	CC.C **C**	201	-0.113
145	...A **C**	1495	-0.104
150	..CC **C**	624	-0.057
161	...A **G**	795	-0.107
191	.TTT **T**	63	-0.277
190	..TT **T**	251	-0.253
192	TTTT **T**	16	-0.211*e
494	T.TT **T**	69	-0.209
495	TT.T **T**	78	-0.192
189	...T **T**	1004	-0.182
496	TTT. **T**	65	-0.168
177	...A **T**	1230	-0.024*b

(b) mtDNA position ranking: %NR P_{95}

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(P95)}$ index
421	AAA. **A**	172	-0.101
141	...T **A**	1373	-0.042
130	..AA **A**	524	-0.039*e
446	CCC. **C**	205	-0.181
152	CCCC **C**	69	-0.132
151	.CCC **C**	224	-0.109
444	C.CC **C**	201	-0.103
145	...A **C**	1495	-0.074
150	..CC **C**	624	-0.041*a
161	...A **G**	795	-0.053
191	.TTT **T**	63	-0.224
190	..TT **T**	251	-0.199
496	TTT. **T**	65	-0.180
494	T.TT **T**	69	-0.143
495	TT.T **T**	78	-0.134
189	...T **T**	1004	-0.123
177	...A **T**	1230	-0.037

(c) position ranking: %NR atypicality

Motif ID	Motif sequence -4 bp to +4 bp	Motif count	$E_{N(KS2)}$ index
137	...G **A**	613	-0.055
129	...A **A**	1604	-0.028
159	.TTT **C**	70	-0.138
454	T.TT **C**	91	-0.118
154	..GG **C**	151	-0.114
441	AAA. **C**	168	-0.107
153	...G **C**	711	-0.079
158	..TT **C**	308	-0.057*a
149	...C **C**	1771	-0.040
161	...A **G**	795	-0.033*a
177	...A **T**	1230	-0.042
181	...C **T**	1440	-0.034

$E_{N(P50)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution median (P_{50}) level;
$E_{N(P95)}$ – relative (base-specific) enrichment index for position rank order by %NR distribution 95th percentile (P_{95}) level;
$E_{N(KS2)}$ – relative (base-specific) enrichment index for position rank order by %NR atypicality, measured as K-S test statistic;
A – adenine, C – cytosine, G – guanine, T – thymine, · – any base type;
Motifs ranked by enrichment relative to position of interest base type, only statistically significant E_N index values shown;
*Significance level: $p < 0.00002$, except: *a $p < 0.0001$, *b $p < 0.0002$, *c $p < 0.0003$, *d $p < 0.0004$, *e $p < 0.0005$*

A1.3 mtDNA sequence context and %NR differences between sequencing directions

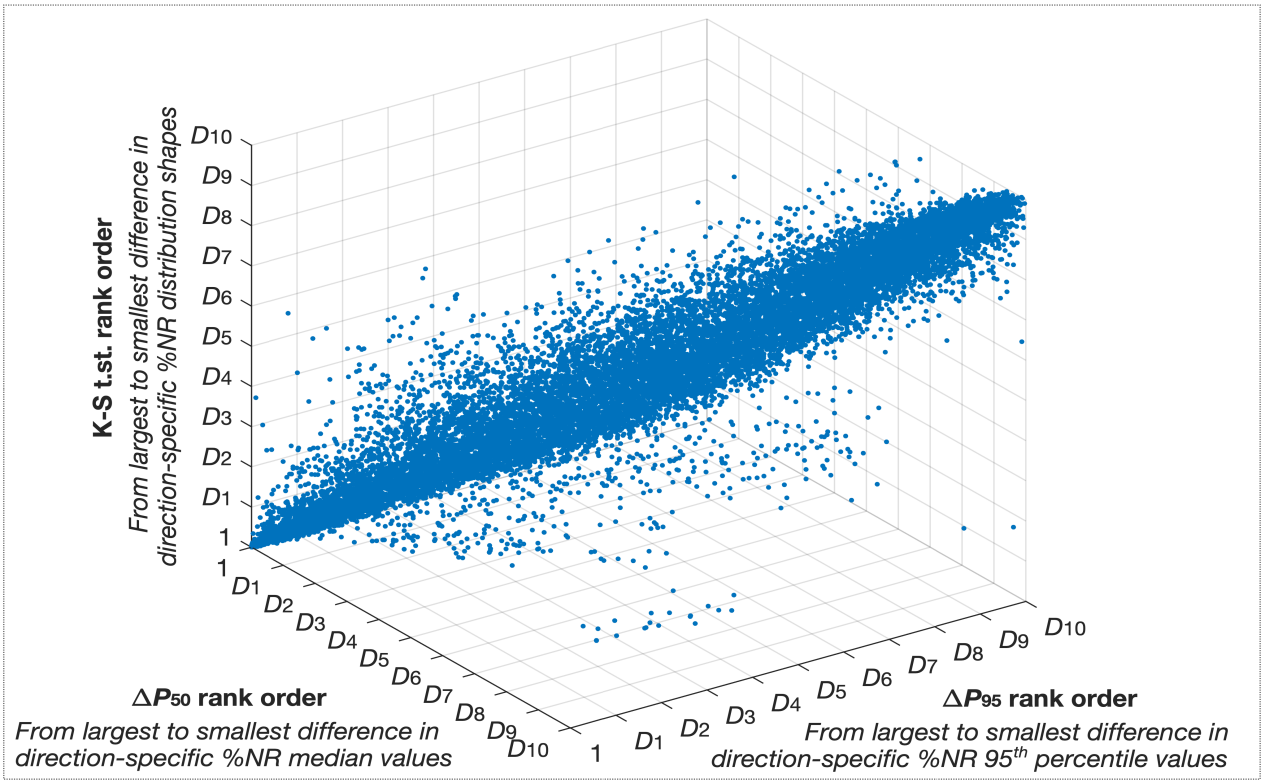
A1.3.1 Comparison between measures of sequencing direction specific %NR behaviour differences

Supplementary figure 3.4 S1 presents the comparison between MT genome position ranking by three different measures of strand-specific %NR behaviour differences: the difference between H and L strand %NR distribution medians (ΔP_{50}), and between 95th percentile values (ΔP_{95}) and the KS2 test statistic describing the overall strand specific %NR distribution shape differences. Figure demonstrates that all three measures largely correlate, although the relationships between differences in %NR 95th percentiles and either of the other two measures are much more noisy compared to the relationship between the differences in %NR medians versus %NR distribution shapes.

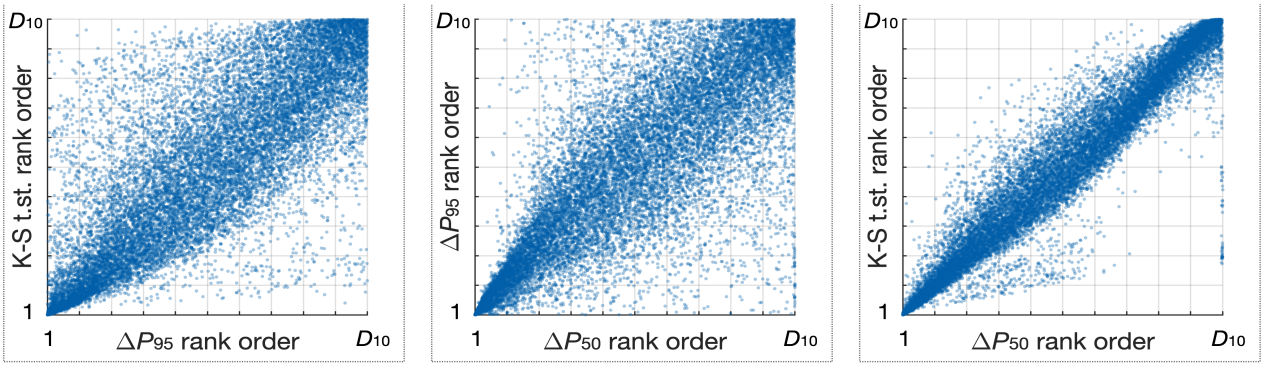
Supplementary figure 3.4 S1: mtDNA positions ranked by three measures of %NR behaviour differences between sequencing directions

Comparison of mtDNA position rank order by the difference in sequencing direction specific population %NR medians or 95th percentiles or the overall %NR distribution shapes

(a) mtDNA position rank order by decreasing K-S test statistic vs ΔP_{50} vs ΔP_{95} values



(b) Pairwise comparison of mtDNA position rank order by K-S test statistic vs ΔP_{50} vs ΔP_{95} values



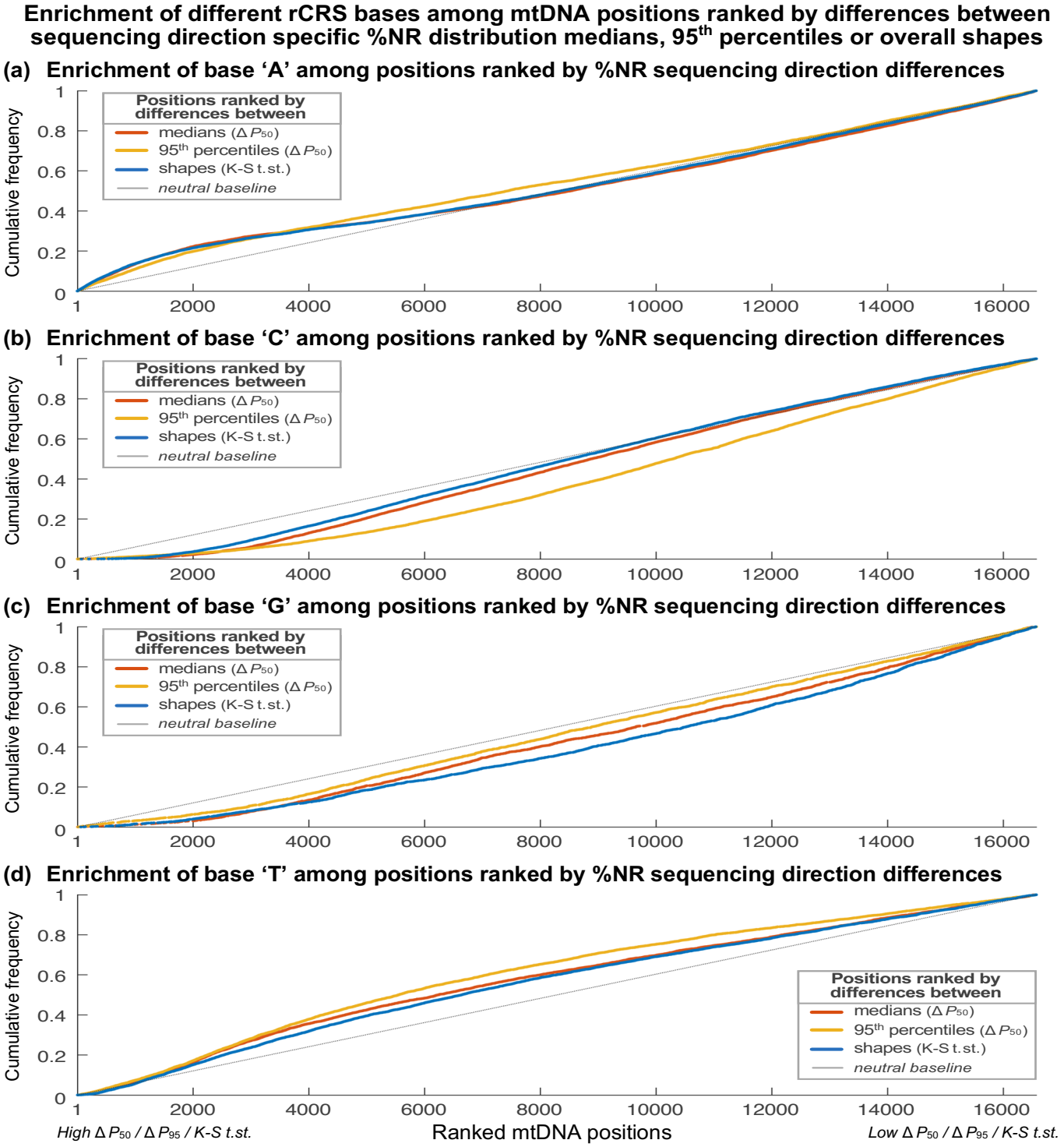
K-S t.st. vs ΔP_{95} **ΔP_{95} vs ΔP_{50}** **K-S t.st. vs ΔP_{50}**

Visual comparison between outcomes of mtDNA position ranking by three measures of strand-specific %NR distribution differences
***A** – Three-way comparison of position ranking by difference between strand-specific %NR medians (ΔP_{50}, x-axis), 95th percentiles (ΔP_{95}, y-axis) and distribution shapes (KS t.st., z-axis). **B** – Pairwise position ranking outcome comparisons. Positions ranked from highest to lowest, with rank = 1 representing largest strand-specific %NR distribution differences. %NR – non-reference read frequency. ΔP_{50} – difference between medians, ΔP_{95} – difference between 95th percentiles, KS t.st. – Kolmogorov-Smirnov test statistic. D_1-D_{10} – position rank order deciles*

A1.3.2 Link between sequencing direction %NR behaviour differences and the base type

Enrichment score based on base type cumulative frequencies across the ranked list of sites can be used to quantify these relationships between base type and directional %NR differences in the same way how base type association with atypical distributions or elevated overall %NR levels were measured earlier. **Supplementary figure 3.4.1 S1** presents the base type enrichments visualisation via cumulative frequency curves with enrichment score values listed in **Supplementary figure 3.4.1 S2.**

Supplementary figure 3.4.1 S1: Different base type enrichment among mtDNA positions ranked by three measures of %NR behaviour differences between sequencing directions



Cumulative frequency of different DNA base types among mtDNA positions ranked by differences between strand-specific %NR distribution medians (red), 95th percentiles (yellow), or overall shapes (blue). X-axis: position rank from highest (left) to lowest (right) by each %NR behaviour or interest. Y-axis: base of interest cumulative frequency. %NR – non-reference read frequency. ΔP_{50} – difference between medians, ΔP_{95} – difference between 95th percentiles, KS t.st. – Kolmogorov-Smirnov test statistic

Supplementary figure 3.4.1 S2: Quantification of the association between the base type and %NR behaviour differences between sequencing directions

Base type enrichment among mtDNA positions ranked by differences between sequencing direction specific %NR distribution medians or 95th percentiles or overall shapes

Base type	Instances found	$E_{(\Delta P50)}$	$E_{(\Delta P95)}$	$E_{(\Delta KS2)}$
A	5124	0.015 ** ↑	0.037 ** ↑	0.019 ** ↑
C	5181	-0.043 ** ↓	-0.102 ** ↓	-0.024 ** ↓
G	2169	-0.070 ** ↓	-0.038 ** ↓	-0.098 ** ↓
T	4094	0.073 ** ↑	0.104 ** ↑	0.059 ** ↑

$E_{(\Delta P50)}$ – enrichment index for positions ranked by directional %NR distribution median (ΔP_{50}) value differences;
$E_{(\Delta P95)}$ – enrichment index for positions ranked by directional %NR distribution 95th percentile (ΔP_{95}) value differences;
$E_{(KS2)}$ – enrichment index for positions ranked by directional %NR distribution shape differences, measured as KS test statistic;
*↑ – positive enrichment, ↓ – negative enrichment; significance level: ** $p < 0.00002$, * $p = 0.0007$, ǂ $p = 0.025$*

Supplementary figures 3.4.1 S1 and S2 demonstrate a significant association between the DNA base type and how much the observed %NR distribution behaviours differ between sequencing directions. Regardless of which measure of %NR directional discrepancy is used to rank positions, all bases show similar enrichment trends. Ts show strongest association with differences between %NR direction-specific distributions, as shown by highest positive enrichments scores. We also see positive but comparatively weaker association between %NR directional differences and As. In contrast positions with rCRS base C or G are negatively associated with direction-specific %NR differences.

These results indicate that we are likely looking at larger sequencing error contribution to our observed %NR landscapes at mtDNA positions with Ts and As compared to Cs and Gs. Seeing both members of the complementary base pairs follow the same enrichment trend is consistent with positions where a specific base pair is found being subject to the same, probably adjacent sequence context dependent process regardless of the orientation of that base pair. This indicates that whatever the process leading to %NR directional differences may be, sequence orientation specific rather than strand specific (i.e. an A on L strand for example are affected in the same way an A on H strand).

It is also interesting to note the differences in enrichment patterns observed between positions where the reference base is A versus T. While Ts show a stronger association to higher %NR differences between sequencing directions overall (larger are under the cumulative frequency curve), As show more extreme enrichment specifically at the very top of ranked positions, although overall the area under the cumulative frequency curve is lower. This suggests that while proportion of positions with reference base A appear to be very strongly associated with directional %NR difference, there must be a large enough proportion of other 'A' sites where differences are smaller, suggesting there is a strong adjacent context component to determining whether or not %NR distributions will be highly different between sequencing directions at any given 'A' position.

Why this is not a case with 'T' positions to the same extent is unclear, but might be due to the mtDNA L and H strand base composition differences which mean that perhaps Ts are less often found in sequence contexts leading to the most extreme %NR behaviours compared to As simply due to what sequence context tend to be more common upstream of A/T base pairs on L vs H strands.

Appendix 2 Supplementary information for Chapter 4

List of supplementary figures associated with Chapter 4

A2.1 Supplementary figures associated with section 4.2 of the main text

4.2.3 S1: Enrichment of variable upstream G content sequence patterns ending in T within sets of highly error prone sequence motifs 231
4.2.4 S1: Sequencing error rate differences between closely related 5bp motifs 232
4.2.4 S2: Sequencing error rate differences between closely related 2bp, 3bp and 4bp motifs 233
4.2.5 S1: GGT' an 'GGGT' pattern motif ratios found associated with high (ΣMER ≥1%) and low (ΣMER<1%) sequencing error rates 234

A2.2 Supplementary figures associated with sections 4.3.1 and 4.3.2 of the main text

4.3.1 S1: Motif inter-sample sequencing error variation: variability between motifs 235
4.3.1 S2: Pearson correlation of motif error proneness and the MER inter-sample variability 236
4.3.1 S3: Visualisation of the relationship between motif ΣMER and the MER variability 236
4.3.1 S4: Sequencing error rates of the same motif observed in different samples: MER population distribution shape variability across the analysed 6bp motifs 238
4.3.2 S1: Motif sequencing error rates measured in the same sample: MER distribution shape variability between 2,035 samples 239

A2.3 Supplementary figures associated with section 4.3.3 of the main text

4.3.3 S1: Possible models for error proneness variation between samples 242
4.3.3 S2: Expected ΔMER and MER_{SC} distribution behaviours with each MER variation model 243
4.3.3 S3: 6bp motif ΔMER and MER_{SC} value distribution variability between individuals 244
4.3.3 S4: Motif error rate scaled rank distance measurement principles, sample HG00361 248
4.3.3 S5: Full MER raw rank and positive MER scaled rank distance distribution examples 249
4.3.3 S6: 6bp motif scaled rank distance distribution variability between samples: distributions of 2035 individuals' positive MER RD_{SC} distribution summary statistics and shape parameters 250
4.3.3 S7: Quantification of 6bp motif MER population stratification outlier behaviour 252
4.3.3 S8: Motif ΣMER level correlation with different population stratification outlier measures in 6bp motifs above different minimum ΣMER thresholds 255
4.3.3 S9: 'AGT', 'GGC' and 'GGT' patterns are most common among 6bp motifs with ΣMER ≥0.2% most frequently seen as outliers relative so sample overall error proneness 257

A2.1 Sequence contexts associated with highest error rates

A2.1.1 Is the pattern 'GGT' a good predictor for high sequencing error rate?

Prevalence of patterns ending in 'T' with different upstream G contents across the analysed motif spaces versus in the highest error motif sets is addressed in **Supplementary figure 4.2.3 S1**. **Panel A** summarises different length sequence patterns ending in 'T' found among the highest error rate motifs. **Panels B and C** provide breakdown of all analysed motifs (A) and identified most error prone motifs (C) matching different G-rich patters ending in 'T'. **Panel D** shows percentage of analysed G-rich patterns ending in T represented by most error-prone motif sets

Supplementary figure 4.2.3 S1: Enrichment of variable upstream G content sequence patterns ending in T within sets of highly error prone sequence motifs

Enrichment of sequence patterns ending with 'T' and with different upstream G contents within the sets of motifs found to be most strongly associated with sequencing errors

(a) Motifs ending with 'T' found across all analysed versus the most error prone motif sets

Motifs found ending with 'T'	Motif spaces analysed						
	N2	N3	N4	N5	N6	N7	N8
All analysed	4 25%	16 25%	64 25%	256 25%	1,023 25.02%	3,901 25.25%	11,086 26.17%
Most error prone	1 of 1 100%	2 of 2 100%	6 of 7 85.7%	22 of 23 95.7%	45 of 49 91.8%	44 of 45 97.8%	55 of 59 93.2%

(b) Total analysed motifs of different lengths ending with a 'T' split by upstream G content

Number of upstream Gs	Motif spaces analysed						
	N2	N3	N4	N5	N6	N7	N8
0	3 75%	9 56.3%	27 42.2%	81 31.6%	243 23.8%	726 18.6%	1,943 17.5%
1	1 25%	6 37.5%	27 42.2%	108 42.2%	405 39.6%	1,392 35.7%	3,418 30.8%
2		1 6.3%	9 14.1%	54 21.1%	269 26.3%	1,130 29.0%	3,086 27.8%
3			1 1.6%	12 4.7%	90 8.8%	504 12.9%	1,773 16.0%
4				1 0.39%	15 1.5%	130 3.3%	686 6.2%
5					1 0.10%	18 0.46%	158 1.4%
6						1 0.03%	21 0.19%
7							1 0.01%

(c) Most error prone motifs of different lengths ending with a 'T' found split by upstream G content

Number of upstream Gs	Motif spaces analysed						
	N2	N3	N4	N5	N6	N7	N8
0	0	0	0	0	0	0	0
1	1 100%	1 50%	1 16.7%	1 4.5%	0	0	0
2		1 50%	4 66.7%	12 54.5%	14 31.1%	3 6.8%	5 9.1%
3			1 16.7%	8 36.4%	20 44.4%	13 29.5%	11 20.0%
4				1 4.5%	10 22.2%	17 38.6%	23 41.8%
5					1 2.2%	10 22.7%	11 20.0%
6						1 2.3%	4 7.3%
7							1 1.8%

(d) Faction of all possible variable G content motifs of different lengths ending with 'T' that is represented within the most error prone motif sets

Number of upstream Gs	Motif spaces analysed						
	N2	N3	N4	N5	N6	N7	N8
0	0%	0%	0%	0%	0%	0%	0%
1	100%	16.7%	3.7%	0.9%	0%	0%	0%
2		100%	44.4%	22.2%	5.2%	0.3%	0.2%
3			100%	66.7%	22.2%	2.6%	0.6%
4				100%	66.7%	13.1%	3.4%
5					100%	55.6%	7.0%
6						100%	19.0%
7							100%

A2.1.2 Sequencing error rate comparison between closely related motifs

Supplementary figures S1 and S2 below compare population ΣMER values of the most error prone motifs 2bp to 5bp motifs with the ΣMER values of different closely related motifs. **Panel A** lists the set of most error prone motifs investigated with their respective ΣMER values. **Panel B** lists ΣMER values of every possible related motif that diverge from the corresponding motif in **panel A** directly above by a single base. Data for top 10 most error prone 5bp motifs shown in **Supplementary figures 4.2.4 S1.** Data for all highly error prone 2bp, 3bp and 4bp motifs shown in **supplementary figure 4.2.4 S2**.

Supplementary figure 4.2.4 S1: Sequencing error rate differences between closely related 5bp motifs

Population level sequencing error rates associated with every possible 5bp long motif that differs only by a single base from the top 10 most error prone 5bp sequences

(a) Top 10 most error prone 5bp motifs associated with highest population level sequencing error rates

Top 5bp motifs	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 8	Motif 9	Motif 10
Motif sequence	CGGGT	GGGGT	AGGGT	CAGGT	GAGGT	CGAGT	AAGGT	AGAGT	TGGGT	GCGGT
Population ΣMER	10.23% **	5.54% **	4.39% **	3.84% **	3.61% **	2.67% **	2.52% **	2.46% **	2.26% **	2.22% **

(b) ΣMER across motifs most closely related to each of the 10 most error prone 5bp sequences

Related motif variant position		Motif 1 CGGGT	Motif 2 GGGGT	Motif 3 AGGGT	Motif 4 CAGGT	Motif 5 GAGGT	Motif 6 CGAGT	Motif 7 AAGGT	Motif 8 AGAGT	Motif 9 TGGGT	Motif 10 GCGGT
Position 0	----A	1.78% *	1.38% **	0.86% **	0.54%	0.61% **	0.4%	0.31%	0.23%	0.63%	0.85%
	----C	0.89% *	0.58%	0.33%	0.12%	0.17%	0.13%	0.31%	0.1%	0.11%	0.23%
	----G	-0.17%	-0.19%	-0.16%	-0.11%	-0.13%	-0.12%	-0.11%	-0.12%	-0.13%	-0.13%
	----T										
Position -1	---A-	0.11%	0.16%	0.08%	0.07%	0.01%	-0.02%	0.04%	-0.06%	0.04%	0.12%
	---C-	0.08%	0.13%	0.01%	-0.05%	0.02%	-0.07%	-0.07%	-0.04%	0%	0.05%
	---G-										
	---T-	0.02%	0.02%	-0.02%	-0.05%	-0.11%	0.00%	-0.07%	-0.07%	-0.04%	0.00%
Position -2	--A--	2.67% **	2.12% **	2.46% **	0.76% *	1.21% **		1.26% **		1.16% **	0.91% **
	--C--	1.05%	1.06% **	0.74%	0.56%	0.4%	1.05%	0.61%	0.74%	0.63%	0.58%
	--G--						10.23% **		4.39% **		
	--T--	0.64%	0.84% **	0.77% **	0.36%	0.71% **	0.64%	0.52% *	0.77% **	0.48% *	0.72% *
Position -3	-A---	3.84% **	3.61% **	2.52% **			0.76% *		1.26% **	1.69% **	3.61% **
	-C---	1.49% **	2.22% **	1.48% **	1.49% **	2.22% **	0.65% *	1.48% **	0.67% *	1.13% *	
	-G---				10.23% **	5.54% **		4.39% **			5.54% **
	-T---	1.53% **	1.69% **	1.65% **	1.53% **	1.69% **	0.55% **	1.65% **	0.81% **	1.14% **	1.69% **
Position -4	A----	4.39% **	4.39% **		2.52% **	2.52% **	2.46% **			4.39% **	1.48% **
	C----		10.23% **	10.23% **		3.84% **		3.84% **	2.67% **	10.23% **	1.49% **
	G----	5.54% **		5.54% **	3.61% **		2.12% **	3.61% **	2.12% **	5.54% **	
	T----	2.26% **	2.26% **	2.26% **	1.69% **	1.69% **	1.16% **	1.69% **	1.16% **		1.13% *

*Colour scale: ΣMER above 1% in red, ΣMER below 1% in blue; Motifs marked with ** and * show statistically significant population ΣMER plus either statistically significant MER in ≥ 90% of samples (**) or ≥ 75% of samples (*) with that motif*

Supplementary figure 4.2.4 S2: Sequencing error rate differences between closely related 2bp, 3bp and 4bp motifs

Population level sequencing error rates associated with every possible 2bp to 4bp motif that differs only by a single base from the most error prone 2bp, 3bp and 4bp sequences

(a) Most error prone 2bp, 3bp and 4bp motifs with highest population level sequencing error rates

Most error prone motifs	4bp motifs							3bp motifs		2bp motif
	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 1	Motif 2	Motif 1
Motif sequence	GGGT	AGGT	GAGT	CGGT	TGGT	GGGA	AAGT	GGT	AGT	GT
Population ΣMER	4.43% **	2.86% **	1.88% **	1.71% **	1.47% **	1.12% **	1.04% **	2.9% **	1.08% **	1.43% **

(b) ΣMER across motifs most closely related to the most error prone 2bp, 3bp and 4bp sequences

Related motif variant position		4bp motifs							3bp motifs			2bp motifs	
		Motif 1 GGGT	Motif 2 AGGT	Motif 3 GAGT	Motif 4 CGGT	Motif 5 TGGT	Motif 6 GGGA	Motif 7 AAGT	Variant position	Motif 1 GGT	Motif 2 AGT	Variant position	Motif 1 GT
Position 0	---A	1.12% **	0.48% **	0.49% **	0.51% *	0.61% **		0.15%	--A	0.67% **	0.23% **	-A	0.39% **
	---C	0.41% *	0.17%	0.12%	0.11%	0.12%	0.41% *	0.15%	--C	0.21% *	0.12% *	-C	0.13% *
	---G	-0.16%	-0.12%	-0.12%	-0.11%	-0.11%	-0.16%	-0.08%	--G	-0.13%	-0.09%	-G	-0.09%
	---T						4.43% **		--T			-T	
Position -1	--A-	0.10%	0.03%	-0.05%	0.08%	0.01%	0.07%	-0.08%	-A-	0.05%	-0.08%	A-	-0.06%
	--C-	0.05%	-0.04%	-0.06%	0.06%	-0.02%	0.12%	-0.13%	-C-	0.01%	-0.12%	C-	-0.07%
	--G-								-G-			G-	
	--T-	-0.01%	-0.09%	-0.04%	-0.01%	-0.06%	0.13%	-0.08%	-T-	-0.05%	-0.07%	T-	-0.07%
Position -2	-A--	1.88% **	1.04% **		0.71% **	0.71% **	1.88% **		A--	1.08% **			
	-C--	0.90% **	0.51% **	0.90% **	0.56% **	0.58% **	0.41% *	0.51% **	C--	0.64% **	0.64% **		
	-G--			4.43% **				2.86% **	G--		2.90% **		
	-T--	0.69% **	0.48% **	0.69% **	0.47% **	0.39% **	0.44% **	0.48% **	T--	0.50% **	0.50% **		
Position -3	A---	2.86% **		1.04% **	2.86% **	2.86% **	0.48% **						
	C---	1.71% **	1.71% **	0.71% **		1.71% **	0.51% *	0.71% **					
	G---		4.43% **		4.43% **	4.43% **		1.88% **					
	T---	1.47% **	1.47% **	0.71% **	1.47% **		0.61% **	0.71% **					

*Colour scale: ΣMER above 1% in red, ΣMER below 1% in blue; Motifs marked with ** and * show statistically significant population ΣMER plus either statistically significant MER in ≥ 90% of samples (**) or ≥ 75% of samples (*) with that motif*

A2.1.3 Predictive power of motif patterns for identification of error sites

Supplementary figure 4.2.5 S1 summarises the differences between proportions our identified 'GGT' or 'GGGT' matching motifs found showing high (ΣMER ≥1%) versus low (ΣMER<1%) population-level error rates. **Panel A** shows high and low error rate 'GGT' and 'GGGT' motif counts and proportion among 8bp motifs, with **panels B and C** focusing on 7bp and 6bp motif spaces respectively.

Supplementary figure 4.2.5 S1: 'GGT' an 'GGGT' pattern motif ratios found associated with high (ΣMER ≥1%) and low (ΣMER<1%) sequencing error rates

Comparison of 'GGT' and 'GGGT' pattern matching motif proportions found associated with high (ΣMER ≥ 1%) or low (ΣMER < 1%) population level sequencing error rates

(a) 8bp motifs that match 'GGT' and 'GGGT' patterns

N8 motif space	'GGT' pattern motifs			'GGGT' pattern motifs			'GGGT' pattern proportion		
	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%
Total found	706	546 77.3%	160 22.7%	192	180 93.8%	12 6.3%	27.2%	33.0%	7.5%
1000+ samples	381	308 80.8%	73 19.2%	116	113 97.4%	3 2.6%	30.4%	36.7%	4.1%
2 to 999 samples	238	183 76.9%	55 23.1%	59	54 91.5%	5 8.5%	24.8%	29.5%	9.1%
1 sample	87	55 63.2%	32 36.8%	17	13 76.5%	4 23.5%	19.5%	23.6%	12.5%
2+ samples	619	491 79.3%	128 20.7%	175	167 95.4%	8 4.6%	28.3%	34.0%	6.3%

(b) 7bp motifs that match 'GGT' and 'GGGT' patterns

N7 motif space	'GGT' pattern motifs			'GGGT' pattern motifs			'GGGT' pattern proportion		
	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%
Total found	247	202 81.8%	45 18.2%	61	59 96.7%	2 3.3%	24.7%	29.2%	4.4%
1000+ samples	196	162 82.7%	34 17.3%	51	49 96.1%	2 3.9%	26.0%	30.2%	5.9%
2 to 999 samples	41	32 78.0%	9 22.0%	9	9 100%	0 0%	22.0%	28.1%	0%
1 sample	10	8 80.0%	2 20.0%	1	1 100%	0 0%	10.0%	12.5%	0%
2+ samples	237	194 81.9%	43 18.1%	60	58 96.7%	2 3.3%	25.3%	29.9%	4.7%

(b) 7bp motifs that match 'GGT' and 'GGGT' patterns

N6 motif space	'GGT' pattern motifs			'GGGT' pattern motifs			'GGGT' pattern proportion		
	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%	All motifs	High ΣMER ≥ 1%	Low ΣMER < 1%
Total found	64	60 93.8%	4 6.3%	16	16 100%	0 0%	25.0%	26.7%	0%
1000+ samples	63	59 93.7%	4 6.3%	16	16 100%	0 0%	25.4%	27.1%	0%
2 to 999 samples	1	1 100%	0 0%	0	- -	- -	0%	0%	-
1 sample	0	- -	- -	0	- -	- -	-	-	-
2+ samples	64	60 93.8%	4 6.3%	16	16 100%	0 0%	25.0%	26.7%	0%

Percentages provided with high/low ΣMER motif counts indicate what proportion of all motifs with that pattern this represents; 'GGGT' pattern proportion – 'GGGT' pattern motif subset as percentage of all 'GGT' pattern motifs

A2.2 Sequencing error rate variability between samples

A2.2.1 Motif error rate variation between individuals and relationship to population ΣMER level

Supplementary figure 4.3.1 S1 summarises our observed motif-associated error rate variability. Sample-level motif error rate interquartile range is used as the measure of variability between samples. Only motifs analysed in at least 2 samples and with population ΣMER>0 (number of such motifs shown in **panel A**) are included in the analyses. **Panel B** lists the number of positive ΣMER motifs of different lengths found with sample-level MER IQR above specific thresholds. **Panel C** lists various descriptive statistics of the observed sample-level MER IQR distributions across the analysed motif sets.

Supplementary figure 4.3.1 S1: Motif inter-sample sequencing error variation: variability between motifs

Overview of motif error rate inter-sample variability observed for sets of different length motifs associated with positive population error rates and analysed in multiple samples

(a) Number of 2bp to 8bp motifs analysed in >1 sample

Motifs found in multiple samples	Motif space						
	N2	N3	N4	N5	N6	N7	N8
All	16	64	256	1,024	4,080	14,920	36,979
With ΣMER >0	9 *56.3%*	41 *64.1%*	156 *59.4%*	588 *57.4%*	2,240 *54.9%*	8,105 *54.3%*	19,856 *53.7%*

Percentages indicate fraction of all motifs analysed in multiple samples represented by motifs with positive ΣMER

(b) Positive ΣMER motifs analysed in >1 sample found with MER IQR above different thresholds

Sample-level MER IQR threshold	Motif space						
	N2	N3	N4	N5	N6	N7	N8
≥ 0.1%	5 *55.6%*	21 *51.2%*	89 *58.6%*	409 *69.6%*	2,031 *90.7%*	7,964 *98.3%*	19,237 *96.9%*
≥ 0.5%	1 *11.1%*	3 *7.32%*	10 *6.58%*	52 *8.84%*	213 *9.51%*	1,055 *13.0%*	3,131 *15.8%*
≥ 1%	0	1 *2.44%*	4 *2.63%*	14 *2.38%*	76 *3.39%*	366 *4.52%*	1,088 *5.48%*
≥ 5%	0	0	0	0	2 *0.09%*	24 *0.30%*	89 *0.45%*
≥ 10%	0	0	0	0	2 *0.09%*	3 *0.04%*	26 *0.13%*
≥ 20%	0	0	0	0	0	0	8 *0.04%*

Percentages indicate fraction of all positive ΣMER motifs found with MER IQR above the specified threshold

(c) Descriptive statistics of MER IQR distribution across positive ΣMER motifs analysed in >1 sample

Sample-level MER IQR statistic	Motif space						
	N2	N3	N4	N5	N6	N7	N8
Mean	0.20%	0.18%	0.20%	0.22%	0.28%	0.36%	0.42%
SD	0.23%	0.24%	0.26%	0.33%	0.58%	0.58%	0.92%
Median	0.11%	0.10%	0.11%	0.13%	0.17%	0.23%	0.25%
IQR	0.08% - 0.14%	0.08% - 0.15%	0.08% - 0.17%	0.09% - 0.20%	0.13% - 0.26%	0.18% - 0.33%	0.19% - 0.37%
Range	0.08% - 0.77%	0.04% - 1.41%	0.05% - 2.09%	0.05% - 4.66%	0.06% - 16.1%	0.01% - 16.0%	$1x10^{-6}$ - 43.3%

Only motifs analysed in at least 2 samples and found to be associated with population ΣMER >0 were included in the analysis; ΣMER – population global motif error rate; MER – sample-level motif error rate; MER IQR – MER interquartile range

Across all motif spaces majority of motifs are associated with MER IQR values falling between 0.1% and 0.5%. although the longer the analysed motifs are, the more instances we find where motifs show higher inter-sample variability. Motif MER IQR central tendency remains fairly similar across 2, 3 and 4 base-pair motifs and then starts to increase for the larger motif spaces. For all motif spaces the MER IQR mean is almost 2 times higher than the MER IQR median, consistent with presence of high MER IQR motifs skewing the overall MER IQR distribution.

The following two figures address the relationship between motif population-level ΣMER levels with the variability between sample-level MER. Only data for motifs with positive population ΣMER that were analysed in ≥2 samples is included. **Supplementary figure 4.3.1 S2** addresses the correlation between motif error rate variability between samples (as expressed by MER IQR) with how high motif error rates are in terms of both the global population-level ΣMER as well as median sample-level MER. Pearson correlation coefficients are shown for each motif space, demonstrating a strong relationship between the overall motif error rate level and the observed inter-sample variability across. This is present across all motif lengths, although the correlation weakens slightly as the motifs become longer.

Supplementary figure 4.3.1 S2: Pearson correlation of motif error proneness and the MER inter-sample variability

Pearson correlation between motif error rate and the inter-sample error rate variability

Sample-level MER IQR correlation with	Motif space						
	N2	N3	N4	N5	N6	N7	N8
Population ΣMER	0.98	0.98	0.98	0.98	0.92	0.87	0.80
Median sample MER	0.97	0.97	0.97	0.96	0.91	0.81	0.76

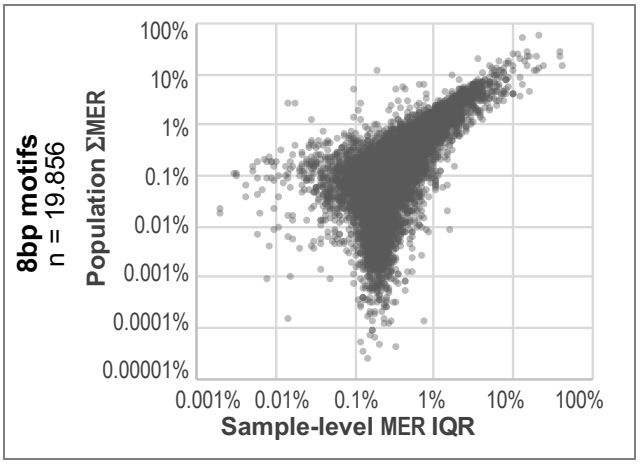
Only motifs analysed in at least 2 samples and found to be associated with population ΣMER >0 were included in the analysis; ΣMER – population global motif error rate; MER – sample-level motif error rate; MER IQR – MER interquartile range

The relationship between 6bp to 8bp motif population-level error rate and the inter-sample MER variability is visualised via scatter plots in **Supplementary figure 4.3.1 S3.**

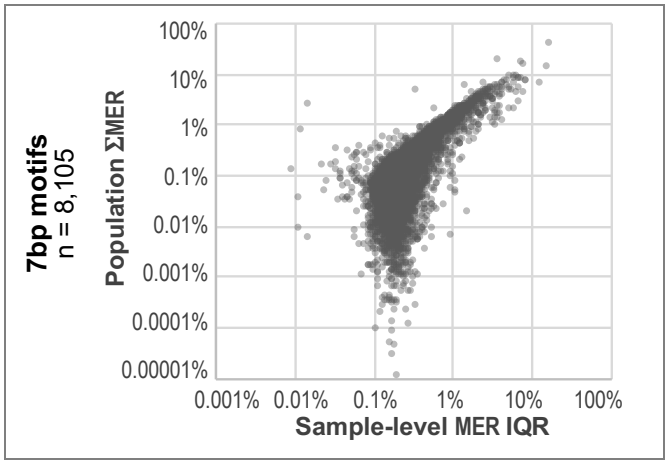
Supplementary figure 4.3.1 S3: Visualisation of the relationship between motif ΣMER and the MER variability

The relationship between motif population-level error rate ΣMER and the observed variability between samples, as described by the sample-level MER interquartile range

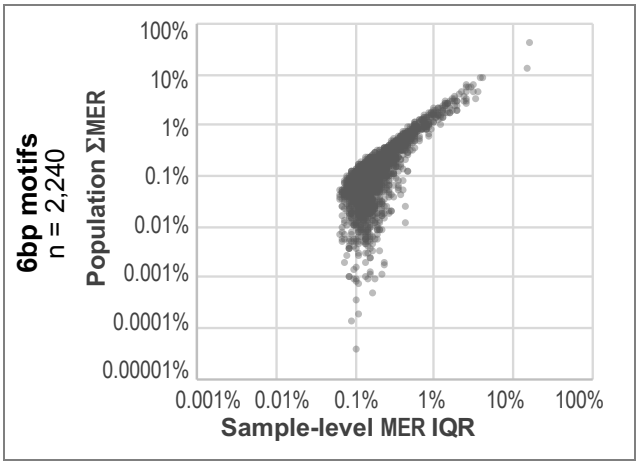
(a) ΣMER vs MER IQR across 8bp motifs



(b) ΣMER vs MER IQR across 7bp motifs



(c) ΣMER vs MER IQR across 6bp motifs



Data shown for 6bp, 7bp and 8bp motif spaces. Motifs associated with ΣMER ≤0 or analysed in fewer than 2 samples are excluded. Data points represent individual motifs. ΣMER – motif population-level motif error rate. MER – sample-level motif error rate. IQR – interquartile range.

Data presented in **Supplementary figure 4.3.1 S2** shows a combination of two different trends, where the relationship between motif error rate level and the inter-sample variability is different between motifs with higher ΣMER compared to those with lower ΣMER. For motifs with ΣMER above approx. 0.1-0.2%, the sample MER IQR values correlate directly with population ΣMER. Below this threshold though any further decrease in ΣMER no longer corresponds to decrease in the MER variability between samples, with IQR remaining around 0.2%. This suggests that in this dataset the ΣMER baseline noise level may be approximately 0.1%-0.2%

This switch in the ΣMER to MER variability relationship around the 0.1%-0.2% ΣMER mark suggests that we are likely starting to see random noise rather than genuine variation between individuals, whereas above this threshold genuine variation between motif error rates in different samples in no longer masked by the noise. The variation we observe between sample MER is generally within one order of magnitude of ΣMER.

So far motif sample-level MER population distribution IQR has been used as the measure of how much variation there is between MER values calculated from different samples. The next figures provide a broader look at the variability between motif sample-level MER distributions via use of box-and-whiskers plots that visualise the overall distributions of different motif or sample MER value distribution descriptive statistics. Figures focus specifically on 6bp long motifs, with this length chosen as a compromise for maximising the considered upstream sequence length while still ensuring that multiple instances of most motifs are analysed within each sample.

A note on negative MER values

In the following figures some MER values appear to be negative. Such negative MER values are an artefact of how motif error rate level is measured, whereby MER represents the a difference between base mismatch rates in forward and reverse sequencing directions relative to motif orientation. While we use a filter cut-off where only motif instances with reverse mismatch rate below 1% are used for MER calculation, this nonetheless means that for some motifs that are not linked to context dependent sequencing errors the reverse mismatch rate can surpass the forward mismatch rate thus leading to a negative MER value.

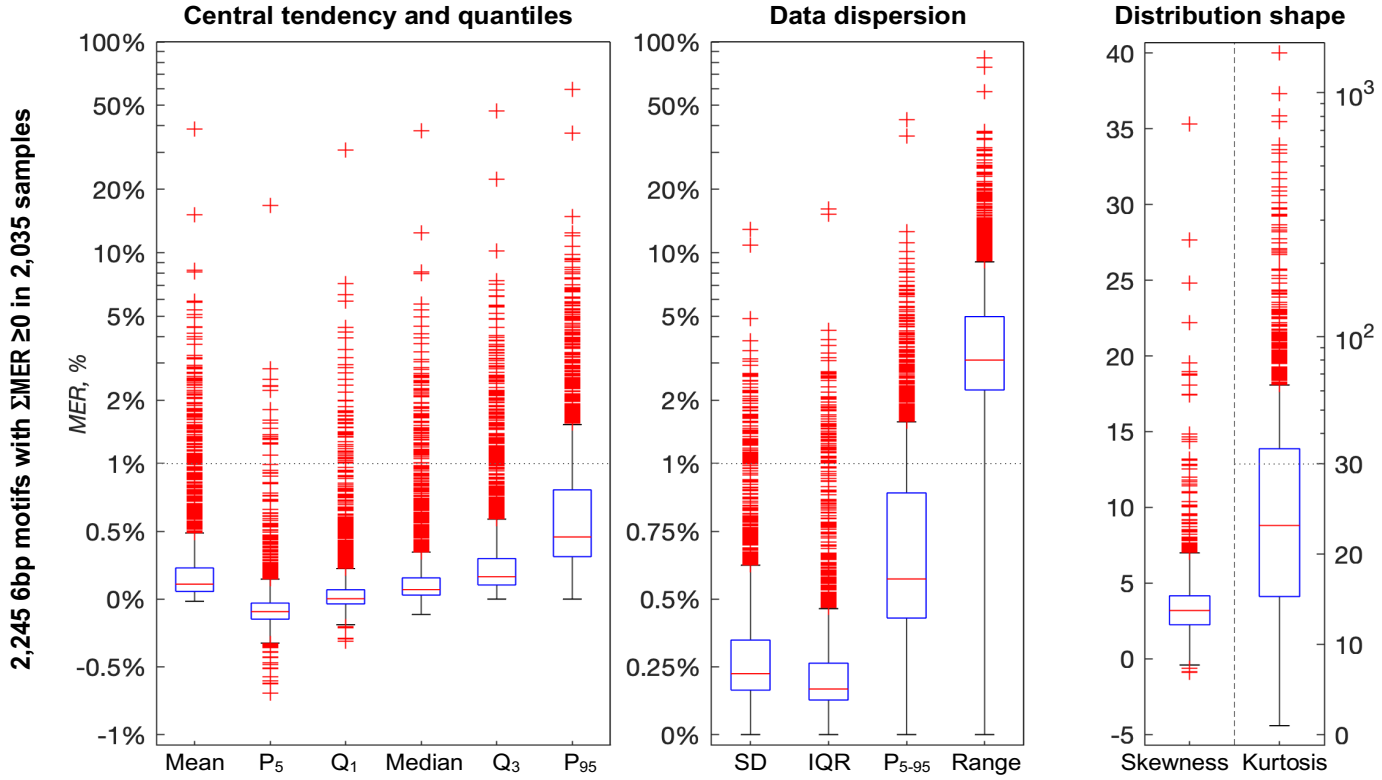
A2.2.2 MER distribution variability between motifs

Supplementary figure 4.3.1 S4 visualises how various motif-specific MER distribution quantile and shape property values are distributed across the 6bp motifs associated with population ΣMER ≥0. **Panel A** shows box-and-whiskers plots representing distributions of various analysed motif MER population distribution summary statistics and shape properties. Each boxplots represents a distribution of 2,245 motif-specific data points. Summary statistics of each box-and-whiskers plot from **panel A** are listed directly below in **panel B**.

Supplementary figure 4.3.1 S4: Sequencing error rates of the same motif observed in different samples: MER population distribution shape variability across the analysed 6bp motifs

Motif sample-level MER population distribution variability between 6bp motifs with ΣMER ≥0

(a) Distributions of analysed motif MER population distribution summary statistics and shape properties



(b) Motif MER population distribution summary statistic and shape property boxplot statistic values

Boxplot statistic	Central tendency and quantiles						Data dispersion				Distribution shape	
	Mean	P_5	Q_1	Median	Q_3	P_{95}	SD	IQR	P_{5-95}	Range	Skewness	Kurtosis
Whisker Hi	0.49%	0.15%	0.23%	0.35%	0.59%	1.53%	0.63%	0.46%	1.58%	9.05%	7.0	63.3
Box Hi - Q_3	0.23%	-0.03%	0.07%	0.16%	0.30%	0.81%	0.35%	0.26%	0.89%	4.98%	4.2	34.6
Median	0.11%	-0.09%	0.00%	0.07%	0.16%	0.46%	0.22%	0.17%	0.57%	3.10%	3.2	23.2
Box Lo - Q_1	0.06%	-0.15%	-0.04%	0.03%	0.10%	0.31%	0.16%	0.13%	0.43%	2.23%	2.3	15.3
Whisker Lo	-0.02%	-0.32%	-0.19%	-0.11%	0%	0%	0%	0%	0%	0%	-0.4	1

Each boxplot represents the distribution of a specific MER population distribution summary statistic or shape property value; MER distribution descriptive statistics: P_5, P_{95} – 5th & 95th percentiles; Q_1, Q_3 –1st & 3rd quartiles; SD – standard deviation; IQR – interquartile range; P_{5-95} – Range between 5th and 95th percentiles

Distribution of all motif-specific MER distribution quantiles are skewed, such that the bulk of data points cluster near 0 with a long tail of outlier motif data points with up to several orders of magnitude higher values compared to the majority of motif data points. Motif-specific sample-level MER value spread measures, such as SD, IQR and range also form skewed distributions, with extreme outlier motifs found to have orders of magnitude more spread-out population MER distributions compared to the majority of motifs. This confirms that different motifs vary widely in terms of how variable the measured motif error rates are between analysed samples.

In terms of motif-specific MER population distribution shape parameters indicate that the vast majority of motif MER distributions are tail-heavy and positively skewed by outlier samples with higher than typical MER values.

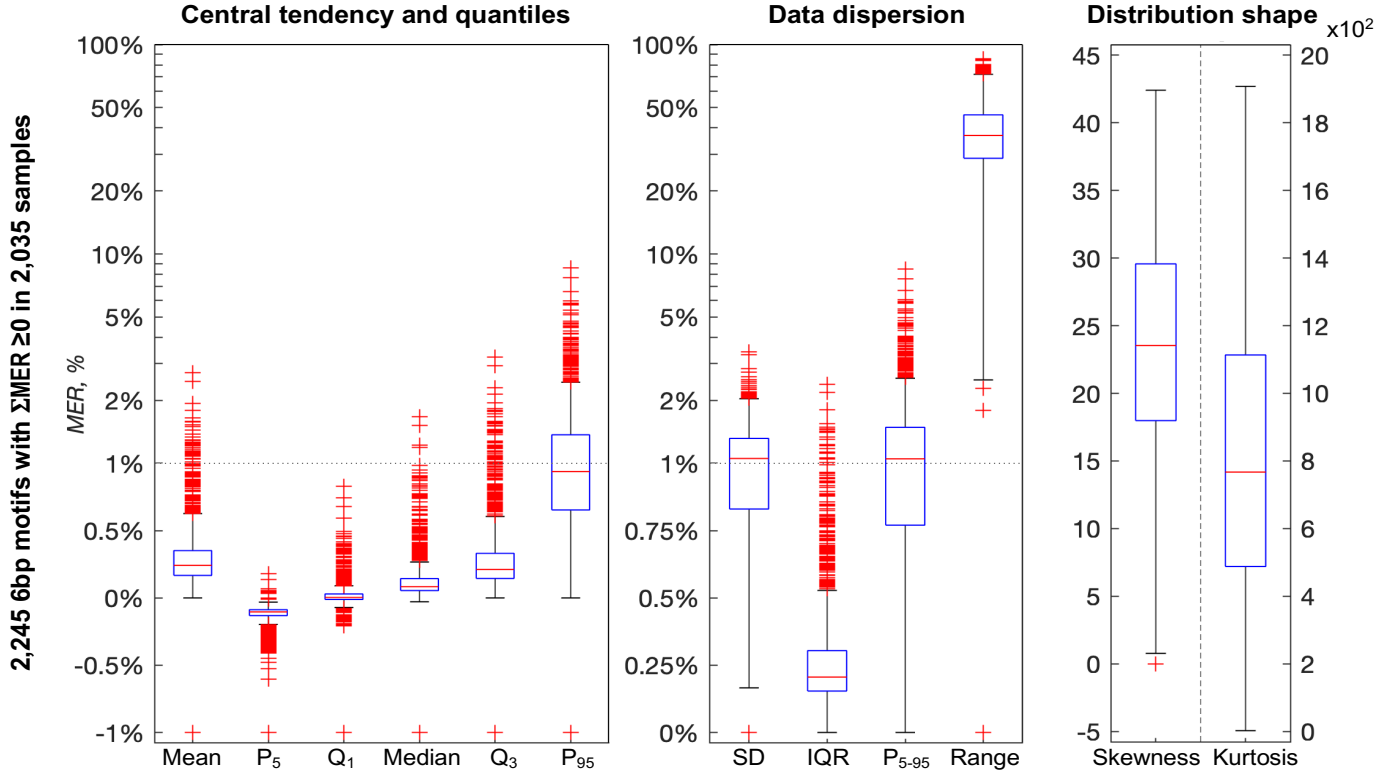
A2.2.3 Sample-level MER distribution variability between individuals

Supplementary figure 5.3.2 S1 below addresses context-dependent sequencing error rate distribution variability between samples rather than between motifs. We again focus on 6bp motif subset with population ΣMER ≥0. **Panel A** shows box-and-whiskers plots representing distributions of various analysed sample MER distribution summary statistics and shape properties. Each boxplots represents a distribution of 2035 sample-specific data points. Summary statistics of each box-and-whiskers plot from **panel A** are listed directly below in **panel B**.

Supplementary figure 4.3.2 S1: Motif sequencing error rates measured in the same sample: MER distribution shape variability between 2,035 samples

6bp non-negative ΣMER motif MER distribution variability between 2,035 individuals

(a) Distributions of individual-specific motif MER distribution summary statistics and shape properties



(b) Individual-specific motif MER distribution summary statistic and shape property boxplot values

Boxplot statistic	Central tendency and quantiles						Data dispersion				Distribution shape	
	Mean	P_5	Q_1	Median	Q_3	P_{95}	SD	IQR	P_{5-95}	Range	Skewness	Kurtosis
Whisker Hi	0.63%	-0.03%	0.09%	0.27%	0.61%	2.44%	2.04%	0.53%	2.55%	72.1%	42.4	1907
Box Hi - Q_3	0.35%	-0.09%	0.03%	0.14%	0.33%	1.37%	1.32%	0.30%	1.49%	46.2%	29.6	1113
Median	0.24%	-0.10%	0.003%	0.08%	0.21%	0.94%	1.06%	0.21%	1.05%	36.8%	23.5	767
Box Lo - Q_1	0.17%	-0.13%	-0.01%	0.05%	0.14%	0.65%	0.83%	0.15%	0.77%	28.6%	18.0	488
Whisker Lo	0%	-0.20%	-0.07%	-0.03%	0%	0%	0%	0%	0%	2.50%	0.8	2.8

Each boxplot represents the distribution of specific summary statistic or shape property values of 2535 individual-specific motif MER distributions; MER distribution descriptive statistics: P_5, P_{95} – 5th & 95th percentiles; Q_1, Q_3 –1st & 3rd quartiles; SD – standard deviation; IQR – interquartile range; P_{5-95} – Range between 5th and 95th percentiles

Individuals are more variable in terms of on their MER distribution 95th percentile values than their 75th percentile (Q_3) or 50th percentile (median) values, indicating that MER values calculated from some samples are much more spread out compared to others even even if the MER distribution medians are not that different between most samples. In terms of the individual MER distribution shapes, compared to the previous figure were we looked at motif-specific MER distributions, the sample-specific MER distributions tend to show much higher skewness, consistent with the low MER values observed for the vast majority of motifs in any given individual, with only a small proportion of much higher high MER motifs per sample being the norm.

As stated earlier, negative MER values are an artefact of MER calculation and are not meaningfully different from 0.

A2.3 Measuring sample 'error proneness'

The results on sequencing error rate profile variability between samples presented in the main text raise several question about the population-level behaviour of context dependent sequencing errors. Our data clearly point to existence of variability between different individual samples in terms of their general 'error proneness', with broadly similar population stratification patterns observed qualitatively across different motifs. The question addressed in this supplementary analyses is how an individual's 'error proneness' can be meaningfully measured quantitatively.

In **figures 4.3.2 F2** and **F3** in the main text we used an approach where individuals are ranked by their observed error rates for each motifs, which allows us to determining each individuals median rank within the populations. While we can use this as an approximation of that individuals error proneness relative to other individuals in the population, this approach has two major drawbacks.

Firstly, any rank based measure of error proneness of a sample informs us about the error proneness of that sample relative to other samples being studied. It does not tell us anything about individuals' absolute error proneness in terms of the actual error rate levels, and therefore is not meaningfully transferable outside of the immediate context of this one dataset. If we were to hypothesise that some universal context-dependent error landscape exists across the motif space regardless of the sequence studied that is then compounded and modified by additional external factors specific to a particular sample or a particular sequencing run, then to untangle this we would need to be able to objectively quantify the absolute error rate landscape behaviour differences between samples.

Secondly, for situations where we do need to look at individuals' relative error proneness within the population, such as investigation of outlier motifs that do not fit population stratification patterns, ranking samples by their median MER rank is still problematic. We have touched upon our observed unusual motif population stratification patterns and have argued that some of the inversed patterns observed (i.e. high median rank samples being ranked lowest at some motifs) are a result of artificially negative sample MER values. This is a major issue for using this simple ranking as a measure for individuals' relative error proneness, because the median rank of an individual would be distorted by aberrantly low ranks of any negative MER motifs. Instead, we would therefore want to find a way to describe an individual's error proneness in a way that accounts for artefact MER values while also allowing for individuals to be directly comparable to each other.

In this supplementary analysis we briefly address possible solutions for each of these highlighted issues below.

A2.3.1 Absolute 'error proneness' of a sample

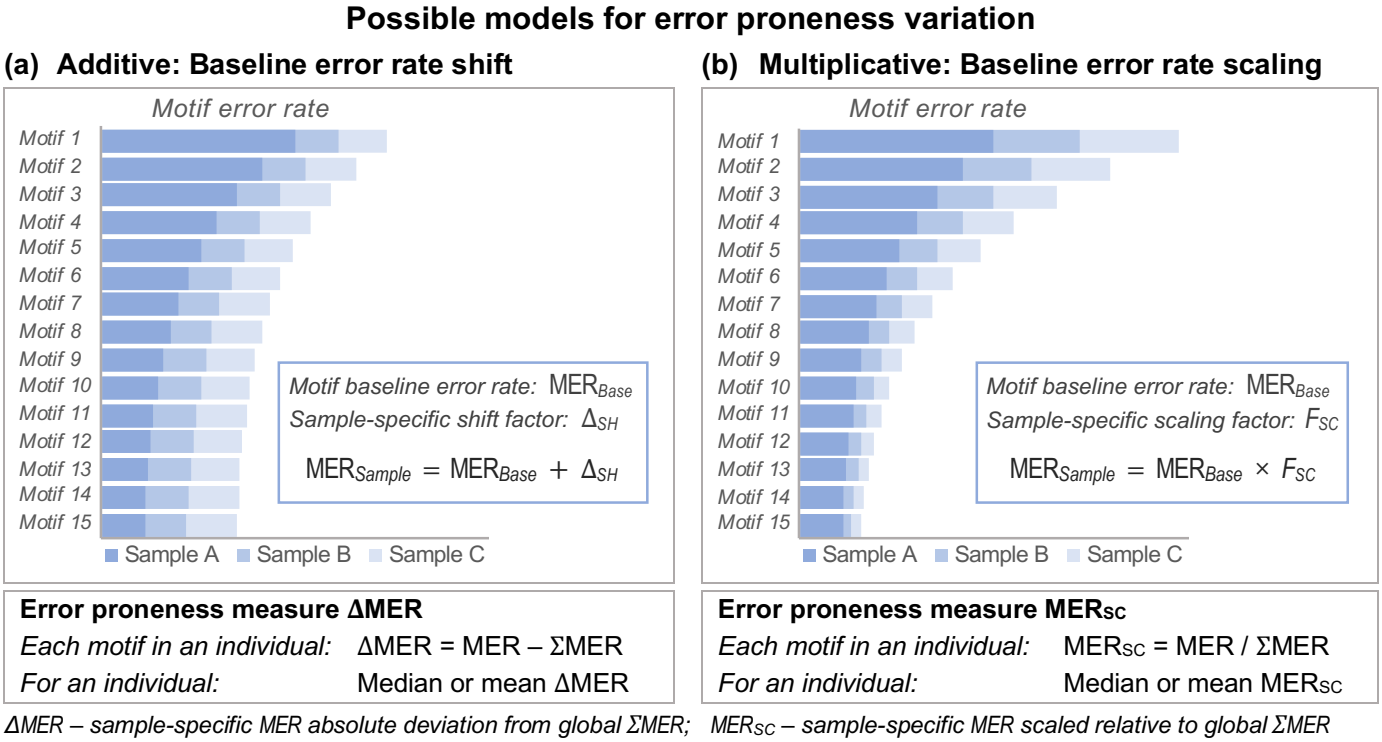
There are several reasons why one might need to measure exactly how strongly affected by context-dependent sequencing errors (CSEs) a particular sample is. In general terms, being able to quantitate the general CSE proneness of a sample allows one to interrogate the error proneness level distribution in the population. Is CSE proneness something that follows normal distribution or are there are asymmetries and why. If there is some CSE proneness 'normal range', this could be useful for flagging up unusual or suspicious samples in an experiment. And having a numeric measure of CSE proneness levels for a particular experiment could also facilitate interpreting results across different experiments or datasets.

In the more specific context of our study, we want to understanding the general sample-level MER 'landscape' behaviour differences between different sequenced individuals. We have observed that many individuals share common patterns of which motifs are associated with CSEs. Yet there is also clear variation in the error rates between samples at individual motif level. We hypothesise that this is a result of variable overall CSE proneness levels between individuals that compounds a more universal baseline CSE 'landscape' (in itself independent from sample-specific factors)

Let us assume that there is some set of universal mechanisms that induce specific levels of CSEs at specific motifs. If we were to take two perfectly identical performed sequencing experiments where the only difference is the DNA source sample, if we were to measure motif error rates from both, in absence of any other compounding factors we should expect the same motifs to be associated with the same error rates. In this hypothetical ideal example the only source of variation between the MER landscapes of the two samples should be random chance. Once we introduce some external factor that for whatever reason modifies one sample's vulnerability to context dependent sequencing errors, we should expect this to be reflected in the affected individual's MER landscape as a global upwards or downwards change of that individual's MER values for any CSE linked motifs.

Now if we imagine more and more samples being added, each affected by its own combination of modifying external factors, we end up with a multitude of error rate landscapes, each slightly different but all still fundamentally sharing the same baseline CSE pattern. Measuring the population level motif error rates, whether in the form of ΣMER or as the sample-level MER value distribution central tendency, should provide be a reasonable approximation of this baseline CSE pattern. Measuring the differences in error proneness between individuals however depends on what effect we expect shift in error proneness to have. There are two broad possibilities, illustrated in **supplementary figure 4.3.2 S2**. The simplest possibility is that error proneness changes have an additive effect resulting in simply shifting an individual's MER values up by a specific amount (**panel A**). Or it could be a more multiplicative effect, such that in the more error prone individuals the baseline MER pattern becomes scaled up (**panel B**). If that was the case, rather than seeing an all-round shift of MER values, we should observe motifs more strongly associated CSEs experiencing biggest MER shifts between individuals compared to motifs less associated with errors.

Supplementary figure 4.3.3 S1: Possible models for error proneness variation between samples



ΔMER – sample-specific MER absolute deviation from global ΣMER; MER_{SC} – sample-specific MER scaled relative to global ΣMER

For a model that assumes simple additive baseline error rate shift, a reasonable estimator of a specific individual's error proneness would be their median ΔMER, expressed for each motif as the deviation between that individual's MER from the population baseline ΣMER. If the model however assumed multiplicative baseline error rate scaling, a better estimate for individual's error proneness would be their median MER_{SC} (i.e. scaled MER), expressed the ratio between that individual's MER and the population baseline ΣMER for each motif.

We can make some predictions of what one should expect an individual's ΔMER and MER_{SC} distributions to look like under each of these models. These predictions are summarised in **supplementary figure 4.3.3 S2** below.

Broadly, if the additive shift model predominates, little ΔMER variation within an individual should be expected with analysed motif ΔMER values in a sample forming a narrow, symmetric distribution, with any ΔMER value variance largely arising from random noise. There should be little to no correlation between MER and ΔMER values within a sample. At the same time we should expect ΔMER values to be highly variable between individuals, reflected as wide and asymmetric distributions of individuals' ΔMER distribution central tendencies and quantiles. Median ΔMER variation between individuals should be comparable to our observed median absolute MER level variation between individuals.

ΔMER is only a useful error proneness measure if all motifs in an individual differ from the baseline by a similar amount. If the effect of individual's error proneness was uneven and more pronounced for high error rate motifs, as assumed by error proneness scaling model, we should expect to see high ΔMER variability within an individual, forming spread out, positively skewed distribution due to a minority of highest ΣMER motifs exhibiting more extreme MER value shift compared to lower ΣMER motifs.

Under the assumption of MER scaling model, MER_{SC} would be a much more appropriate estimate of an individual's error proneness, and unlike ΔMER we would expect MER_{SC} values to be much more similar within an individual, reflecting the factor by which that individual's MER landscape is scaled above or below the general ΣMER baseline.

Supplementary figure 4.3.3 S2: Expected ΔMER and MER_{SC} distribution behaviours with each MER variation model

Expected ΔMER and MER_{SC} distribution behaviours with each MER variation model

(a) Expected ΔMER distribution behaviour within a sample

Model 1: baseline error rate shift	Model 2: baseline error rate scaling
ΔMER central tendency directly describes CSE proneness of a sample - *In absence of noise same ΔMER value for all motifs* - *No correlation between MER and ΔMER values*	**ΔMER does not directly measure sample CSE proneness** - *Different ΔMER value for each motif* - *ΔMER directly correlates with motif MER* - *ΔMER central tendency correlates to sample proneness to CSEs*
Little ΔMER variability within a sample - *Narrow symmetric distribution with little dispersion and low IQR / P_{5-95} range / total range* - *Any ΔMER variability within sample represents noise*	**Highly variable ΔMER within a sample** - *Spared out positively skewed distribution with high dispersion and large IQR / P_{5-95} range / total rang* - *Skewness and kurtosis comparable to MER distribution*
ΔMER dispersion reflects noise present in a sample - *Wider dispersion => more noisy MER data* - *Inverse relationship between ΔMER dispersion and how well MER pattern in that sample reflects population ΣMER landscape* - *ΔMER dispersion not related to sample proneness to CSEs => no correlation expected between ΔMER dispersion and ΔMER median*	**ΔMER dispersion related to sample CSE proneness** - *Wider ΔMER dispersion => stronger CSE effect* - *Direct correlation between ΔMER central tendency and ΔMER dispersion* - *No relationship expected between individual's ΔMER dispersion and how well motif rank order by MER in that individual matches motif rank order by population ΣMER*

(b) Expected MER_{SC} distribution behaviour within a sample

Model 1: baseline error rate shift	Model 2: baseline error rate scaling
MER_{SC} does not directly measure sample CSE proneness - *Different MER_{SC} value for each motif* - *Inverse correlation between motif MER_{SC} and MER values* - *MER_{SC} central tendency correlates to sample proneness to CSEs*	**MER_{SC} central tendency directly describes CSE proneness of a sample** - *In absence of noise same MER_{SC} value for all motifs* - *No correlation between MER and MER_{SC} values*
Highly variable MER_{SC} within a sample - *Spared out positively skewed distribution with high dispersion and large IQR / P_{5-95} range / total range* - *Skewness and kurtosis more extreme than for MER distribution*	**Little MER_{SC} variability within a sample** - *Symmetric, narrow distribution with little dispersion and low IQR / P_{5-95} range / total range* - *Any MER_{SC} variability within sample represents noise*
MER_{SC} dispersion related to sample CSE proneness - *Wider MER_{SC} dispersion => stronger CSE effect* - *Direct correlation between MER_{SC} central tendency and MER_{SC} dispersion*	**MER_{SC} dispersion reflects noise present in a sample** - *Wider dispersion => more noisy MER data* - *Inverse relationship between MER_{SC} dispersion and how well MER pattern in that sample reflects population ΣMER landscape*

CSE – context-dependent sequencing errors; ΣMER – population global motif error rate; MER – sample-specific motif error rate; ΔMER – sample MER absolute deviation from global ΣMER; MER_{SC} – scaled sample MER expressed as a ratio to global ΣMER

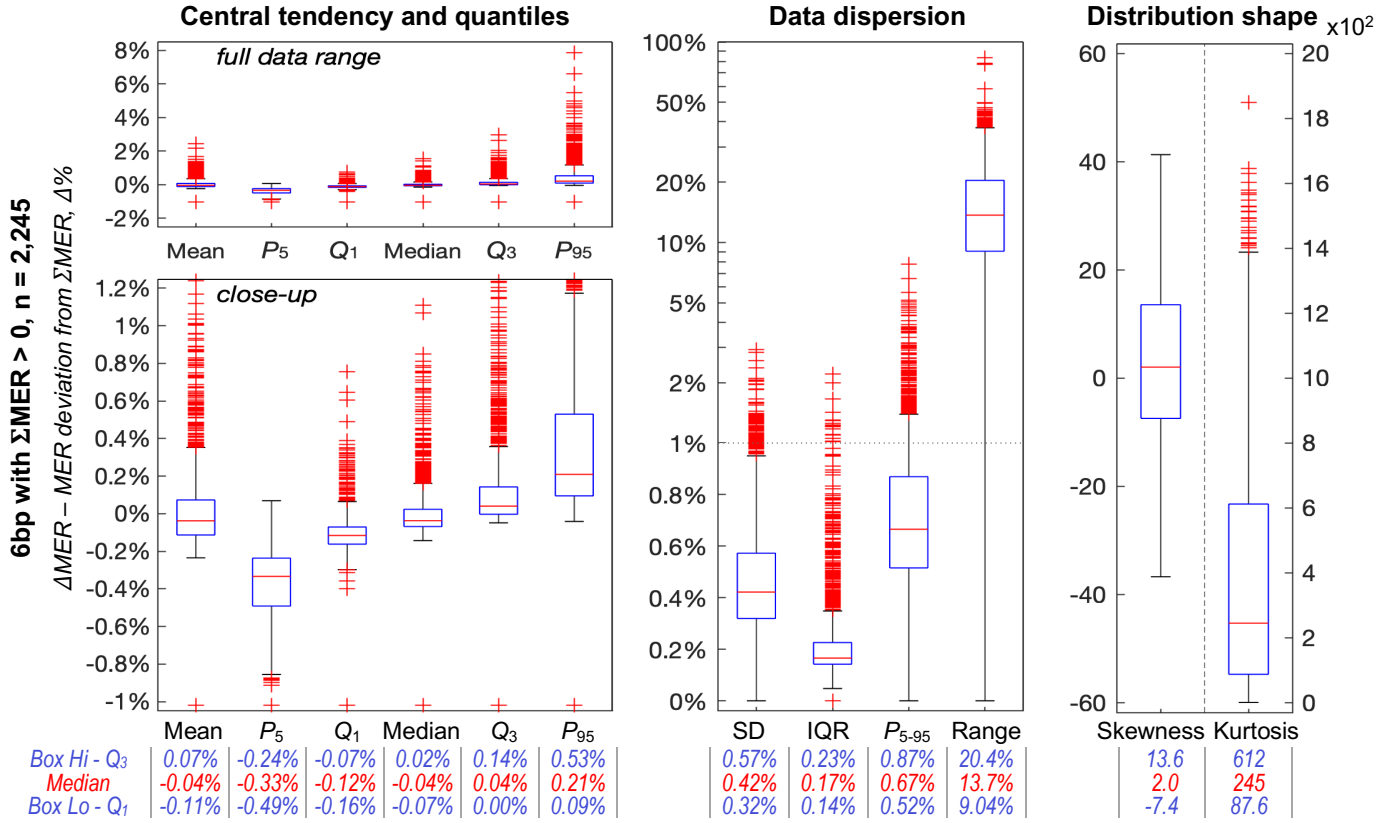
One complication associated with using MER_{SC} is that for motifs with negative ΣMER individuals with high MER values would have large but negative MER_{SC} values. To avoid this and any potential artefacts the simplest option is to exclude motifs with negative ΣMER from analysis altogether. When only positive ΣMER motifs are considered, we should expect sample median MER_{SC} values across the population to be highly variable. It is reasonable to expect median MER_{SC} distribution to be positively skewed with the bulk of individuals found near MER_{SC} = 1 and a tail formed by the minority of highly error prone outlier samples with large positive MER_{SC} values.

To assess whether either of the two CSE proneness variability models might be more probable in the context of our dataset, **Supplementary figure 4.3.3 S3** below presents population distributions of various summary statistics and shape parameters of the 6bp motif set ΔMER and MER_{SC} sample-specific distributions. Only 6bp motifs with ΣMER ≥0 are included in the analysis, with each boxplot representing the distribution of some ΔMER or MER_{SC} summary statistic values in 2035 analysed samples. **Panel A** presents ΔMER data while **panel B** shows data for MER_{SC}. The exact values for three main quantiles are listed directly beneath each box-and-whiskers plot (median – red, Q1 and Q3 – blue).

Supplementary figure 4.3.3 S3: 6bp motif ΔMER and MER_{SC} value distribution variability between individuals

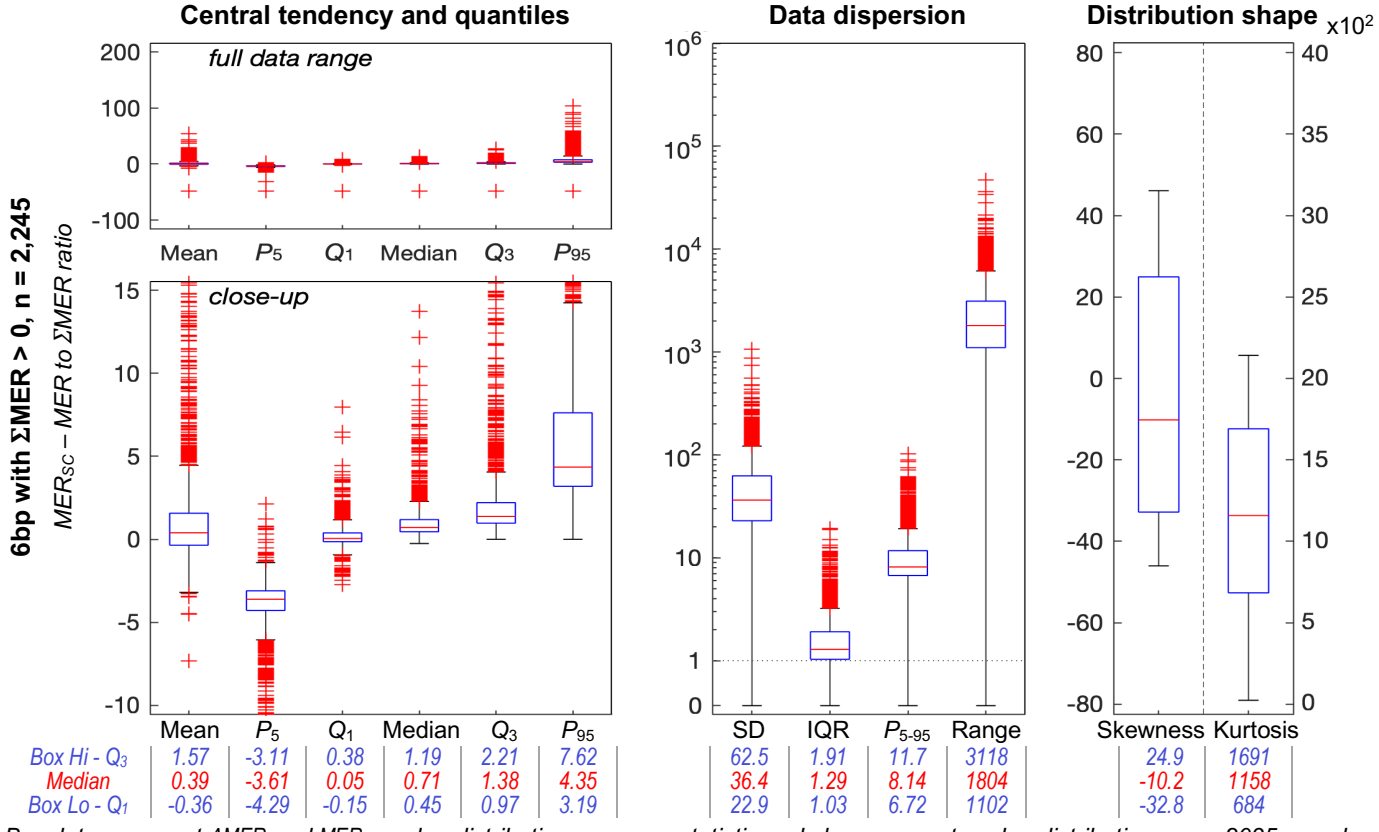
6bp motif ΔMER and MER_{SC} value distribution variability between population individuals

(a) Distributions of individuals' motif ΔMER distribution summary statistics and shape properties



	Mean	P_5	Q_1	Median	Q_3	P_{95}	SD	IQR	$P_{5\text{-}95}$	Range	Skewness	Kurtosis
Box Hi - Q_3	0.07%	-0.24%	-0.07%	0.02%	0.14%	0.53%	0.57%	0.23%	0.87%	20.4%	13.6	612
Median	-0.04%	-0.33%	-0.12%	-0.04%	0.04%	0.21%	0.42%	0.17%	0.67%	13.7%	2.0	245
Box Lo - Q_1	-0.11%	-0.49%	-0.16%	-0.07%	0.00%	0.09%	0.32%	0.14%	0.52%	9.04%	-7.4	87.6

(b) Distributions of individuals' motif ΔMER_{SC} distribution summary statistics and shape properties



	Mean	P_5	Q_1	Median	Q_3	P_{95}	SD	IQR	$P_{5\text{-}95}$	Range	Skewness	Kurtosis
Box Hi - Q_3	1.57	-3.11	0.38	1.19	2.21	7.62	62.5	1.91	11.7	3118	24.9	1691
Median	0.39	-3.61	0.05	0.71	1.38	4.35	36.4	1.29	8.14	1804	-10.2	1158
Box Lo - Q_1	-0.36	-4.29	-0.15	0.45	0.97	3.19	22.9	1.03	6.72	1102	-32.8	684

Boxplots represent ΔMER and MER_{SC} value distribution summary statistic and shape property value distributions over 2035 samples; ΔMER – absolute deviation of motif MER in an individual from the population global ΣMER level; MER_{SC} – scaled motif MER, expressed as a ratio of individuals' MER to population global ΣMER level; Distribution summary statistics: P_5, P_{95} – 5th and 95th percentiles; Q_1, Q_3 –1st and 3rd quartiles; SD – standard deviation; IQR – interquartile range; $P_{5\text{-}95}$ - 5th to 95th percentile range

Data presented in **Supplementary figure 4.3.3 S3 panel A** shows that ΔMER distribution medians are generally similar across the population (median: -0.04%; IQR: -0.07% to 0.02%), which indicates that for the majority of samples MER landscape is quite close to or just under the baseline. We do see a smaller proportion of outliers above this, the most extreme being an individual with median MER ~2% above the population baseline. Other ΔMER distribution quantiles, especially 5th or 95th percentiles, are much more variable between individuals however, indicating variable MER value spread between samples. Dispersion measure distributions confirm this, with difference of ~2 orders of magnitude observed between the distributions of ΔMER distribution interquartile range (median 0.14%, IQR 0.17%-0.23%) versus total range (median 13.7%, IQR 9.0%-20.4%). Skewness, while fairly symmetrically distributed around 0 with only minor bias towards being positive, is also highly variable between samples, as is kurtosis, further highlighting the variable ΔMER distribution shapes between individuals.

ΔMER quantile, especially 5th and 95th percentile distribution indicate that in the majority of samples there are sets of motifs both with MER above and below population ΣMER baseline level. This is contrary to our expectation that in any given sample most if not all MER values would be offset from the population baseline in the same direction. This not being the case probably speaks to a significant noise component.

If individuals' MER variability differences followed some approximation of the model of simple up/down shift from baseline, we should have largely expected to see sample ΔMER distributions with variable central tendencies but similar dispersion levels, evidenced by similar positively skewed distributions of different quantiles and little variance in overall distribution shape parameters. Our actual observed ΔMER behaviours however, even accounting for a substantial noise component, are clearly quite different. Therefore we can be fairly confident that the model where individual's MER patterns reflect a simple upwards or downwards shift from population baseline is not correct.

The increase in the inter-sample variability in terms of ΔMER distribution higher quantiles compared to central tendency would be more in line with the expectations of the MER landscape baseline scaling model. If individual's MER patterns were representing a scaled rather than shifter version of some baseline pattern, we would be expecting the absolute differences for the most CSE prone motifs to be largest, which would lead to more variation between individuals' highest MER motif values, and we could also reasonably expect to see more variability between individuals' MER distribution shape overall parameters.

Our observed ΔMER behaviours are not inconsistent with some type of error proneness baseline scaling model, it is also important to address the behaviour of MER_{SC}, shown in **Supplementary figure 4.3.3 S3 panel B**.

MER_{SC} distribution medians are generally fairly similar across the population (median 0.71, IQR 0.45-1.19), with the vast majority of sample median $MER_{SC} < 5$, but we see higher variability of other MER_{SC} distribution quantiles, especially 95th percentile, indicating that there are differences in how variable MER_{SC} values within samples are. We see the biggest jump in the observed inter-sample variability levels for distribution of MER_{SC} 95th percentile values compared to 75th percentile values, with MER_{SC} 95th percentile value distribution also being the most skewed across the population of all quantiles, suggesting that the largest differences between individuals' MER_{SC} distributions lie in the tails. MER_{SC} total range distribution (median 1804, IQR 1102-3118) that is several orders of magnitude more spread out than even 5th to 95th percentile range distribution (median 8.1, IQR 6.7-11.7) confirms this. Large variability in MER_{SC} distribution shapes between individuals is further evidenced by shape parameter distributions, both of which are very wide with kurtosis tending to be extremely large (median 1158, IQR 684-1692) and negative skewness being somewhat more common (median -10.2, IQR -32.8-24.9).

An important observation relates to the overall MER_{SC} range distribution between individuals. MER_{SC} value total range tends to be very high with examples of motifs with over 1000 times higher MER than ΣMER found in over 75% of population individuals. MER_{SC} this high is only mathematically possible if motif population ΣMER is well below

0.1%. This suggests that a lot of the extreme variation between sample MER_{SC} distributions must be found at motifs with very low baseline error rates, and could well be a result of random noise between sample MER values becoming dominant once the motif baseline error prate falls below our resolution threshold. So although on first examination several of the observed MER_{SC} distribution behaviours appear to be contradictory to expectations from some form of baseline CSE scaling model, the noise consideration is important and provides a plausible explanation for some of the discrepancies.

Thus while our results are not clear cut, we can with reasonable confidence dismiss simple CSE baseline shift model as the explanation for MER landscape variability between individuals, as it is more likely that the differences between the overall error level patterns between individuals are better described using some type of scaling based measure.

In addition to ΔMER and MER_{SC} population behaviours illustrated by data shown in **Supplementary figure 4.3.3 S3**, another important variable to consider is the relationship between the motif population ΣMER and the sample-level MER value dispersion. Under the additive shift model ΔMER in an individual is assumed to be independent of motif baseline error rate. Thus there is no reason to expect a correlation between motif sample-level MER value dispersion and the population global ΣMER level. Whereas with the scaling model the variable independent of motif ΣMER is MER_{SC}, while ΔMER is expected to be vary linearly with motif baseline ΣMER. This would lead to the motif ΔMER population variation being amplified more for motifs with higher ΣMER, which should be reflect as a measurable correlation between sample MER value dispersion and motif ΣMER.

ΣMER vs MER dispersion relationship in our dataset was already addressed in a different context earlier in **Supplementary figures 4.3.1 S3** and **S4,** which show a strong positive correlation between motif ΣMER and sample-level MER IQR across all motif spaces (Pearson r = 0.92 for 6bp motifs), with the relationship being clearest for motifs with ΣMER above around 0.2%. This result further supports a conclusion that some form of scaling rather than simple up/down shift is a more probable model for describing CSE proneness variation patterns between individuals, and a scaling factor relative to the baseline, while not perfect, could be a reasonable measure for how error prone a particular sample is, especially if some minimum motif ΣMER threshold was applied and only motif with baseline error rates above noise threshold were used for determining individual's error proneness overall level.

A2.3.2 Describing individual's relative 'error proneness'

Not all situations where we need to assess a particular sample's context-dependent sequencing error proneness require that the absolute differences in error proneness between samples are known. One situation where simply the relative ranking of samples could be needed would be if we were interested in finding outlier motifs for which the population stratification by MER values does not match that observed for most other motifs.

We already addressed why simply ranking individuals by their MER values at any given motif and using median ranks as the overall population stratification measure is problematic for this purpose. One alternative could be to use some other numeric error proneness measure such as ΔMER or MER_{SC} for assessing population stratification differences, but this has its own issues, including the fact that both these measures probably offer only a very simplistic approximation of the actual error proneness behaviours, plus without much more in-depth investigation there is no way to account for noise levels that may or may not differ between different between samples or even motifs. Furthermore, any results would need to be somehow normalised between motifs due to difference baseline motif error rates. So instead a rank based approach more similar to population ranking by median MER levels but one that correctly handles artefacts introduced by negative MER values would be more appropriate. One such possible is detailed below.

This uses an adjusted population ranking measure we shall refer to as scaled Rank Distance or RD_{SC}, which for each motif analysed within an individual is calculated as follows:

Step 1 – Raw rank: Rank all individuals in which the motif was analysed by that motif MER value from highest (rank 1) to lowest (rank 2035 assuming motif found in found in all individuals).

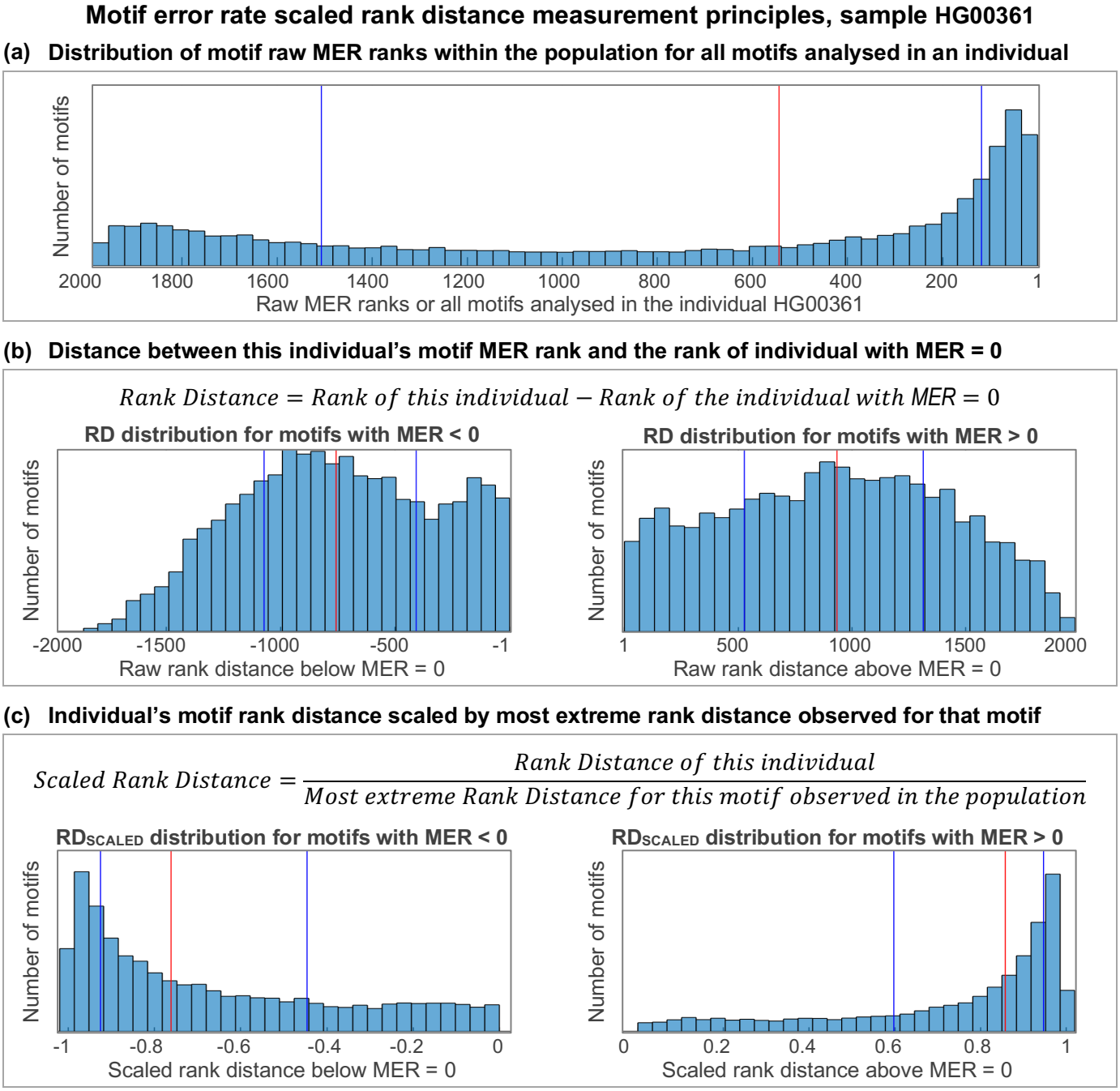
Step 2 – Rank Distance (RD): Determine where within the ranked population the sample with MER = 0 would be ranked, and count how many ranks away from that position each ranked individual is found and sign indicating. We shall refer to this measure as Rank Distance or RD. RD is positive for samples with MER>0 and negative for samples with MER<0. Sample with MER=0 if such exists has RD=0.

Step 3 – Scaled Rank Distance (RD_{SC}): Divide all positive RD values by the largest observed positive RD value and all negative RD values by the largest negative RD value. The result is Scaled Rank Distance or RD_{SC}, with range from -1 to 1. The sample with largest observed positive MER has RD_{SC} = 1 and the sample with largest negative observed MER (nearest -1%) has RD_{SC} = -1. Sample with MER=0 if such exists has RD_{SC} = 0

The resulting scaled rank distance provides a normalised measure of how different an individual's rank within population is from the rank of sample with MER=0 for any given motif. RD_{SC} values are independent of how many individuals with the particular motif were analysed or found to have positive versus negative MER values, which makes RD_{SC} values comparable between different motifs (unlike raw RD values). This means we can look at the shape of the RD_{SC} distribution across all motifs analysed in a particular sample as an unbiased descriptor of that specific individual's relative CSE error proneness behaviour compared to the rest of the population. And because the sign of RD_{SC} for any given motif shows whether that individual's motif MER is above or below zero, the relative rank distance patterns of positive MER motifs can easily be analysed completely separately from negative MER motifs

An example of the raw MER rank, absolute rank distance and scaled rank distance distributions across all analysed motifs in a specific individual (sample H00361) are shown in **Supplementary figure 4.3.3 S4. Panel A** presents histogram of the raw motif rank distribution across all analysed motifs regardless of motif MER values. The absolute rank distance distributions (**panel B**) and scaled rank distance distributions (**panel C**) are shown split into upper (positive MER, above MER=0 rank) and lower (negative MER, below MER=0 rank) distributions. In each panel the red vertical line indicates the distribution median with vertical blue lines marking interquartile range.

Supplementary figure 4.3.3 S4: Motif error rate scaled rank distance measurement principles, sample HG00361



Vertical lines mark distribution median (red) and IQR (blue)

We observed earlier that for many motifs a subset of samples exists with high median MER ranks that for that particular motif rank unusually low. **Supplementary figure 4.3.3 S4** demonstrates that sample HG00361 is one such example, with the bimodal nature of MER rank distribution, shown in **panel A**, highlighting why motif raw rank central tendency may not be the best descriptor of error proneness in this case

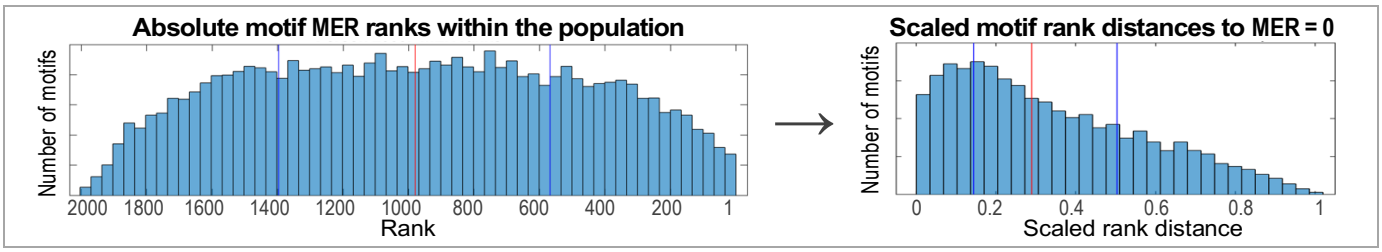
We speculated that since negative MER values are an artefact of how MER is calculated and result when background error rates in a sample exceed the context dependent error rate associated with a particular motif, sample rank patterns for artificially negative MER values should mirror the patterns of positive MER motifs. **Supplementary figure 4.3.3 S4** shows that for sample H00361 this is indeed the case – once motif scaled rank distances are calculated and positive values are split from negative ones, symmetry between this individual's upper and lower RD_{SC} distributions becomes evident. While this observation of symmetry provides an important proof of principle, negative MER values are an artefact, and it would be most appropriate to exclude negative MER motifs from any analysis requiring assessment of an individual's error proneness. Only the of individuals' upper RD_{SC} distribution behaviours should be assessed instead.

Supplementary figure 4.3.3 S5 shows 5 further individuals' MER raw rank (all motifs) and scaled upper rank distance (motifs with MER≥0 only) histograms demonstrating the variability of distribution shapes observed in our dataset. Similarly to previous figure, vertical lines mark distribution quartiles with median shown in red and IQR in blue.

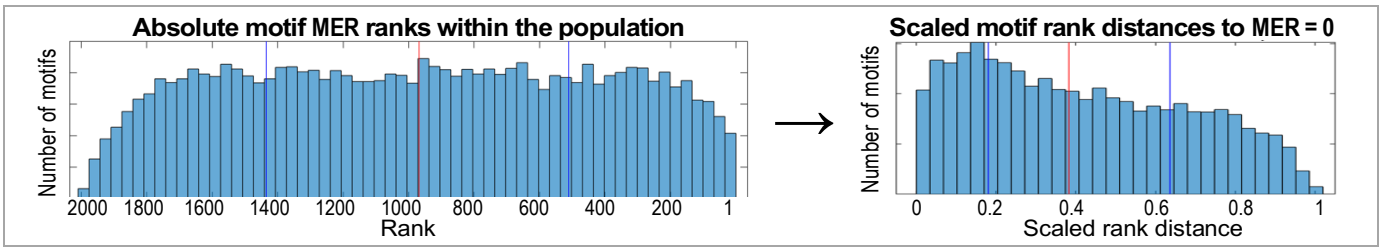
Supplementary figure 4.3.3 S5: Full MER raw rank and positive MER scaled rank distance distribution examples

Select individuals' full MER raw rank and positive MER scaled rank distance distributions

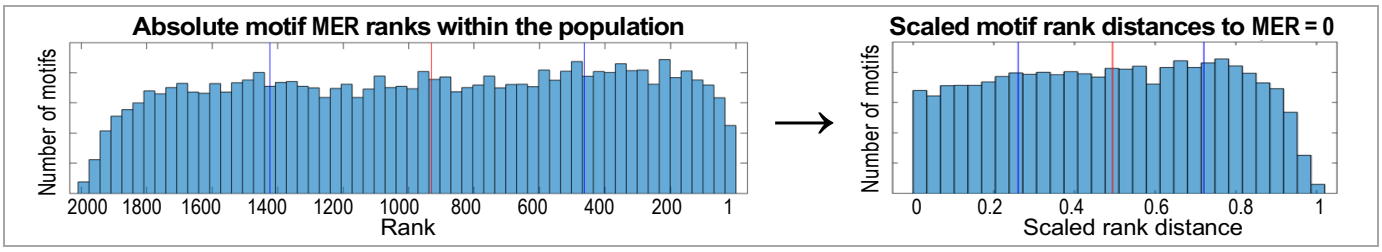
(a) Individual NA21108 – low error proneness



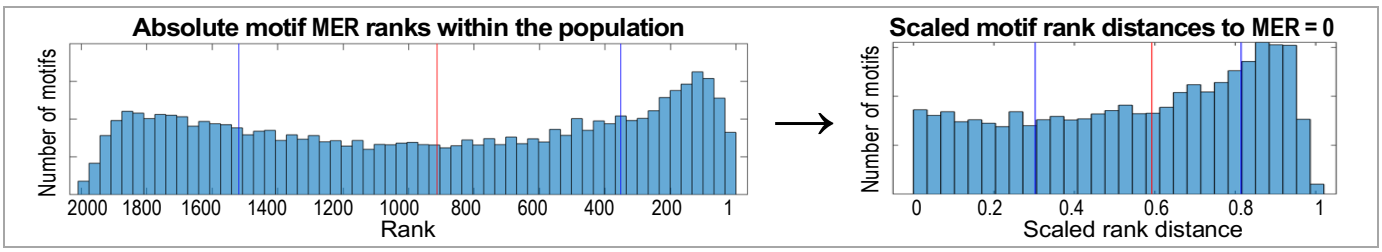
(b) Individual NA19184



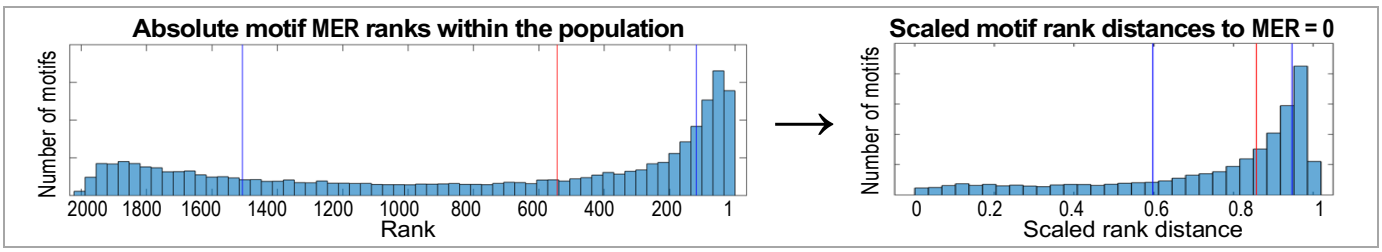
(c) Individual HG01369



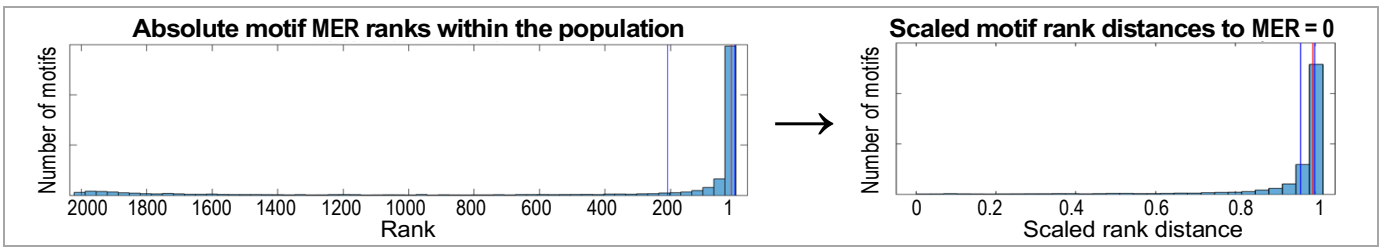
(d) Individual NA19059



(e) Individual HG00361



(f) Individual NA19309 – very high error proneness

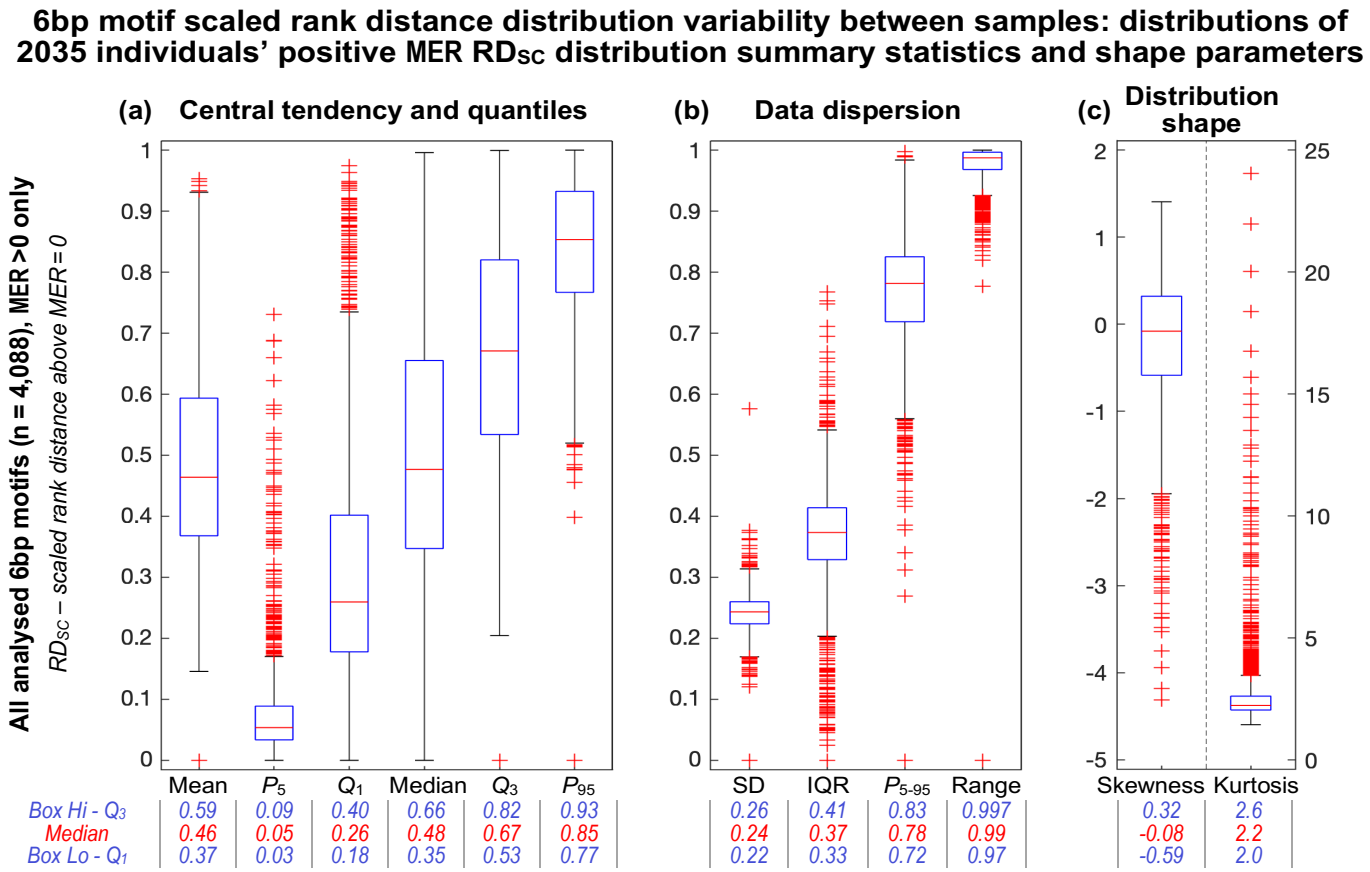


Scaled rank distance determined for motifs with positive MER only; Vertical lines mark distribution median (red) and IQR (blue)

Supplementary figure 4.3.3 S5 demonstrates that when we look at MER raw ranks, distribution bimodality is common. But once the negative MER motifs are removed and the remaining motif MER ranks are expressed as scaled distances above MER=0, the secondary distribution peak disappears. Scaled rank distance behaviour in a n individual thus becomes much easier to describe with simpler parameters compared to raw rank distribution behaviour, making RD_{SC} a more convenient measure of relative error proneness of an individual.

Supplementary figure 4.3.3 S6 below provides a visual summary of the overall variability observed between population individuals in terms of their MER RD_{SC} distributions across all 4088 analysed 6bp motifs. Only upper rank distances are considered, with any motif MER values below 0 excluded. Each boxplot represents population-wide distribution of a specific summary statistic or shape parameter describing each individual's MER RD_{SC} distribution.

Supplementary figure 4.3.3 S6: 6bp motif scaled rank distance distribution variability between samples: distributions of 2035 individuals' positive MER RD_{SC} distribution summary statistics and shape parameters



Boxplots represent MER scaled upper rank distance RD_{SC} distribution summary statistic and shape property value distributions over 2035 samples; RD_{SC} – Scaled rank distance, expressed as sample MER rank distance above rank of sample with MER=0, scaled as a fraction of largest observed MER rank distance for that motif. Distribution summary statistics: P_5, P_{95} – 5th and 95th percentiles; Q_1, Q_3 –1st and 3rd quartiles; SD – standard deviation; IQR – interquartile range; P_{5-95} - 5th to 95th percentile range

On a population scale, individual's RD_{SC} distribution medians follow a symmetric distribution with the mean of a median of 0.48 and IQR of 0.35-0.44. Individual's RD_{SC} distribution means are also distributed fairly symmetrically with a very similar median of 0.46 and IQR of 0.37-0.59. RD_{SC} distribution dispersion measure distributions show that there will nearly always be the entire spectrum of possible motif RD_{SC} values found in an individual, but individual vary in how peaked vs spread out their MER RD_{SC} distributions are as demonstrated by much wider IQR than range distribution. Kurtosis and skewness distribution indicates that extremely peaked distributions, such as demonstrated by sample NA19309 in previous figure are very much the minority. Furthermore, data presented shows that there are no individuals with peaked, highly positively skewed MER RD_{SC} distributions, yet many samples exist whose distributions are highly negatively skewed, confirming that while a minority of individuals who rank as some of the highest in the population across nearly all motifs exist, the opposite is not true.

Overall, based on this RD_{SC} distribution variability data, any one of several RD_{SC} distribution parameters could be argued to be the most useful single measure for individual's relative error proneness, and whether it is the RD_{SC} distribution mean, median or even skewness that is the best option may differ depending on the exact intended application or the research question being addressed.

A2.3.3 Motif likelihood of being a stratification outlier in terms of MER rank

One of the unanswered questions we raised earlier comes from visual observation of population stratification pattern differences between motifs observed in **figures 4.3.2 F1** and **F2** in the main text, which suggested that while across most motifs individual sample rank order follow a fairly similar pattern, for some of the very highest ΣMER motifs this sample rank order is more jumbled up. This was however an entirely qualitative observation, and thus the question is – can we quantitatively measure to what extent a particular motif behaves as an outlier in terms of population stratification by MER values, and can we test whether higher ΣMER correlates with a higher 'outlier-ness' of that motif in terms of stratification pattern by MER.

To answer this, we can re-phrase the problem. The motif scaled MER rank distance RD_{SC} discussed in the previous section provides a convenient measure where relative to the rest of the population each individual's MER value of a specific motif is ranked. And if each analysed sequence has a certain intrinsic error proneness associated with it, we would expect that in absence of noise, most motifs MER values measured in one sample would fall in the same position in the overall population rank order, and thus that individual sample would have the same RD_{SC} values across most motifs. Of course with a considerable noise component present, as is the case in our data, we are not looking at single RD_{SC} values but rather RD_{SC} distributions that are associated with each individual, with specific examples of these shown earlier in **Supplementary figure 4.3.3 S5**.

If there were motifs associated with uncharacteristically high or low MER for a given individual, RD_{SC} values of such motifs should fall further away from that individuals' RD_{SC} distribution central tendency compared to motifs with error rates more characteristic to that sample. So if we were to look at each analysed motif and determine where within different individuals' RD_{SC} distributions that motif tends to typically fall, if we then found a particular motif that would always fall within individuals' RD_{SC} distribution tails, this would be a strong indicator that that particular motif is an outlier with an unusual population sample ranking order. Thus for each motif we can use its typical placement within individuals' RD_{SC} distribution as a quantitative measure for how well that motif conforms to the typical MER population stratification pattern.

Of several possible ways to identify distribution outliers, here we shall use an outlier definition based on the distance outside the distribution IQR, whereby any data points are defined as outliers if their values fall outside the distribution middle two quartile boundaries by more than 1.5 times the IQR width. This is a frequently used outlier definition when handling non-normally distributed data as well as a common method for defining whisker lengths in box-and-whisker data visualisations, including all B&W plots presented in this thesis. Here and further we shall refer to this measure as IQR distance, or specifically upper IQR distance when specifying data point distance above Q_3 value and lower IQR distance when specifying data point distance below Q_1 value.

In this section we test our hypothesis that if we measure how often a motif happens to be an outlier within an individual's RD_{SC} distribution in terms of its IQR distance outside sample RD_{SC} distribution central tendency to quartile boundary, and identify in what proportion of samples motif matches outlier definition of IQR distance >1.5, the motifs found to have the largest outlier fraction among samples should match high ΣMER motifs observed qualitatively as population stratification outliers. Our results are presented in **Supplementary figure 4.3.3 S7** below.

Supplementary figure 4.3.3 S7: Quantification of 6bp motif MER population stratification outlier behaviour

Quantification of 6bp motif MER population stratification outlier behaviour

(a) Top 30 motifs most frequently identified in analysed individuals as MER stratification outliers

Rank by outliers	Rank by ΣMER	Motif sequence	Population ΣMER	Significant sample %	Number of samples where motif is an outlier			Rank by outlier sample proportion			Rank by median IQR distance outside Q_1/Q_3		
					Total	Lower	Upper	Total	Lower	Upper	Total	Lower	Upper
1	91	AGGGGC	1.26%	68.5%	89	78	11	3	1	130	4	48	4
2	**1**	**GCGGGT**	**38.4%**	**99.9%**	**67**	**54**	**13**	19	22	62	133	596	13
3	26	CATGGT	2.77%	76.2%	64	60	4	41	4	516	20	4	148
4	151	AGCGGC	0.82%	7.2%	61	49	12	36	46	76	445	707	335
5	1,091	CCCCGC	0.10%	6.7%	60	27	33	5	190	1	186	475	71
6	10	CCGAGT	4.55%	87.2%	60	54	6	58	26	330	506	1,273	282
7	31	GGCTCG	2.58%	72.9%	55	47	8	66	65	188	527	962	188
8	541	ATCCCA	0.22%	2.2%	54	46	8	54	38	188	266	396	246
9	24	AAGAGT	2.93%	89.7%	54	54	0	100	11	964	220	117	407
10	1,843	CACCGC	0.03%	0.4%	50	28	22	14	85	4	288	652	227
11	**3**	**CGGGGT**	**8.57%**	**97.5%**	**49**	**43**	**6**	174	79	348	1,172	1,437	1,153
12	**8**	**AGGGGT**	**5.19%**	**99.6%**	**49**	**49**	**0**	177	33	964	171	148	285
13	249	TTCCCG	0.52%	18.4%	48	41	7	59	47	183	977	1,091	923
14	394	CCGCCT	0.32%	18.1%	47	32	15	80	269	31	661	957	399
15	1,207	ACACAC	0.09%	1.6%	46	29	17	96	308	26	486	756	522
16	146	CCCGCT	0.84%	25.7%	45	30	15	57	186	24	3	244	1
17	124	CTCCCG	0.94%	34.3%	44	31	13	75	215	37	556	1,141	221
18	68	CGGAGT	1.57%	52.2%	44	43	1	256	102	960	597	1,180	256
19	2,420	TGTGCT	-0.01%	0%	40	26	14	2	29	3	244	301	169
20	1,053	TCTCGC	0.11%	0.05%	40	34	6	125	111	237	92	441	43
21	469	TCCGAC	0.27%	0.1%	39	37	2	273	172	692	662	1,738	308
22	20	CAGAGT	3.05%	87.2%	38	38	0	405	233	964	679	1,789	195
23	**27** *(19)*	**GAGAGT**	**2.71%**	**91.9%**	**38**	**38**	**0**	424	171	964	646	1,196	529
24	1,962	CCACGC	0.02%	0%	37	29	8	37	42	70	56	209	32
25	268	TCTCTG	0.48%	20.2%	37	36	1	49	18	780	522	605	500
26	860	CGAATC	0.14%	0%	37	28	9	115	246	74	112	710	39
27	**4**	**GCAGGT**	**8.11%**	**98.8%**	**37**	**34**	**3**	443	391	562	1,500	2,462	424
28	1,252	CGACAC	0.09%	0.05%	36	30	6	44	43	138	40	68	83
29	924	ACACCG	0.12%	1.6%	36	24	12	51	177	22	22	195	31
30	1,159	CGATGC	0.10%	0%	36	29	7	91	136	114	130	392	107

*Highlighted **in bold** – members of 49 most error prone 6bp motif subset (ΣMER ≥1, significance % ≥90%)*

(b) Other motifs among 49 most error prone 6bp motif subset earlier visually flagged as outliers

Rank by outliers	Rank by ΣMER	Motif sequence	Population ΣMER	Significant sample %	Number of samples where motif is an outlier			Rank by outlier sample proportion			Rank by median IQR distance outside Q_1/Q_3		
					Total	Lower	Upper	Total	Lower	Upper	Total	Lower	Upper
43	28 *(20)*	GGGAGT	2.70%	91.6%	33	33	0	593	342	964	576	1,690	228
93	39 *(29)*	AGGAGT	2.24%	94.9%	28	28	0	838	508	964	1,220	2,025	686
92	53 *(36)*	TGGAGT	1.76%	92.0%	28	28	0	831	493	964	757	1,628	445
961	82 *(43)*	TTTGGT	1.34%	92.3%	14	13	1	1,639	1276	963	1,570	1,523	1,806
2,426	94 *(46)*	ATGAGT	1.18%	91.2%	6	6	0	2,307	1997	964	1,946	2,676	1,141

(c) Correlation between motif population ΣMER level and its tendency to be a stratification outlier

Correlation with motif ΣMER	Total outlier count	Sample % where motif is outlier			Motif IQR distance outside Q_1/Q_3		Motif lower IQR distance below Q_1		Motif upper IQR distance above Q_3	
		Total outliers	Lower outliers	Upper outliers	Mean	Median	Mean	Median	Mean	Median
Pearson's *r*	0.13	-0.07	-0.06	-0.02	-0.11	-0.07	-0.16	-0.19	-0.003	0.03
p-value	*<1 x10^{-5}*	*0.0001*	*0.001*	*0.19*	*<1 x10^{-5}*	*4.1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*0.85*	*0.09*
Spearman's *ρ*	-0.27	-0.67	-0.62	-0.33	-0.61	-0.40	-0.72	-0.68	-0.27	-0.17
p-value	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*	*<1 x10^{-5}*

Analysis includes 3026 motifs identified in ≥2000 samples and MER>0 found in ≥500 samples; RD_{SC} – scaled MER rank distance; Outlier motifs – motifs found in sample MER RD_{SC} distribution >1.5 IQR lengths below Q_1 (lower outlier) or above Q_3 (upper outlier); IQR distance – distance of motif RD_{SC} value from individual's RD_{SC} overall distribution Q_1 or Q_3, expressed in IQR lengths; Lower IQR distance calculated for motifs with RD_{SC} below sample median (positive – motif below Q_1, negative – above Q_1) while upper IQR distance calculated for motifs with RD_{SC} above sample median (positive – motif above Q_3, negative – below Q_3)

The motifs listed in **Supplementary figure 4.3.3 S7 panel A** were identified as stratification outliers in the largest number of samples. **Panel B** lists 5 additional motifs we flagged as potential stratification outliers visually earlier, but which fall outside the top 30 by number of outliers found based on IQR distance. We also tested a hypothesis that the stratification pattern 'outlier-ness' of a motif is linked to its ΣMER level by measuring the correlation between motif ΣMER and either its sample outlier fraction or it's IQR distance central tendency across all analysed samples, with results presented in **panel C**.

Whether a motif is an outlier is assessed based on each individual's distributions of MER upper RD_{SC} values (i.e. individual's scaled ranking among samples with positive MER only), thus any motifs found in less than 2000 individuals in total or with less than 500 positive MER values found across the analysed samples were discarded. **Supplementary figure 4.3.3 S7** presents analysis of the remaining 3026 motifs. For each motif analysed in a sample we determined its RD_{SC} IQR distance from that sample's RD_{SC} distribution nearest quartile. For motifs with RD_{SC} values above individual's median upper IQR distance is measured to Q_3 value (i.e. how far above the IQR upper boundary the value is), whereas for motifs with RD_{SC} values below individual's median lower IQR distance is measured to Q_1 value (i.e. how far below IQR lower boundary the value is). Thus for both upper and lower IQR distances a positive value means motif falls outside of that individual's RD_{SC} distribution middle 2 quartiles, while negative value means motif RD_{SC} is found within the distribution central 50%.

In total from the 3026 analysed motifs we found that 3009 motifs match our outlier definition of being 1.5 IQR lengths outside an individual's RD_{SC} distribution Q_1 to Q_3 range in at least one sample. Of these, 1364 motifs (45.3%) were classed as outliers in between 1 and 9 samples. 1282 motifs (42.6%) matched outlier definition in 10 to 19 samples. Another 343 motifs (11.4%) were found as outliers in between 20 and 39 samples with only 10 motifs (0.3%) found to be outliers in more than 40 samples. The largest number of samples in which any motif was found to be an outlier was 89. Only 17 of the analysed motifs were not to be outliers in any samples.

In addition to listing the total number samples where each of the top 30 motifs are outliers overall, **Panel A** also specifically lists the number of lower outliers found >1.5 IQRs below Q_1 value and the upper outliers found >1.5 IQRs above Q_3 value. **Panel A** also lists motif rank positions sorted by the fraction of analysed samples in which that motif is an outlier, again both overall as well as separately specifically as the upper and lower outlier fractions.

The number of samples in which each of the top 30 motifs fits our outlier definition of 1.5 IQR lengths outside an individual's RD_{SC} distribution interquartile range is listed in **panel A** as both the total number of outliers, as well as split into specifically the number of lower outliers 1.5 IQRs below Q_1 value and the upper outliers 1.5 IQRs above Q_3 value. In addition to raw outlier counts, **panel A** also lists motif rank positions sorted by the fraction of analysed samples in which that motif is an outlier, both overall as well as separately as specifically the upper and lower outlier fractions.

Lastly, for each motif we also analysed its overall measured IQR distance distribution across all samples. This can also be split into distributions of only the upper IQR distances (i.e. only from samples where motif falls below sample median) and lower IQR distances (i.e. only from samples where motif is above sample median). As there are many motifs not found to be outliers in any samples We used the medians of the overall IQR distance distribution as well as medians of the separate lower and upper IQR distance distributions as another way to rank motifs. These ranks for each of the top 30 motifs are listed in the three rightmost columns in **panel A**.

Panel A shows that of the top 30 motifs identified as stratification outliers in the largest number of samples, only 5 belong to the subset of 49 most significant highly error prone 6bp motifs identified in our study with ΣMER ≥ 1% and population significance fraction ≥90%. While all of these 5 motifs were indeed flagged as apparent stratification outliers visually previously (see earlier discussion), so were another 5 motifs which did not make top 30 by outlier

count, although 3 did rank within the top 100. These 5 additional visually identified stratification outlier motifs are listed in **panel B** for comparison.

Sets of motifs listed in **Supplementary figure 4.3.3 S7** show that while indeed the measures used here for quantifying outlier behaviour do indeed flag up motifs we had previously marked as population stratification abnormalities based on qualitative observation, these are not the only motifs that exhibit stratification outlier behaviour based on quantitative measures. Among the top 30 motifs we see a very large ΣMER variation between -0.01% to 38.4%. Similarly MER population significance fraction among these 30 motifs ranges from 0% to 99.9%. Although our original hypothesis was that there is a link between motif ΣMER and how likely the motif is to be an outlier in how the population individuals rank by their respective MER values, the fact that we see very large ΣMER variation among these top 30 most likely be population stratification outlier motifs casts doubt on the relationship being quite this simple.

This hypothesis is further tested in **panel B**, where we attempt to quantify the overall correlation between motif population error rate and the likelihood of motif being a stratification outlier. Panel presents Pearson and Spearman correlation coefficients for motif ΣMER relationship to motif sample outlier fraction, both overall as well as by upper and lower outliers separately, as well as for the correlation between ΣMER and motif IQR distance overall distribution central tendency measures, again both for the overall IQR distances as well as for upper and lower IQR distance distributions separately.

The only variable tested that shows positive correlation with ΣMER is the raw number of outliers found, with Pearson correlation coefficient r=0.13 which is extremely weak. Spearman rank correlation coefficient ρ between the same two variables is =-0.27, indicating a very slightly stronger but negative correlation instead. When the raw outlier count is normalised relative to the total number of samples with positive MER analysed, we see a stronger inverse correlation between ΣMER and the total outlier sample fraction ($r=-0.07$, $\rho=-0.67$, $p<0.05$).

When motif IQR distance distribution central tendency is considered instead of the outlier count per se, ΣMER again shows an inverse correlation with both the IQR distance mean ($r=-0.11$, $\rho=-0.61$, $p<0.05$) and the median ($r=-0.07$, $\rho=-0.40$, $p<0.05$), albeit the correlation is stronger with the mean. If the measured IQR distances are split into upper and lower distances, both lower IQR distance distribution central tendency measures show very similar strong inverse correlation to ΣMER (mean: $r=-0.16$, $\rho=-0.71$, $p<0.05$; median: $r=-0.19$, $\rho=-0.68$, $p<0.05$). In contrast, the upper IQR distance distribution central tendency is much less strongly correlated to ΣMER with only Spearman p value being significant ($\rho=-0.0.27$, $p<0.05$).

Overall for all correlations tested, variable ranks are more strongly correlated then the absolute values as indicated by higher Spearman ρ magnitude compared to Pearson r.

The correlation results are somewhat contradictory to the impression of potentially a direct relationship that data presented in Panels A and B suggest. Motifs with ΣMER ≥1% make up 36.7% of the top 30 motifs with highest outlier counts listed in panel A. Among the top 100 motifs MER≥1% is found in 18% of cases, whereas among all 3026 analysed motifs, motifs with MER≥1% make up only 3%. This certainly was suggestive of a correlation. However when we rank motifs by normalised sample fraction instead, there are only 4 motifs with MER≥1% among the top 100, only 1 of which is among the top 30, confirming that once the number of analysed samples is accounted for the apparent direct correlation between ΣMER and the motif stratification outlier likelihood disappears. And the data presented in **panel C** suggests that we may be in fact looking at an inverse relationship between motif ΣMER level and how well that motif sample-level MER tends to matches the individuals' overall MER landscapes.

One of the potential reasons why lower error rate motifs may be more likely to appear as outliers in a larger proportion of individuals relative to each individuals baseline error rate landscape could be a result of the background noise present in our data. The lower is a motif error rate, the more compressed the population MER range becomes. So even if most individuals MER values still stratified same as for majority of motifs, the same amount of noise in MER value detection would have a potential of moving an individual's MER rank by more places compared to higher ΣMER motifs where the population MER values are more spread out. This could lead to more low error rate motifs being more likely to appear as outliers in terms of sample rank order than motifs with higher ΣMER.

Supplementary figure 4.3.3 S8 shows correlation coefficients between motif ΣMER level and the same 10 measures of motif stratification outlier behaviour as listed in **Supplementary figure 4.3.3 S7 panel C**, except within motif subsets with a defined minimum ΣMER. ΣMER minimum thresholds of 0.1%, 0.2%, 0.3%, 0.5% and 1% were selected based on previous observations from **Supplementary figure 4.3.3 S7** which indicated largest impact of random noise for MER values below ~0.2% threshold. Pearson correlation coefficients are shown in **panel A** and Spearman correlation coefficients are listed in **panel B**. P-values higher than 1.0×10^{-4} are shown below the relevant Pearson *r* or Spearman *ρ*. Where no value is listed, p is $<1.0 \times 10^{-4}$.

Supplementary figure 4.3.3 S8: Motif ΣMER level correlation with different population stratification outlier measures in 6bp motifs above different minimum ΣMER thresholds

Comparison of correlation between motif ΣMER level and different population stratification outlier measures in 6bp motif subsets with minimum ΣMER ≥ 0.1%, 0.2%, 0.3%, 0.5% and 1%

(a) Pearson correlation coefficients for 6bp motif sets with different minimum ΣMER thresholds

Pearson correlation with motif ΣMER *in motif subset with*	Total outlier count	Sample % where motif is outlier: Total outliers	Lower outliers	Upper outliers	Motif IQR distance outside IQR: Mean	Median	Motif lower IQR distance below Q_1: Mean	Median	Motif upper IQR distance above Q_3: Mean	Median
ΣMER **≥0.1%** *n = 1084*	0.26	0.19	0.13	0.19	0.10	0.07	0.04 *p: 0.0186*	-0.02 *p: 0.3375*	0.16	0.17
ΣMER **≥0.2%** *n = 586*	0.32	0.28	0.23	0.26	0.22	0.18	0.16	0.11	0.27	0.24
ΣMER **≥0.3%** *n = 410*	0.35	0.33	0.29	0.30	0.28	0.22	0.22	0.18	0.32	0.27
ΣMER **≥0.5%** *n = 251*	0.41	0.40	0.35	0.36	0.39	0.31	0.33	0.27	0.43	0.36
ΣMER **≥1 %** *n = 114*	0.42	0.44	0.64	0.37	0.45	0.36	0.36	0.31	0.53	0.42

P-values when >0.0001 shown below the corresponding correlation coefficient; when not specified, p value <0.0001

(b) Spearman rank correlation coefficients for motif sets with different minimum ΣMER thresholds

Spearman correlation with motif ΣMER *in motif subset with*	Total outlier count	Sample % where motif is outlier: Total outliers	Lower outliers	Upper outliers	Motif IQR distance outside IQR: Mean	Median	Motif lower IQR distance below Q_1: Mean	Median	Motif upper IQR distance above Q_3: Mean	Median
ΣMER **≥0.1%** *n = 1084*	-0.13	-0.21	-0.07	-0.21	-0.32	-0.24	-0.46	-0.47	-0.10	-0.04 *p: 0.0470*
ΣMER **≥0.2%** *n = 586*	-0.08	-0.12	-0.02 *p: 0.3422*	-0.13	-0.14	-0.04 *p: 0.0424*	-0.29	-0.19	0.06 *p: 0.0014*	0.10
ΣMER **≥0.3%** *n = 410*	-0.03 *p: 0.1451*	-0.06 *p: 0.0006*	0.05 *p: 0.0099*	-0.07 *p: 0.0003*	-0.05 *p: 0.0033*	0.02 *p: 0.3082*	-0.19	-0.07	0.10	0.10
ΣMER **≥0.5%** *n = 251*	0.23	0.21	0.14	0.20	0.21	0.22	0.11	0.15	0.31	0.27
ΣMER **≥1 %** *n = 114*	0.39	0.39	0.16	0.38	0.39	0.37	0.32	0.35	0.41	0.35

P-values when >0.0001 shown below the corresponding correlation coefficient; when not specified, p value <0.0001

Supplementary figure 4.3.3 S8 shows that when motifs with extremely low population ΣMER are excluded, the correlation observed between ΣMER and the different motif stratification outlier behaviour measures changes. Spearman rank correlation which indicated a significant inverse correlation between outlier behaviour and ΣMER when measured all motifs shifts to showing a significant positive correlation among motif subset with ΣMER ≥ 0.5%. Pearson correlation coefficients indicates a positive correlation between ΣMER values and motif stratification outlier

behaviour even earlier, with positive albeit very weak (r=0.19, p=3.8x10^{-26}) correlation between ΣMER and the motif relative outlier fraction among even the motifs with ΣMER≥0.1%, with *r* increased to 0.28 for motifs with ΣMER >0.2%, and to 0.4 for motifs with ΣMER≥0.5%.

Increase in both Spearman and Pearson correlation coefficients with increasing minimum ΣMER indicate that regardless of whether absolute values or ranks are concerned, the higher is the ΣMER threshold used in selecting the motif subset, the stronger is the detected positive correlation motif ΣMER values to any of our selected stratification outlier tendency measures. No specific ΣMER boundary emerges above which the correlation strength between ΣMER and outlier behaviour would increase no further (at least not within the range of the five ΣMER thresholds tested). So this data does not indicate that there would be some obvious easily identifiable boundary of presumed noise level at which the correlation between ΣMER to motif outlier tendency would clearly flip between inverse to direct, instead we see more of a continuum.

Overall, results presented in this section provide some quantitative evidence supporting our earlier visual observations that population stratification in terms of how samples rank by their motif specific error rates is more chaotic for the very high error rate motifs. We have evidence of a positive correlation between motif error rate and the probability of that motif being found more towards the tails of individuals' relative population rank position distributions, which means that higher ΣMER motifs are more likely to be stratification outliers in terms of how samples are ranked by their respective MER values compared to most motifs than lower ΣMER motifs, but this only demonstrably true among motifs with population error rate above ∼0.3% to ∼0.5% threshold. Motifs with lower context dependent error rate association than that do not follow the same correlation, likely due to background noise levels.

A2.3.4 Sequence patterns commonly found among outlier motifs

While we do not analyse specific sequence context association with different stratification outlier behaviours in detail, we do briefly address candidate pattern 'AGT' discussed earlier. This pattern was visually identified as a potential stratification outlier in an earlier section, with 4 instances found among the top 49 most error prone 6bp motifs. To check whether we detect a link between this 'AGT' pattern and the population stratification quantitative measured discussed in this section, we sorted 586 6bp motifs which have ΣMER ≥0.2% (threshold chosen due to noise level considerations addressed in previous figure) by their normalised total sample fraction and extracted the set of 100 highest ranked motifs. We found that among these top 100 motifs 'AGT' is in fact the most common pattern, with 12 out of 100 motifs matching this pattern. The second and third most common patterns are 'GGC' with 10 motifs found and 'GGT' with 8 motifs.

A brief summary of motifs found matching these top 3 patterns is provided in **Supplementary figure 4.3.3 S9**. **Panel A** compares prevalence of each of these three patterns among all 6bp motifs versus set of 586 motifs with ΣMER ≥0.2 versus the subset of top 100 highest ranked outlier motifs. Panels B, C and D address the differences between motif subsets matching each of the three patterns found among the wider set of 586 motifs versus those found specifically among the top 100 highest ranked motifs. In **panel B** median ΣMER is compared between pattern matches among the two sets, with **panel C** comparing median motif outlier fractions and **panel D** comparing median motif IQR distance medians between pattern matches found in the two motif sets. For each of these measures we show both the ratio and the absolute differences between the two subset medians.

Supplementary Figure 4.3.3 S9: 'AGT', 'GGC' and 'GGT' patterns are most common among 6bp motifs with ΣMER ≥0.2% most frequently seen as outliers relative so sample overall error proneness

Motifs matching top 3 most common patterns 'AGT', 'GGC' and 'GGT' found among 100 highest ranked 6bp motifs with ΣMER ≥0.2% by population stratification outlier fraction

(a) AGT, GGC and GGT pattern prevalence among all motifs vs top 100 motifs by outlier fraction

Motif 3' pattern	TOTAL All 3026 analysed 6bp motifs		SET 1 586 motifs with ΣMER ≥0.2%		SET 2 Top 100 motifs by outlier fraction		Pattern prevalence differences between sets: SET 1 from TOTAL		SET 2 from TOTAL		SET 2 from SET 1	
	Matches	%	Matches	%	Matches	%	*ratio*	Δ%	*ratio*	Δ%	*ratio*	Δ%
AGT	64	2.1%	64	10.9%	12	12%	↑ *5.2*	*+8.8%*	↑ *5.7*	*+9.9%*	↑ *1.1*	*+1.1%*
GGC	62	2.0%	19	3.2%	10	10%	↑ *1.6*	*+1.2%*	↑ *4.9*	*+8.0%*	↑ *3.1*	*+6.8%*
GGT	63	2.1%	63	10.8%	8	8%	↑ *5.2*	*+8.7%*	↑ *3.8*	*+5.9%*	↓ *0.7*	*-2.8%*

100 set 2 motifs are a subset of 586 set 1 motifs; % - pattern match prevalence within set; Δ% – prevalence difference between sets

(b) AGT, GGC and GGT pattern matching motif subset median ΣMER levels

Motif 3' pattern	SUBSET 1 Pattern matches among all 586 motifs with ΣMER ≥ 0.2%		SUBSET 2 Pattern matches among top 100 motifs with highest outlier fraction		SUBSET 2 vs SUBSET 1 Difference between subset ΣMER level medians	
	ΣMER median	IQR	ΣMER median	IQR	*ratio*	Δ
AGT	0.87%	0.63% – 1.27%	2.33%	1.50% – 2.77%	↑ *2.7*	*+1.46%*
GGC	0.32%	0.24% – 0.44%	0.40%	0.28% – 0.78%	↑ *1.3*	*+0.09%*
GGT	1.98%	1.33% – 2.89%	5.74%	4.59% – 8.23%	↑ *2.9*	*+3.76%*

(c) AGT, GGC and GGT pattern matching motif average sample outlier fractions

Motif 3' pattern	SUBSET 1 Pattern matches among all 586 motifs with ΣMER ≥ 0.2%		SUBSET 2 Pattern matches among top 100 motifs with highest outlier fraction		SUBSET 2 vs SUBSET 1 Difference between subset outlier fraction medians	
	Outlier % median	IQR	Outlier % median	IQR	*ratio*	Δ
AGT	0.30%	0.20% – 0.72%	1.60%	1.37% – 1.98%	↑ *5.3*	*+1.30%*
GGC	1.01%	0.68% – 1.57%	1.57%	1.36% – 1.72%	↑ *1.6*	*+0.56%*
GGT	0.26%	0.15% – 0.52%	2.13%	1.33% – 2.61%	↑ *8.2*	*+1.87%*

(d) AGT, GGC and GGT pattern matching motif average median IQR distances outside Q_1/Q_3

Motif 3' pattern	SUBSET 1 Pattern matches among all 586 motifs with ΣMER ≥ 0.2%		SUBSET 2 Pattern matches among top 100 motifs with highest outlier fraction		SUBSET 2 vs SUBSET 1 Difference between subset median IQR distance medians	
	M.IQR dist. median	IQR	M.IQR dist. median	IQR	*ratio*	Δ
AGT	-0.021	-0.039 – 0.008	0.052	0.018 – 0.061	↑ *-2.5*	*+0.072*
GGC	0.010	-0.034 – 0.043	0.043	0.025 – 0.062	↑ *4.3*	*+0.033*
GGT	-0.033	-0.052 – -0.011	0.036	0.009 – 0.102	↑ *-1.1*	*+0.069*

ΣMER – population error rate; IQR – interquartile range; ratio – set 2 median divided by set 1 median; Δ – set 2 median minus set 1 median; Outlier fraction – % of samples where motif IQR distance >1.5; IQR distance – motif distance outside of individual's scaled rank distance distribution interquartile range boundaries; M.IQR dist. – motif median IQR distance across samples with that motif

Motifs that match patterns 'AGT' and 'GGT' are associated with higher sequencing error rates in general than motifs matching 'GGC' pattern and as a result 100% of analysed 'AGT' and 'GGT' motifs are represented within the set of 586 motifs with ΣMER ≥ 0.2%, compared to only 19 out of 62 analysed 'GGC' motifs which represents only a slight 'GGC' prevalence increase in the higher ΣMER set. But when we look at the top 100 motifs by outlier fraction, we see an increase in prevalence of 'AGT' and 'GGC' patterns but a decrease in 'GGT' pattern. 'AGT' motifs make up 10.9% of all motifs with ΣMER ≥ 0.2%, and among the top 100 motifs this increases to 12%. 'GGC' pattern prevalence increase among the top 100 mots is more extreme, going from 3.2% among all motifs with ΣMER ≥0.2% to 10% among the top 100 motifs, representing 'GGC' prevalence increase by +6.8% compared to only +1.1% increase for 'AGT' motifs. Overall 'AGT' pattern is 1.1 times more prevalent among top 100 motifs, and 'GGC' pattern is 3.1 times more prevalent in the top 100 compared top prevalence among all motifs with ΣMER ≥0.2%.

The third pattern 'GGT' actually shows a decrease in frequency of -2.8% among the top 100 highest outlier % motifs compared to the wider set of all 586 motifs above the 0.2% ΣMER threshold, indicating that 'GGT' is probably less correlated to motif outlier behaviour than the other two patterns. 'GGT' motifs also show much wider variability in

terms of both the actual outlier fraction and the RD_{SC} IQR distance distribution (see IQR values) compared to the other two patterns.

Within the wider 586 motif set, median ΣMER level for 'GGT' pattern motifs is 2.0%, while for 'AGT' motifs this is 0.8% and for 'GGC' motifs it is 0.3%. Among the top 100 motifs, the median ΣMER for 'GGT' pattern motifs is 5.7% (2.9 times higher), for 'AGT' motifs 2.3% (2.7 times higher) and for 'GGC' median 0.4% (1.3 times higher). This suggests that even though we showed earlier that ΣMER is correlated with likelihood that a motif is an outlier, this does not seem to be the main factor, and actually the 'GGT' pattern motifs which are associated with highest median ΣMER show the lowest relative prevalence increase among the top 100 outlier motifs, whereas the pattern 'GGC' which is associated with much lower ΣMER levels is actually the pattern demonstrating biggest increase in prevalence among the very top outlier motifs. Pattern 'AGT' is slightly more frequent among top outlier motifs but does not exhibit nearly as high a frequency increase as 'GGC' does.

Data presented in **Supplementary figure 4.3.3 S9** suggests that probably a link exists between motif sequence pattern and what population stratification order results from error rates measured for that motif. In depth analysis of motif sequence pattern relationship to population stratification behaviours is beyond the scope of this study, however our preliminary results, while by no means exhaustive, point to sequence patterns ending in 'AGT' and 'GGC' are candidates likely linked to deviations in sample error rate rank order.

In conclusions, the data presented in this supplementary analyses section provides an example of how sample-level relative error proneness analysis approach such as this could be a useful tool for investigating sequencing error rate landscapes and their variability. Results from this kind of motif population stratification behaviour quantification could be basis for further clustering analysis searching for patterns of which motifs follow which kinds of stratification behaviours, and whether there are sequence commonalities associated with more or less randomness in context-dependent sequencing error rates variation between samples. Among other potential applications, finding sequence contexts that are least likely to produce error rate measurements that are deviant from the baseline expected CSE proneness of that specific sample could perhaps provide a set of 'most informative' contexts one should preferentially look at when assessing background error rate differences between sequencing runs.

Appendix 3 Supplementary information for Chapter 5

List of supplementary figures associated with Chapter 4

A3.1 Supplementary figures associated with section 5.2 of the main text

5.2.1 S1: Overview of the analysed dataset and the heteroplasmic variant filtration strategy 261
5.2.1 S2: Sample size analysed with different heteroplasmy identification methods 262
5.2.1 S3: Mitochondrial genome regions with lower number of analysed samples 263
5.2.1 S4: Overview of the analysed mitochondrial genome fraction across the population 264
5.2.1 S5: Shared and unique heteroplasmies identified at different filtration stringencies 265
5.2.1 S6: Increased filtration stringency effect on the reduction of putative heteroplasmies 266
5.2.1 S7:Heteroplasmy incidence in mitochondrial genome found at different filtration stringency levels: comparison between two heteroplasmy identification methods 267
5.2.1 S8: Overview of heteroplasmic sample counts per position found using two heteroplasmy identification methods across different filtration stringency levels 268
5.2.1 S9: Heteroplasmy incidence in population found at different filtration stringency levels: comparison between two heteroplasmy identification methods 269
5.2.1 S10: Overview of heteroplasmic position counts per sample found using two heteroplasmy identification methods across different filtration stringency levels 270
5.2.2 S1: Mean minor allele frequency across all identified putative heteroplasmies 271
5.2.2 S2: Average heteroplasmy level across the subset of shared putative heteroplasmies identified with both standard and consensus methods 272
5.2.2 S3: Mean MAF measured by standard method across shared vs unique heteroplasmies 273
5.2.2 S4: MT genome positions with unique putative heteroplasmies found at filter 7 274
5.2.3 S1: Identified heteroplasmy base frequency shift away from MT genome baseline 275
5.2.3 S2: Comparison between observed and expected heteroplasmy base pair frequencies 276
5.2.4 S1: AT pairs found in MT genome in different orientations and sequence contexts 278
5.2.4 S2: Impact of CSEs at GGT-like motifs on the expected heteroplasmy A/T ratio 279
5.2.4 S3: Actual observed A/T and C/G ratios in the putative heteroplasmy subsets 280

A3.2 Supplementary figures associated with section 5.3 of the main text

5.3.1 S1: Landscapes of mitochondrial heteroplasmy reports across 20 previous studies 283
5.3.2 S1: Relative base type frequency deviation from mitochondrial genome baseline frequencies found in position subsets with different numbers of previously reported heteroplasmies 284
5.3.3 S1: Overlap between putative heteroplasmies identified using standard method and the sets of heteroplasmic mtDNA positions reported in literature previously 285
5.3.3 S2: Correlation between heteroplasmy counts reported in previous literature with the heteroplasmic sample fraction identified in this study using standard method 286
5.3.3 S3: Correlation between heteroplasmy counts between 6 largest studies and this study 287

A3.3 Supplementary figures associated with section 5.4 of the main text

5.4.1 S1: Number of heteroplasmies detected at each mtDNA position by consensus method 289
5.4.1 S2: Number of heteroplasmies detected at each mtDNA position by standard method 290
5.4.1 S3: Heteroplasmy prevalence at each mtDNA position found with standard method 291
5.4.1 S4: Positions with largest discrepancy between total heteroplasmy counts versus normalised heteroplasmic sample fractions at filter levels 5 (MAF ≥0.1%) and 7 (MAF≥1%) 292
5.4.1 S5: Distribution of mtDNA positions with different number of heteroplasmies 293
5.4.2 S1: 142 positions with ≥100 heteroplasmies found at filter 5 with standard method 295
5.4.2 S2: 88 positions with ≥10 heteroplasmies found at filter 7 with standard method 297
5.4.2 S3: Sites with prevalent heteroplasmy ≥0.1% and ≥1% MAF by consensus method 298
5.4.2 S4: Sites with prevalent heteroplasmy ≥0.1% and ≥1% MAF by standard method 298

List of supplementary figures for Chapter 5 (cont.)

A3.4 Supplementary figures associated with section 5.5 of the main text

5.5.1 S1: Mean MAF detected across all analysed samples using standard method 300
5.5.1 S2: Mean MAF detected across all analysed samples using consensus method 301
5.5.1 S3: Heteroplasmy incidence above different MAF thresholds across MT genome 302
5.5.2 S1: Mean MAF across identified heteroplasmies using standard method 304
5.5.2 S2: Overview of the heteroplasmy mean MAF level landscape across the MT genome 305
5.5.2 S3: Median MAF across identified heteroplasmies using consensus method 307
5.5.2 S 4: Median MAF across identified heteroplasmies using standard method 308
5.5.2 S5: Overview of the heteroplasmy median MAF level landscape across the MT genome 309
5.5.2 S6: Average heteroplasmy fractions above different MAF thresholds 310
5.5.2 S7: Distributions of position-specific MAF population distribution summary statistics and shape properties at filter levels 5 and 7 identified using standard method 311
5.5.3 S1: Heteroplasmic position base composition – standard method 312
5.5.3 S2: Position base type and the average heteroplasmy MAF level – standard method 312
5.5.4 S1: Relationship between the heteroplasmies found per site and the position-specific median MAF measured by standard method across heteroplasmies with MAF>0.1% 313
5.5.4 S2: Overview of 1,499 mtDNA positions with median MAF ≥1% by standard method 314
5.5.4 S3: Sequence context around high heteroplasmy positions with median MAF ≥ 1% 315
5.5.4 S4: High heteroplasmy positions 10+ heteroplasmies and median MAF≥1% 316
5.5.4 S5: Sequence context of 191 very high heteroplasmy positions with median MAF ≥10% 317
5.5.5 S1: Heteroplasmic positions with samples above 0.1%, 0.5%, 1%, 5% and 10% MAF 318
5.5.5 S2: Adjacent context of high heteroplasmy sites (5+ samples with MAF ≥1%, 5%, 10%) 319
5.5.6 S1: Observed MAF comparison with heteroplasmy report frequency in literature 320
5.5.6 S2: Correlation between previous heteroplasmy report frequency and our observed average MAF levels or sample counts above different MAF thresholds 321
5.5.6 S3: MAF levels at 107 positions with highest heteroplasmy incidence at filter 5 323
5.5.6 S4: MAF levels at 88 positions with highest heteroplasmy incidence at filter 7 324

A3.5 Supplementary figures associated with section 5.6 of the main text

5.6.1 S1: MAF observed at the heteroplasmic sites found in different genomic contexts 325
5.6.1 S2: Heteroplasmic site proportions with different levels of heteroplasmy prevalence 325
5.6.1 S3: Heteroplasmy discovery rates at different genomic contexts in the literature 326
5.6.2 S1: Variability in literature: heteroplasmy distribution between codon positions 327
5.6.3 S1: Base composition of heteroplasmies in different genomic contexts 328

A3.6 Supplementary figures addressing comparison between previously known versus novel heteroplasmies

5.7.1 S1: Heteroplasmies with MAF ≥0.1% found at previously known versus novel positions 330
5.7.1 S2: Heteroplasmies with MAF ≥1% found at previously known versus novel positions 331
5.7.2 S1: Base composition of previously known versus novel heteroplasmic positions 333
5.7.3 S1: MAF and heteroplasmy prevalence at previously known versus novel sites 336
5.7.4 S1: Heteroplasmic variants found at previously known vs novel heteroplasmic sites 338
5.7.4 S2: Transitions and transversions at previously known vs novel heteroplasmic sites 340

A3.1 Heteroplasmy measurement method comparison

A3.1.1 Putative heteroplasmies detected at different allele filtration stringencies

Supplementary figure 5.2.1 S1 provides an overview of the analysed dataset and the proposed variant allele filtration strategy.

Supplementary figure 5.2.1 S1: Overview of the analysed dataset and the heteroplasmic variant filtration strategy

Overview of the analysed dataset and putative heteroplasmic variant filtration strategy

(a) Original dataset and the initial filtration

Original dataset	3202 samples x 16568 positions = 53,050,736 sites	
Dataset after initial filtration steps*	**Standard method**	**Consensus method**
Analysed sites	51,883,365	51,882,772 (-593)
as % of original dataset	97.80%	97.80%

** Initial filtration excludes sites with < 500 reads per sequencing direction or with deletions in consensus sequence*

(b) Criteria applied during putative heteroplasmic site filtration

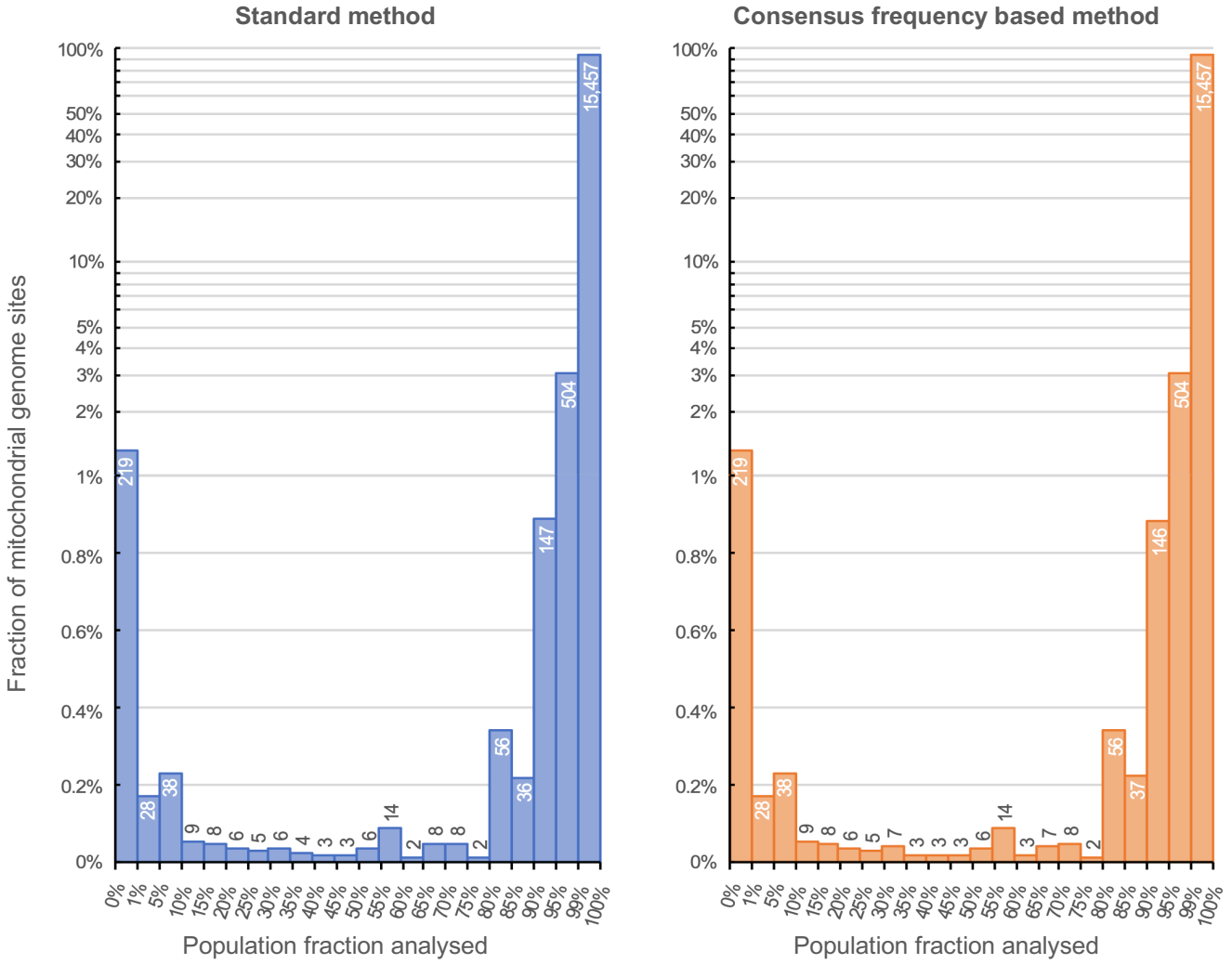
Filter	Minimum allele read count threshold	Minimum allele frequency threshold
Filter 1	*1 read total*	*none*
Filter 2	*1 read in each sequencing direction*	*none*
Filter 3	*5 reads in each sequencing direction*	*none*
Filter 4	*5 reads in each sequencing direction*	*0.1% total*
Filter 5	*5 reads in each sequencing direction*	*0.1% in each sequencing direction*
Filter 6	*5 reads in each sequencing direction*	*1% total*
Filter 7	*5 reads in each sequencing direction*	*1% in each sequencing direction*

Supplementary figure 5.2.1 S2 summarises the distribution of proportions of mitochondrial genome with different levels of population coverage with and without excess read filtration.

Supplementary figure 5.2.1 S2: Sample size analysed with different heteroplasmy identification methods

Overview of the analysed sample size with different heteroplasmy identification methods

(a) Analysed sample size distribution across different mtDNA positions



(b) Number of sites analysed in different percentages of samples

Analysed in % of samples	Number of sites		Percentage of MT genome	
	Standard method	Consensus method	Standard method	Consensus method
0%	166	166	1.0%	1.0%
<10%	119	119	0.7%	0.7%
10% - 50%	44	44	0.3%	0.3%
50% - 75%	38	38	0.2%	0.2%
75% - 90%	94	95	0.6%	0.6%
>90%	2293	2294	13.8%	13.8%
100%	13815	13813	83.4%	83.4%

(c) Overview of the number of individual samples analysed per site

Analysed per site	Number of samples		Analysed population fraction	
	Standard method	Consensus method	Standard method	Consensus method
Mean	3131.35	3131.32	97.8%	97.8%
SD	433.54	433.64	13.5%	13.5%
Median	3202	3202	100%	100%
Minimum	0	0	0%	0%
Maximum	3202	3202	100%	100%

SD – standard deviation

Only a small proportion on mtDNA sites were found with severely reduced analysed sample numbers. These lower coverage regions are summarised in **Supplementary figure 5.2.1 S3**.

Supplementary figure 5.2.1 S3: Mitochondrial genome regions with lower number of analysed samples

Mitochondrial genome regions with lower number of analysed samples: region locations, overlap with functional sequence elements and analysed sample fractions

Region	Region positions	Analysed sample %	Region size	Overlap with MT functional elements	Region average analysed sample %
Region 1	**1 - 157**		**157**	**Control region**	Standard method: 2.93%
Sub regions:	1 - 100	0%	100	*CR:7S-like, CR:OH57, CR:OH, CR:HVS2, ATT*	Consensus method: 2.91%
	101 - 104	<10%	4		
	105	0%	1		
	106 - 142	<10%	37		
	143 - 157	<50%	15		
Region 2	**302 - 315**		**14**	**Control region**	Standard method: 12.91%
Sub regions:	302 - 309	<50%	8	*CR:TFY, CR:OH, CR:CSB2, CR:HVS2, ATT*	Consensus method: 12.87%
	310 - 315	<10%	6		
Region 3 *	**3106 - 3109**		**3**	**16S ribosomal RNA**	Standard method: 0.01%
Sub regions:	3106	<10%	1		Consensus method: 0.01%
	3108 - 3109	0%	2		
Region 4**	**16416 - 16569**		**154**	**Control region**	Standard method: 5.49%
Sub regions:	16416 - 16436	<50%	21	*CR:7S-like, CR:mt3L, ATT*	Consensus method: 5.46%
	16437 - 16465	<10%	29		
	16466 - 16523	0%	58		
	16524 - 16565	<10%	42		
	16566 - 16569	0%	4		

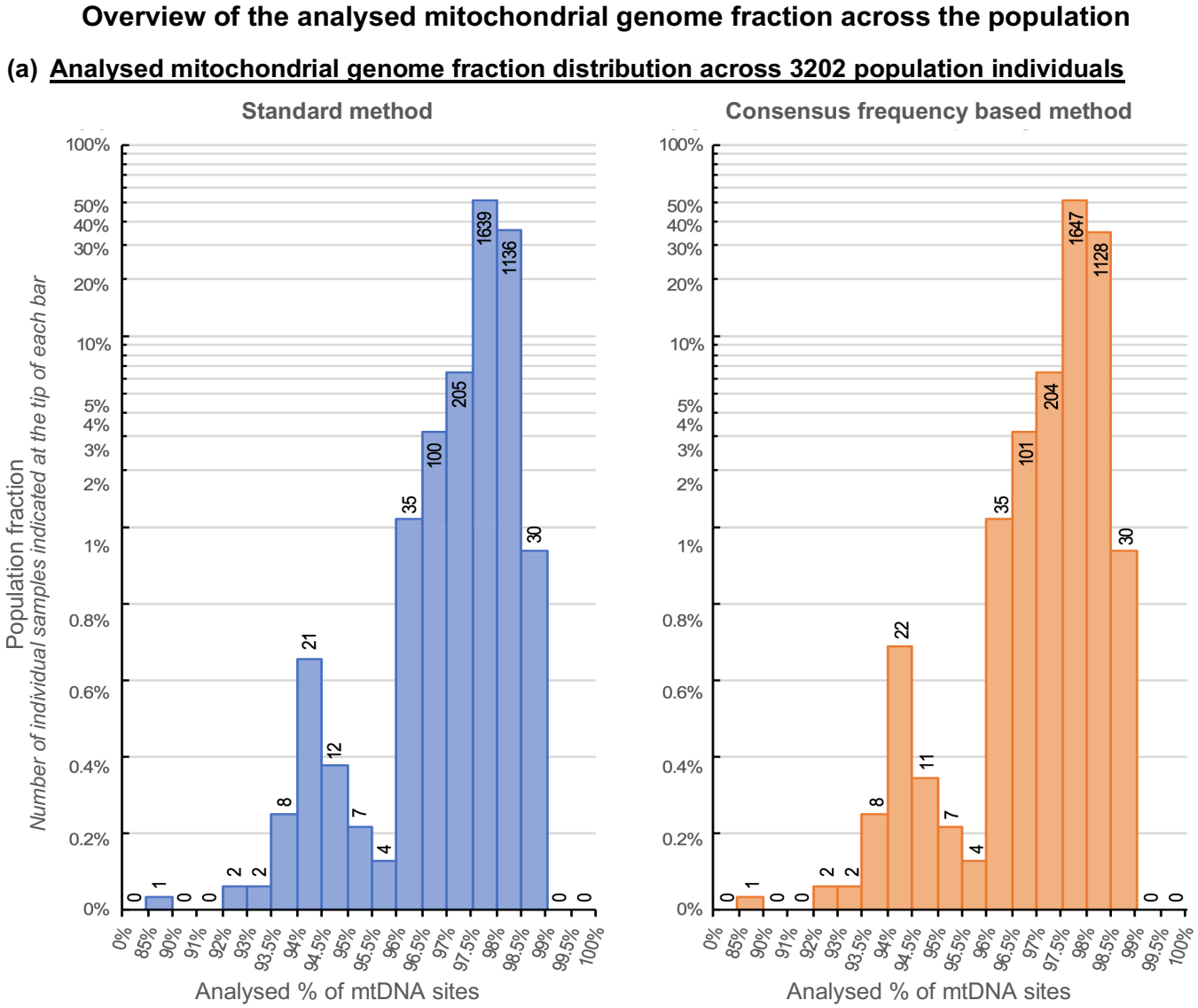
** Position 3107 in region 3 is a deletion artefact retained to preserve historic position numbering*

*** Split between regions 4 and 1 is artificial, arising from position numbering convention*

Functional sequence elements: CR:7S-like – 7S DNA (16106-191), CR:HVS2 – Hypervariable segment 2 (57-372), CR:OH – H-strand origins (110-441), CR:OH57 – H-strand origin (57), CR:TFY – mtTF1 binding site (276-303), CR:mt3L – L-strand control element (16499-16506), ATT – membrane attachment site (15925-499)

Supplementary figure 5.2.1 S4 visualises the analysed sample mtDNA sequence fraction distribution before and after excess read removal. It demonstrates a very good and similar genome coverage across all samples.

Supplementary figure 5.2.1 S4: Overview of the analysed mitochondrial genome fraction across the population



(a) Analysed mitochondrial genome fraction distribution across 3202 population individuals

(b) Number of samples with different analysed mitochondrial genome percentages

Analysed genome fraction	Number of samples		Population fraction	
	Standard method	Consensus method	Standard method	Consensus method
< 89%	0	0	0%	0%
89% - 96%	57	57	1.78%	1.78%
96% - 97%	135	136	4.22%	4.25%
97% - 98%	1844	1851	57.59%	57.81%
98% - 99%	1166	1158	36.41%	36.16%
>99%	0	0	0%	0%

(c) Overview of the number of mitochondrial genome sites analysed per sample

Analysed per sample	Number of sites		Analysed sequence fraction	
	Standard method	Consensus method	Standard method	Consensus method
Mean	16203.42	16203.24	97.80%	97.80%
SD	100.35	100.46	0.61%	0.61%
Median	16223	16223	97.92%	97.92%
Minimum	14779	14777	89.20%	89.19%
Maximum	16392	16392	98.94%	98.94%

SD – standard deviation

Supplementary figure 5.2.1 S5 lists the putative heteroplasmies identified at different allele filtration stringency levels with standard and consensus methods.

Supplementary figure 5.2.1 S5: Shared and unique heteroplasmies identified at different filtration stringencies

Shared and unique putative heteroplasmies identified at different filtration stringency levels

Filter	Standard method			Consensus method		
	Total	Shared	Unique	Total	Shared	Unique
Filter 1	39,314,429	8,835,935 *(22.5%)*	30,478,494 *(77.5%)*	8,835,935	8,835,935 *(100%)*	0 *(0%)*
Filter 2	8,836,025	2,852,091 *(32.3%)*	5,983,934 *(67.7%)*	2,852,091	2,852,091 *(100%)*	0 *(0%)*
Filter 3	197,188	138,981 *(70.5%)*	58,207 *(29.5%)*	138,981	138,981 *(100%)*	0 *(0%)*
Filter 4	180,091	120,754 *(67.1%)*	59,337 *(32.9%)*	120,754	120,754 *(100%)*	0 *(0%)*
Filter 5	151,354	120,697 *(79.7%)*	30,657 *(20.3%)*	120,754	120,697 *(99.95%)*	57 *(0.05%)*
Filter 6	17,935	13,403 *(74.7%)*	4,532 *(25.3%)*	13,403	13,403 *(100%)*	0 *(0%)*
Filter 7	13,403	13,390 *(99.9%)*	13 *(0.1%)*	13,403	13,390 *(99.9%)*	13 *(0.1%)*

Shared – putative heteroplasmies identified by both methods; Unique – putative heteroplasmies identified by one method only

Supplementary figure 5.2.1 S6 quantifies how much each allele filtration step reduces the total putative heteroplasmy count compared to the previous filter.

Supplementary figure 5.2.1 S6: Increased filtration stringency effect on the reduction of putative heteroplasmies

Increased filtration stringency effect on the reduction of putative heteroplasmy counts

(a) Increasingly stringent allele filtration criteria used in putative heteroplasmic site filtration

Stringency level	Minimum criteria for supporting variant allele
Filter 1	*Variant allele must be supported by 1 read*
Filter 2	*Require at least one variant read to be found in each sequencing direction*
Filter 3	*Require minimum 5 variant reads to be found in each sequencing direction*
Filter 4	*Set the minimum required variant allele frequency to be 0.1% of all reads*
Filter 5	*Apply minimum 0.1% allele frequency requirement to each sequencing direction separately*
Filter 6	*Increase the required minimum allele frequency tenfold from 0.1% to 1% of all reads*
Filter 7	*Apply the increased minimum 1% allele frequency requirement to each sequencing direction*

(b) Factor by which filter stringency increase decreases the number of detected putative heteroplasmies

Filter stringency increase	Standard method Total	Standard method Unique	Standard method Unique %	Consensus method total	Found by both methods
Filter 1 > 2	↓ 4.4	↓ 5.1	↓ 1.1	↓ 3.1	↓ 3.1
Filter 2 > 3	↓ 44.8	↓ 102.8	↓ 2.3	↓ 20.5	↓ 20.5
Filter 3 > 4	↓ 1.1	↓ 0.98 (↑ 1.02)	↓ 0.90 (↑ 1.12)	↓ 1.2	↓ 1.2
Filter 4 > 5	↓ 1.2	↓ 1.9	↓ 1.6	1	1
Filter 5 > 6	↓ 8.4	↓ 6.8	↓ 0.80 (↑ 1.25)	↓ 9.0	↓ 9.0
Filter 6 > 7	↓ 1.3	↓ 348.6	↓ 260.5	1	↓ 1.0

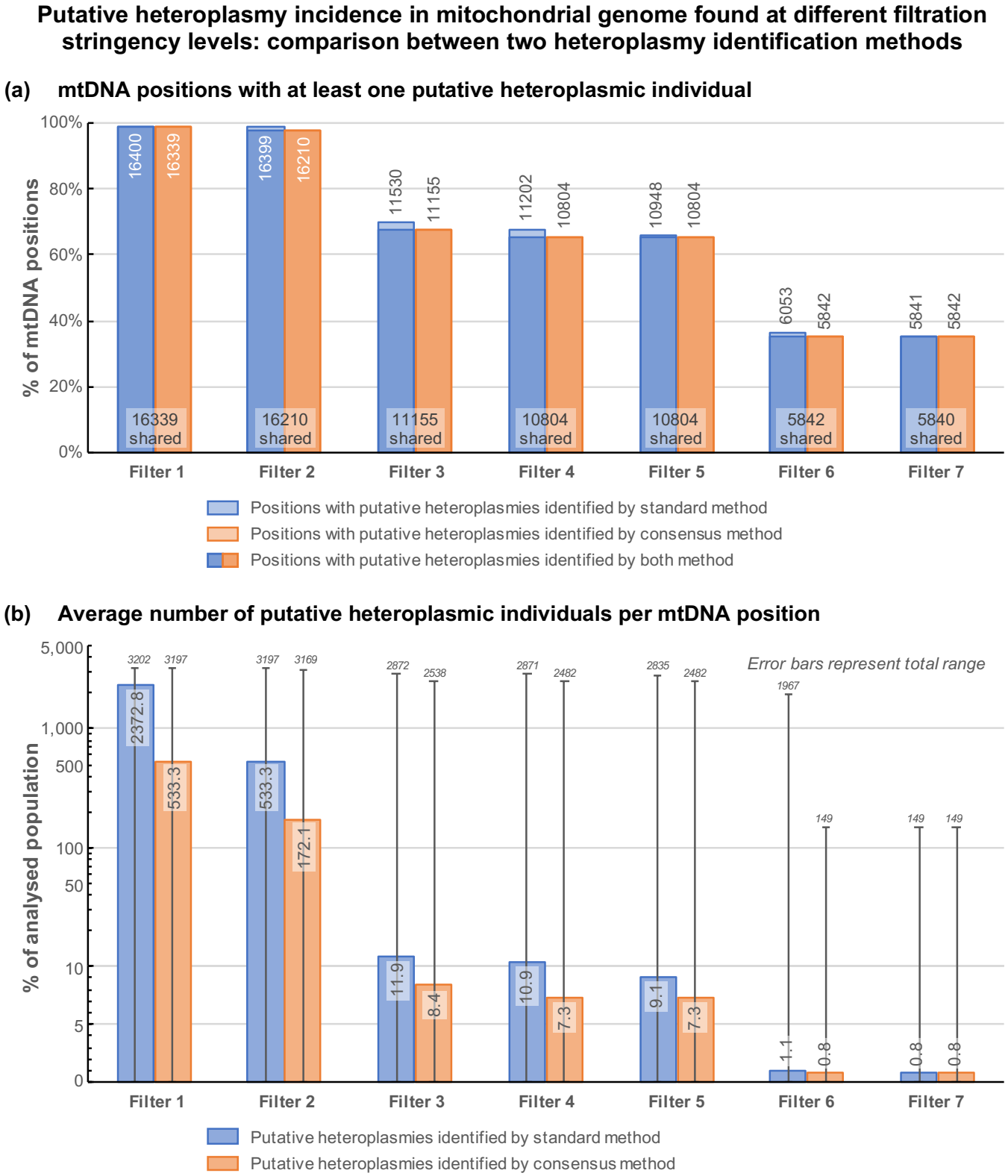
More stringent filter application reduces the number of total putative heteroplasmies detected. The biggest relative decrease takes place when multiple reads requirement in both sequencing directions is imposed on alleles (filter 3). The second most impactful allele filtration criteria is the minimum required allele frequency increased to 1% overall, with the putative heteroplasmy count being reduced 8.4 times for standard method and 9.0 times for consensus method. If we consider the unique putative heteroplasmies identified by standard method only as more likely to be rising from sequencing artefacts compared to those corroborated by consensus method, the changes in the unique putative heteroplasmy proportion between different filtration stringency levels are informative about the effectiveness of different filter criteria on specifically removing artefact component of the data. Based on this, data in **Supplementary figure 5.2.1 S6** demonstrates that the more robust strategy for minimising likely artefact contribution should be to apply minimum allele frequency requirements to read data from each sequencing direction separately (filter levels 5 and 7), as opposed to across all reads in both directions (filter levels 4 and 6),

Note on filter levels 4 and 6 with consensus method: For consensus method criteria applied at filters steps 4 and 6 are functionally identical to those of steps 5 and 7 respectively, as the consensus allele frequencies in both sequencing directions are by default the same when excess reads have been excluded.

Number of mtDNA positions with detected heteroplasmies

Supplementary figure 5.2.1 S7 shows the number of mtDNA positions at which at least one putative heteroplasmy was detected in the population (**panel A**), and on average how many individuals with putative heteroplasmies were found per position (**panel B**) across the 7 filtration stringency levels with standard and consensus methods.

Supplementary figure 5.2.1 S7:Heteroplasmy incidence in mitochondrial genome found at different filtration stringency levels: comparison between two heteroplasmy identification methods



Supplementary figure 5.2.1 S8 compares central tendencies and dispersion measures of putative heteroplasmy counts found per mtDNA position across the mitochondrial genome by standard versus consensus methods.

Supplementary figure 5.2.1 S8: Overview of heteroplasmic sample counts per position found using two heteroplasmy identification methods across different filtration stringency levels

Overview of putative heteroplasmic sample counts per position found using two heteroplasmy identification methods across different filtration stringency levels

(a) Putative heteroplasmies per position identified using standard method

	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Mean	2372.8	533.3	11.9	10.9	9.1	1.1	0.8
Median	2396	434	2	2	1	0	0
SD	529.6	401.1	88.2	86.2	75.8	18.4	2.2
IQR	616	305	6	6	5	1	1
Range	3202	3197	2872	2871	2835	1967	149

(b) Putative heteroplasmies per position identified using consensus frequency based method

	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Mean	533.3	172.1	8.4	7.3	7.3	0.8	0.8
Median	434	92	2	1	1	0	0
SD	401.1	291.6	61.3	57.7	57.7	2.2	2.2
IQR	305	110	6	5	5	1	1
Range	3197	3169	2539	2482	2482	149	149

(c) Difference between putative heteroplasmy counts per position between two methods

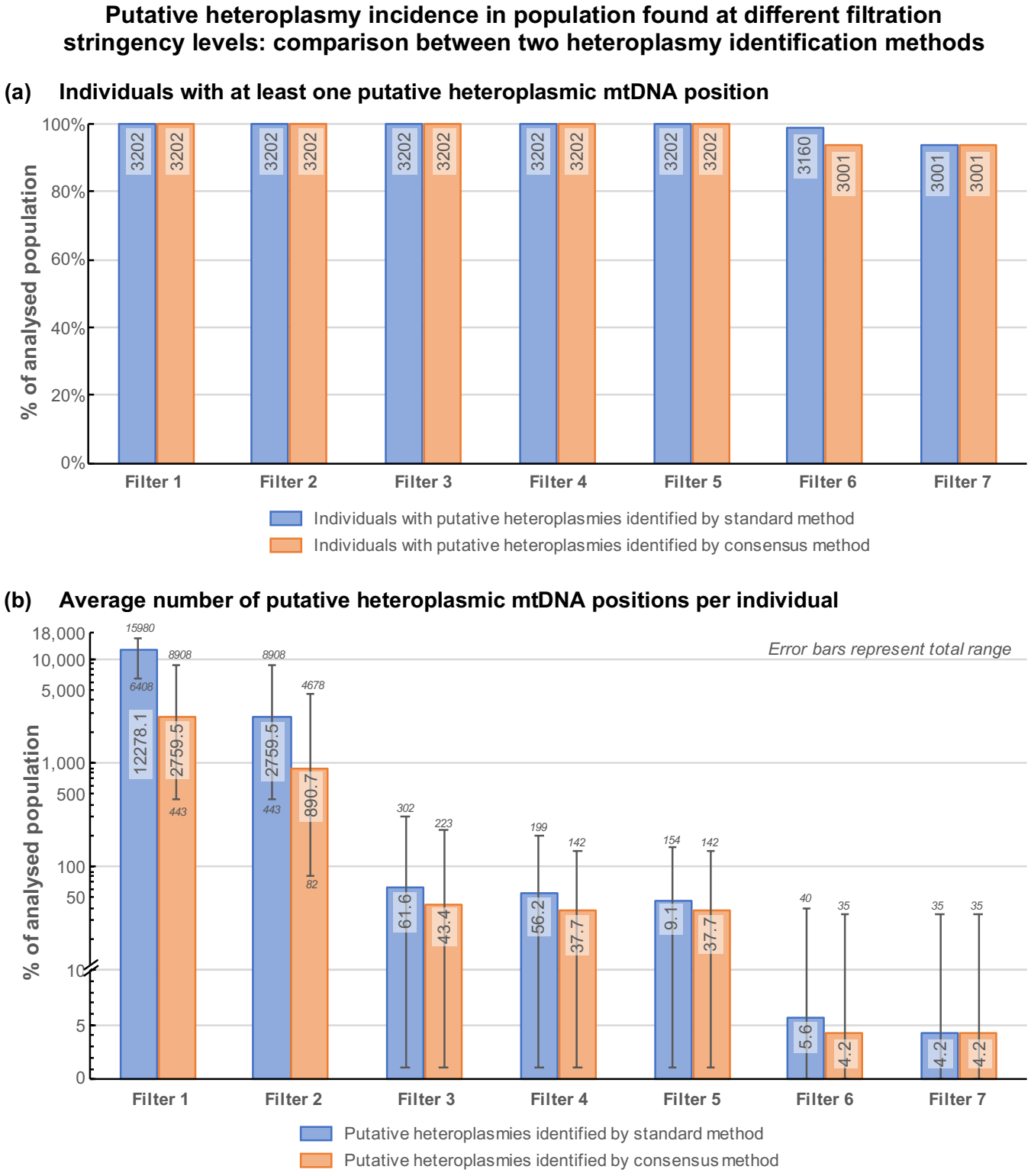
	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Δ mean	1839.5	361.2	3.5	3.6	1.8	0.3	0
Δ median	1875	328	0	0	0	0	0
Δ SD	395.0	186.1	42.7	42.9	30.1	18.3	0.1
Δ IQR	318	197	1	1	0	0	0
Δ range	3084	2023	1953	1953	1472	1965	6

SD – standard deviation; IQR – interquartile range; Δ – difference between methods, calculated by subtracting putative heteroplasmies found by consensus method minus from heteroplasmies found by consensus method

Number of individuals with detected heteroplasmies

Supplementary figure 5.2.1 S9 shows the number of individuals with at least one putative heteroplasmy found (**panel A**), and on average how many positions with putative heteroplasmies were found per individual (**panel B**) across the 7 filtration stringency levels with standard and consensus methods.

Supplementary figure 5.2.1 S9: Heteroplasmy incidence in population found at different filtration stringency levels: comparison between two heteroplasmy identification methods



Supplementary figure 5.2.1 S10 compares central tendencies and dispersion measures of the putative heteroplasmy counts observed per sample by standard versus consensus methods

Supplementary figure 5.2.1 S10: Overview of heteroplasmic position counts per sample found using two heteroplasmy identification methods across different filtration stringency levels

Overview of putative heteroplasmic position counts per sample found using two heteroplasmy identification methods across different filtration stringency levels

(a) Putative heteroplasmies identified per sample using standard method

	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Mean	12278.1	2759.5	61.6	56.2	47.3	5.6	4.2
Median	12311.5	2530	56	52	44	5	4
SD	1502	1161.9	31.6	25.7	20.9	3.3	3
IQR	2186.8	1485.5	38	32	26	4	4
Range	9572	8465	301	198	153	40	35

(b) Putative heteroplasmies identified per sample using consensus frequency based method

	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Mean	2759.5	890.7	43.4	37.7	37.7	4.2	4.2
Median	2530	760.5	38	34	34	4	4
SD	1161.9	507.1	24.6	19	19	3	3
IQR	1485.5	579	29	23	23	4	4
Range	8465	4596	222	141	141	35	35

(c) Difference between putative heteroplasmy counts per sample between two methods

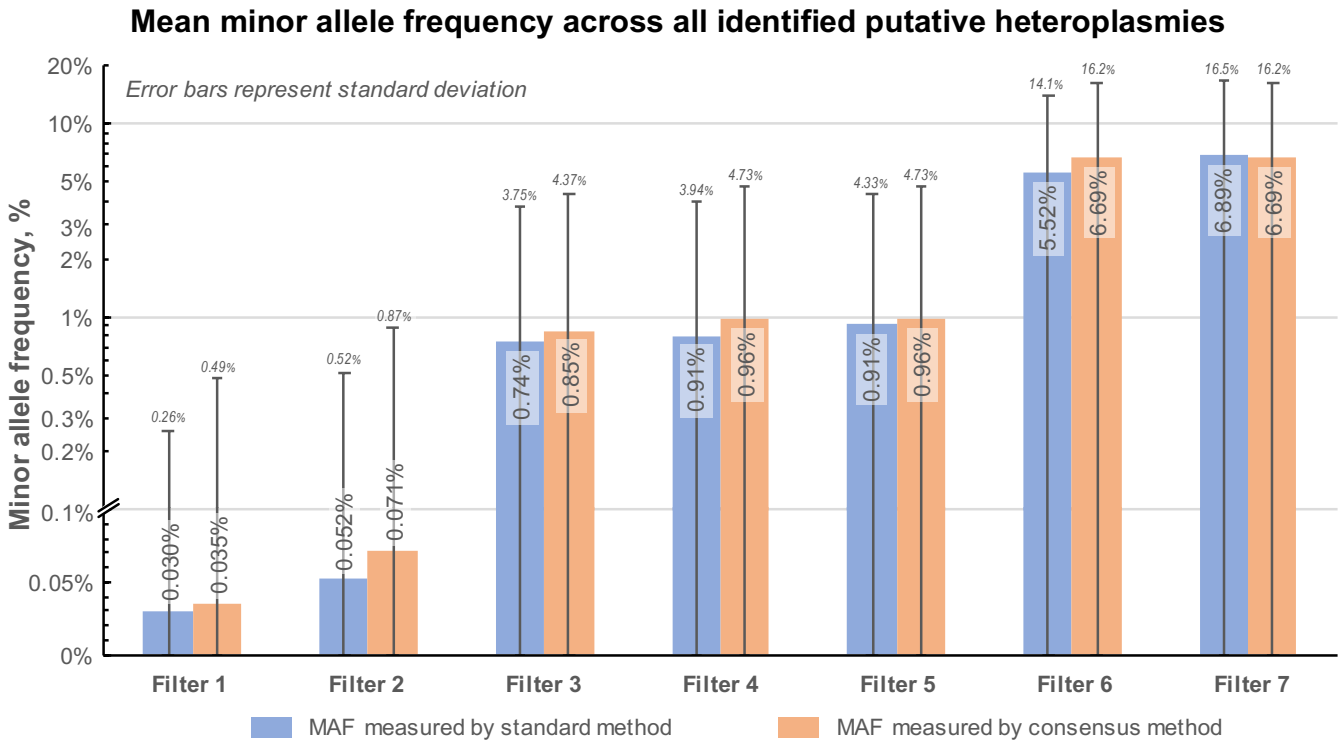
	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Δ mean	9518.6	1868.8	18.2	18.5	9.6	1.4	0
Δ median	9661.5	1762	16	17	9	1	0
Δ SD	545.7	676.2	9.3	9.6	4.9	1.1	0.1
Δ IQR	678	914	11	12	6	1	0
Δ range	4961	4215	79	81	36	17	2

SD – standard deviation; IQR – interquartile range; Δ – difference between methods, calculated by subtracting putative heteroplasmies found by consensus method minus from heteroplasmies found by consensus method

Average MAF across putative heteroplasmies

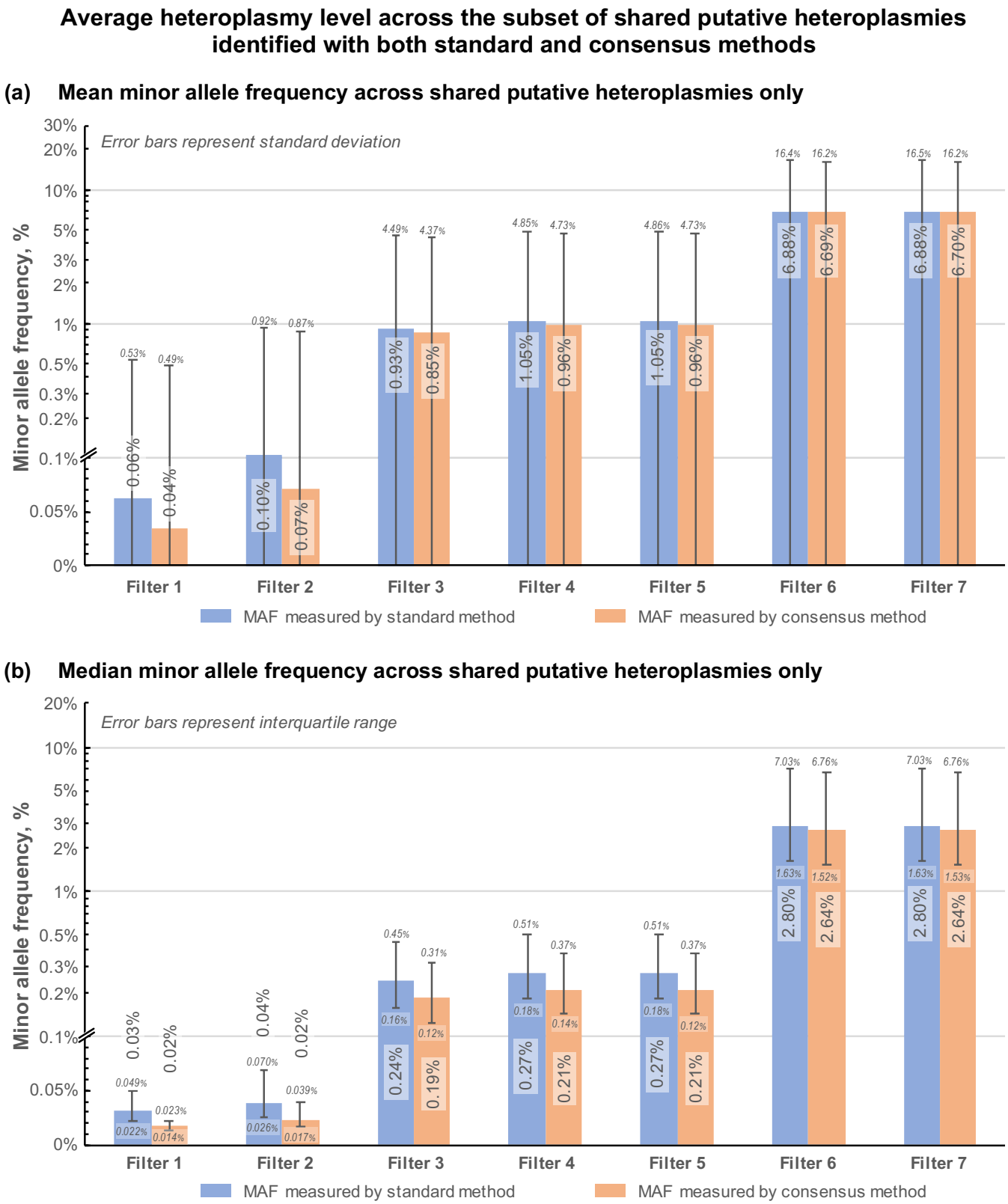
Supplementary figure 5.2.2 S1 visualises the mean allele frequency levels detected per site across all analysed putative heteroplasmies found at each filtration stringency level.

Supplementary figure 5.2.2 S1: Mean minor allele frequency across all identified putative heteroplasmies



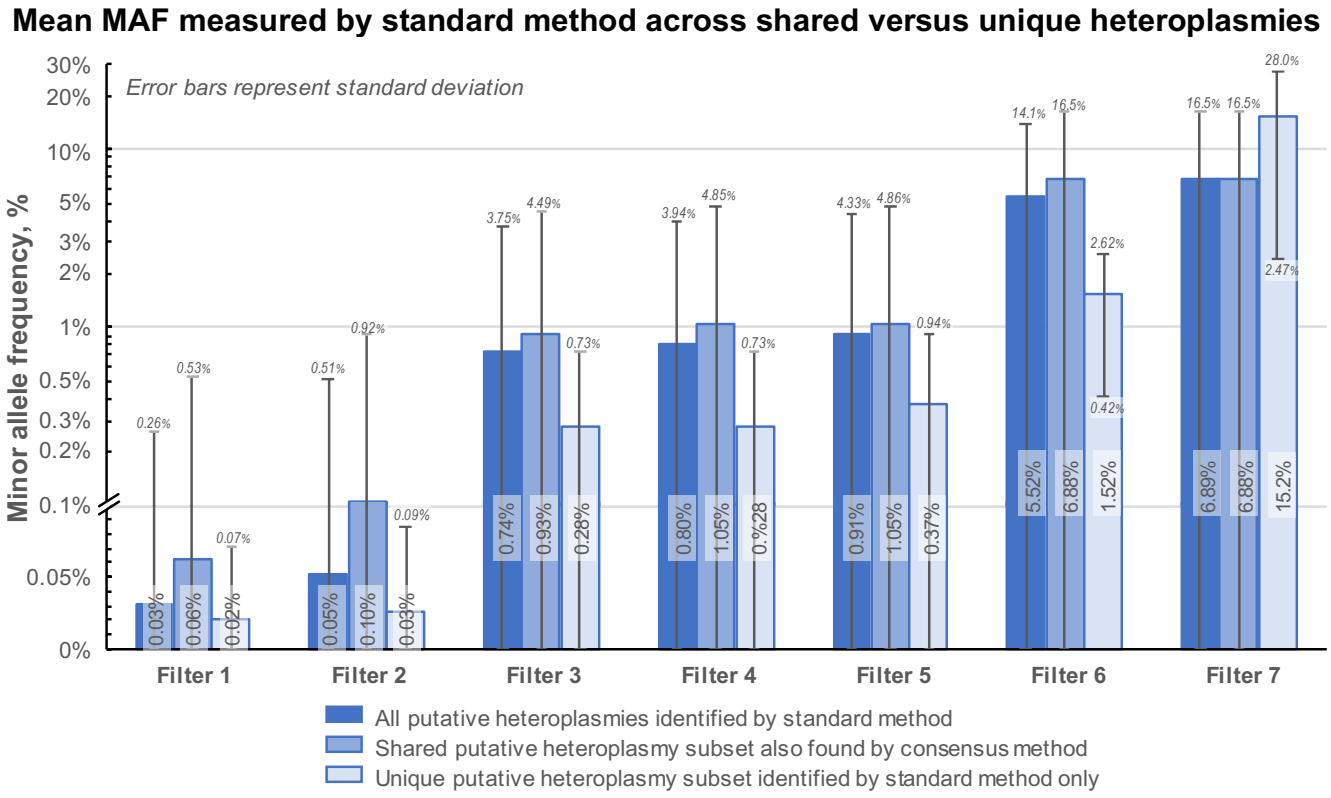
Supplementary figure 5.2.2 S2 shows the mean and median minor allele frequency levels detected across the shared putative heteroplasmy subsets identified with both the standard and the consensus methods.

Supplementary figure 5.2.2 S2: Average heteroplasmy level across the subset of shared putative heteroplasmies identified with both standard and consensus methods



Supplementary figure 5.2.2 S3 compares the mean minor allele frequency measured using standard method across putative heteroplasmies corroborated by consensus versus those uniquely identified to standard method.

Supplementary figure 5.2.2 S3: Mean MAF measured by standard method across shared vs unique heteroplasmies



Supplementary figure 5.2.2 S4 lists unique heteroplasmies at filter level 7 that were identified by only one of the two methods.

Supplementary figure 5.2.2 S4: MT genome positions with unique putative heteroplasmies found at filter 7

Mitochondrial genome positions with unique putative heteroplasmies found at filter 7

(a) mtDNA positions with unique putative heteroplasmies detected by standard method

7 position	rCRS base	Total samples analysed	Putative heteroplasmies detected		
			Total	Shared	Unique
146	T	458 *(14.3%)*	3 *(0.66%)*	2 *(0.44%)*	1 *(0.22%)*
152	T	832 *(26.0%)*	11 *(1.32%)*	10 *(1.20%)*	1 *(0.12%)*
309	C	606 *(18.9%)*	2 *(0.33%)*	0 *(0%)*	2 *(0.33%)*
310	T	16 *(0.5%)*	10 *(62.50%)*	5 *(31.25%)*	5 *(31.25%)*
6069	G	3199 *(99.9%)*	3 *(0.09%)*	2 *(0.06%)*	1 *(0.03%)*
16192	C	2695 *(84.2%)*	41 *(1.52%)*	39 *(1.45%)*	2 *(0.07%)*
16215	A	3197 *(99.8%)*	5 *(0.16%)*	4 *(0.13%)*	1 *(0.03%)*

(b) mtNDA positions with unique putative heteroplasmies detected by consensus method

13 position	rCRS base	Total samples analysed	Putative heteroplasmies detected		
			Total	Shared	Unique
185	G	3067 *(95.8%)*	5 *(0.16%)*	4 *(0.13%)*	1 *(0.03%)*
1320	G	3202 *(100%)*	2 *(0.06%)*	1 *(0.03%)*	1 *(0.03%)*
3605	T	3202 *(100%)*	2 *(0.06%)*	1 *(0.03%)*	1 *(0.03%)*
4551	T	3150 *(98.4%)*	1 *(0.03%)*	0 *(0%)*	1 *(0.03%)*
5296	T	3202 *(100%)*	3 *(0.09%)*	2 *(0.06%)*	1 *(0.03%)*
5919	T	3202 *(100%)*	1 *(0.03%)*	0 *(0%)*	1 *(0.03%)*
6054	G	3200 *(99.9%)*	5 *(0.16%)*	4 *(0.13%)*	1 *(0.03%)*
7161	G	2977 *(93.0%)*	6 *(0.20%)*	5 *(0.17%)*	1 *(0.03%)*
7943	T	3202 *(100%)*	3 *(0.09%)*	2 *(0.06%)*	1 *(0.03%)*
7986	G	3202 *(100%)*	4 *(0.12%)*	3 *(0.09%)*	1 *(0.03%)*
9862	T	3202 *(100%)*	3 *(0.09%)*	2 *(0.06%)*	1 *(0.03%)*
12172	A	3202 *(100%)*	10 *(0.31%)*	9 *(0.28%)*	1 *(0.03%)*
14040	G	3202 *(100%)*	5 *(0.16%)*	4 *(0.12%)*	1 *(0.03%)*

A3.1.2 Putative heteroplasmy versus base type

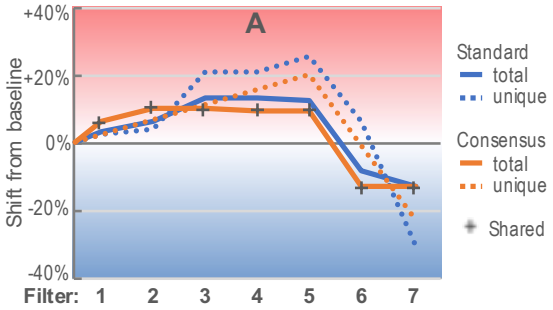
Supplementary figure 5.2.3 S1: Identified heteroplasmy base frequency shift away from MT genome baseline

Overview of base frequency shifts away from MT genome overall base frequencies in the subsets of putative heteroplasmies identified at different filtration stringency levels

(a) A frequency shift in putative heteroplasmy subsets away from 30.9% genome baseline frequency

Site subset	Filter stringency level 1	2	3	4	5	6	7
Standard total	+3.0%	+6.1%	+13.2%	+13.3%	+12.8%	-8.3%	-13.3%
Standard unique	+2.1%	+4.0%	+20.9%	+21.1%	+25.8%	+6.6%	-30.9%
Consensus total	+6.1%	+10.4%	+9.9%	+9.5%	+9.5%	-13.3%	-13.3%
Consensus unique	n/a	n/a	n/a	n/a	+20.0%	n/a	-23.2%
Shared	+6.1%	+10.4%	+9.9%	+9.5%	+9.5%	-13.3%	-13.3%

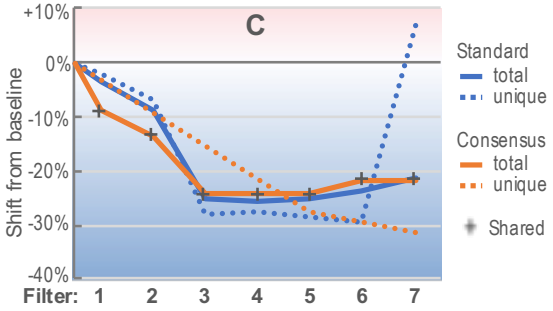
Visualisation of A frequency shift



(b) C frequency shift in putative heteroplasmy subsets away from 31.3% genome baseline frequency

Site subset	Filter stringency level 1	2	3	4	5	6	7
Standard total	-3.6%	-8.9%	-25.3%	-25.5%	-25.2%	-23.6%	-21.5%
Standard unique	-2.1%	-6.7%	-27.9%	-27.8%	-28.6%	-29.5%	+7.2%
Consensus total	-8.9%	-13.5%	-24.3%	-24.4%	-24.4%	-21.6%	-21.6%
Consensus unique	n/a	n/a	n/a	n/a	-27.8%	n/a	-31.3%
Shared	-8.9%	-13.5%	-24.2%	-24.4%	-24.4%	-21.6%	-21.6%

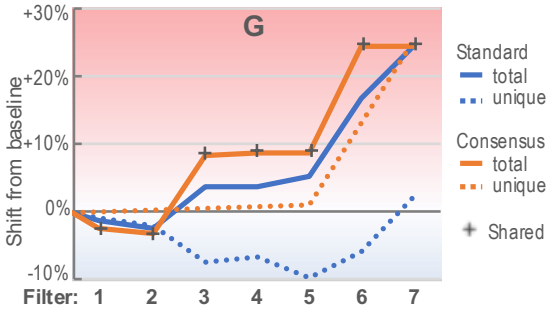
Visualisation of C frequency shift



(c) G frequency shift in putative heteroplasmy subsets away from 13.1% genome baseline frequency

Site subset	Filter stringency level 1	2	3	4	5	6	7
Standard total	-1.3%	-2.6%	+3.8%	+3.8%	+5.1%	+16.9%	+24.5%
Standard unique	-0.9%	-2.2%	-7.3%	-6.6%	-9.6%	-5.7%	+2.3%
Consensus total	-2.6%	-3.3%	+8.5%	+8.9%	+8.9%	+24.6%	+24.6%
Consensus unique	n/a	n/a	n/a	n/a	0.9%	n/a	+25.4%
Shared	-2.6%	-3.3%	+8.5%	+8.9%	+8.9%	+24.6%	+24.6%

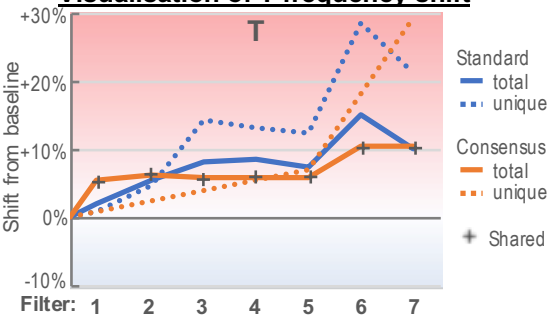
Visualisation of G frequency shift



(d) T frequency shift in putative heteroplasmy subsets away from 24.7% genome baseline frequency

Site subset	Filter stringency level 1	2	3	4	5	6	7
Standard total	+1.9%	+5.4%	+8.3%	+8.4%	+7.3%	+14.9%	+10.3%
Standard unique	+0.9%	+4.9%	+14.2%	+13.3%	+12.4%	+28.7%	+21.4%
Consensus total	+5.4%	+6.4%	+5.8%	+6.0%	+6.0%	+10.3%	+10.3%
Consensus unique	n/a	n/a	n/a	n/a	+6.9%	n/a	+29.1%
Shared	+5.4%	+6.4%	+5.8%	+6.0%	+6.0%	+10.3%	+10.3%

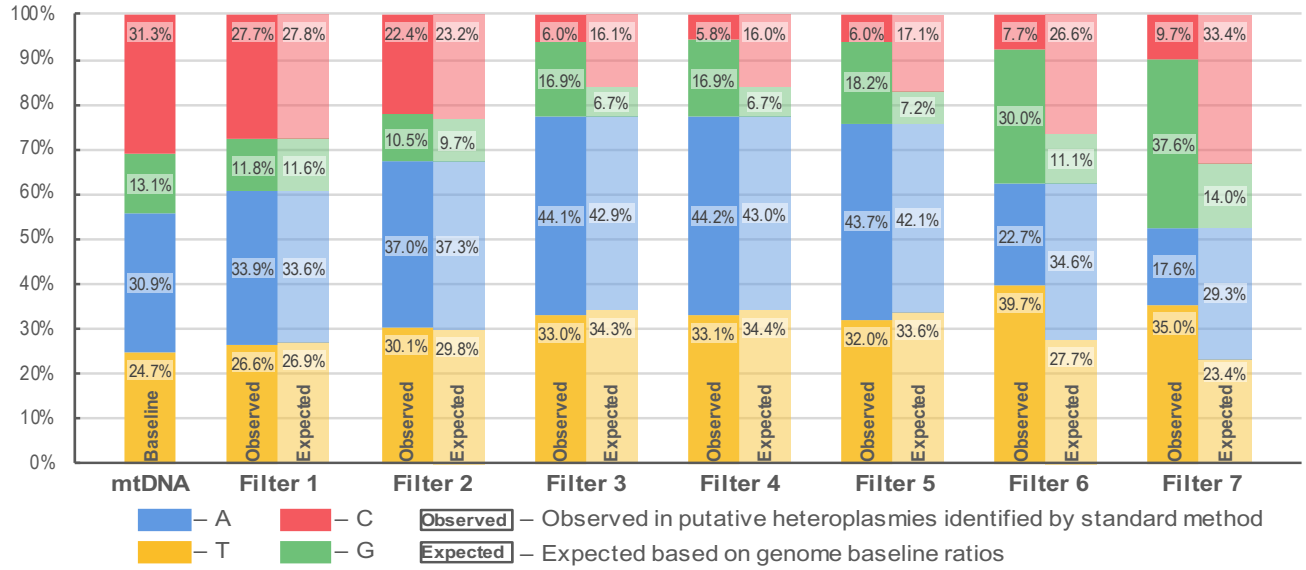
Visualisation of T frequency shift



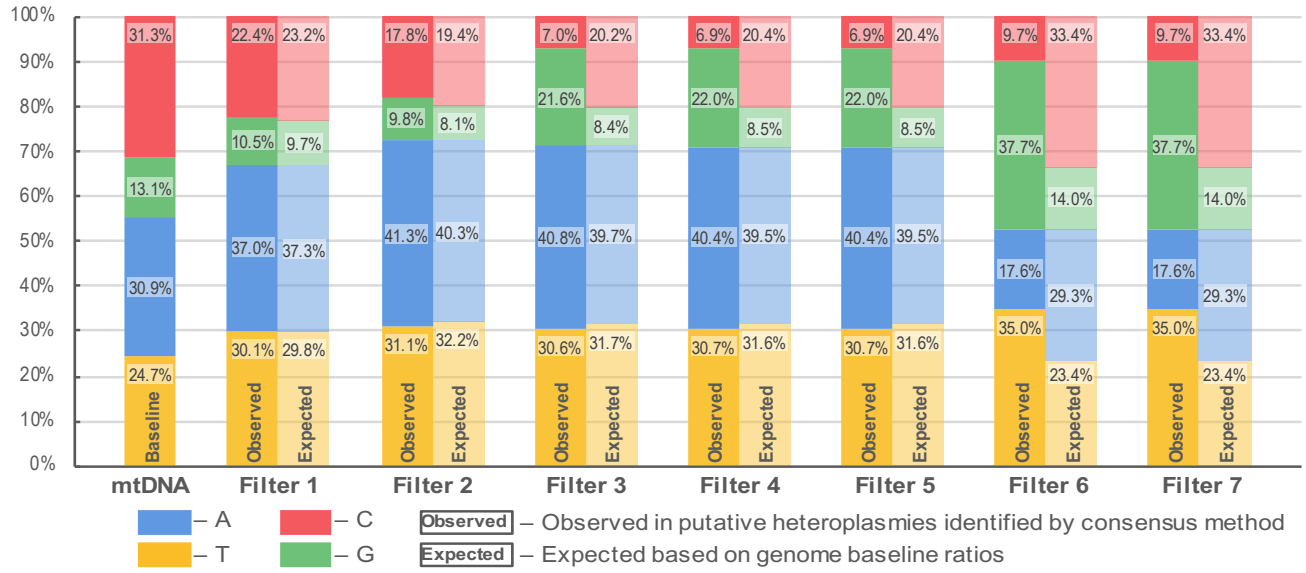
Supplementary figure 5.2.3 S2: Comparison between observed and expected heteroplasmy base pair frequencies

Comparison between observed and expected base pair frequencies in putative heteroplasmy subsets identified across different filtration stringency levels

(a) Observed and expected base frequencies in putative heteroplasmy set found by standard method



(b) Observed and expected base frequencies in putative heteroplasmy set found by consensus method



A3.1.3 Modelling CSE impact on detected heteroplasmy major allele base composition

Any genome site contains a base pair in one of two possible orientations, with one base on L strand and its complementary base on H strand. Therefore any process that is biasing a specific base towards being identified as a putative heteroplasmy entirely based on the identify of that one base, affects the entire base pair.

From any mechanism that increases the probability of a specific base type being identified as a putative heteroplasmy entirely based on the identity of that one base, we expect an equal effect on the complementary base type, and the ratio between the complementary bases to remain stable – i.e. for every G flagged up as putative heteroplasmy there should be on average 2.4 Cs also flagged up, and for every T there should also be on average 1.25 As.

But this is only if the mechanism modifying the likelihood of a base being flagged as putative heteroplasmy is dependent only on the base identity itself. Situation changes if the adjacent sequence context also functions as a modifier for the likelihood of a site being identified as putative heteroplasmy. And this can be either by changing its likelihood of being affected by true or by an error that doesn't get filtered out, both in principle would have the same effect.

With the different base composition of mtDNA strands where L strand is richer in As and Cs and H strand is richer in Ts and Gs, the likelihood of any base existing in a particular adjacent sequence environment differs between H and L strands. Thus if a particular sequence context affected how likely a specific base type was subject to either true mutation or a sequencing error, and if that context was present in different abundance on the two strands, it would affect base pairs in one orientation more than the other, biasing the ratio between two complementary base types detected as putative heteroplasmies.

Since we see the A/T and C/G ratios in our putative heteroplasmy sets diverging from those of the genome overall, we can speculate that not only base type, but also the adjacent sequence context is in fact impacting the probabilities of sites being identified as putative heteroplasmies.

To investigate the interaction between base identity and sequence context effects, we can model how different true heteroplasmy or error rates of certain sequence contexts could bias the ratios of detected putative heteroplasmies.

The most obvious pattern associated with context dependent sequencing error is a T preceded by a G rich motif, and the more Gs there are, the higher the error rate. We know the probabilities of any given T existing in different G rich sequence contexts, which allows us to calculate how such Ts being erroneously flagged up as heteroplasmies would affect the resulting T to A ratio in the identified putative heteroplasmy set.

There are total of 9218 AT base pairs in MT genome. The ratio between these two possible orientations is 1.25. When we search for occurrences of any T preceded by some variation of G rich motif (summarised in table), we see many more such occurrences when the AT pair is oriented so that T is on the H strand (A_LT_H orientation) compared to T on L strand (A_HT_L orientation). The ratio of AT pairs found in two possible orientations (A_HT_L vs A_LT_H) ranges from 4.4 for a T preceded by minimum 2 Gs within the 5 upstream positions ("[2/5G]T" motif) up to 22.9 for a T preceded by at least 4 Gs within the 5 upstream positions ("[4/5G]T" motif). For comparison, the ration between A_HT_L vs A_LT_H base pair orientations in the whole MT genome is 1.25.

Supplementary figure 5.2.4 S1: AT pairs found in MT genome in different orientations and sequence contexts

AT pairs found in mitochondrial genome in different orientations and sequence contexts

(a) All AT pairs found in mitochondrial genome

AT pair orientation	Shown in rCRS as	Instances found	% of all AT pairs	A_LT_H to A_HT_L ratio
any orientation	A or T	9218	100%	
A_LT_H	A	5124	55.6%	1.2516
A_HT_L	T	4094	44.4%	

(b) AT pairs found within specific G-rich upstream contexts

Sequence context	Instances found	% of all AT pairs	Found in each orientation A_LT_H		A_HT_L		Ratio	Strand-specific frequency % of T_L	% of T_H	Ratio
[2/3G] **T**	1410	15.3%	1200	*85.1%*	210	*14.9%*	5.71	5.1%	23.4%	4.57
GGG **T**	217	2.4%	199	*91.7%*	18	*8.3%*	11.06	0.4%	3.9%	8.83
[2/5G] **T**	3030	32.9%	2470	*81.5%*	560	*18.5%*	4.41	13.7%	48.2%	3.52
[3/5G] **T**	1091	11.8%	972	*89.1%*	119	*10.9%*	8.17	2.9%	19.0%	6.53
[4/5G] **T**	263	2.9%	252	*95.8%*	11	*4.2%*	22.91	0.3%	4.9%	18.30
[1/4G] G **T**	1372	14.9%	1168	*85.1%*	204	*14.9%*	5.73	5.0%	22.8%	4.57
[2/4G] G **T**	652	7.1%	588	*90.2%*	64	*9.8%*	9.19	1.6%	11.5%	7.34
[3/4G] G **T**	203	2.2%	193	*95.1%*	10	*4.9%*	19.30	0.2%	3.8%	15.42
[5/10G] **T**	1068	11.6%	1013	*94.9%*	55	*5.1%*	18.42	1.3%	19.8%	14.72
[4/9G] G **T**	540	5.9%	514	*95.2%*	26	*4.8%*	19.77	0.6%	10.0%	15.80

So if indeed we were expecting to see more sequencing errors at Ts preceded by Gs, AT pairs in one orientation would be affected more by sequencing errors than those in the opposite orientation. And if these sequencing errors were then mistaken as putative heteroplasmies, we would see an overrepresentation of putative heteroplasmies at Ts in A_LT_H orientation.

We previously showerdthat GGT-like motifs can be associated with highest increase in sequencing error rates. Hence we have selected this group of motifs for modelling their possible effect on skewing A/T ratio of the identified putative heteroplasmies. Frequencies of some of these motifs are summarised in the table above.

We can model differences between the likelihoods of two possible base pair orientations being flagged up as putative heteroplasmies would alter the expected ratio between the two complementary base types in the resulting identified putative heteroplasmy set. To simplify the model, we shall assume that

- True heteroplasmy rate is constant and effects all Ts equally,
- No other sequencing errors besides selected motif sites are present,
- Every site where the selected motif is present is misidentified as heteroplasmic, and
- Sequencing errors at true heteroplasmic site don't prevent identification of heteroplasmy.

Under these assumptions, we can show that the expected observed A to T ratio in the identified putative heteroplasmy set, $R^{AT}{}_{obs}$, can be expressed as the following relationship

$$R^{AT}{}_{obs} = R^{AT} \times \frac{\%het + \%err^H(1-\%het)}{\%het + \%err^L(1-\%het)}$$

where R^{AT} is the overall MT genome A to T ratio, $\%het$ is the true heteroplasmy rate, and $\%err^H$ and $\%err^L$ are the strand-specific T error rates on H and L strands respectively. For full description of the model, see materials and methods. When error rates on L and H strands are the same, the observed putative heteroplasmy A to T ratio will match that of the MT genome overall. The relationships between the variables meant that changes in H strand error rate show linear correlation to A/T ratio, with an increase in H strand error rate producing an increase in the observed A to T ratio. L strand error rate relationship to the observed A to T ratio is inverse and non-linear, with the

A to T ratio inflation effect decreasing in magnitude as L error rate rises. Changes in true heteroplasmy rate modulate the effects of strand specific error rates, with higher true heteroplasmy levels dampening the effects from the strand specific error rates and reducing the amplitude of the possible A/T ratio variation.

We are interested in whether our observed putative heteroplasmy subset A to T ratio shifts could be consistent with some GGT-like sequence motifs being disproportionately affected by sequencing errors and being misidentified as heteroplasmies. We looked at 10 motif groups with variable number of Gs required to be found upstream of a T, and calculated the putative heteroplasmy set A to T ratios we could expect to observe should such motif sites be consistently misidentified as heteroplasmies. As the true heteroplasmy level in our dataset is unknown, we calculated these expected A to T ratios for a range of possible true heteroplasmy rates. GGT-like motif groups are summarised in **supplementary figure 5.4.2 S2** which shows expected A to T ratios for each group plotted over the range of possible true heteroplasmy rates.

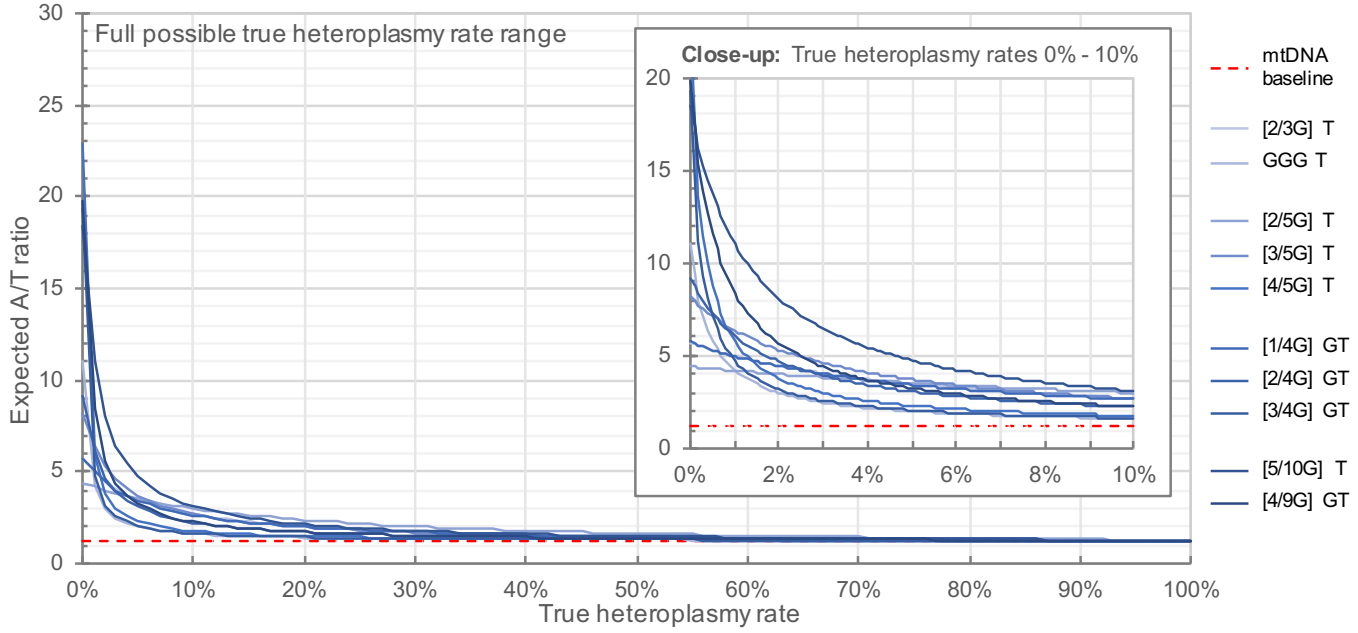
Supplementary figure 5.2.4 S2: Impact of CSEs at GGT-like motifs on the expected heteroplasmy A/T ratio

Modelled impact of context dependent sequencing errors at GGT-like motifs on the expected A/T ratio of the identified putative heteroplasmy sets

(a) Potential error prone sequences and their associated putative heteroplasmy subset A/T ratio changes

Upstream context	Count on strand H	Count on strand L	Strand-specific frequency H	Strand-specific frequency L	Strand-specific frequency Ratio	A/T ratio at different heteroplasmy rates 0.1%	1%	5%	10%	50%
[–*any*–] **T**	5124	4094	100%	100%	1	1.25	1.25	1.25	1.25	1.25
[2/3G] **T**	1200	210	23.4%	5.1%	4.57	5.63	4.98	3.45	2.66	1.47
GGG **T**	199	18	3.9%	0.4%	8.83	9.24	4.22	2.01	1.62	1.29
[2/5G] **T**	2470	560	48.2%	13.7%	3.52	4.39	4.19	3.53	2.99	1.63
[3/5G] **T**	972	119	19.0%	2.9%	6.53	7.94	6.38	3.71	2.69	1.45
[4/5G] **T**	252	11	4.9%	0.3%	18.3	17.03	5.80	2.30	1.76	1.31
[1/4G]G **T**	1168	204	22.8%	5.0%	4.57	5.64	4.97	3.43	2.64	1.46
[2/4G]G **T**	588	64	11.5%	1.6%	7.34	8.71	6.07	3.07	2.23	1.37
[3/4G]G **T**	193	10	3.8%	0.2%	15.42	14.05	4.77	2.05	1.64	1.30
[5/10G] **T**	1013	55	19.8%	1.3%	14.72	17.23	11.05	4.74	3.10	1.48
[4/9G]G **T**	514	26	10.0%	0.6%	15.8	17.25	8.40	3.25	2.25	1.37

(b) GGT-type motif associated expected A/T ratios across a range of possible true heteroplasmy rates



Supplementary figure 5.4.2 S3 lists the actual complementary base ratios observed in our dataset of different putative heteroplasmy sets identified across a range of filtration stringency levels.

Supplementary figure 5.2.4 S3: Actual observed A/T and C/G ratios in the putative heteroplasmy subsets

Actual observed A/T and C/G ratios in putative heteroplasmy subsets identified with standard and consensus methods across different filtration stringency levels

(a) A/T ratios observed in our identified putative heteroplasmy subsets across different stringency filters

Subset	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Standard total	↑1.27	↓1.23	↑1.33	↑1.33	↑1.36	↓0.57	↓0.50
Standard unique	↑1.29	↓1.18	↑1.33	↑1.37	↑1.53	↓0.70	↓ 0
Consensus total	↓1.23	↑1.33	↑1.34	↑1.31	↑1.31	↓0.50	↓0.50
Consensus unique	-	-	-	-	↑1.61	-	↓0.14
Shared	↓1.23	↑1.33	↑1.34	↑1.31	↑1.31	↓0.50	↓0.50

↑↓ – *Arrows indicates A/T ratio value change direction from the mitochondrial genome baseline A/T ratio of 1.2516*

(b) C/G ratios observed in our identified putative heteroplasmy subsets across different stringency filters

Subset	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Standard total	↓2.35	↓2.12	↓0.35	↓0.34	↓0.33	↓0.26	↓0.26
Standard unique	↑2.40	↓2.26	↓0.59	↓0.54	↓0.77	↓0.24	↑2.50
Consensus total	↓2.12	↓1.81	↓0.33	↓0.31	↓0.31	↓0.26	↓0.26
Consensus unique	-	-	-	-	↓0.25	-	↓ 0
Shared	↓2.12	↓1.81	↓0.33	↓0.31	↓0.31	↓0.26	↓0.26

↑↓ – *Arrows indicates C/G ratio value change direction from the mitochondrial genome baseline C/G ratio of 2.3887*

(c) Total number of identified putative heteroplasmic As and Ts

A + T	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Standard all	23,798,420	5,928,486	152,092	139,287	114,616	11,176	7,055
Standard unique	17,869,987	3,863,019	52,866	53,397	28,773	4,119	6
Consensus all	5,928,436	2,065,470	99,229	85,893	85,893	7,060	7,060
Consensus unique	0	0	0	0	47	0	8
Shared	5,928,436	2,065,470	99,229	85,893	85,846	7,060	7,052

(d) Total number of identified putative heteroplasmic Cs and Gs

C + G	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Standard all	15,516,009	2,907,539	45,096	40,804	36,738	6,759	6,348
Standard unique	12,608,507	2,120,915	5,341	5,940	1,884	413	7
Consensus all	2,907,499	786,621	39,752	34,861	34,861	6,343	6,343
Consensus unique	0	0	0	0	10	0	5
Shared	2,907,499	786,621	39,752	34,861	34,851	6,343	6,338

Comparing the actual observed putative heteroplasmy AT ratios to those predicted by our model, we see that the observed ratios are much smaller than those predicted, and the predicted and observed values become comparable only at very high (and therefore unlikely) true heteroplasmy levels.

It should be noted however, that the simplified model used here in predicting A/T ratios has several limitations. Not only are we assuming a 100% error rate at every potential error-prone motif site, but we are also looking at only a single mechanism in isolation, which creates a very artificial system. In reality, we should expect interaction of a range of different processes interacting and modulating one another's effects.

Nevertheless, this simplified model informs us about the underlying processes in several ways:

Firstly, we have shown that increasing true heteroplasmy rate dampens any context specific site selection effects. But this could be generalised to any process that is independent of the sequence context, and any randomness in the system should have the same dampening effect obscuring the effect of context dependent mechanisms.

While we do expect the putative heteroplasmy sets identified with lowest allele filtration stringency with standard method at filter 1 to include the largest number of misidentified sites arising through context dependent error mechanisms, this set is also most likely to contain largest number of random and non-specific errors dampening any of the context-specific effects on the observed complementary base ratios. As a result, we wouldn't expect this set to show large deviations from baseline A to T and C to G ratios.

Secondly, the number of detected putative heteroplasmies per person at highest allele filtration stringency levels is much lower than the number sites with some form of potentially error prone motifs described in previous chapters. By filter level 5 less than 50 sites per person on average are flagged as putative heteroplasmies, and this is reduced even further to just over 4 sites per person on average by filter 7. This means only a very small proportion of all possible positions with potentially error prone motifs could possibly be contributing to the detected putative heteroplasmy set in these more stringently filtered datasets in any one individual. Pooling of population data should therefore be considered in further investigation of sequencing errors effects.

Thirdly, at higher filtration stringency and especially when minimum allele frequency is required, we are likely looking at much lower true heteroplasmy rates than in the earlier stages. This in turn means any putative heteroplasmy detection bias based on the surrounding sequence context (whether it be real heteroplasmy or misidentification of errors) should have a more pronounced effect on the resulting complementary base pair ratios in these later filters compared to the earlier least stringent filtration steps.

Lastly, if the premise behind our proposed consensus frequency based heteroplasmy analysis method is valid, we should be successfully excluding larger proportion of incorrectly sequenced reads using consensus method compared to standard method. Consequently, the set of putative heteroplasmies found only by standard method but not consensus method should have a higher error to true heteroplasmy ratio compared to the set corroborated by both methods, and the resulting lower true heteroplasmy rate in this standard unique should mean reduced dampening of any context-dependent error on complementary base ratios.

Looking at our actual data at the A/T ratios, we see a very slight increase in A to T ratio in putative heteroplasmy set at filter levels 3 to 5, before the ratio shows a large drop below genome baseline for the final 2 filter levels. For the subset of sites identified by both methods highest A/T ratio reached is 1.34 at filter 3 that decreases to 1.31 for filters 4 and 5. In comparison, genome baseline ratio is 1.25. Over the subsequent two filtration steps we see some further A/T ratio increase within the subset of putative heteroplasmies unique to standard method only, reaching a peak value of 1.53 at filter 5 before a drastic drop for filter 6. While the A/T ratio remains slightly higher for standard unique subset at filter 6 compared to subset found by both methods, 0.7 compared to 0.5, this is well below the genome baseline of 1.25.

Not all subsets have enough sites to make meaningful observations about base pair orientation ratios. Figure 5c and 5d shows number of AT pairs and CG pairs in each subset. For standard method, there are very few unique sites found at filter 7 (only 6 AT pairs and 7 CG pairs). And for consensus method, any unique sites at all are found only at filter levels 5 and 7, and in both cases it is a very small number of sites (47 and 8 AT pairs and 10 and 5 CG pairs for filters 5 and 7 respectively). These putative heteroplasmy counts in these subsets are too low to speculate about their potential orientation ratio significance.

We chose to model strand-specific error rate difference impact specifically on AT pair orientation ratios because our previous work on context specific sequencing errors suggest T as the base type most affected by such errors. The high error rates observed at GGGT-type motifs in combination with large differences in G content between H and L strands makes A/T pair orientation within the putative heteroplasmy set a good test case for investigating context-dependent sequencing error effects on heteroplasmy identification.

Overall, there is little evidence of string effect from GGGT-type context specific sequencing errors on skewing the resulting A to T ratios in identified putative heteroplasmy sets. The slightly higher A to T ratio of the putative heteroplasmies unique to standard method at the intermediate filtration steps could be consistent with some small proportion of such error sites being present with standard method, but there is little to suggest a substantive impact on sites identified by both methods.

In fact much more obvious is the A/T ratio skew in the opposite direction at the two most stringent filtration levels, with the frequency of Ts far exceeding the expected levels relative to As, suggestive of some unknown mechanism biasing sites with an AT pairs in A_HT_L orientation much more strongly towards being found heteroplasmic compared to the alternative A_LT_H orientation.

Similarly, if we look at C to G ratio, we also see strong orientation bias emerging even earlier at filter 3, with C_HG_L orientation being much more likely to flagged as heteroplasmic compared to the alternative C_LG_H orientation, leading to an excess of Gs compared to Cs.

Previous work on context dependent sequencing errors does not show evidence possible sequence motifs strongly associated with incorrect sequencing of Gs or Cs, leading one to conclude that whatever mechanism leads to our observed CG pair orientation bias out putative heteroplasmy sets must be of a different origin.

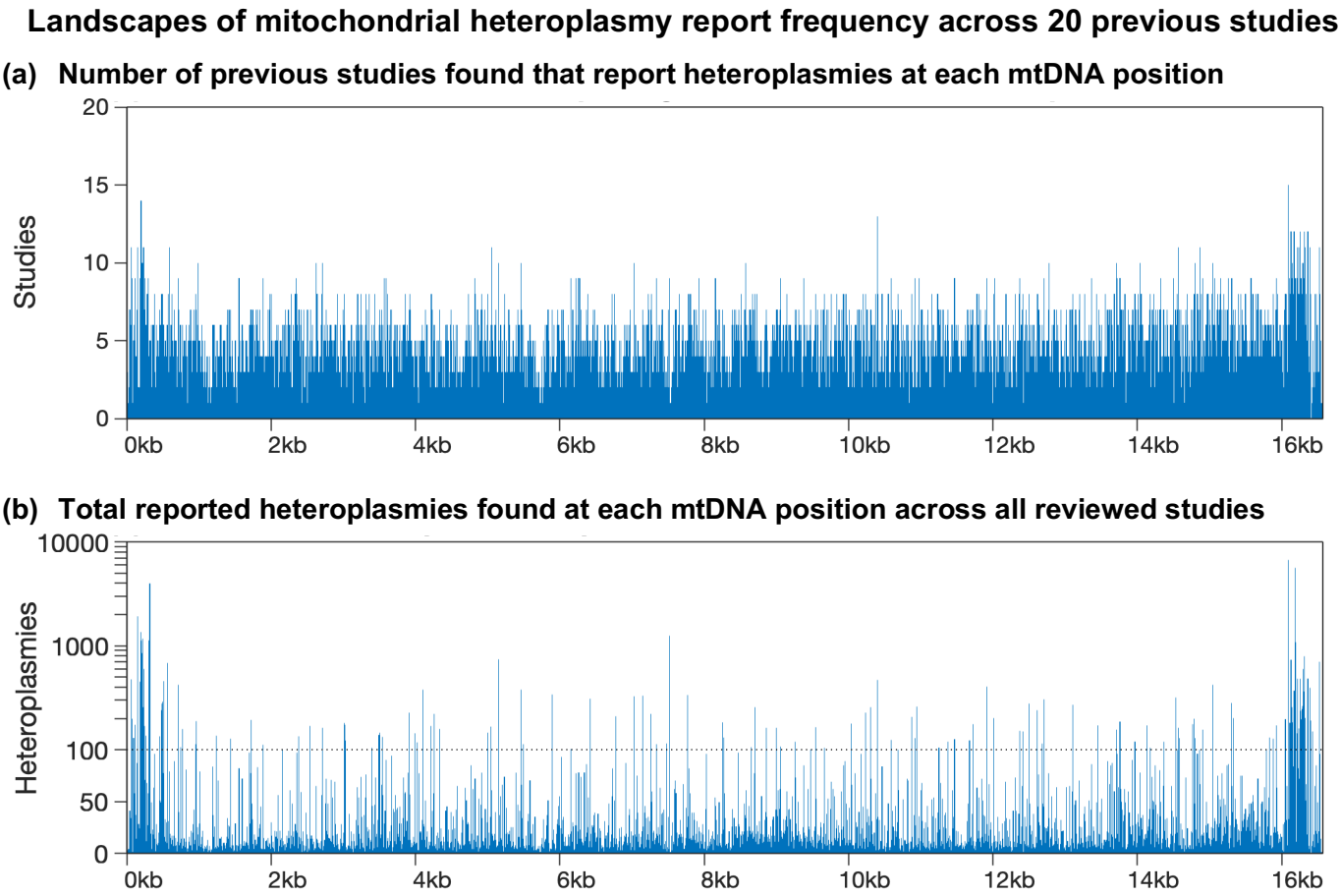
Interestingly, the subset of putative heteroplasmies showing least deviation from genome C/G ratio is the sites identified uniquely by standard method but not my consensus method, which could perhaps be an indication of the importance of the C/G pair orientation within the adjacent context for presence of true heteroplasmy.

A3.2 Comparison with literature

A3.2.1 Heteroplasmy report prevalence in literature

Supplementary figure 5.3.1 S1 visualises previous heteroplasmy report frequency in literature across the MT genome, both in terms of the number of previous studies reporting site as heteroplasmic as well as total heteroplasmies reported across all reviewed studies at any given site

Supplementary figure 5.3.1 S1: Landscapes of mitochondrial heteroplasmy reports across 20 previous studies



Base composition of heteroplasmies reported in literature previously

Supplementary figure 5.3.2 S1 visualises the relative position base type frequency change from mtDNA baseline among subsets of MT genome positions with variable heteroplasmy report frequency in the literature. **Panel A** compares base type relative frequency shift between sets of positions with and without any known heteroplasmies. **Panels B and C** show base frequency shift for groups for positions based on reported heteroplasmy frequency, either by number of studies or by number of samples. **Panel D** provides a comparison with base type frequency shift observed in our identified sets of most stringently filtered putative heteroplasmies found by consensus frequency based method at minimum MAF thresholds of 0.1% and 1%.

Supplementary figure 5.3.2 S1: Relative base type frequency deviation from mitochondrial genome baseline frequencies found in position subsets with different numbers of previously reported heteroplasmies

Relative base type frequency deviation from mitochondrial genome baseline frequencies found in position subsets with different numbers of previously reported heteroplasmies

(a) Positions with and without known heteroplasmies previously reported in literature

Positions reported in literature as	Positions	Base type relative frequency deviation from mtDNA baseline A	C	G	T
Heteroplasmic	11,180	-1.84%	-3.35%	+2.79%	+2.40%
Not heteroplasmic	5,388	+3.82%	+6.94%	-5.78%	-4.98%

(b) Known heteroplasmies grouped by number of studies position has been reported in

Group	Papers	Positions	Base type relative frequency change A	C	G	T
Unique	1	4,011	+1.66%	+5.40%	-3.37%	-3.69%
Rare	2 to 3	4,346	-1.61%	-1.61%	+0.69%	+2.53%
Intermediate	4 to 5	1,783	-6.36%	-17.08%	+10.69%	+12.75%
Common	6 to 9	992	-8.25%	-21.39%	+22.09%	+7.55%
Very common	10 plus	48	-14.26%	-8.35%	+13.99%	+8.62%

(c) Known heteroplasmies grouped by number of individual samples position has been reported in

Group	Samples	Positions	Base type relative frequency change A	C	G	T
Unique	1	2,751	-0.03%	+3.41%	-2.04%	-1.34%
Very rare	2 to 9	6,016	-1.24%	-0.52%	+1.09%	+0.67%
Rare	10 to 99	2,207	-5.28%	-19.26%	+12.60%	+11.95%
Intermediate	100 to 999	193	-6.06%	-3.81%	+12.30%	-2.43%
Most common	1000 plus	13	-15.54%	-31.27%	+2.29%	+44.52%

(d) Putative heteroplasmies identified in this study with MAF ≥0.1% and MAF ≥1%

Putative heteroplasmies identified by consensus method	% of sites analysed	Base type relative frequency change A	C	G	T
Filter 5: MAF ≥0.1%	0.23%	+9.46%	-24.38%	+8.89%	+6.03%
Filter 7: MAF ≥1%	0.026%	-13.29%	-21.61%	24.57%	10.33%

A3.2.2 Comparison between heteroplasmies reported in this study with those reported in literature

Supplementary figure 5.3.3 S1 addresses the overlap between heteroplasmies identified in the present study using the standard method and those reported in the literature previously.

Supplementary figure 5.3.3 S1: Overlap between putative heteroplasmies identified using standard method and the sets of heteroplasmic mtDNA positions reported in literature previously

Overlap between putative heteroplasmies identified using standard method and the sets of heteroplasmic mtDNA positions reported in literature previously

(a) mtDNA positions with and without putative heteroplasmies found using standard method

Standard method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Positions with one or more putative heteroplasmies	16,400 *(99.98%)*	16,339 *(99.61%)*	11,530 *(70.29%)*	11,202 *(68.29%)*	10,948 *(66.74%)*	6,053 *(36.90%)*	5,841 *(35.61%)*
Positions without any putative heteroplasmies	3 *(0.02%)*	64 *0.39%)*	4,873 *(29.71%)*	5,201 *(31.71%)*	5,455 *(33.26%)*	10,350 *(63.10%)*	10,562 *(64.39%)*

Percentages indicate fraction of the total 16,403 mtDNA positions analysed in this study

(b) Putative heteroplasmies found at known heteroplasmic positions reported in literature previously

Standard method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Number of positions	11,095	11,063	8,786	8,608	8,457	5,221	5,066
% of all putative heteroplasmies	67.65%	67.71%	76.20%	76.84%	77.25%	86.25%	86.73%
% of all previously reported sites	99.97%	99.68%	79.17%	77.56%	76.20%	47.04%	45.65%

(c) Putative heteroplasmies found at novel positions not previously reported in literature

Standard method	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Number of positions	5,305	5,276	2,744	2,594	2,491	832	775
% of all putative heteroplasmies	32.35%	32.29%	23.80%	23.16%	22.75%	13.75%	13.27%
% of all previously unknown sites	100%	99.45%	51.72%	48.90%	46.96%	15.68%	14.61%

(d) Representation of sites with variable previous heteroplasmy report frequency in literature among positions identified as heteroplasmic using standard method in this study

Group of sites from literature	Reported in studies	Percentage of each group found heteroplasmic in this study						
		Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
Unique	1	99.95%	99.55%	63.95%	61.44%	59.47%	26.07%	24.79%
Rare	2 to 3	100%	99.86%	81.12%	79.57%	78.20%	45.28%	43.78%
Intermediate	4 to 5	99.94%	99.61%	96.79%	96.23%	95.55%	73.87%	72.02%
Common	6 to 9	100%	99.59%	99.29%	99.19%	98.98%	88.62%	88.11%
Very common	10 plus	100%	100%	97.83%	97.83%	97.83%	97.83%	95.65%

Supplementary figure 5.3.3 S2 addresses the correlation between heteroplasmies identified using standard method and the previous heteroplasmy report frequency in literature.

Supplementary figure 5.3.3 S2: Correlation between heteroplasmy counts reported in previous literature with the heteroplasmic sample fraction identified in this study using standard method

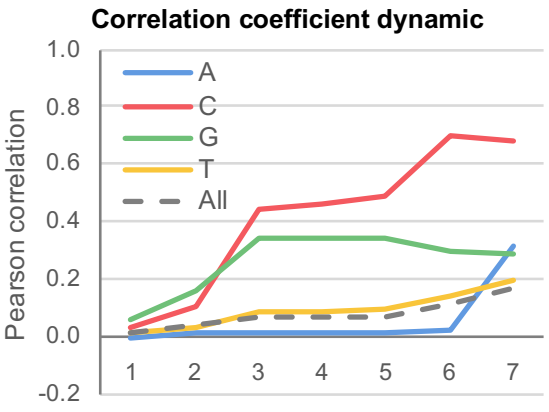
Correlation between heteroplasmy counts reported in previous literature with the heteroplasmic sample fraction identified in this study using standard method

(a) Break-down by heteroplasmy report frequency in previous literature

Group of sites from literature	Number of studies	Pearson correlation coefficient at different allele filtration stringencies						
		Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6	Filter 7
All	**any**	0.019	0.039	0.073	0.071	0.073	0.114	0.173
Unique	1	0.017	0.050	0.079	0.081	0.052	0.002	0.005
Rare	2 to 3	-0.021	-0.015	0.004	0.004	0.004	0.007	0.016
Intermediate	4 to 5	0.017	0.002	0.087	0.089	0.103	0.152	0.215
Common	6 to 9	0.027	0.046	0.043	0.041	0.043	0.233	0.242
Very common	10 plus	-0.03	0.016	0.251	0.275	0.345	0.880	0.888

(b) Break-down by heteroplasmic position base type

Filter level	Correlation by base type				Overall correlation
	A	C	G	T	
Filter 1	-0.003	0.03	0.06	0.02	0.02
Filter 2	0.02	0.10	0.16	0.03	0.04
Filter 3	0.02	0.44	0.34	0.09	0.07
Filter 4	0.02	0.46	0.35	0.09	0.07
Filter 5	0.02	0.49	0.35	0.10	0.07
Filter 6	0.03	0.70	0.30	0.14	0.11
Filter 7	0.31	0.69	0.29	0.19	0.17



Supplementary figure 5.3.3 S3 compares how well heteroplasmy landscape identified in the present study matched the findings from some of the largest previous studies individually.

Supplementary figure 5.3.3 S3: Correlation between heteroplasmy counts between 6 largest studies and this study

Correlation found between heteroplasmic sample counts between 6 largest studies across heteroplasmic positions shared between pairs of studies and with putative heteroplasmy fractions identified at MAF thresholds of 0.1%and 1% in this study

(a) Pearson correlation between heteroplasmic sample fractions reported at shared positions

Bolze 2020 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	**8413**	1	1	1	1	1
Laricchia 2022	4253	0.11	0.02	0.15	0.33	0.71
Calabrese 2022	1952	0.07	0.16	0.40	0.09	0.04
Wei 2019	963	0.40	0.42	0.49	0.23	0.43
Calabrese 2014*	2665	0.10	0.11	0.16	0.30	0.10
Ding 2015	594	0.31	0.18	0.40	0.12	0.52
Standard F5	6287	0.04	0.04	0.39	0.17	0.03
Standard F7	3606	0.07	0.07	0.38	0.17	0.06
Consensus F5	6223	0.04	0.05	0.33	0.16	0.04
Consensus F7	3606	0.10	0.09	0.39	0.17	0.08

Laricchia 2022 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	4253	0.11	0.02	0.15	0.33	0.71
Laricchia 2022	**5150**	1	1	1	1	1
Calabrese 2022	1438	0.01	0.004	0.06	0.18	0.03
Wei 2019	765	0.11	0.02	0.12	0.70	0.68
Calabrese 2014*	2061	0.18	0.40	0.10	0.55	0.33
Ding 2015	496	0.10	0.05	0.13	0.23	0.78
Standard F5	4294	0.02	0.07	0.26	0.50	0.07
Standard F7	2753	0.03	0.09	0.73	0.44	0.09
Consensus F5	4257	0.02	0.08	0.23	0.45	0.08
Consensus F7	2752	0.04	0.12	0.71	0.44	0.12

Calabrese 2022 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	1952	0.07	0.16	0.40	0.09	0.05
Laricchia 2022	1438	0.01	0.004	0.06	0.18	0.03
Calabrese 2022	**2334**	1	1	1	1	1
Wei 2019	379	0.08	0.13	0.22	0.06	0.03
Calabrese 2014*	1037	0.06	0.05	0.04	0.30	0.03
Ding 2015	227	0.06	0.05	0.15	0.03	0.02
Standard F5	1969	0.03	0.02	0.11	0.18	0.02
Standard F7	1296	0.01	0.002	0.21	0.11	0.002
Consensus F5	1952	0.03	0.02	0.09	0.18	0.02
Consensus F7	1296	0.02	0.00	0.21	0.12	0.003

Wei 2019 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	963	0.40	0.42	0.49	0.23	0.44
Laricchia 2022	765	0.11	0.02	0.12	0.70	0.68
Calabrese 2022	379	0.08	0.13	0.22	0.06	0.03
Wei 2019	**1140**	1	1	1	1	1
Calabrese 2014*	571	0.13	0.07	0.06	0.40	0.13
Ding 2015	170	0.33	0.11	0.21	0.17	0.58
Standard F5	1024	0.06	0.04	0.12	0.40	0.04
Standard F7	766	0.07	0.06	0.22	0.33	0.06
Consensus F5	1016	0.06	0.05	0.09	0.36	0.05
Consensus F7	766	0.09	0.08	0.23	0.33	0.08

Calabrese 2014 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	2665	0.10	0.11	0.16	0.30	0.10
Laricchia 2022	2061	0.18	0.40	0.10	0.55	0.33
Calabrese 2022	1037	0.06	0.05	0.04	0.30	0.03
Wei 2019	571	0.13	0.07	0.06	0.40	0.13
Calabrese 2014*	**3519**	1	1	1	1	1
Ding 2015	399	0.18	0.17	0.12	0.17	0.26
Standard F5	3319	0.16	0.28	0.27	0.50	0.28
Standard F7	2662	0.52	0.68	0.28	0.37	0.68
Consensus F5	3306	0.17	0.29	0.25	0.48	0.29
Consensus F7	2661	0.52	0.68	0.28	0.37	0.68

Ding 2015 vs	Sites shared	All sites	By base type A	C	G	T
Bolze 2020	594	0.31	0.18	0.40	0.12	0.52
Laricchia 2022	496	0.10	0.05	0.13	0.23	0.78
Calabrese 2022	227	0.06	0.05	0.15	0.03	0.02
Wei 2019	170	0.33	0.11	0.21	0.17	0.58
Calabrese 2014*	399	0.18	0.17	0.12	0.17	0.26
Ding 2015	**687**	1	1	1	1	1
Standard F5	618	0.10	0.05	0.65	0.76	0.05
Standard F7	462	0.05	0.06	0.18	0.14	0.06
Consensus F5	618	0.12	0.06	0.67	0.76	0.06
Consensus F7	462	0.07	0.09	0.19	0.14	0.09

(b) Compared study heteroplasmy counts and base type distribution shifts from mtDNA baseline

Study	MAF threshold	Sample size	Heteroplasmies reported Rate	Samples	Positions	Heteroplasmic positions by base type A	C	G	T
Bolze 2020	1%	195,983	0.024	< 40%	8,413	2,681 *(+0.9%)*	2,681 *(-1.9%)*	2,681*(-0.02%)*	2,681 *(+1.0%)*
Laricchia 2022	10%	56,434	0.024	< 40%	5,150	1,534 *(-1.1%)*	1,150 *(-8.9%)*	922 *(+4.8%)*	1,544 *(+5.3%)*
Calabrese 2022	5%	6,214	0.184	66.50%	2,334	665 *(-2.4%)*	497 *(-10.0%)*	428 *(+5.2%)*	744 *(+7.2%)*
Wei 2019	1%	3,052	0.038	45.10%	1,140	292 *(-5.3%)*	156 *(-17.6%)*	297 *(+13.0%)*	395 *(+9.9%)*
Calabrese 2014*	1%	2,419	0.386	86.90%	3,519	809 *(-7.9%)*	593 *(-14.4%)*	908 *(+12.7%)*	1,209 *(+9.6%)*
Ding 2015	4%	2,077	0.043	42.30%	687	171 *(-6.0%)*	126 *(-12.9%)*	178 *(+12.8%)*	212 *(+6.1%)*

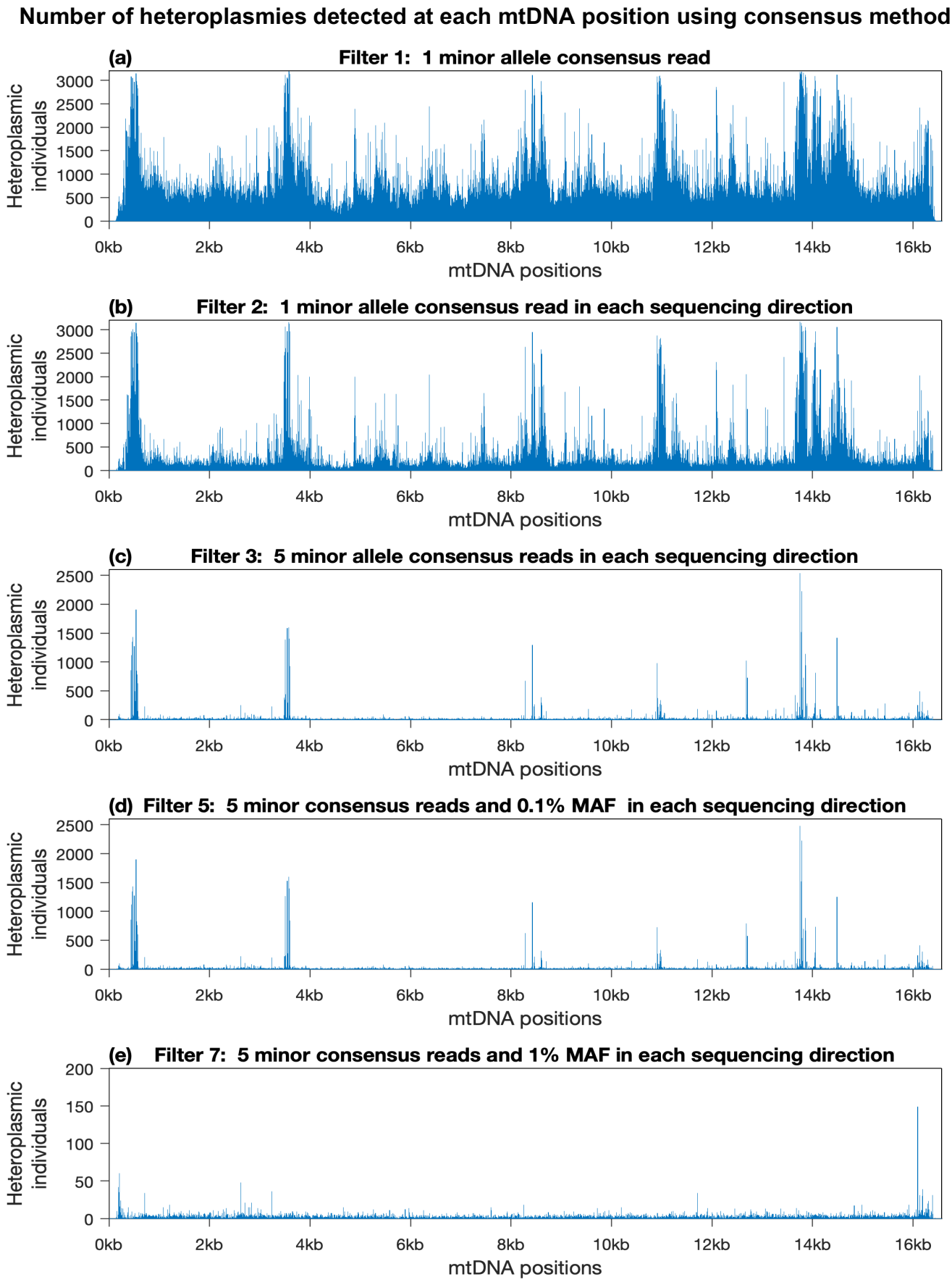
A3.3 Heteroplasmy prevalence landscapes

A3.3.1 Observed heteroplasmy prevalence landscapes across the mitochondria genome

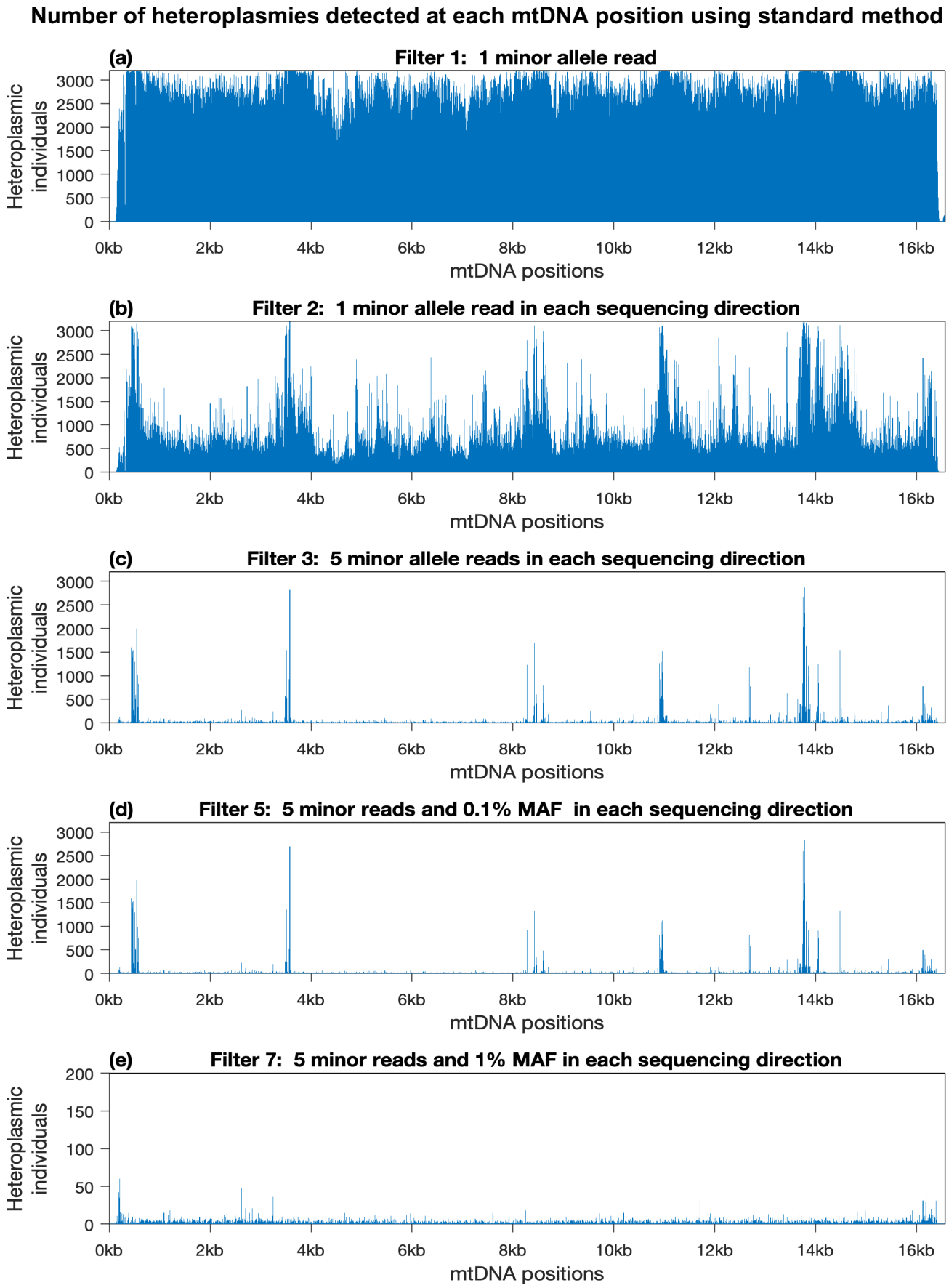
Supplementary figures 5.4.1 S2 and **5.4.1 S2** visualise the landscapes of total heteroplasmy counts per mtDNA position detected with consensus and standard methods respectively.

Unlike the absolute heteroplasmy count, the heteroplasmy prevalence is expressed as the proportion of heteroplasmies found among all analysed samples at each mtDNA site. **Supplementary figures 5.4.1 S3** shows the landscape of heteroplasmy prevalence in the population observed by standard method.

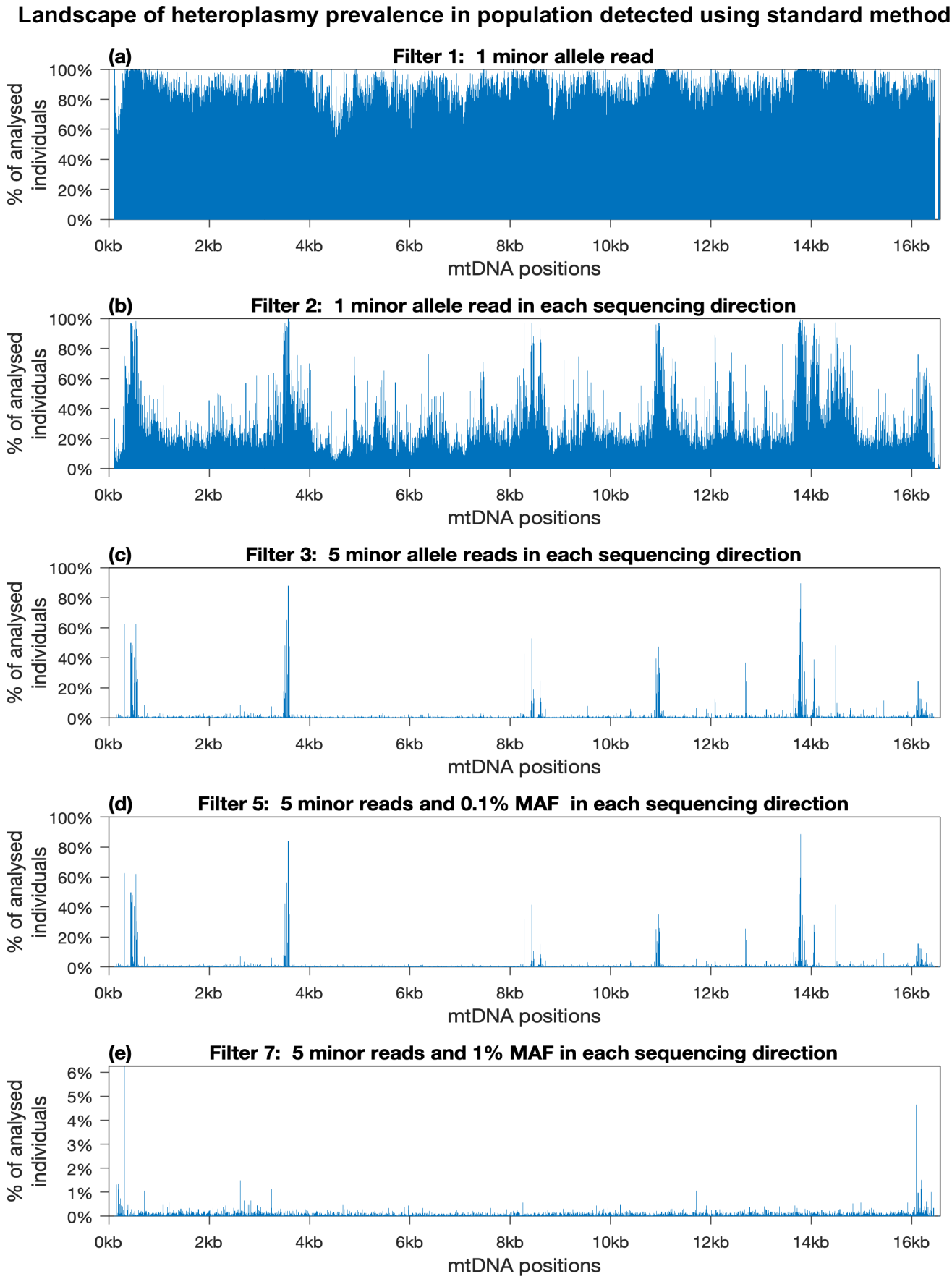
Supplementary figure 5.4.1 S1: Number of heteroplasmies detected at each mtDNA position by consensus method



Supplementary figure 5.4.1 S2: Number of heteroplasmies detected at each mtDNA position by standard method



Supplementary figure 5.4.1 S3: Heteroplasmy prevalence at each mtDNA position found with standard method



Heteroplasmy prevalence – proportion of analysed samples with detectable heteroplasmy at a given MT genome position

Supplementary figure 5.4.1 S4 lists 20 positions with largest discrepancy between the total number of detected heteroplasmies and what proportion of all analysed samples this represents. Positions resulting in the latest deviations are found in areas with reduced sequence coverage where smaller number of samples were analysed. The most extreme difference between the number of heteroplasmies found and what proportion analysed samples this represents is found at the known alignment problem position 310.

Supplementary figure 5.4.1 S4: Positions with largest discrepancy between total heteroplasmy counts versus normalised heteroplasmic sample fractions at filter levels 5 (MAF ≥0.1%) and 7 (MAF≥1%)

Positions with largest discrepancy between total heteroplasmy counts versus normalised heteroplasmic sample fractions at filter levels 5 (MAF ≥0.1%) and 7 (MAF≥1%)

(a) 20 positions with largest difference between heteroplasmy fraction vs raw count at MAF≥0.1%

	Position	rCRS base	Standard method – filter 5					Consensus method – filter 5				
			N_{TOTAL}	N_{HET}	HF	HF_{3202}	ΔHF	N_{TOTAL}	N_{HET}	HF	HF_{3202}	ΔHF
1	**310**	T	16	10	62.50%	0.31%	62.19%	11	5	45.45%	0.16%	45.30%
2	**523**	A	1881	530	28.18%	16.55%	11.62%	1881	315	16.75%	9.84%	6.91%
3	517	A	1876	391	20.84%	12.21%	8.63%	1876	238	12.69%	7.43%	5.25%
4	519	A	1876	402	21.43%	12.55%	8.87%	1876	235	12.53%	7.34%	5.19%
5	521	A	1876	413	22.01%	12.90%	9.12%	1876	230	12.26%	7.18%	5.08%
6	**16189**	T	2604	315	12.10%	9.84%	2.26%	2603	304	11.68%	9.49%	2.18%
7	8279	T	2889	920	31.84%	28.73%	3.11%	2889	623	21.56%	19.5%	2.11%
8	**152**	T	832	19	2.28%	0.59%	1.69%	825	13	1.58%	0.41%	1.17%
9	515	A	1872	64	3.42%	2.00%	1.42%	1872	45	2.40%	1.41%	1.00%
10	**16192**	C	2695	147	5.45%	4.59%	0.86%	2693	128	4.75%	4.00%	0.76%
11	**16188**	C	2621	66	2.52%	2.06%	0.46%	2621	64	2.44%	2.00%	0.44%
12	**146**	T	458	4	0.87%	0.12%	0.75%	453	2	0.44%	0.06%	0.38%
13	**16187**	C	2615	60	2.29%	1.87%	0.42%	2614	49	1.87%	1.53%	0.34%
14	**16438**	G	291	2	0.69%	0.06%	0.62%	286	1	0.35%	0.03%	0.32%
15	**143**	G	368	2	0.54%	0.06%	0.48%	366	1	0.27%	0.03%	0.24%
16	**16184**	C	2611	25	0.96%	0.78%	0.18%	2611	25	0.96%	0.78%	0.18%
17	**16183**	A	2634	25	0.95%	0.78%	0.17%	2634	24	0.91%	0.75%	0.16%
18	561	A	3195	827	25.88%	25.83%	0.06%	3185	761	23.89%	23.8%	0.13%
19	8280	A	2906	117	4.03%	3.65%	0.37%	2906	39	1.34%	1.22%	0.12%
20	439	A	3189	1586	49.73%	49.53%	0.20%	3189	859	26.94%	26.8%	0.11%

(b) 20 positions with largest difference between heteroplasmy fraction vs raw count at MAF≥1%

	Position	rCRS base	Standard method – filter 7					Consensus method – filter 7				
			N_{TOTAL}	N_{HET}	HF	HF_{3202}	ΔHF	N_{TOTAL}	N_{HET}	HF	HF_{3202}	ΔHF
1	**310**	T	16	10	62.50%	0.31%	62.19%	11	5	45.45%	0.16%	45.30%
2	**152**	T	832	11	1.32%	0.34%	0.98%	825	10	1.21%	0.31%	0.90%
3	**146**	T	458	3	0.66%	0.09%	0.56%	453	2	0.44%	0.06%	0.38%
4	**16438**	G	291	1	0.34%	0.03%	0.31%	286	1	0.35%	0.03%	0.32%
5	**143**	G	368	1	0.27%	0.03%	0.24%	366	1	0.27%	0.03%	0.24%
6	**16192**	C	2695	41	1.52%	1.28%	0.24%	2693	39	1.45%	1.22%	0.23%
7	**16189**	T	2604	30	1.15%	0.94%	0.22%	2603	30	1.15%	0.94%	0.22%
8	**16183**	A	2634	14	0.53%	0.44%	0.09%	2634	14	0.53%	0.44%	0.09%
9	305	C	844	1	0.12%	0.03%	0.09%	844	1	0.12%	0.03%	0.09%
10	**16187**	C	2615	11	0.42%	0.34%	0.08%	2614	11	0.42%	0.34%	0.08%
11	**16184**	C	2611	9	0.34%	0.28%	0.06%	2611	9	0.34%	0.28%	0.06%
12	**16188**	C	2621	6	0.23%	0.19%	0.04%	2621	6	0.23%	0.19%	0.04%
13	16390	G	3083	31	1.01%	0.97%	0.04%	3079	31	1.01%	0.97%	0.04%
14	189	A	3119	42	1.35%	1.31%	0.03%	3118	42	1.35%	1.31%	0.04%
15	7501	T	2658	4	0.15%	0.12%	0.03%	2658	4	0.15%	0.12%	0.03%
16	16399	A	2839	6	0.21%	0.19%	0.02%	2832	6	0.21%	0.19%	0.02%
17	16180	A	2705	4	0.15%	0.12%	0.02%	2704	4	0.15%	0.12%	0.02%
18	516	C	1876	1	0.05%	0.03%	0.02%	1876	1	0.05%	0.03%	0.02%
19	**523**	A	1881	1	0.05%	0.03%	0.02%	1881	1	0.05%	0.03%	0.02%
20	7521	G	2647	3	0.11%	0.09%	0.02%	2647	3	0.11%	0.09%	0.02%

N_{TOTAL} – total number of samples analysed; N_{HET} – number of samples with putative heteroplasmies; HF – putative heteroplasmic fraction of analysed samples; HF_{3202} – putative heteroplasmic fraction if all 3202 samples had been analysed; ΔHF – difference between HF and HF_{3202} values;
*Positions shown **in bold** are common to both groups;*

A3.3.2 Heteroplasmy prevalence comparison between standard and consensus methods

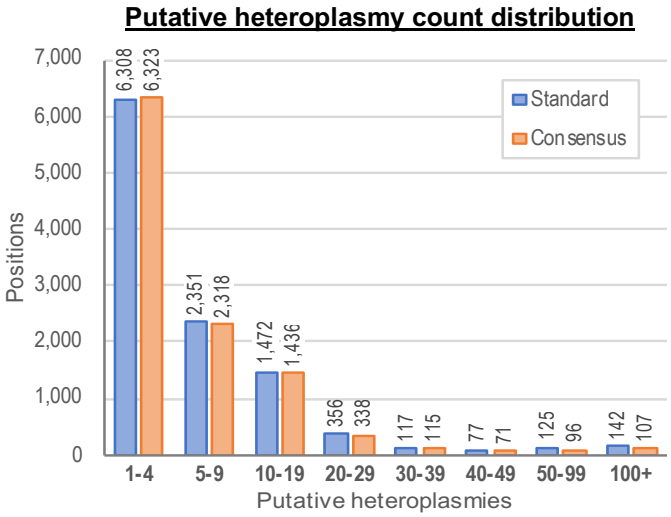
Supplementary figure 5.4.1 S5 compares distributions of identified heteroplasmy counts per site at filter levels 5 and 7 between standard and consensus methods, demonstrating a decrease in the number of heteroplasmies found per site by consensus method, especially at the lower MAF≥0.1% threshold.

Supplementary figure 5.4.1 S5: Distribution of mtDNA positions with different number of heteroplasmies

Distribution of mtDNA positions with different number of putative heteroplasmic samples identified at filter levels 5 and 7 with standard and consensus methods

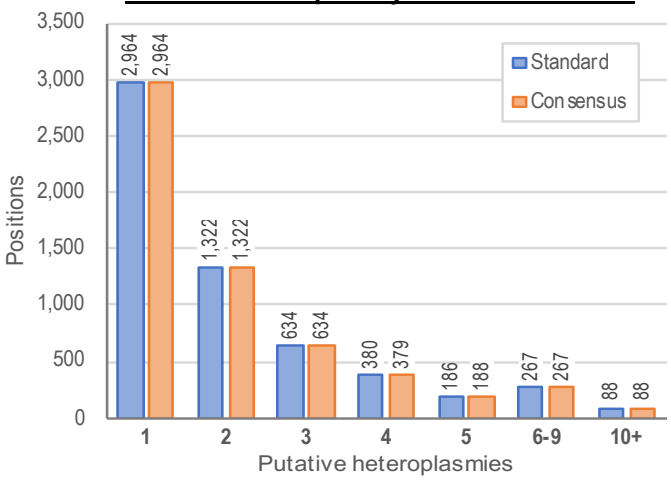
(a) Putative heteroplasmy counts per position found at filter 5 with MAF ≥0.1%

Filter 5 MAF ≥ 0.1%		Putative heteroplasmic positions Standard method		Consensus method	
Number of putative heteroplasmies found	1 - 4	6,308	*57.6%*	6,323	*58.5%*
	5 - 9	2,351	*21.5%*	2,318	*21.5%*
	10 - 19	1,472	*13.4%*	1,436	*13.3%*
	20 - 29	356	*3.3%*	338	*3.1%*
	30 - 39	117	*1.1%*	115	*1.1%*
	40 - 49	77	*0.7%*	71	*0.7%*
	50 - 99	125	*1.1%*	96	*0.9%*
	100+	142	*1.3%*	107	*1.0%*
Total		10,948		10,804	



(b) Putative heteroplasmy counts per position found at filter 7 with MAF ≥1%

Filter 7 MAF ≥ 1%		Putative heteroplasmic positions Standard method		Consensus method	
Number of putative heteroplasmies found	1	2,964	*50.7%*	2,964	*50.7%*
	2	1,322	*22.6%*	1,322	*22.6%*
	3	634	*10.9%*	634	*10.9%*
	4	380	*6.5%*	379	*6.5%*
	5	186	*3.2%*	188	*3.2%*
	6 - 9	267	*4.6%*	267	*4.6%*
	10+	88	*1.5%*	88	*1.5%*
Total		5,841		5,842	



A3.3.3 Positions with highest heteroplasmy prevalence detected by standard method

Supplementary figure 5.4.2 S1 lists the set of 142 mtDNA positions with 100+ heteroplasmies with MAF≥0.1% detected by standard method. This includes all 107 high heteroplasmy incidence sites also detected by consensus method shown in **Figure 5.4.2 F1** in the main text. **Supplementary figure 5.4.2 S2** lists the set of 88 positions with 10+ heteroplasmies with MAF>1% detected by standard method.

Supplementary figure 5.4.2 S1: 142 positions with ≥100 heteroplasmies found at filter 5 with standard method

142 positions with ≥100 putative heteroplasmies found at filter 5 with standard method

Position	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥ 0.1%	*n* with MAF ≥ 1%
204	T	CTAAAG**T**	TAACAC**A**	127	60
439	A	CCCCCC**A**	GTTAGT**T**	1586	0
440	A	CCCCCA**A**	TGTTAG**T**	200	0
443	A	CCAACT**A**	ATGTGT**T**	597	0
446	A	ACTAAC**A**	ATAATG**T**	285	0
448	A	TAACAC**A**	AAATAA**T**	1379	0
464	A	CCTCCC**A**	TGGGAG**T**	1512	0
470	A	ACTCCC**A**	GTAGTA**T**	1539	0
471	T	CTCCCA**T**	AGTAGT**A**	275	0
475	A	CATACT**A**	GATTAG**T**	118	0
478	A	ACTACT**A**	TGAGAT**T**	219	0
492	A	CAATAC**A**	GGGGGT**T**	113	0
503	A	CCGCCC**A**	GTAGGA**T**	1289	1
512	A	CTACCC**A**	GTGTGC**T**	501	0
517	A	CAGCAC**A**	TGTGTG**T**	391	0
519	A	GCACAC**A**	GGTGTG**T**	402	0
521	A	ACACAC**A**	GCGGTG**T**	413	0
523	A	ACACAC**A**	CAGCGG**T**	530	1
538	A	AACCCC**A**	GGGGTA**T**	1984	1
550	A	CGAACC**A**	TTTGGT**T**	583	0
554	A	CCAACC**A**	GGGGTT**T**	980	0
561	A	AACCCC**A**	TGTCTT**T**	827	0
567	A	AAAGAC**A**	GGGGGG**T**	233	0
574	A	CCCCCC**A**	AAACTG**T**	750	0
709	G	ATCCCC**G**	CTGGAA**C**	217	34
2623	A	ATAGGG**A**	TACAGG**T**	227	48
2702	G	AGGCGG**G**	GTTATG**C**	110	21
3244	G	GGCAGA**G**	ACCGGG**C**	201	36
3489	A	GCCCCT**A**	GGGTTT**T**	253	0
3501	A	CGCCAC**A**	GGTAGA**T**	245	0
3508	A	TCTACC**A**	GGGTGA**T**	1354	0
3518	A	CCCTCT**A**	GTGATG**T**	246	0
3547	A	CTCACC**A**	GAGCGA**T**	1800	2
3561	A	TCTACT**A**	GGTTCA**T**	520	0
3572	T	CCCCCC**T**	ATGGGG**A**	156	0
3577	A	CTCCCC**A**	TGGGTA**T**	2695	0
3578	T	TCCCCA**T**	TTGGGT**A**	1661	1
3583	A	ATACCC**A**	GGGGGT**T**	2360	0
3584	A	TACCCA**A**	AGGGGG**T**	338	1
3601	A	AACCTC**A**	CTAGGT**T**	1125	0
8279	T	CCCCTC**T**	GGGGGT**A**	920	3
8280	A	CCCTCT**A**	AGGGGG**T**	117	0
8425	A	TACACT**A**	GAGGAA**T**	138	0
8426	T	ACACTA**T**	TGAGGA**A**	1325	1
8459	A	AACACA**A**	GGTAGT**T**	125	0
8463	A	CAAACT**A**	AGGTGG**T**	339	1
8469	T	ACCACC**T**	GGAGGT**A**	114	2
8470	A	CCACCT**A**	GGGAGG**T**	181	2
8597	T	TGATCA**T**	AATAGA**A**	487	1
8603	T	TTCTAT**T**	GGGGGA**A**	267	0
8613	A	CCCTCT**A**	GATCAA**T**	192	0
8701	A	ATGATA**A**	GTATGG**T**	140	6
9540	T	CCCCAA**T**	CTCCTA**A**	101	1
10398	A	GACTGA**A**	ATTCGG**T**	140	7
10873	T	CATCCC**T**	TAGTAG**A**	111	3
10906	A	CAACCT**A**	GCTAAA**T**	805	1
10915	T	TAGCTG**T**	TGGGGA**A**	305	0
10921	A	TTCCCC**A**	AAAGGT**T**	548	0
10927	T	AACCTT**T**	GGAGGA**A**	105	1
10941	T	ACCCCC**T**	GTTGTT**A**	301	1
10942	A	CCCCCT**A**	GGTTGT**T**	1079	0
10945	A	CCTAAC**A**	GGGGGT**T**	1039	0
10946	A	CTAACA**A**	GGGGGG**T**	498	0
10953	T	CCCCCC**T**	ATTAGG**A**	114	2
10956	T	CCCTCC**T**	AGTATT**A**	145	0
10957	A	CCTCCT**A**	TAGTAT**T**	418	0
10959	T	TCCTAA**T**	GTTAGT**A**	1119	2
10962	T	TAATAC**T**	GTAGTT**A**	106	0
10963	A	AATACT**A**	GGTAGT**T**	831	3
10967	A	CTAACT**A**	GTCAGG**T**	271	2
10977	T	GACTCC**T**	AGGGGT**A**	753	0
10978	A	ACTCCT**A**	GAGGGG**T**	364	2
10985	A	CCCCTC**A**	TGATTG**T**	161	0
10987	A	CCTCAC**A**	CATGAT**T**	269	0
10988	A	CTCACA**A**	CCATGA**T**	237	1
11711	G	ATAATC**G**	CGTGGG**C**	177	34
11914	G	AACCAC**G**	GGAGAA**C**	132	14
12082	A	ACACCT**A**	GGGGGA**T**	126	1
12090	T	CCCCCA**T**	AGGAGA**A**	107	0
12684	G	TAATCA**G**	GAAGAA**C**	818	0
12705	C	ACTCAT**C**	TAGGAA**G**	580	1
13105	A	GCAGGA**A**	AGAAGA**T**	125	8
13276	A	CTCATA**A**	TAACTA**T**	129	1
13433	T	AAACCA**T**	AGAGGT**A**	292	3
13650	C	CTTCCC**C**	AAGGGT**G**	314	1
13681	A	AACCCC**A**	GTAGGG**T**	117	5
13696	A	AACCCC**A**	GTTTAA**T**	207	0
13738	A	TTTCTC**A**	TAGTAA**T**	136	0
13739	T	TTCTCA**T**	TTAGTA**A**	132	1
13744	A	ATTACT**A**	TGTTGT**T**	330	0
13747	A	ACTAAC**A**	AAATGT**T**	619	0
13750	A	AACAAC**A**	GGGAAA**T**	2540	0
13751	T	ACAACA**T**	GGGGAA**A**	664	0
13752	T	CAACAT**T**	GGGGGA**A**	2591	1
13753	T	AACATT**T**	CGGGGG**A**	145	0
13761	A	CCCCGC**A**	GGGGGA**T**	791	0
13762	T	CCCGCA**T**	AGGGGG**A**	1554	2
13768	T	TCCCCC**T**	TTTGGA**A**	835	1
13769	T	CCCCCT**T**	GTTTGG**A**	227	0
13772	A	CCTTCC**A**	GTTGTT**T**	867	0
13776	A	CCAAAC**A**	GATTGT**T**	359	0
13777	A	CAAACA**A**	GGATTG**T**	329	0
13779	A	AACAAC**A**	GGGGAT**T**	1761	0
13780	A	ACAACA**A**	GGGGGA**T**	842	2
13781	T	CAACAA**T**	AGGGGG**A**	1910	4
13789	T	CCCCTC**T**	TTAGGT**A**	2835	2
13790	A	CCCTCT**A**	TTTAGG**T**	618	3
13794	A	CTACCT**A**	GAGTTT**T**	125	1

*** Figure continues on the next page ***

*** Figure continued from the previous page ***

Position (cont. 2)	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥ 0.1%	*n* with MAF ≥ 1%
13799	T	TAAAAC**T**	GCTGTG**A**	128	1
13803	A	ACTCAC**A**	GAGGGC**T**	177	0
13819	T	GTCACT**T**	CTAGGA**A**	1109	0
13853	A	TCAACT**A**	GTTAGG**T**	147	0
13857	A	CTACCT**A**	GTTGGT**T**	917	0
13861	A	CTAACC**A**	GTTTGT**T**	694	0
13864	A	ACCAAC**A**	TAAGTT**T**	130	0
13868	T	ACAAAC**T**	ATTTTA**A**	562	0
13887	A	CCCACT**A**	TGTGCA**T**	144	0
13894	T	TGCACA**T**	AATAAA**A**	153	2
14045	A	AGCACC**A**	GAGATT**T**	144	0
14046	A	GCACCA**A**	GGAGAT**T**	247	0
14047	A	CACCAA**A**	TGGAGA**T**	305	2
14053	A	ATCTCC**A**	TGGAGG**T**	907	5
14059	A	ACCTCC**A**	TGATGA**T**	750	0
14487	T	CCATCA**T**	GGGGGA**A**	1325	2
14513	A	AAAACT**A**	GTTTAA**T**	121	0

Position (cont. 3)	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥ 0.1%	*n* with MAF ≥ 1%
14783	T	ATAAAA**T**	TAATTA**A**	120	1
15043	G	CATCGG**G**	GCCTCG**C**	141	4
15301	G	CATCTT**G**	GAAGGG**C**	165	3
15443	C	GGCTTA**C**	AGAGAA**G**	295	0
16093	T	ATGTAT**T**	GTACGA**A**	240	149
16111	C	TGCCAG**C**	ATGGTG**G**	112	6
16126	T	ATATTG**T**	TACCGT**A**	123	5
16129	G	TTGTAC**G**	TGGTAC**C**	498	31
16170	A	AAACCC**A**	GTGGAT**T**	278	1
16171	A	AACCCA**A**	TGTGGA**T**	400	0
16189	T	ACCCCC**T**	ATGGGG**A**	315	30
16192	C	CCCTCC**C**	AGCATG**G**	147	41
16223	C	TCAACC**C**	AGTTGA**G**	152	2
16269	A	TCACCC**A**	TCCTAG**T**	164	5
16278	C	AGGATA**C**	TTGTTG**G**	114	5
16293	A	CTACCC**A**	TAAGGG**T**	295	20
16311	T	ACATAG**T**	TTATGT**A**	220	23

Groups of nearby positions forming visually identifiable peaks at filter 5 highlighted in blue;
Upstream GGGT motifs highlighted in red; GGGT-like motifs with 3 Gs out of 4 upstream positions shown in orange

Supplementary figure 5.4.2 S2: 88 positions with ≥10 heteroplasmies found at filter 7 with standard method

88 positions with ≥10 putative heteroplasmies found at filter 7 with standard method

Position	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥ 0.1%	*n* with MAF ≥ 1%
152	T	TCATCC **T**	AATAAT **A**	19	11
189	A	AGGCGA **A**	AGTATG **T**	79	42
195	T	ACATAC **T**	TTAGTA **A**	66	35
200	A	CTTACT **A**	ACACTT **T**	26	12
204	T	CTAAAG **T**	TAACAC **A**	127	60
214	A	TTAATT **A**	ATTAAT **T**	19	11
215	A	TAATTA **A**	CATTAA **T**	41	16
234	A	GGACAT **A**	TATTAT **T**	52	24
251	G	TTGAAT **G**	TGCAGA **C**	33	14
279	T	CAGACA **T**	GTTATG **A**	28	13
310	T	CCCCCC **T**	CGGGGG **A**	10	10
385	A	AGCCTA **A**	ATCTGG **T**	45	15
709	G	ATCCCC **G**	CTGGAA **C**	217	34
1082	A	TTAGAT **A**	GTGGGG **T**	67	15
1169	G	CTCAAA **G**	CAGGTC **C**	44	12
1201	A	TAGAGG **A**	ACAGGC **T**	57	18
1435	A	CAGTAA **A**	TCTTAG **T**	19	10
1598	G	TTGGAC **G**	CTGGTT **C**	48	11
2075	T	TCTAAA **T**	AAGGGG **A**	19	11
2174	G	TAGTAG **G**	TTTAGG **C**	23	10
2275	T	GACCAA **T**	TGATAG **A**	28	12
2333	G	TCCTCC **G**	CTTATG **C**	75	14
2492	G	TACCCC **G**	AACAGG **C**	53	11
2623	A	ATAGGG **A**	TACAGG **T**	227	48
2702	G	AGGCGG **G**	GTTATG **C**	110	21
2778	T	CACAGG **T**	TTTAGG **A**	42	15
2805	A	TTAAAA **A**	CCGAAA **T**	42	15
2833	A	AGCAGA **A**	GTTGGG **T**	85	21
2960	T	CTAGAG **T**	ATATGG **A**	49	14
3022	G	GCAGCC **G**	TAATAG **C**	63	12
3242	G	ATGGCA **G**	CGGGCT **C**	48	11
3244	G	GGCAGA **G**	ACCGGG **C**	201	36
3421	G	CCCAAC **G**	CTACAA **C**	25	10
4665	G	GCAACC **G**	TGGATG **C**	52	15
5262	G	AAATGG **G**	TAATGG **C**	54	12
5585	G	GTAACA **G**	CCTTAG **C**	33	11
5894	A	TACCTC **A**	TGGGGG **T**	51	12
5970	G	TTATTC **G**	ATGCGC **C**	66	14
6260	G	AGTGGA **G**	TCCGGC **C**	41	11
7598	G	CATGCA **G**	CTTGCG **C**	35	15
7609	T	AGTAGG **T**	TTGTAG **A**	39	11
8251	G	AATAGG **G**	TACGGG **C**	68	18
9055	G	GGAAGC **G**	GGGTGG **C**	43	10
9716	T	ACTGGG **T**	ATAGAG **A**	34	14

Position (cont.)	rCRS base	Upstream context L strand	Upstream context H strand	*n* with MAF ≥ 0.1%	*n* with MAF ≥ 1%
9868	G	TCATCC **G**	AGTTGG **C**	46	10
9947	G	AGATGT **G**	TCAAAC **C**	37	11
10197	G	TCCCCC **G**	CGCGGG **C**	58	15
11225	G	CTAGTA **G**	GGGAGC **C**	24	10
11456	G	CTTGCC **G**	GTACTG **C**	43	10
11711	G	ATAATC **G**	CGTGGG **C**	177	34
11914	G	AACCAC **G**	GGAGAA **C**	132	14
12457	G	ATTGTC **G**	TGGATG **C**	34	13
12771	G	CTGAGA **G**	TACGCC **C**	48	12
13063	G	ACCCCA **G**	CTGAGA **C**	54	13
13369	T	TCCGGG **T**	TGATGG **A**	58	13
13676	A	AAAATA **A**	GTGGGG **T**	76	14
13722	A	AAGCCT **A**	TGCGAA **T**	46	11
13879	T	ATAAAA **T**	GTGGGG **A**	49	10
14831	G	ATCTCC **G**	ATCATG **C**	60	17
14905	G	AGCCAT **G**	GTAGTG **C**	52	10
14985	G	TCATCC **G**	AGGTAG **C**	70	18
15200	G	CTATCC **G**	GGATGG **C**	37	11
15884	G	AAATGG **G**	GACAGG **C**	56	12
15924	A	TTGTAA **A**	CTCCGG **T**	93	18
16065	G	AGTATT **G**	GTGAGT **C**	43	12
16086	T	AACCGC **T**	ATACAT **A**	58	15
16092	T	TATGTA **T**	TACGAA **A**	54	15
16093	T	ATGTAT **T**	GTACGA **A**	240	149
16124	T	GAATAT **T**	CCGTAC **A**	67	11
16129	G	TTGTAC **G**	TGGTAC **C**	498	31
16172	T	ACCCAA **T**	ATGTGG **A**	78	11
16183	A	ATCAAA **A**	AGGGGG **T**	25	14
16187	C	AAACCC **C**	GGGGAG **G**	60	11
16189	T	ACCCCC **T**	ATGGGG **A**	315	30
16192	C	CCCTCC **C**	AGCATG **G**	147	41
16256	C	CCAAAG **C**	GGGGTG **G**	62	13
16261	C	GCCACC **C**	GGTGAG **G**	58	13
16266	C	CCCTCA **C**	TAGTGG **G**	55	10
16291	C	ACCTAC **C**	AGGGTG **G**	64	20
16292	C	CCTACC **C**	AAGGGT **G**	71	12
16293	A	CTACCC **A**	TAAGGG **T**	295	20
16298	T	CACCCT **T**	ACTGTT **A**	41	12
16301	C	CCTTAA **C**	TGTACT **G**	71	10
16309	A	GTACAT **A**	ATGTAC **T**	69	16
16311	T	ACATAG **T**	TTATGT **A**	220	23
16327	C	CATTTA **C**	TGTACG **G**	44	12
16362	T	TTCTCG **T**	ATGGGG **A**	61	10
16390	G	GATAGG **G**	AGGGAC **C**	97	31

Positions with 100+ putative heteroplasmies with MAF >0.1% highlighted in green; GGGT motifs highlighted in red

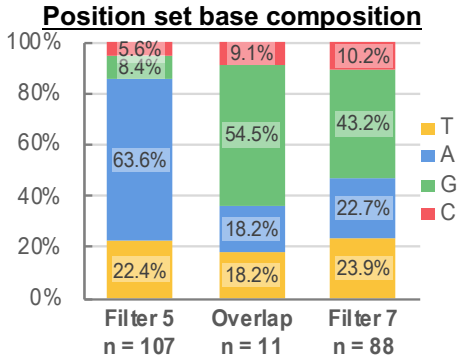
Supplementary figures 5.4.2 S3 and 5.4.2 S4 address the overlap between position subsets with highest prevalence of heteroplasmies above 0.1% MAF and 1% MAF thresholds identified by consensus and standard methods respectively. Site base composition and adjacent sequence contexts are addressed in **panels A and B** respectively.

Supplementary figure 5.4.2 S3: Sites with prevalent heteroplasmy ≥0.1% and ≥1% MAF by consensus method

Overlap between sets of mtDNA position with ≥100 heteroplasmies with MAF≥0.1% and positions with ≥10 heteroplasmies with MAF≥1% identified using consensus method

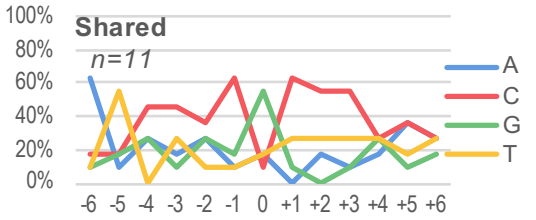
(a) Base composition of the positions with most putative heteroplasmies by both measures

Consensus method		Positions found	By rCRS base type A	C	G	T
Filter 5:	100+ samples with MAF ≥ 0.1%	**107**	68 *63.6 %*	6 *5.6%*	9 *8.4%*	24 *22.4%*
Filter 7:	10+ samples with MAF ≥ 1%	**88**	20 *22.7%*	9 *10.2%*	38 *43.2%*	1 *23.9%*
Overlap:	100+ with MAF ≥ 0.1% & 10+ with MAF ≥ 1%	**11**	2 *18.2%*	1 *9.1%*	6 *54.5%*	2 *18.2%*
as % of filter 5 set with 100+ samples		*9.2%*	*2.9%*	*16.7%*	*66.7%*	8.3%
as % of filter 7 set with 10+ samples		*14.8%*	*10.0%*	*11.1%*	*15.8%*	9.5%



(b) Adjacent region base composition around 11 overlap positions

	L strand upstream -6	-5	-4	-3	-2	-1	0	L strand downstream +1	+2	+3	+4	+5	+6
A	64%	9%	27%	18%	27%	9%	18.2%	0%	18%	9%	18%	36%	27%
C	18%	18%	45%	45%	36%	64%	9.1%	64%	55%	55%	27%	36%	27%
G	9%	18%	27%	9%	27%	18%	54.5%	9%	0%	9%	27%	9%	18%
T	9%	55%	0%	27%	9%	9%	18.2%	27%	27%	27%	27%	18%	27%



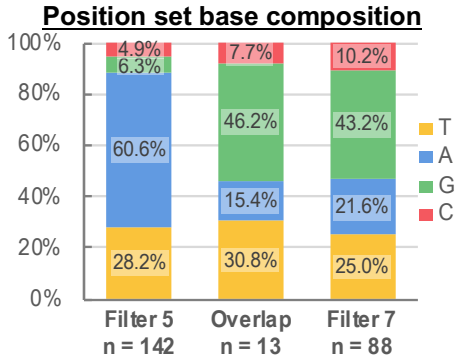
Supplementary figure 5.4.2 S4: Sites with prevalent heteroplasmy ≥0.1% and ≥1% MAF by standard method

Overlap between sets of mtDNA position with ≥100 heteroplasmies with MAF≥0.1% and positions with ≥10 heteroplasmies with MAF≥1% identified using standard method

(a) Base composition of the positions with most putative heteroplasmies by both measures

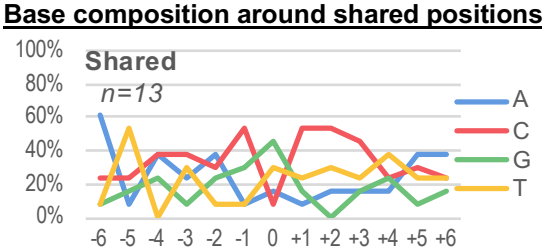
Standard method		Positions found	By rCRS base type A	C	G	T
Filter 5:	100+ samples with MAF ≥ 0.1%	**142**	86 *60.6%*	7 *4.9%*	9 *6.3%*	40 *28.2%*
Filter 7:	10+ samples with MAF ≥ 1%	**88**	19 *21.6%*	9 *10.2%*	38 *43.2%*	22 *25.0%*
Overlap:	100+ with MAF ≥ 0.1% & 10+ with MAF ≥ 1%	**13**	2 *15.4%*	1 *7.7%*	6 *46.2%*	4 *30.8%*
as % of filter 5 set with 100+ samples		*9.2%*	*2.3%*	*14.3%*	*66.7%*	*10.0%*
as % of filter 7 set with 10+ samples		*14.8%*	*10.5%*	*11.1%*	*15.8%*	*18.2%*

mtDNA overall base composition: A 30.9%, C 31.3%, G 13.1%, T 24.7%



(b) Adjacent region base composition around 13 overlap positions

	L strand upstream -6	-5	-4	-3	-2	-1	0	L strand downstream +1	+2	+3	+4	+5	+6
A	62%	8%	38%	23%	38%	8%	15.4%	8%	15%	15%	15%	38%	38%
C	23%	23%	38%	38%	31%	54%	7.7%	54%	54%	46%	23%	31%	23%
G	8%	15%	23%	8%	23%	31%	46.2%	15%	0%	15%	23%	8%	15%
T	8%	54%	0%	31%	8%	8%	30.8%	23%	31%	23%	38%	23%	23%

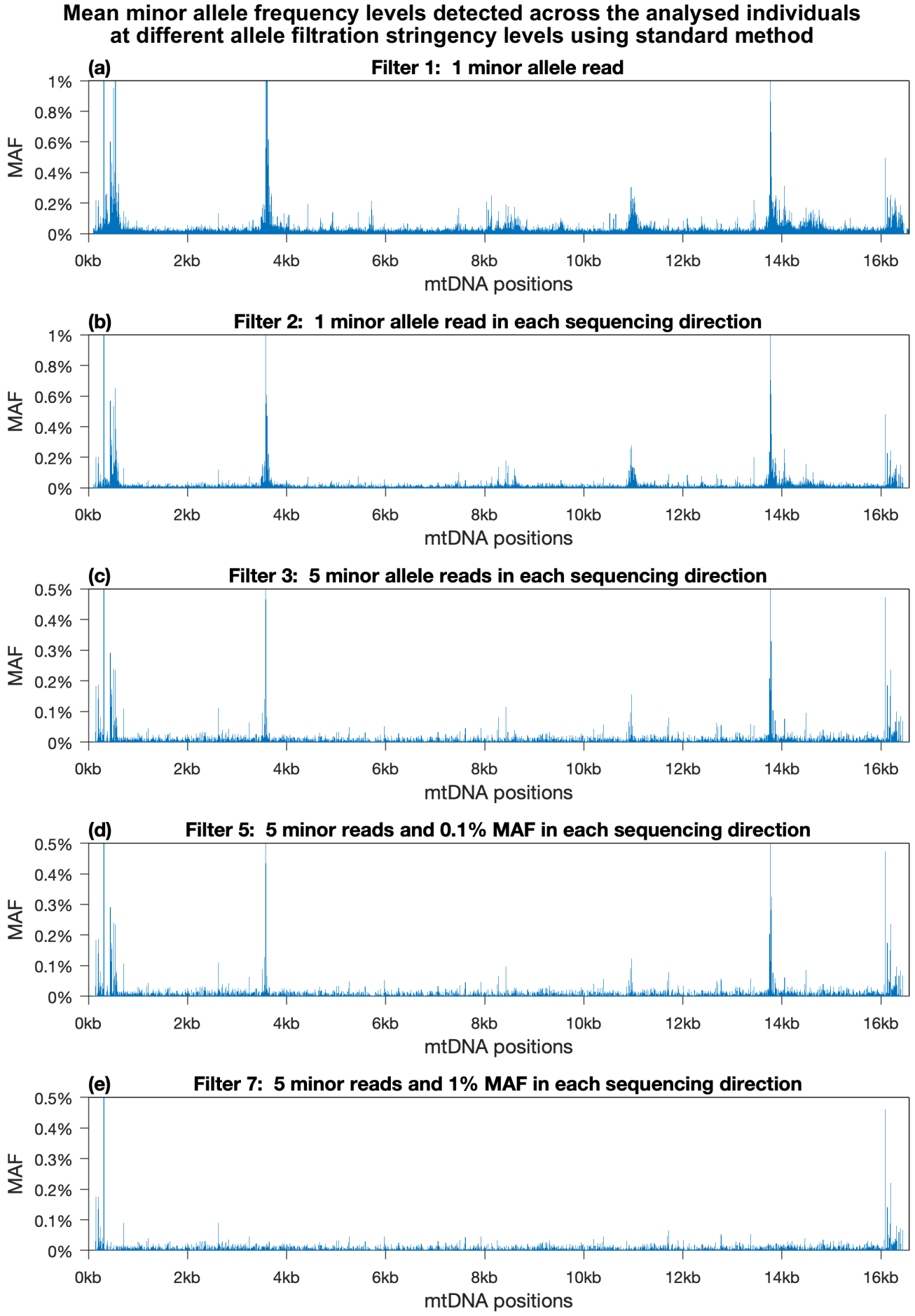


A3.4 Heteroplasmy levels

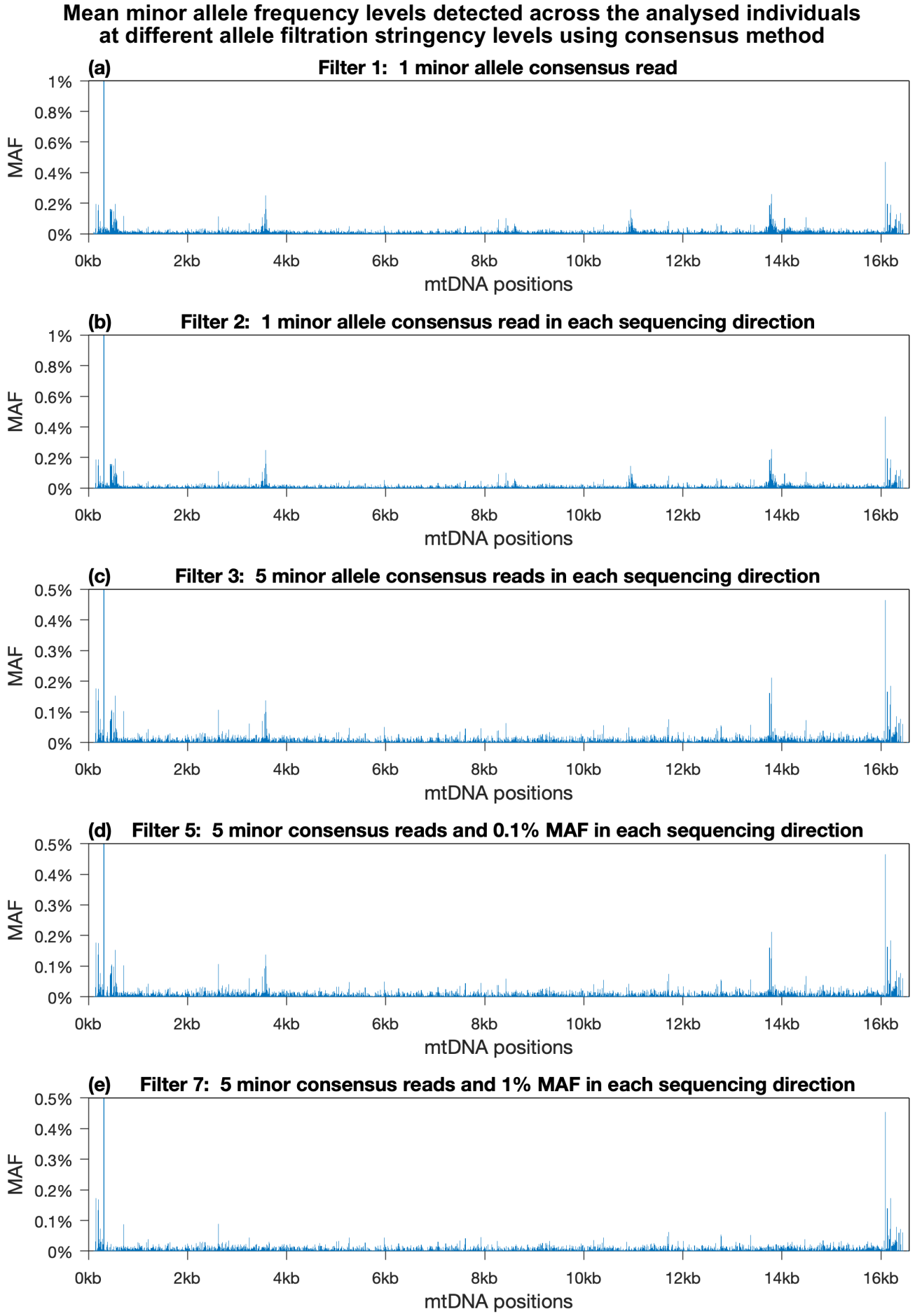
A3.4.1 Average MAF in population

Supplementary figures 5.5.1 S1 and **5.5.1 S2** show the landscapes of population mean MAF measured at each MT genome position using consensus and standard methods respectively at five allele filtration stringency levels. In both cases, MAF is averaged as the mean across all analysed individuals at any given mtDNA position.

Supplementary figure 5.5.1 S1: Mean MAF detected across all analysed samples using standard method



Supplementary figure 5.5.1 S2: Mean MAF detected across all analysed samples using consensus method



Sample fraction with MAF above different thresholds

Supplementary Figure 5.5.1 S3 summarises our observed average analysed sample counts (**Panel A**) and proportions (**Panel B**) found at each mtDNA positions to exceed different MAF thresholds at different allele filtration stringencies, showing that on average only a very small fraction of samples at any given site have high MAF levels. Coefficients of variation (**Panel C**) indicate that standard method results in a higher variability between positions in terms of sample fraction found above MAF thresholds of 0.5% or higher when compared to consensus method. This could be consistent with a large contribution to MAF distribution variability between positions coming from excess reads which we removed when calculating allele consensus frequencies used for MAF measurement with consensus frequency based method.

Supplementary figure 5.5.1 S3: Heteroplasmy incidence above different MAF thresholds across MT genome

Average incidence of putative heteroplasmies above 0.1%, 0.5%, 1%, 5% and 10% MAF thresholds found across mtDNA positions at different allele filtration stringency levels

(a) Average number of samples found per position with MAF above different thresholds

MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	538.7 *SD 399.5*	173.9 *SD 292.6*	8.47 *SD 61.6*			2396.9 *SD 474.8*	538.7 *SD 399.5*	12.0 *SD 88.7*		
MAF ≥ 0.1%	12.8 *SD 103.1*	12.2 *SD 99.9*	7.36 *SD 58.0*	7.36 *SD 58.0*		73.8 *SD 340.6*	35.8 *SD 205.7*	11.0 *SD 86.7*	9.23 *SD 76.2*	
MAF ≥ 0.5%	1.40 *SD 3.99*	1.40 *SD 3.94*	1.39 *SD 3.85*	1.39 *SD 3.85*		5.38 *SD 92.9*	3.42 *SD 55.8*	2.22 *SD 32.3*	2.15 *SD 29.5*	
MAF ≥ 1%	0.82 *SD 2.22*	0.82 *SD 2.21*	0.82 *SD 2.21*	0.82 *SD 2.21*	0.82 *SD 2.21*	2.35 *SD 59.5*	1.49 *SD 29.9*	1.09 *SD 18.5*	1.06 *SD 15.5*	0.82 *SD 2.21*
MAF ≥ 5%	0.26 *SD 0.93*	0.26 *SD 0.93*	0.26 *SD 0.93*	0.26 *SD 0.93*	0.26 *SD 0.93*	0.30 *SD 3.53*	0.27 *SD 1.12*	0.27 *SD 0.95*	0.27 *SD 0.95*	0.27 *SD 0.95*
MAF ≥ 10%	0.146 *SD 0.62*	0.146 *SD 0.62*	0.146 *SD 0.62*	0.146 *SD 0.62*	0.146 *SD 0.62*	0.151 *SD 0.64*	0.151 *SD 0.64*	0.151 *SD 0.64*	0.151 *SD 0.64*	0.151 *SD 0.64*

Averaged are calculated across the entire mitochondrial genome; SD – standard deviation; Position 310 excluded

(b) Average fraction of analysed samples found per position with MAF above different thresholds

MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	16.9% *SD 12.5%*	5.45% *SD 9.22%*	0.27% *SD 1.94%*			75.6% *SD 13.2%*	16.9% *SD 12.5%*	0.38% *SD 2.79%*		
MAF ≥ 0.1%	0.41% *SD 3.26%*	0.39% *SD 3.16%*	0.23% *SD 1.82%*	0.23% *SD 1.82%*		2.34% *SD 10.7%*	1.13% *SD 6.49%*	0.35% *SD 2.73%*	0.29% *SD 2.4%*	
MAF ≥ 0.5%	0.04% *SD 0.13%*	0.04% *SD 0.13%*	0.04% *SD 0.13%*	0.04% *SD 0.13%*		0.17% *SD 2.91%*	0.11% *SD 1.74%*	0.07% *SD 1.01%*	0.07% *SD 0.92%*	
MAF ≥ 1%	0.03% *SD 0.07%*	0.03% *SD 0.07%*	0.03% *SD 0.07%*	0.03% *SD 0.07%*	0.03% *SD 0.07%*	0.07% *SD 1.86%*	0.05% *SD 0.93%*	0.03% *SD 0.58%*	0.03% *SD 0.49%*	0.03% *SD 0.07%*
MAF ≥ 5%	0.008% *SD 0.03%*	0.008% *SD 0.03%*	0.008% *SD 0.03%*	0.008% *SD 0.03%*	0.008% *SD 0.03%*	0.009% *SD 0.11%*	0.009% *SD 0.04%*	0.009% *SD 0.03%*	0.009% *SD 0.03%*	0.008% *SD 0.03%*
MAF ≥ 10%	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*	0.005% *SD 0.02%*

Values represent fraction of total all analysed samples per position; SD – standard deviation; Position 310 excluded

(c) Coefficient of variation of analysed sample fraction with MAF above different thresholds

MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	0.74	1.69	7.25			0.17	0.74	7.36		
MAF ≥ 0.1%	7.97	8.15	7.86	7.86		4.58	5.74	7.87	8.23	
MAF ≥ 0.5%	2.96	2.94	2.89	2.89		17.17	16.24	14.45	13.63	
MAF ≥ 1%	2.74	2.74	2.74	2.74	2.74	25.21	19.91	16.82	14.54	2.75
MAF ≥ 5%	3.64	3.64	3.64	3.64	3.64	11.81	4.15	3.63	3.63	3.64
MAF ≥ 10%	4.34	4.34	4.34	4.34	4.34	4.37	4.37	4.37	4.37	4.37

Coefficient of variation calculated as the ratio of standard deviation to the mean;

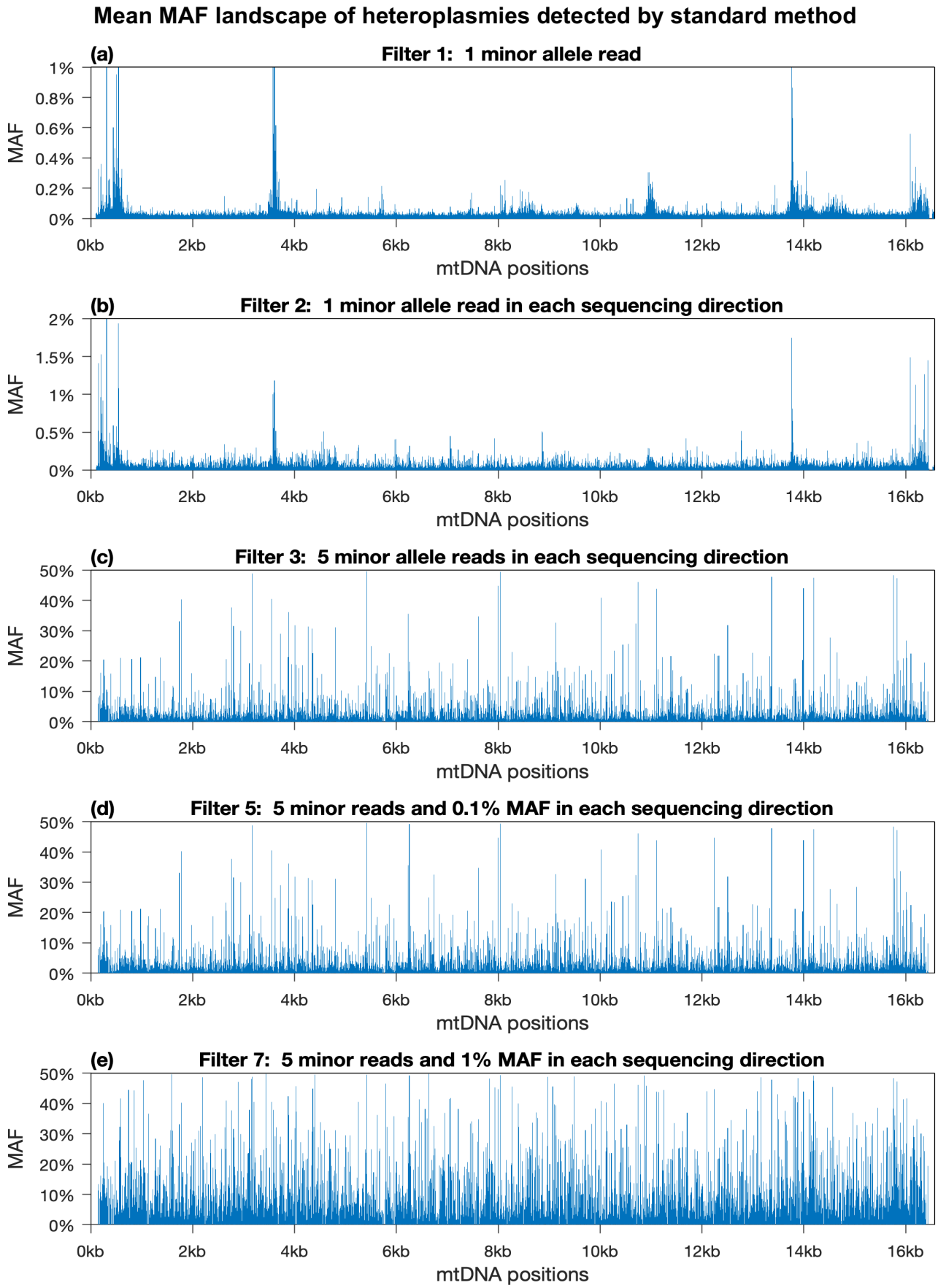
A3.4.2 Average MAF of the identified heteroplasmies

Mean MAF landscapes

Supplementary figure 5.5.2 S1 shows the mean MAF of the detected heteroplasmies at each mtDNA position using standard method at different allele filtration stringency levels. For the corresponding data based on consensus method see **Figure 5.5.2 F1** in the main text.

Supplementary figure 5.5.2 S2 provides an overview of the overall position-specific heteroplasmy mean MAF distribution. **Panel A** shows number of positions with heteroplasmy mean MAF falling in different ranges, and **panel B** provides distribution central tendency and spread measures.

Supplementary figure 5.5.2 S1: Mean MAF across identified heteroplasmies using standard method



MAF – minor allele frequency; MAF mean calculated across the heteroplasmic sample subset at each MT genome position

Supplementary figure 5.5.2 S2: Overview of the heteroplasmy mean MAF level landscape across the MT genome

Overview of the putative heteroplasmy mean MAF level landscape across the MT genome

(a) Number of positions found with different heteroplasmic sample subset mean MAF levels

Mean MAF across het. samples	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
< 0.01%	6 *0.04%*	0	0			6 *0.04%*	0	0		
0.01% - 0.05%	14,182 *86.8%*	11,013 *67.9%*	14 *0.13%*			15,209 *92.7%*	13,189 *80.7%*	19 *0.16%*		
0.05% - 0.1%	1,373 *8.40%*	2,255 *13.9%*	461 *4.13%*			835 *5.09%*	2,016 *12.3%*	407 *3.53%*		
0.1% - 0.5%	762 *4.66%*	2,522 *15.6%*	5,145 *46.1%*	4,977 *46.1%*		331 *2.02%*	1,100 *6.73%*	5,469 *47.4%*	4,845 *44.3%*	
0.5% - 1%	10 *0.06%*	335 *2.07%*	2,104 *18.9%*	2,156 *20.0%*		11 *0.07%*	21 *0.13%*	2,239 *19.4%*	2,308 *21.1%*	
1% - 5%	5 *0.03%*	79 *0.49%*	2,770 *24.8%*	2,955 *27.4%*	3,595 *61.5%*	7 *0.04%*	12 *0.07%*	2,806 *24.3%*	3,098 *28.3%*	3,520 *60.3%*
5% - 10%	1 *0.01%*	6 *0.04%*	427 *3.83%*	450 *4.17%*	1,141 *19.5%*	1 *0.01%*	1 *0.01%*	378 *3.28%*	442 *4.04%*	1,176 *20.1%*
> 10%	0	0	234 *2.10%*	266 *2.46%*	1,106 *18.9%*	0	0	212 *1.84%*	255 *2.33%*	1,145 *19.6%*

Percentages indicate fraction of all MT genome positions with any putative heteroplasmies found at that filter level

(b) Descriptive statistics of heteroplasmic subset mean MAF level distribution

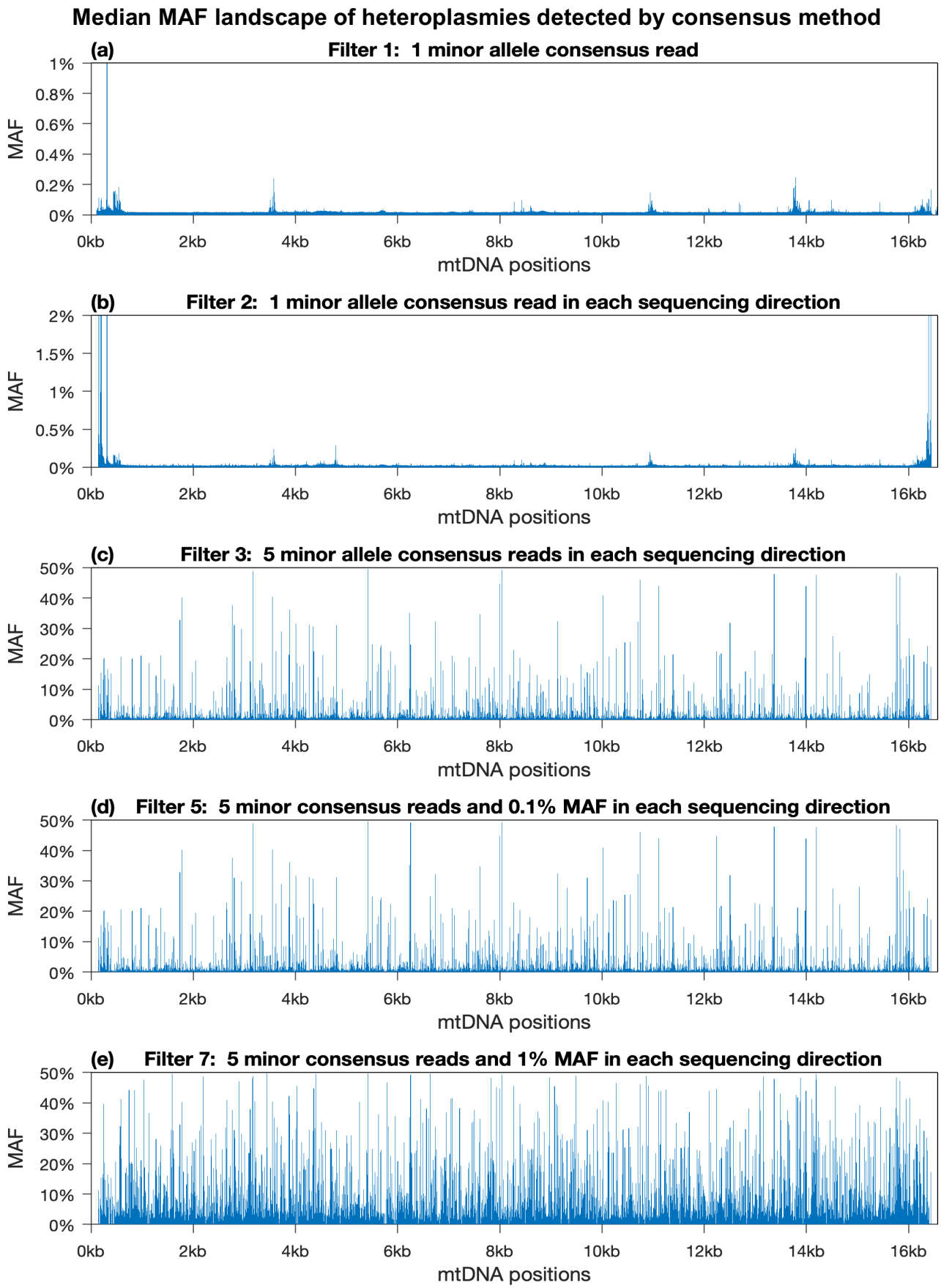
Mean MAF statistic	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5
Mean	0.03%	0.08%	1.45%	1.61%	6.57%	0.03%	0.04%	1.35%	1.60%	6.73%
SD	0.06%	0.22%	3.20%	3.41%	7.73%	0.08%	0.09%	3.02%	3.34%	7.76%
Median	0.02%	0.03%	0.50%	0.56%	3.55%	0.02%	0.03%	0.48%	0.59%	3.69%
IQR	0.01%	0.05%	1.07%	1.22%	6.21%	0.01%	0.02%	0.99%	1.20%	6.29%
Range	5.77%	8.82%	49.59%	49.52%	48.82%	6.92%	9.04%	49.50%	49.43%	48.78%

Median MAF landscapes

Supplementary figures 5.5.2 S3 and **5.5.2 S4** show the median MAF of the heteroplasmies detected at each mtDNA position using consensus and standard methods respectively at different allele filtration stringency levels

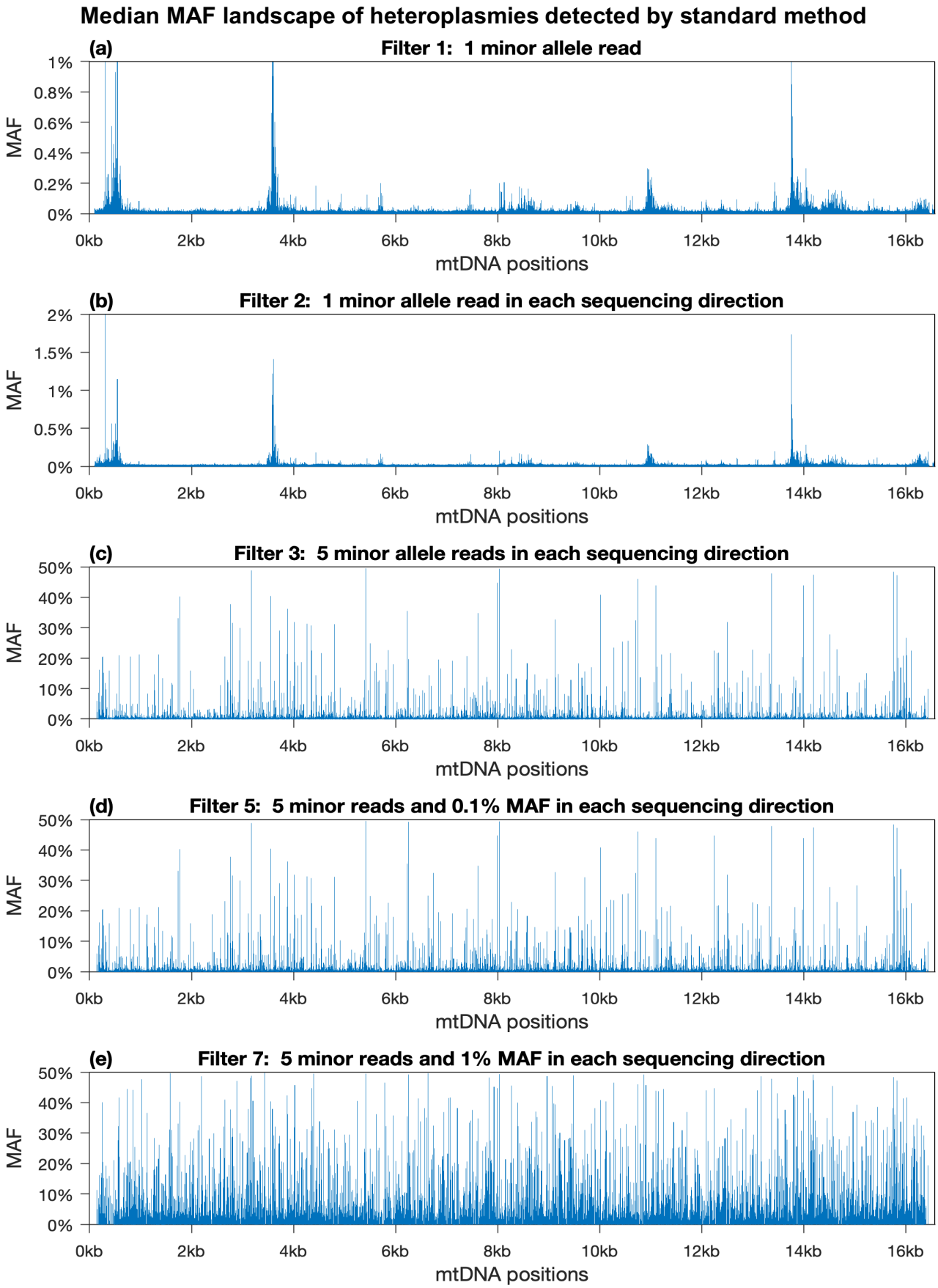
Supplementary figure 5.5.2 S5 provides an overview of the overall position-specific heteroplasmy median MAF distribution. **Panel A** shows number of positions with heteroplasmy median MAF falling in different ranges, and **panel B** provides distribution central tendency and spread measures.

Supplementary figure 5.5.2 S3: Median MAF across identified heteroplasmies using consensus method



MAF – minor allele frequency; MAF median calculated across the heteroplasmic sample subset at each MT genome position

Supplementary figure 5.5.2 S 4: Median MAF across identified heteroplasmies using standard method



MAF – minor allele frequency; MAF median calculated across the heteroplasmic sample subset at each MT genome position

Supplementary figure 5.5.2 S5: Overview of the heteroplasmy median MAF level landscape across the MT genome

Overview of the putative heteroplasmy median MAF level landscape across the MT genome

(a) Number of positions found with different heteroplasmic sample subset median MAF levels

Median MAF across het. samples	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
< 0.01%	6 *0.04%*	0	0			515 *3.14%*	0	0		
0.01% - 0.05%	16,204 *99.2%*	15,775 *97.3%*	17 *0.15%*			15,166 *92.5%*	15,706 *96.1%*	19 *0.16%*		
0.05% - 0.1%	101 *0.62%*	299 *1.84%*	600 *5.38%*			460 *2.80%*	428 *2.62%*	533 *4.62%*		
0.1% - 0.5%	27 *0.17%*	123 *0.76%*	7,988 *71.6%*	7,806 *72.3%*		241 *1.47%*	192 *1.18%*	8,471 *73.5%*	7,715 *70.5%*	
0.5% - 1%	0	5 *0.03%*	1,256 *11.3%*	1,521 *14.1%*		11 *0.07%*	8 *0.05%*	1,305 *11.3%*	1,734 *15.8%*	
1% - 5%	0	4 *0.02%*	971 *8.70%*	1,115 *10.3%*	4,057 *69.4%*	6 *0.04%*	4 *0.02%*	918 *7.96%*	1,148 *10.5%*	3,983 *68.2%*
5% - 10%	1 *0.01%*	4 *0.02%*	150 *1.34%*	171 *1.58%*	896 *15.3%*	1 *0.01%*	0	132 *1.14%*	167 *1.53%*	939 *16.1%*
> 10%	0	0	173 *1.55%*	191 *1.77%*	889 *15.2%*	0	1 *0.01%*	152 *1.32%*	184 *1.68%*	919 *15.7%*

Percentages indicate fraction of all MT genome positions with any putative heteroplasmies found at that filter level

(b) Descriptive statistics of heteroplasmic subset median MAF level distribution

Median MAF statistic	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5
Mean	0.02%	0.03%	0.84%	0.97%	5.90%	0.02%	0.03%	0.77%	0.97%	6.06%
SD	0.06%	0.13%	2.91%	3.13%	7.70%	0.09%	0.09%	2.72%	3.06%	7.73%
Median	0.02%	0.02%	0.25%	0.30%	2.82%	0.02%	0.02%	0.26%	0.33%	2.97%
IQR	0.002%	0.01%	0.30%	0.35%	4.40%	0.01%	0.01%	0.27%	0.35%	4.50%
Range	7.91%	8.82%	49.59%	49.52%	48.82%	9.77%	10.34%	49.50%	49.43%	48.78%

(c) Difference from heteroplasmic subset mean MAF level distribution descriptive statistics

MAF descriptive statistic difference	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 1	Filter 2	Filter 3	Filter 4	Filter 5
Δ Mean	-0.016%	-0.060%	-0.61%	-0.64%	-0.67%	-0.006%	-0.018%	-0.58%	-0.63%	-0.67%
Δ SD	-0.002%	-0.085%	-0.29%	-0.28%	-0.03%	+0.017%	-0.005%	-0.30%	-0.28%	-0.03%
Δ Median	-0.004%	-0.011%	-0.24%	-0.27%	-0.73%	-0.005%	-0.008%	-0.22%	-0.27%	-0.72%
Δ IQR	-0.012%	-0.042%	-0.77%	-0.86%	-1.81%	-0.004%	-0.015%	-0.71%	-0.84%	-1.79%
Δ Range	+2.15%	+0.001%	0%	0%	0%	+2.86%	+1.30%	0%	0%	0%

A3.4.3 Observed heteroplasmy MAF variability

Supplementary figure 5.5.2 S6 compares what proportion of our identified heteroplasmies per site on average reach different MAF thresholds with standard and consensus methods across different allele filtration stringencies.

Supplementary figure 5.5.2 S7 addresses the variability between heteroplasmic positions in terms of the MAF distributions of heteroplasmies identified by standard method above 0.1% (**panel A**) and 1% (**panel B**) MAF thresholds.

Supplementary figure 5.5.2 S6: Average heteroplasmy fractions above different MAF thresholds

Average putative heteroplasmy fractions above 0.1%, 0.5%, 1%, 5% and 10% MAF thresholds across heteroplasmic position set at different allele filtration stringency levels

(a) Average number of samples with MAF above different thresholds found per heteroplasmic position

MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	540.8 *SD 398.8*	175.9 *SD 293.7*	12.5 *SD 74.4*			2397.2 *SD 474.1*	540.8 *SD 398.8*	17.1 *SD 105.4*		
MAF ≥ 0.1%	12.9 *SD 103.3*	12.3 *SD 100.5*	10.8 *SD 70.1*	11.2 *SD 71.2*		73.8 *SD 340.6*	35.9 *SD 206.1*	15.6 *SD 103.0*	13.8 *SD 92.9*	
MAF ≥ 0.5%	1.41 *SD 3.99*	1.42 *SD 3.96*	2.04 *SD 4.53*	2.11 *SD 4.58*		5.38 *SD 92.9*	3.43 *SD 55.9*	3.16 *SD 38.4*	3.23 *SD 36.1*	
MAF ≥ 1%	0.82 *SD 2.22*	0.83 *SD 2.23*	1.20 *SD 2.60*	1.24 *SD 2.63*	2.29 *SD 3.22*	2.35 *SD 59.5*	1.50 *SD 30.0*	1.56 *SD 22.1*	1.59 *SD 19.0*	2.29 *SD 3.22*
MAF ≥ 5%	0.26 *SD 0.93*	0.26 *SD 0.93*	0.38 *SD 1.10*	0.39 *SD 1.12*	0.73 *SD 1.44*	0.30 *SD 3.53*	0.27 *SD 1.12*	0.38 *SD 1.12*	0.40 *SD 1.14*	0.75 *SD 1.48*
MAF ≥ 10%	0.15 *SD 0.62*	0.15 *SD 0.63*	0.22 *SD 0.74*	0.22 *SD 0.76*	0.41 *SD 0.99*	0.15 *SD 0.64*	0.15 *SD 0.64*	0.22 *SD 0.76*	0.23 *SD 0.77*	0.42 *SD 1.02*

Averages are calculated across sites with at least one filter-passing putative heteroplasmy; SD – standard deviation

(b) Average fraction of putative heteroplasmies found with MAF above different thresholds

MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	100%	100%	100%			100%	100%	100%		
MAF ≥ 0.1%	1.38% *SD 4.29%*	4.78% *SD 9.48%*	88.1% *SD 22.0%*	100%		2.48% *SD 10.8%*	3.08% *SD 9.54%*	89.7% *SD 20.9%*	100%	
MAF ≥ 0.5%	0.31% *SD 0.92%*	1.32% *SD 3.37%*	30.0% *SD 31.2%*	33.4% *SD 32.4%*		0.19% *SD 2.98%*	0.43% *SD 2.47%*	29.1% *SD 30.4%*	34.8% *SD 32.5%*	
MAF ≥ 1%	0.18% *SD 0.72%*	0.79% *SD 2.33%*	18.4% *SD 26.3%*	20.5% *SD 27.7%*	100%	0.09% *SD 1.97%*	0.23% *SD 1.58%*	17.4% *SD 25.2%*	20.8% *SD 27.6%*	100%
MAF ≥ 5%	0.06% *SD 0.59%*	0.26% *SD 1.37%*	5.93% *SD 16.1%*	6.58% *SD 17.1%*	31.3% *SD 38.6%*	0.02% *SD 0.50%*	0.07% *SD 0.69%*	5.39% *SD 15.0%*	6.44% *SD 16.7%*	32.3% *SD 39.0%*
MAF ≥ 10%	0.03% *SD 0.14%*	0.14% *SD 0.90%*	3.28% *SD 12.0%*	3.64% *SD 12.8%*	17.3% *SD 31.3%*	0.01% *SD 0.34%*	0.04% *SD 0.48%*	2.99% *SD 11.3%*	3.56% *SD 12.5%*	17.9% *SD 31.7%*

Values represent fraction of total putative heteroplasmic samples found per position; SD – standard deviation

(c) Coefficient of variation of putative heteroplasmic sample fractions above different MAF thresholds

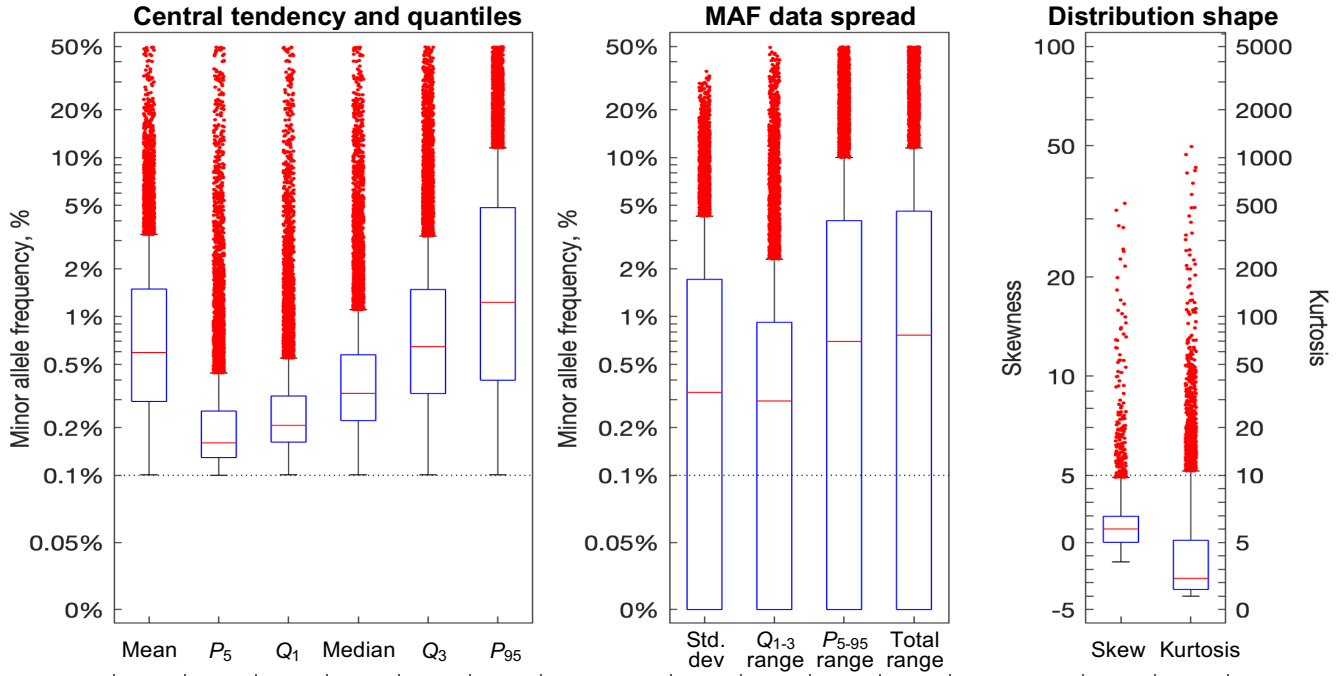
MAF threshold	Consensus method					Standard method				
	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7	Filter 1	Filter 2	Filter 3	Filter 5	Filter 7
MAF ≥ 0%	0	0	0			0	0	0		
MAF ≥ 0.1%	3.10	1.99	0.25	0		4.37	3.10	0.23	0	
MAF ≥ 0.5%	2.98	2.56	1.04	0.97		15.6	5.75	1.05	0.93	
MAF ≥ 1%	3.89	2.96	1.43	1.35	0	22.3	6.85	1.45	1.33	0
MAF ≥ 5%	9.60	5.29	2.72	2.60	1.23	31.4	10.5	2.78	2.59	1.21
MAF ≥ 10%	4.39	6.31	3.66	3.52	1.80	38.0	13.0	3.77	3.52	1.77

Coefficient of variation calculated as the ratio of standard deviation to the mean

Supplementary figure 5.5.2 S7: Distributions of position-specific MAF population distribution summary statistics and shape properties at filter levels 5 and 7 identified using standard method

Overview of the distributions of position-specific MAF population distribution summary statistics and shape properties at filter levels 5 and 7 identified using standard method

(a) MAF distribution features of 10,948 putative heteroplasmic positions at filter 5 (MAF ≥ 0.1%)

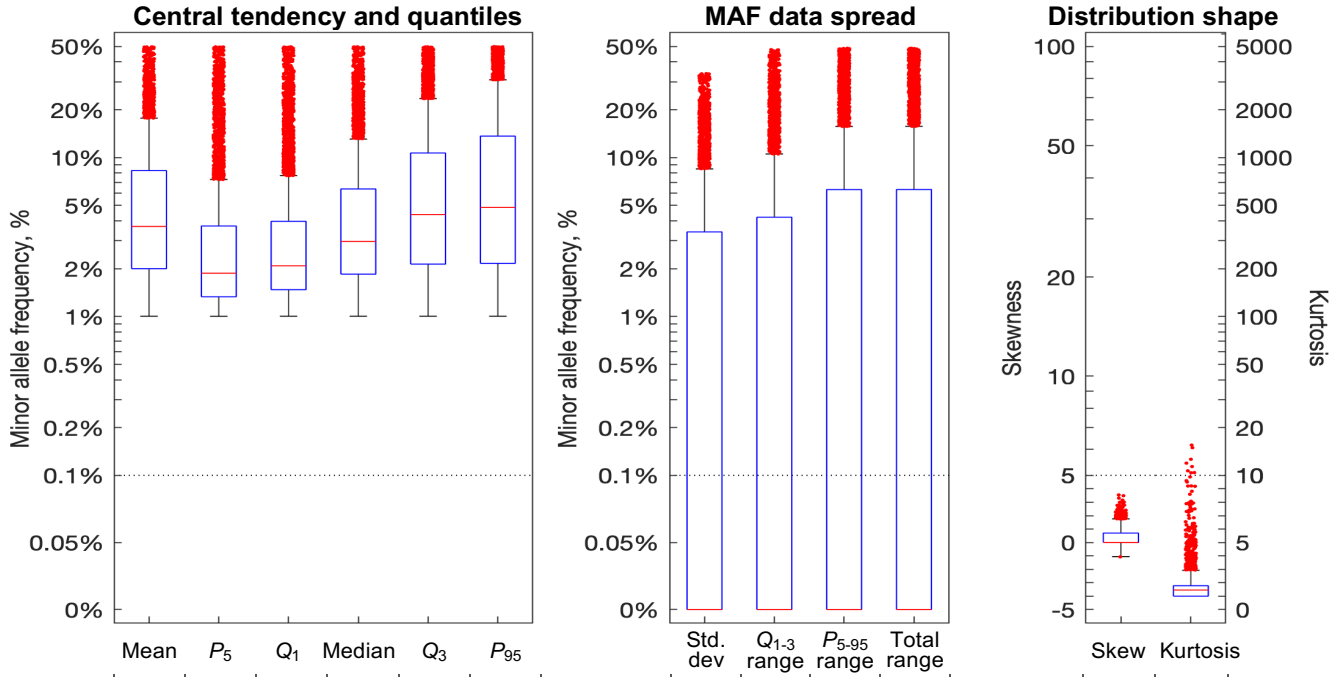


	Mean	P_5	Q_1	Median	Q_3	P_{95}
Whisker Hi	3.28%	0.44%	0.55%	1.10%	3.21%	11.5%
Box Hi - Q_3	1.49%	0.25%	0.32%	0.57%	1.48%	4.84%
Median - Q_2	0.59%	0.16%	0.21%	0.33%	0.65%	1.23%
Box Lo - Q_1	0.29%	0.13%	0.16%	0.22%	0.33%	0.40%
Whisker Lo	0.10%	0.10%	0.10%	0.10%	0.10%	0.10%
Outliers	1227	1514	1379	1377	1342	1416

Std. dev	$Q_{1\text{-}3}$ range	$P_{5\text{-}95}$ range	Total range
4.28%	2.30%	10.0%	11.5%
1.71%	0.92%	4.01%	4.60%
0.33%	0.29%	0.70%	0.76%
0%	0%	0%	0%
0%	0%	0%	0%
1460	1279	1459	1539

Skew	Kurtosis	
4.85	10.7	*Whisker Hi*
1.95	5.17	*Box Hi - Q_3*
1.01	2.33	*Median - Q_2*
0.01	1.50	*Box Lo - Q_1*
-1.45	1	*Whisker Lo*
178	796	*Outliers*

(b) MAF distribution features of 5,841 putative heteroplasmic positions at filter 7 (MAF ≥ 1%)



	Mean	P_5	Q_1	Median	Q_3	P_{95}
Whisker Hi	17.7%	7.28%	7.71%	13.1%	23.5%	30.9%
Box Hi - Q_3	8.29%	3.71%	3.97%	6.35%	10.7%	13.7%
Median - Q_2	3.69%	1.87%	2.08%	2.97%	4.38%	4.86%
Box Lo - Q_1	2.00%	1.33%	1.48%	1.85%	2.14%	2.16%
Whisker Lo	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%
Outliers	499	817	792	695	594	567

Std. dev	$Q_{1\text{-}3}$ range	$P_{5\text{-}95}$ range	Total range
8.49%	10.5%	15.7%	15.7%
3.41%	4.22%	6.29%	6.30%
0%	0%	0%	0%
0%	0%	0%	0%
0%	0%	0%	0%
828	783	843	844

Skew	Kurtosis	
1.76	2.93	*Whisker Hi*
0.70	1.78	*Box Hi - Q_3*
0	1.46	*Median - Q_2*
0	1	*Box Lo - Q_1*
-1.06	1	*Whisker Lo*
102	293	*Outliers*

A3.4.4 Relationship between MAF and the sequence context

MAF and the base type

Supplementary figure 5.5.3 S1. **Panel A** shows the relative base composition of the set of sites where we detect heteroplasmies using standard method at different allele filtration stringency levels. Colour scale indicates base composition deviation from MT genome bassline. **Panel B** provides an overview of what proportion of all analysed positions the heteroplasmies were found at.

Supplementary figure 5.5.3 S1: Heteroplasmic position base composition – standard method

Heteroplasmic position base composition – standard method

(a) Base composition of heteroplasmic position sets identified using the standard method

	Positions	A	C	G	T
Filter 1	16,400	31.0% *(-0.0004%)*	31.3% *(-0.01%)*	13.0% *(+0.002%)*	24.7% *(+0.005%)*
Filter 2	16,339	31.0% *(+0.02%)*	31.3% *(-0.03%)*	12.9% *(-0.03%)*	24.7% *(+0.04%)*
Filter 3	11,530	26.9% *(-4.1%)*	22.7% *(-8.7%)*	17.6% *(+4.7%)*	32.8% *(+8.2%)*
Filter 5	10,948	26.0% *(-5.0%)*	21.5% *(-9.8%)*	18.5% *(+5.5%)*	33.9% *(+9.3%)*
Filter 7	5,841	20.6% *(-10.4%)*	11.8% *(-19.6%)*	27.9% *(+14.9%)*	39.7% *(+15.0%)*

Values in brackets indicate base type frequency change from base composition of 16,403 analysed mtDNA positions

(b) Heteroplasmic fraction of all analysed mtDNA positions

	Heteroplasmic fraction	Heteroplasmic fraction of analysed positions by base type			
		A	C	G	T
Filter 1	99.98%	99.98%	99.96%	100%	100%
Filter 2	99.6%	99.7%	99.5%	99.4%	99.8%
Filter 3	70.3%	60.9%	50.8%	95.7%	93.5%
Filter 5	66.7%	56.0%	45.8%	95.2%	91.8%
Filter 7	35.6%	23.6%	13.4%	76.7%	57.3%

Supplementary figure 5.5.3 S2 summarises the average heteroplasmic position mean MAF (**A**) and median MAF (**B**) levels observed across subsets of heteroplasmic positions identified by standard method, grouped by the position base type, with the colour scale indicating the relative differences between base-specific average MAF and the overall heteroplasmic position set average MAF.

Supplementary figure 5.5.3 S2: Position base type and the average heteroplasmy MAF level – standard method

Position base type and the average heteroplasmy MAF level – standard method

(a) Average position-specific mean heteroplasmy MAF observed by standard method

Standard method	Overall average	Average median by base type							
		A		C		G		T	
Filter 1	0.028%	0.032%	*(+0.004%)*	0.018%	*(-0.010%)*	0.027%	*(-0.001%)*	0.037%	*(+0.009%)*
Filter 2	0.04%	0.04%	*(-0.004%)*	0.03%	*(-0.015%)*	0.07%	*(+0.027%)*	0.05%	*(+0.009%)*
Filter 3	1.35%	1.39%	*(+0.03%)*	1.35%	*(-0.003%)*	1.30%	*(-0.05%)*	1.36%	*(+0.004%)*
Filter 5	1.60%	1.67%	*(+0.07%)*	1.59%	*(-0.02%)*	1.50%	*(-0.10%)*	1.61%	*(+0.01%)*
Filter 7	6.73%	7.36%	*(+0.63%)*	7.21%	*(+0.47%)*	6.15%	*(-0.58%)*	6.68%	*(-0.06%)*

Values in brackets show base-type specific average mean MAF difference from the overall heteroplasmic position mean MAF average

(b) Average position-specific median heteroplasmy MAF observed by standard method

Standard method	Overall average	Average median by base type							
		A		C		G		T	
Filter 1	0.022%	0.027%	*(+0.004%)*	0.014%	*(-0.008%)*	0.015%	*(-0.007%)*	0.031%	*(+0.008%)*
Filter 2	0.025%	0.028%	*(+0.003%)*	0.020%	*(-0.005%)*	0.022%	*(-0.003%)*	0.030%	*(+0.005%)*
Filter 3	0.77%	0.97%	*(+0.20%)*	1.15%	*(+0.37%)*	0.40%	*(-0.37%)*	0.55%	*(-0.22%)*
Filter 5	0.97%	1.21%	*(+0.24%)*	1.38%	*(+0.41%)*	0.52%	*(-0.45%)*	0.77%	*(-0.20%)*
Filter 7	6.06%	6.93%	*(+0.87%)*	7.05%	*(+1.00%)*	5.01%	*(-1.06%)*	6.06%	*(-0.0003%)*

Values in brackets show base-type specific average median MAF difference from the overall heteroplasmic position median MAF average

Relationship between MAF, heteroplasmy prevalence and sequence composition

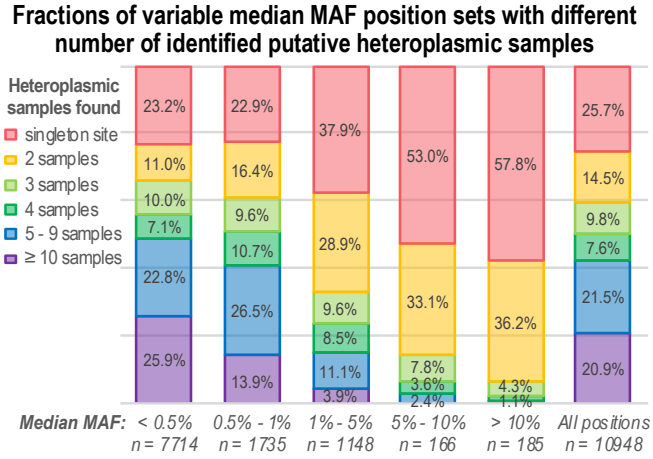
Supplementary figure 5.5.4 S1 addresses how sites with different heteroplasmy counts identified by standard method are distributed among position subsets with different median MAF. The same data obtained by consensus method is shown in **Figure 5.5.4 F1** in the main text.

Supplementary figure 5.5.4 S1: Relationship between the heteroplasmies found per site and the position-specific median MAF measured by standard method across heteroplasmies with MAF>0.1%

Relationship between the number of heteroplasmies found per site and the position-specific median MAF measured by standard method across heteroplasmies with MAF>0.1%

(a) Positions with different putative heteroplasmic sample counts found with standard method

Position median MAF	Total sites	Number of samples with MAF ≥ 0.1% 1	2	3	4	5 - 9	10+
0.1% - 0.5%	7714	1787 *23.2%*	848 *11.0%*	774 *10.0%*	544 *7.1%*	1760 *22.8%*	2001 *25.9%*
0.5% - 1%	1735	398 *22.9%*	284 *16.4%*	167 *9.6%*	185 *10.7%*	459 *26.5%*	242 *13.9%*
1% - 5%	1148	435 *37.9%*	332 *28.9%*	110 *9.6%*	98 *8.5%*	128 *11.1%*	45 *3.9%*
5% - 10%	166	88 *53.0%*	55 *33.1%*	13 *7.8%*	6 *3.6%*	4 *2.4%*	0
≥ 10%	185	107 *57.8%*	67 *36.2%*	8 *4.3%*	2 *1.1%*	0	1 *0.5%*
Total het. positions	10948	2815 *25.7%*	1586 *14.5%*	1072 *9.8%*	835 *7.6%*	2351 *21.5%*	2289 *20.9%*



Percentages indicate fraction of the total number of putative heteroplasmic positions listed in 'Total sites' column

(b) Counts of total vs singleton only putative heteroplasmies with different MAF levels at filter 5

MAF range	Consensus method All heteroplasmies	Singletons only	Standard method All heteroplasmies	Singletons only
0.1% - 0.5%	98,002 *81.2%*	1,836 *65.1%*	116,011 *76.6%*	1,787 *63.5%*
0.5% - 1%	9,344 *7.7%*	360 *12.8%*	17,944 *11.9%*	398 *14.1%*
1% - 5%	9,157 *7.6%*	416 *14.7%*	12,993 *8.6%*	435 *15.5%*
5% - 10%	1,850 *1.5%*	95 *3.4%*	1,924 *1.3%*	88 *3.1%*
≥ 10%	2,401 *2.0%*	114 *4.0%*	2,482 *1.6%*	107 *3.8%*
Total het. count	120,754	2,821	151,354	2,815

Percentages indicate fractions of the respective total found putative heteroplasmy counts listed in 'Total het. count' row

Supplementary figure 5.5.4 S2 summarises the base composition and MAF level distribution of the 1499 high heteroplasmy sites with median MAF >1% measured by standard method. The same type of data obtained by consensus method is shown in **Figure 5.5.4 F2** in the main text.

Supplementary figure 5.5.4 S2: Overview of 1,499 mtDNA positions with median MAF ≥1% by standard method

MAF distribution descriptive statistics and base composition of 1,499 mtDNA positions with putative heteroplasmy median MAF ≥ 1% found at filter 5 with standard method

(a) MAF descriptive statistic averages for sets of sites with variable putative heteroplasmy counts

Het. count	High MAF positions	High MAF positions by rCRS base type A	C	G	T	Avg. values of MAF descriptive statistics Mean	SD	Median	IQR
Any	1499	476	475	109	439	5.5%	3.5%	4.9%	3.2% - 7.6%
1	630 (42.0%)	227 (47.7%)	297 (62.5%)	13 (11.9%)	93 (21.2%)	6.5%	n/a	6.5%	n/a
2	454 (30.3%)	153 (32.1%)	141 (29.7%)	23 (21.1%)	137 (31.2%)	4.9%	5.7%	4.9%	0.9% - 8.9%
3	131 (8.7%)	41 (8.6%)	23 (4.8%)	18 (16.5%)	49 (11.2%)	5.4%	6.4%	3.3%	1.2% - 10.2%
4	106 (7.1%)	19 (4.0%)	9 (1.9%)	15 (13.8%)	63 (14.4%)	4.4%	5.7%	2.4%	0.6% - 8.2%
5	58 (3.9%)	11 (2.3%)	1 (0.2%)	10 (9.2%)	36 (8.2%)	4.3%	5.8%	2.0%	0.5% - 7.0%
6	37 (2.5%)	8 (1.7%)	1 (0.2%)	8 (7.3%)	20 (4.6%)	5.1%	7.3%	2.0%	0.4% - 7.4%
7	16 (1.1%)	2 (0.4%)	1 (0.2%)	6 (5.5%)	7 (1.6%)	5.3%	8.5%	1.6%	0.4% - 6.5%
8	9 (0.6%)	3 (0.6%)	0	1 (0.9%)	5 (1.1%)	4.4%	6.8%	1.6%	0.4% - 6.0%
9	12 (0.8%)	3 (0.6%)	1 (0.2%)	2 (1.8%)	6 (1.4%)	4.5%	7.3%	1.4%	0.4% - 5.2%
10+	46 (3.1%)	9 (1.9%)	1 (0.2%)	13 (11.9%)	23 (5.2%)	4.3%	6.5%	1.6%	0.6% - 4.5%

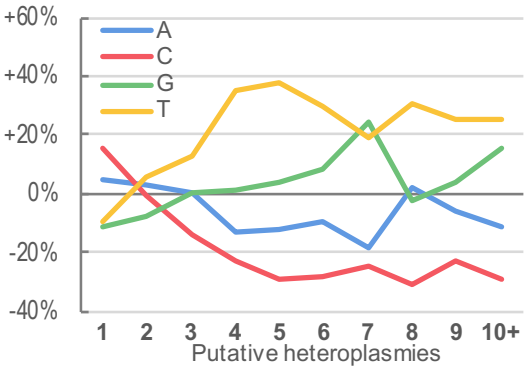
Percentages in brackets indicate relative proportions of total putative heteroplasmic positions of that base type

(b) Base composition of high MAF position subsets with different putative heteroplasmy counts

Het. count	High MAF position subset base composition A	C	G	T
Any	31.8% (+0.8%)	31.7% (+0.4%)	7.3% (-5.8%)	29.3% (+4.6%)
1	36.0% (+5.1%)	47.1% (+15.9%)	2.1% (-11.0%)	14.8% (-9.9%)
2	33.7% (+2.8%)	31.1% (-0.2%)	5.1% (-8.0%)	30.2% (+5.5%)
3	31.3% (+0.4%)	17.6% (-13.7%)	13.7% (+0.6%)	37.4% (+12.7%)
4	17.9% (-13.0%)	8.5% (-22.8%)	14.2% (+1.1%)	59.4% (+34.7%)
5	19.0% (-12.0%)	1.7% (-29.5%)	17.2% (+4.1%)	62.1% (+37.4%)
6	21.6% (-9.3%)	2.7% (-28.6%)	21.6% (+8.5%)	54.1% (+29.3%)
7	12.5% (-18.4%)	6.3% (-25.0%)	37.5% (+24.4%)	43.8% (+19.0%)
8	33.3% (+2.4%)	0% (-31.3%)	11.1% (-2.0%)	55.6% (+30.8%)
9	25.0% (-5.9%)	8.3% (-22.9%)	16.7% (+3.6%)	50.0% (+25.3%)
10+	19.6% (-11.4%)	2.2% (-29.1%)	28.3% (+15.2%)	50.0% (+25.3%)

Percentages in brackets show deviation from mtDNA baseline frequencies

Base composition deviation from MT baseline frequencies



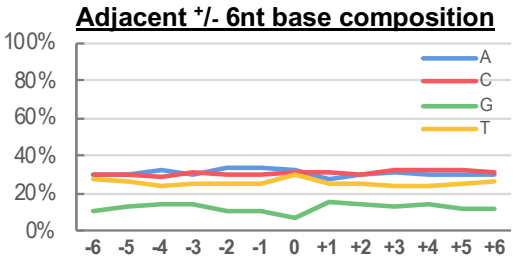
Supplementary figure 5.5.4 S3 visualises the adjacent sequence compositions around 1477 positions where high median heteroplasmy levels (median MAF ≥1%) were detected using the consensus method.

Supplementary figure 5.5.4 S3: Sequence context around high heteroplasmy positions with median MAF ≥ 1%

Adjacent sequence context base composition around 1477 high putative heteroplasmy positions with median MAF ≥ 1% found at filter 5 with consensus method

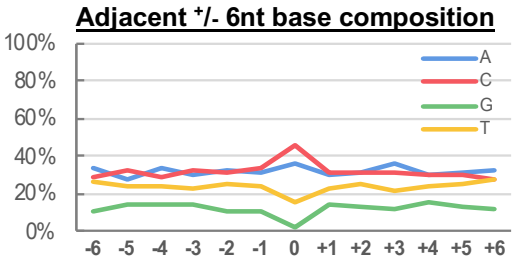
(a) Adjacent context base composition around all 1477 positions with median MAF ≥ 1%

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	-0.3%	-0.5%	+1.7%	-0.9%	+3.1%	+2.7%	+1.1%	-2.7%	-1.0%	0%	-1.4%	-0.5%	-1.3%
C	-0.9%	-0.7%	-2.6%	-0.5%	-1.6%	-1.1%	-0.3%	+0.2%	-1.0%	+1.4%	+1.3%	+0.7%	-0.1%
G	-1.9%	-0.4%	+1.7%	+0.7%	-1.9%	-2.7%	-5.7%	+2.4%	+1.3%	-0.3%	+0.7%	-0.7%	-0.8%
T	+3.0%	+1.5%	-0.8%	+0.6%	+0.4%	+1.1%	+4.9%	+0.1%	+0.7%	-1.1%	-0.6%	+0.5%	+2.3%



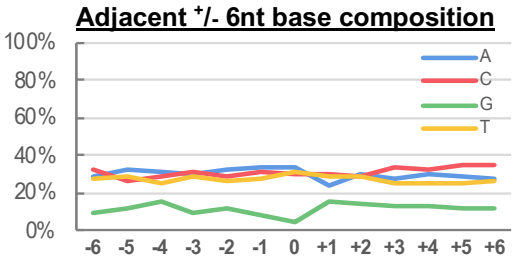
(b) Base composition around 625 high MAF singleton putative heteroplasmy sites only

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	+2.5%	-2.8%	+3.0%	-0.8%	+1.7%	+0.4%	+5.7%	-0.4%	+0.9%	+5.4%	-1.3%	+0.4%	+1.1%
C	-2.0%	+1.7%	-2.8%	+0.7%	+0.2%	+2.5%	+15%	+0.2%	-0.6%	-0.6%	-0.7%	-1.0%	-3.1%
G	-2.2%	+1.3%	+1.0%	+1.6%	-2.7%	-2.4%	-11%	+1.6%	-0.5%	-1.6%	+2.3%	-0.5%	-0.6%
T	+1.7%	-0.2%	-1.2%	-1.5%	+0.7%	-0.6%	-9.7%	-1.5%	+0.1%	-3.3%	-0.2%	+1.0%	+2.6%



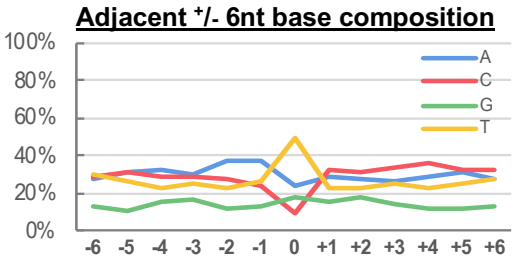
(c) Base composition around 442 high MAF sites with 2 putative heteroplasmies

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	-1.7%	+2.1%	-0.2%	-0.4%	+1.7%	+2.6%	+2.3%	-6.5%	-1.3%	-3.1%	-1.3%	-2.2%	-3.1%
C	+1.8%	-4.6%	-2.1%	-0.3%	-2.3%	-0.5%	-1.0%	-1.2%	-3.0%	+2.9%	+0.6%	+3.1%	+3.3%
G	-3.4%	-1.1%	+1.8%	-3.1%	-1.6%	-5.4%	-8.1%	+3.0%	+0.7%	0%	+0.5%	-1.1%	-1.8%
T	+3.3%	+3.6%	+0.4%	+3.8%	+2.2%	+3.3%	+6.7%	+4.7%	+3.6%	+0.2%	+0.2%	+0.2%	+1.5%



(d) Base composition around 410 high MAF sites with at least 3 putative heteroplasmies

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	-3.1%	+0.3%	+1.8%	-1.4%	+6.9%	+6.1%	-7.3%	-2.1%	-3.6%	-4.8%	-1.7%	-0.2%	-3.1%
C	-2.0%	-0.1%	-2.7%	-2.5%	-3.7%	-7.1%	-22%	+1.7%	+0.4%	+2.9%	+5.1%	+0.7%	+0.7%
G	+0.3%	-2.1%	+2.5%	+3.5%	-1.1%	-0.2%	+4.5%	+3.0%	+4.7%	+1.3%	-1.4%	-0.7%	-0.2%
T	+4.8%	+1.9%	-1.5%	+0.4%	-2.0%	+1.1%	+25%	-2.5%	-1.5%	+0.7%	-2.0%	+0.2%	+2.6%



Percentages show base frequency deviation from MT genome overall composition of 30.9% A, 31.3% C, 13.1% G and 24.7% T

Supplementary figure 5.5.4 S4 lists 48 positions with high average MAF (median MAF ≥1% by consensus method) where heteroplasmy is also most prevalent (10 or more heteroplasmies were detected). This subset of positions represents the most commonly observed high MAF heteroplasmic sites.

Supplementary figure 5.5.4 S4: High heteroplasmy positions 10+ heteroplasmies and median MAF≥1%

Most common high heteroplasmy positions with median MAF above 1% and at least 10 identified heteroplasmic samples with MAF ≥0.1%

	Position	rCRS base	Upstream context L strand	Upstream context H strand	Samples with MAF above 0.1%	0.5%	1%	5%	10%	MAF level Median	MAF level IQR	Prev. lit. reports	Genomic context
1:	152	T	TCATCC **T**	AATAAT **A**	13	13	10	5	3	4.02%	1.44% - 12.0%	*11 (1915)*	CR
2:	1642	G	ACTTAG **G**	AAATCT **C**	11	9	7	5	3	2.78%	0.70% - 13.6%	*5 (10)*	TV
3:	199	T	ACTTAC **T**	CACTTT **A**	16	14	9	2	2	2.39%	0.71% - 3.83%	*9 (273)*	CR
4:	16093	T	ATGTAT **T**	GTACGA **A**	237	170	149	69	45	2.22%	0.35% - 7.25%	*15 (6656)*	CR
5:	2510	T	AAAACA **T**	GAGGTG **A**	10	7	7	2	1	2.08%	0.23% - 2.90%	*7 (14)*	RNR2
6:	748*	G	TCAAAA **G**	TTGTTC **C**	13	9	7	1	1	1.90%	0.27% - 3.35%	*3 (15)*	RNR1
7:	195	T	ACATAC **T**	TTAGTA **A**	53	50	35	17	11	1.86%	0.79% - 7.90%	*14 (1348)*	CR
8:	16368*	T	TCCCCA **T**	TCATCC **A**	12	11	8	2	1	1.85%	0.61% - 3.84%	*8 (91)*	CR
9:	14841	A	GATGAA **A**	CCGAAG **T**	10	5	5	2	2	1.69%	0.27% - 3.96%	*7 (14)*	CYB
10:	214	A	TTAATT **A**	ATTAAT **T**	17	15	11	3	3	1.65%	0.63% - 2.99%	*10 (844)*	CR
11:	1215	T	CTGTAA **T**	TTATCG **A**	10	7	6	2	0	1.58%	0.19% - 2.69%	*3 (4)*	RNR1
12:	16183	A	ATCAAA **A**	AGGGGG **T**	24	17	14	8	2	1.57%	0.35% - 5.32%	*9 (5617)*	CR
13:	3736	G	TATGAA **G**	GGGTGA **C**	15	9	9	2	1	1.54%	0.17% - 3.76%	*6 (42)*	ND1
14:	234	A	GGACAT **A**	TATTAT **T**	41	33	24	13	8	1.50%	0.58% - 7.97%	*11 (601)*	CR
15:	189	A	AGGCGA **A**	AGTATG **T**	64	61	42	15	9	1.47%	0.86% - 3.00%	*12 (633)*	CR
16:	200*	A	CTTACT **A**	ACACTT **T**	19	19	12	2	1	1.45%	0.68% - 2.29%	*9 (346)*	CR
17:	15834	T	CAATCC **T**	AGGATT **A**	13	8	7	1	0	1.45%	0.26% - 2.01%	*4 (7)*	CYB
18:	3143	T	GAGAAA **T**	GGCCTT **A**	13	7	7	2	2	1.38%	0.22% - 2.30%	*4 (6)*	RNR2
19:	12005	T	ACACAA **T**	AGCCCC **A**	15	8	8	3	3	1.37%	0.18% - 4.11%	*7 (19)*	ND4
20:	14547	T	TCAGAA **T**	GTTATT **A**	14	10	8	4	2	1.36%	0.34% - 5.42%	*4 (8)*	ND6
21:	6718	G	ACATAG **G**	ACCATA **C**	11	7	7	3	2	1.35%	0.26% - 5.38%	*6 (9)*	CO1
22:	12589	T	GACTAC **T**	TGGAGA **A**	11	9	7	1	1	1.30%	0.66% - 1.84%	*1 (1)*	ND5
23:	204*	T	CTAAAG **T**	TAACAC **A**	98	88	60	22	13	1.29%	0.79% - 4.12%	*14 (1117)*	CR
24:	194	C	AACATA **C**	TAGTAA **G**	14	14	8	4	3	1.25%	0.77% - 7.12%	*10 (212)*	CR
25:	7925	G	GACTAC **G**	GTCCGC **C**	17	10	9	4	4	1.23%	0.21% - 5.05%	*9 (25)*	CO2
26:	16391	G	ATAGGG **G**	AAGGGA **C**	14	14	8	2	1	1.23%	0.68% - 2.39%	*8 (226)*	CR
27:	12667	G	AACTCA **G**	TTGGGT **C**	17	10	9	4	2	1.22%	0.29% - 4.15%	*4 (10)*	ND5
28:	374	A	ACCCTA **A**	CTGGTG **T**	15	10	8	3	1	1.20%	0.39% - 3.91%	*6 (42)*	CR
29:	2075	T	TCTAAA **T**	AAGGGG **A**	19	11	11	1	1	1.20%	0.21% - 2.59%	*6 (30)*	RNR2
30:	15317	G	ATTGCA **G**	CTAGGG **C**	10	7	5	1	1	1.20%	0.28% - 2.04%	*6 (21)*	CYB
31:	941	G	CGTAAA **G**	AACACT **C**	10	7	5	2	1	1.18%	0.24% - 2.37%	*3 (5)*	RNR1
32:	251*	G	TTGAAT **G**	TGCAGA **C**	25	19	14	5	2	1.13%	0.53% - 3.90%	*9 (58)*	CR
33:	279	T	CAGACA **T**	GTTATG **A**	26	20	13	5	4	1.11%	0.54% - 3.00%	*5 (10)*	CR
34:	8743	G	ATACTA **G**	AGGATA **C**	14	8	8	3	1	1.11%	0.20% - 2.27%	*8 (68)*	ATP6
35:	11154	T	TGGCTA **T**	GTGATG **A**	10	6	6	2	1	1.10%	0.23% - 2.56%	*3 (5)*	ND4
36:	2371	T	TGACAA **T**	CTGTTA **A**	11	7	6	3	2	1.09%	0.31% - 5.76%	*3 (10)*	RNR2
37:	11361	T	ACACAA **T**	AAAGCT **A**	12	8	6	4	3	1.09%	0.13% - 12.2%	*5 (5)*	ND4
38:	8485	G	ACCAAA **G**	TATGGG **C**	10	6	5	0	0	1.08%	0.21% - 2.44%	*5 (26)*	ATP8 [w]
39:	249*	A	AATTGA **A**	CAGACA **T**	14	11	7	3	2	1.06%	0.55% - 3.60%	*7 (31)*	CR
40:	16390*	G	GATAGG **G**	AGGGAC **C**	57	53	31	7	6	1.05%	0.72% - 2.60%	*11 (395)*	CR
41:	1435	A	CAGTAA **A**	TCTTAG **T**	19	13	10	1	1	1.04%	0.21% - 1.93%	*7 (12)*	RNR1
42:	4752*	T	TACTCA **T**	TTAATG **A**	18	13	9	1	1	1.03%	0.33% - 1.49%	*5 (12)*	ND2

** positions with median MAF ≥1% by consensus method only; Previous reports in literature shown an 'studies (heteroplasmies reported)'*
Genomic context: control region , rRNA , tRNA , non-coding , protein ; [w] – codon 3rd wobble position in protein coding genes

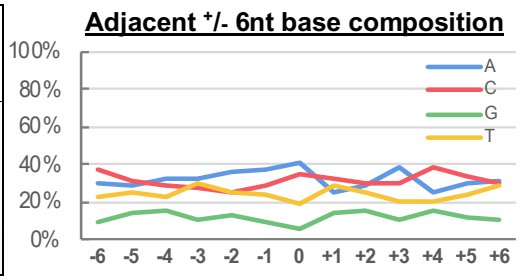
Supplementary figure 5.5.4 S5 visualises the adjacent sequence compositions around 191 positions where very high average heteroplasmy levels (median MAF ≥10%) were detected using the consensus method.

Supplementary figure 5.5.4 S5: Sequence context of 191 very high heteroplasmy positions with median MAF ≥10%

Adjacent sequence context base composition around 191 very high putative heteroplasmy positions with median MAF ≥ 10% found at filter 5 with consensus method

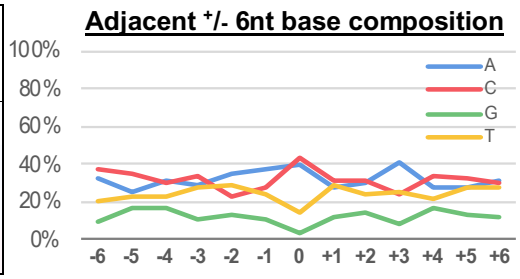
(a) Adjacent context base composition changes around all positions with median MAF ≥ 10%

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	-1.1%	-2.1%	+1.0%	+1.0%	+4.7%	+6.8%	+9.9%	-5.8%	-2.1%	+7.8%	-5.8%	-0.6%	0%
C	+6.4%	-0.4%	-2.0%	-4.0%	-6.1%	-2.5%	+3.8%	+0.7%	-1.4%	-1.4%	+6.9%	+2.8%	-0.9%
G	-3.1%	+1.6%	+2.6%	-2.1%	+0.5%	-3.1%	-7.9%	+1.0%	+2.6%	-2.1%	+2.6%	-1.6%	-2.6%
T	-2.2%	+0.9%	-1.7%	+5.1%	+0.9%	-1.2%	-5.9%	+4.1%	+0.9%	-4.3%	-3.8%	-0.6%	+3.6%



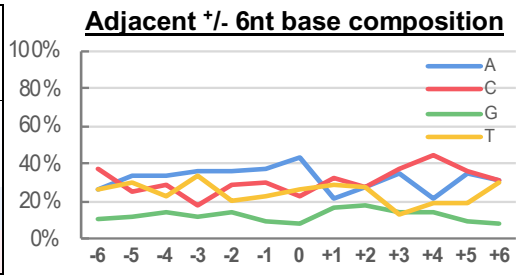
(b) Base composition changes around the subset of 114 singleton positions

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	+1.5%	-5.5%	-0.2%	-2.0%	+4.2%	+6.8%	+8.5%	-3.7%	-1.1%	+10%	-3.7%	-3.7%	-0.2%
C	+6.4%	+3.8%	-1.4%	+2.1%	-8.5%	-3.2%	+12%	+0.3%	+0.3%	-6.7%	+2.9%	+1.2%	-1.4%
G	-3.4%	+3.6%	+3.6%	-2.6%	+0.1%	-2.6%	-9.6%	-0.8%	+0.9%	-4.3%	+3.6%	+0.1%	-0.8%
T	-4.5%	-1.9%	-1.9%	+2.5%	+4.2%	-1.0%	-11%	+4.2%	-0.1%	+0.7%	-2.8%	+2.5%	+2.5%



(c) Base composition changes around the subset of 77 non-singleton positions

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	-5.0%	+2.8%	+2.8%	+5.4%	+5.4%	+6.7%	+12%	-8.8%	-3.7%	+4.1%	-8.8%	+4.1%	+0.2%
C	+6.4%	-6.6%	-2.7%	-13%	-2.7%	-1.4%	-7.9%	+1.2%	-4.0%	+6.4%	+13%	+5.1%	-0.1%
G	-2.7%	-1.4%	+1.2%	-1.4%	+1.2%	-4.0%	-5.3%	+3.8%	+5.1%	+1.2%	+1.2%	-4.0%	-5.3%
T	+1.3%	+5.2%	-1.3%	+9.1%	-3.9%	-1.3%	+1.3%	+3.9%	+2.6%	-12%	-5.2%	-5.2%	+5.2%



Percentages show base frequency deviation from MT genome overall composition of 30.9% A, 31.3% C, 13.1% G and 24.7% T

Positions associated with largest number of high MAF heteroplasmies

Supplementary figure 5.5.5 S1 addresses the positions with highest number of observed high MAF samples identified by consensus method. **Panel A** shows how many MT genome positions were found with different number of samples above 0.1%, 0.5%, 1%, 5% and 10% MAF thresholds. **Panel B** lists 20 positions where we found at least 5 samples with MAF>10%.

Supplementary figure 5.5.5 S1: Heteroplasmic positions with samples above 0.1%, 0.5%, 1%, 5% and 10% MAF

Heteroplasmic positions with samples above 0.1%, 0.5%, 1%, 5% and 10% MAF thresholds

(b) Number of positions with different sample counts found with MAF above specific thresholds

Samples with MAF above threshold	Heteroplasmy MAF threshold									
	≥ 0.1% *Filter 5 lowest MAF limit*		≥ 0.5%		≥ 1% *'High heteroplasmy'*		≥ 5%		≥ 10% *'Very high heteroplasmy'*	
≥ 1	10,804	*(100%)*	7,380	*(100%)*	5,843	*(100%)*	2,824	*(100%)*	1,808	*(100%)*
≥ 2	7,983	*(73.9%)*	4,335	*(49.3%)*	2,879	*(49.3%)*	786	*(27.8%)*	344	*(19.0%)*
≥ 3	6,393	*(59.2%)*	2,845	*(26.7%)*	1,559	*(26.7%)*	266	*(9.42%)*	**90**	*(4.98%)*
≥ 4	5,311	*(49.2%)*	1,901	*(15.8%)*	922	*(15.8%)*	113	*(4.00%)*	42	*(2.32%)*
≥ 5	4,481	*(41.5%)*	1,274	*(9.29%)*	**543**	*(9.29%)*	**57**	*(2.02%)*	**20**	*(1.11%)*
≥ 10	2,163	*(20.0%)*	286	*(1.51%)*	88	*(1.51%)*	11	*(0.39%)*	5	*(0.28%)*
≥ 50	203	*(1.88%)*	13	*(0.03%)*	2	*(0.03%)*	1	*(0.04%)*	0	
≥ 100	107	*(0.99%)*	4	*(0.02%)*	1	*(0.02%)*	0		0	
≥ 500	34	*(0.31%)*	0		0		0		0	
≥ 1000	16	*(0.15%)*	0		0		0		0	

Percentages indicate fraction of total positions where any number of samples above that specific MAF threshold were found; Sets of positions highlighted in red are selected for further investigation

(b) Top 20 positions with 5 or more very high heteroplasmy samples found with MAF ≥ 10%

Position	rCRS base	Upstream context L strand	Upstream context H strand	Samples with MAF above 0.1%	0.5%	1%	5%	10%	Heteroplasmy MAF distribution Median	IQR	Mean	SD	Prev. lit. reports	Genomic context
16093*	T	ATGTAT **T**	GTACGA **A**	237	170	149	69	45	2.22%	0.35% - 7.25%	6.28%	9.83%	*15 (6656)*	CR
16192	C	CCCTCC **C**	AGCATG **G**	128	58	39	21	16	0.39%	0.19% - 2.01%	3.88%	9.02%	*9 (928)*	CR
16129	G	TTGTAC **G**	TGGTAC **C**	413	48	31	17	14	0.16%	0.12% - 0.26%	1.26%	5.45%	*12 (730)*	CR
204*	T	CTAAAG **T**	TAACAC **A**	98	88	60	22	13	1.29%	0.79% - 4.12%	5.72%	11.13%	*14 (1117)*	CR
195*	T	ACATAC **T**	TTAGTA **A**	53	50	35	17	11	1.86%	0.79% - 7.90%	8.15%	12.29%	*14 (1348)*	CR
189*	A	AGGCGA **A**	AGTATG **T**	64	61	42	15	9	1.47%	0.86% - 3.00%	4.72%	8.15%	*12 (633)*	CR
709	G	ATCCCC **G**	CTGGAA **C**	205	50	34	12	9	0.23%	0.15% - 0.49%	1.59%	5.66%	*9 (420)*	RNR1
234*	A	GGACAT **A**	TATTAT **T**	41	33	24	13	8	1.50%	0.58% - 7.97%	5.99%	8.89%	*11 (601)*	CR
2623	A	ATAGGG **A**	TACAGG **T**	224	83	48	12	8	0.32%	0.17% - 0.72%	1.52%	4.42%	*10 (65)*	RNR2
16311	T	ACATAG **T**	TTATGT **A**	77	35	23	12	8	0.46%	0.29% - 1.53%	3.54%	8.21%	*12 (788)*	CR
16092	T	TATGTA **T**	TACGAA **A**	53	21	15	8	6	0.31%	0.16% - 1.47%	4.33%	10.36%	*11 (367)*	CR
16390*	G	GATAGG **G**	AGGGAC **C**	57	53	31	7	6	1.05%	0.72% - 2.60%	4.19%	9.44%	*11 (395)*	CR
5970	G	TTATTC **G**	ATGCGC **C**	63	23	14	7	6	0.39%	0.21% - 0.80%	2.54%	7.19%	*7 (17)*	CO1
10398	A	GACTGA **A**	ATTCGG **T**	139	19	7	7	6	0.20%	0.14% - 0.32%	1.26%	5.01%	*13 (466)*	ND3
16184	C	TCAAAA **C**	GAGGGG **G**	25	11	9	6	6	0.41%	0.25% - 4.44%	5.77%	11.08%	*3 (110)*	CR
16148	C	CTTGAC **C**	ACAGGT **G**	43	11	8	6	6	0.21%	0.13% - 0.59%	3.97%	10.32%	*10 (183)*	CR
16189	T	ACCCCC **T**	ATGGGG **A**	304	188	30	8	5	0.54%	0.39% - 0.65%	1.05%	3.19%	*6 (1082)*	CR
3244	G	GGCAGA **G**	ACCGGG **C**	199	68	36	7	5	0.26%	0.16% - 0.70%	0.97%	2.49%	*7 (30)*	TL1
16362	T	TTCTCG **T**	ATGGGG **A**	29	23	10	5	5	0.83%	0.54% - 1.41%	6.95%	14.05%	*12 (484)*	CR
11690	G	TTCACC **G**	CTGCGC **C**	29	15	8	5	5	0.51%	0.16% - 1.57%	5.58%	12.40%	*6 (11)*	ND4

*IQR – interquartile range; SD – standard deviation; *Positions with heteroplasmy median MAF ≥1%; Previous reports as 'papers (samples)' Genomic context: control region, rRNA, tRNA, non-coding, protein; [w] – codon 3rd wobble position in protein coding genes*

Supplementary figure 5.5.5 S2 visualises the adjacent sequence contexts around high MAF position sets where we find at least 5 samples with MAF ≥ 1% (**A**, n=543), ≥5% (**B**, n=57) and ≥10% (**C**, n=20) using the consensus method.

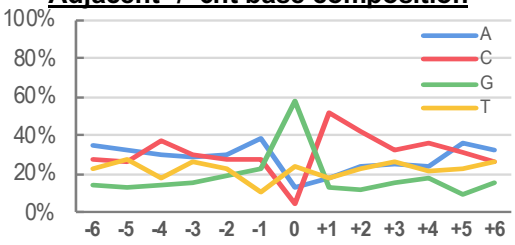
Supplementary figure 5.5.5 S2: Adjacent context of high heteroplasmy sites (5+ samples with MAF ≥1%, 5%, 10%)

Adjacent sequence context of high heteroplasmy (5+ samples with MAF ≥1%, 5% or 10%) sites

(a) Base composition around 543 positions with 5 or more samples with MAF ≥ 1%

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	+3.5%	+2.0%	-1.1%	-2.0%	-0.9%	+7.7%	-17%	-13%	-6.4%	-5.1%	-6.8%	+5.7%	+1.5%
C	-3.1%	-4.9%	+6.1%	-1.4%	-3.3%	-4.2%	-27%	+20%	+10%	+1.3%	+5.2%	-0.1%	-4.6%
G	+0.9%	+0.4%	+1.3%	+2.0%	+5.9%	+10%	+45%	-0.6%	-1.7%	+2.2%	+4.4%	-3.3%	+1.8%
T	-1.3%	+2.5%	-6.3%	+1.4%	-1.7%	-14%	-1.0%	-6.5%	-2.2%	+1.6%	-2.8%	-2.2%	+1.3%

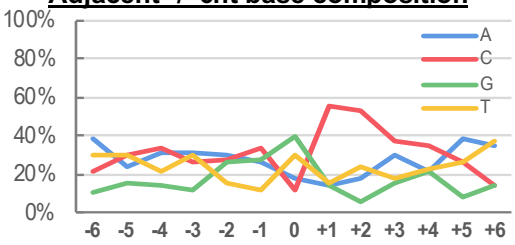
Adjacent +/- 6nt base composition



(b) Base composition around 57 positions with 5 of more samples with MAF ≥ 5%

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	+7.7%	-6.4%	+0.7%	+0.7%	-1.1%	-4.6%	-13%	-17%	-13%	-1.1%	-9.9%	+7.7%	+4.2%
C	-10%	-1.4%	+2.1%	-5.0%	-3.2%	+2.1%	-19%	+25%	+21%	+5.6%	+3.8%	-5.0%	-17%
G	-2.6%	+2.7%	+0.9%	-0.8%	+13%	+15%	+27%	+0.9%	-7.8%	+2.7%	+8.0%	-4.3%	+0.9%
T	+5.1%	+5.1%	-3.7%	+5.1%	-8.9%	-12%	+5.1%	-8.9%	-0.1%	-7.2%	-1.9%	+1.6%	+12%

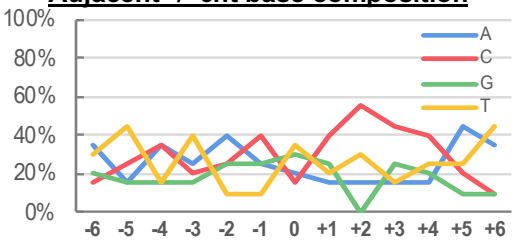
Adjacent +/- 6nt base composition



(c) Base composition around 20 positions with 5 or more samples with MAF ≥ 10%

rCRS base	Base frequency deviation from mtDNA baseline												
	Upstream positions							Downstream positions					
	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	+4.1%	-16%	+4.1%	-5.9%	+9.1%	-5.9%	-11%	-16%	-16%	-16%	-16%	+14%	+4.1%
C	-16%	-6.3%	+3.7%	-11%	-6.3%	+8.7%	-16%	+8.7%	+24%	+14%	+8.7%	-11%	-21%
G	+6.9%	+1.9%	+1.9%	+1.9%	+12%	+12%	+17%	+12%	-13.1%	+12%	+6.9%	-3.1%	-3.1%
T	+5.3%	+20%	-9.7%	+15%	-15%	-15%	+10%	-4.7%	+5.3%	-9.7%	+0.3%	+0.3%	+20%

Adjacent +/- 6nt base composition



Percentages show base frequency deviation from MT genome overall composition of 30.9% A, 31.3% C, 13.1% G and 24.7% T

A3.4.5 Overlap between MAF levels and previous heteroplasmy reports in literature

Supplementary figures below address how well heteroplasmy report frequency in literature match our observed MAF levels in our dataset. In both figures we compare the literature data against the positions with highest MAF in our data based on one of 13 heteroplasmy level measures. These heteroplasmy measures used are the median and mean MAF, as well as the number of samples found with MAF ≥0.1%, 0.5%, 1%, 5% and 10% thresholds. **Supplementary figure 5.5.6 S1** addresses the overlap between positions with different heteroplasmy report frequency in previous reports in literature with highest heteroplasmy sites in our data by either the average MAF or the number of high MAF samples. **Supplementary figure 5.5.6 S2** addresses the correlation between previous heteroplasmy report frequency in literature and the different heteroplasmy levels we used for shortlisting the 13 sets of highest heteroplasmy sites addressed in **S1**. **Panel A** shows correlations between these variables and number of papers reporting heteroplasmy at the sites if each group, whereas **panel B** shows the correlation coefficients with the total number of heteroplasmies reported at a site across all studies.

Supplementary figure 5.5.6 S1: Observed MAF comparison with heteroplasmy report frequency in literature

Comparison between putative heteroplasmic position MAF levels found at filter 5 with consensus method and previous heteroplasmy report frequency in literature

(a) Number of papers that different groups of sites have been previously reported in

	Group of putative heteroplasmic positions	Number of sites	Sites previously reported in different number of papers						Papers reported in			
			Any	1	2 to 3	4 to 5	6 to 9	10+	mean	SD	median	IQR
1:	All putative heteroplasmic positions	10,804	77.4%	21.4%	31.0%	15.6%	9.0%	0.4%	2.3	2.1	2	3
2:	With median MAF ≥1%	1,477	79.4%	23.2%	33.4%	14.9%	7.2%	0.7%	2.3	2.1	2	2
3:	2+ het. samples, median MAF ≥1%	852	85.4%	19.2%	34.4%	19.5%	11.2%	1.2%	2.8	2.3	2	3
4:	3+ het. samples, median MAF ≥1%	410	90.7%	14.9%	31.0%	24.1%	18.5%	2.2%	3.5	2.6	3	3
5:	5+ het. samples, median MAF ≥1%	180	92.8%	8.3%	24.4%	28.3%	26.7%	5.0%	4.4	3.0	4	4
6:	10+ het. samples, median MAF ≥1%	42	100%	2.4%	11.9%	21.4%	42.9%	21.4%	7.1	3.3	7	4
7:	With median MAF ≥10%	191	85.3%	16.2%	38.7%	19.9%	10.5%	0%	2.7	2.0	2	3
8:	2+ het. samples, median MAF ≥10%	77	93.5%	18.2%	37.7%	22.1%	15.6%	0%	3.1	2.1	3	2
9:	3+ het. samples, median MAF ≥10%	11	90.9%	9.1%	9.1%	36.4%	36.4%	0%	4.6	2.5	5	2.5
10:	5+ het. samples with MAF ≥1%	543	99.1%	1.8%	16.8%	28.2%	45.9%	6.4%	5.7	2.5	6	3
11:	5+ het. samples with MAF ≥5%	57	100%	0%	3.5%	7.0%	54.4%	35.1%	8.6	2.7	8	3
12:	5+ het. samples with MAF ≥10%	20	100%	0%	5.0%	0%	30.0%	65.0%	10.2	3.1	11	3.5
13:	3+ het. samples with MAF ≥10%	90	100%	1.1%	8.9%	12.2%	52.2%	25.6%	7.5	3.0	7	3.75

(b) Number of individuals that different groups of sites have been previously reported in

	Group of putative heteroplasmic positions	Number of sites	Sites previously reported in different number of samples						Samples reported in			
			Any	1	2 to 9	10 to 99	100 to 999	1000+	mean	SD	median	IQR
1:	All putative heteroplasmic positions	10,804	77.4%	14.9%	41.7%	19.2%	1.6%	0.1%	13.1	104.8	3	7
2:	With median MAF ≥1%	1,477	79.4%	17.3%	48.0%	12.0%	1.6%	0.5%	24.5	263.2	2	4
3:	2+ het. samples, median MAF ≥1%	852	85.4%	14.6%	50.0%	17.4%	2.6%	0.9%	40.0	345.6	3	7
4:	3+ het. samples, median MAF ≥1%	410	90.7%	11.2%	47.6%	25.9%	4.6%	1.5%	65.5	454.8	5	10
5:	5+ het. samples, median MAF ≥1%	180	92.8%	6.1%	43.3%	31.1%	8.9%	3.3%	137.9	680.3	7	17.25
6:	10+ het. samples, median MAF ≥1%	42	100%	2.4%	19.0%	47.6%	19.0%	11.9%	495.1	1345.6	25.5	251.3
7:	With median MAF ≥10%	191	85.3%	14.7%	59.2%	11.0%	0.5%	0%	5.9	12.2	3	4
8:	2+ het. samples, median MAF ≥10%	77	93.5%	16.9%	61.0%	14.3%	1.3%	0%	8.0	15.5	3	5
9:	3+ het. samples, median MAF ≥10%	11	90.9%	9.1%	45.5%	36.4%	0%	0%	16.5	20.5	7	19
10:	5+ het. samples with MAF ≥1%	543	99.1%	1.1%	29.7%	54.5%	12.3%	1.5%	95.5	409.9	17	50
11:	5+ het. samples with MAF ≥5%	57	100%	0%	3.5%	45.6%	36.8%	14.0%	514.5	1162.8	110	447
12:	5+ het. samples with MAF ≥10%	20	100%	0%	0%	20.0%	60.0%	20.0%	821.6	1428.8	475	658.3
13:	3+ het. samples with MAF ≥10%	90	100%	0%	14.4%	51.1%	27.8%	6.7%	274.5	764.3	59	179.5

Supplementary figure 5.5.6 S2: Correlation between previous heteroplasmy report frequency and our observed average MAF levels or sample counts above different MAF thresholds

Correlation between previous heteroplasmy report frequency in literature and our average identified MAF levels or sample counts with MAF above certain thresholds at putative heteroplasmic positions found at filter 5 by consensus method

(a) Correlation between MAF level measures and number of studies sites are previously reported in

Group of putative heteroplasmic positions	Number of sites	MAF level		Samples with MAF above				
		median	mean	0.1%	0.5%	1%	5%	10%
1: All putative heteroplasmic positions	10,804	0.02	0.15	0.10	0.41	0.44	0.40	0.37
2: With median MAF ≥1%	1,477	0.09	0.18	0.41	0.42	0.40	0.40	0.39
3: 2+ het. samples, median MAF ≥1%	852	0.06	0.19	0.42	0.44	0.42	0.41	0.40
4: 3+ het. samples, median MAF ≥1%	410	0.10	0.21	0.46	0.49	0.46	0.45	0.44
5: 5+ het. samples, median MAF ≥1%	180	0.00	0.16	0.53	0.57	0.52	0.49	0.49
6: 10+ het. samples, median MAF ≥1%	42	0.29	0.46	0.64	0.70	0.64	0.62	0.62
7: With median MAF ≥10%	191	0.14	0.16	0.25	0.30	0.28	0.28	0.21
8: 2+ het. samples, median MAF ≥10%	77	0.16	0.23	0.28	0.35	0.32	0.34	0.24
9: 3+ het. samples, median MAF ≥10%	11	0.11	0.51	0.04	0.17	0.22	0.46	n/a
10: 5+ het. samples with MAF ≥1%	543	0.01	0.10	0.41	0.38	0.40	0.41	0.39
11: 5+ het. samples with MAF ≥5%	57	-0.09	0.15	0.25	0.32	0.48	0.50	0.50
12: 5+ het. samples with MAF ≥10%	20	0.56	0.30	0.06	0.14	0.46	0.51	0.48
13: 3+ het. samples with MAF ≥10%	90	0.13	-0.05	0.35	0.39	0.49	0.49	0.47

(b) Correlation between MAF level measures and number of samples sites are previously reported in

Group of putative heteroplasmic positions	Number of sites	MAF level		Samples with MAF above				
		median	mean	0.1%	0.5%	1%	5%	10%
1: All putative heteroplasmic positions	10,804	0.00	0.03	0.06	0.43	0.54	0.56	0.52
2: With median MAF ≥1%	1,477	-0.03	0.00	0.68	0.66	0.69	0.69	0.65
3: 2+ het. samples, median MAF ≥1%	852	-0.03	0.01	0.68	0.66	0.69	0.70	0.66
4: 3+ het. samples, median MAF ≥1%	410	-0.01	0.03	0.75	0.73	0.76	0.78	0.73
5: 5+ het. samples, median MAF ≥1%	180	0.08	0.08	0.75	0.72	0.75	0.78	0.74
6: 10+ het. samples, median MAF ≥1%	42	0.34	0.20	0.73	0.70	0.73	0.77	0.73
7: With median MAF ≥10%	191	-0.06	-0.05	0.21	0.16	0.19	0.21	0.17
8: 2+ het. samples, median MAF ≥10%	77	0.03	0.05	0.22	0.11	0.18	0.22	0.17
9: 3+ het. samples, median MAF ≥10%	11	0.22	0.32	0.03	-0.07	0.08	0.36	n/a
10: 5+ het. samples with MAF ≥1%	543	0.23	0.19	0.33	0.55	0.68	0.71	0.66
11: 5+ het. samples with MAF ≥5%	57	0.24	0.27	0.22	0.48	0.65	0.72	0.65
12: 5+ het. samples with MAF ≥10%	20	0.67	0.31	0.29	0.65	0.91	0.96	0.96
13: 3+ het. samples with MAF ≥10%	90	0.27	0.12	0.41	0.67	0.85	0.90	0.89

All correlations are expressed with Pearson correlation coefficients

A3.4.6 MAF at the positions with highest heteroplasmy prevalence in population

Supplementary figures 5.5.6 S3 and S4 address the observed frequencies of high MAF heteroplasmies at positions where highest overall heteroplasmy incidence was observed with consensus method, discussed in the earlier section 5.4. **Supplementary figures 5.5.6 S3** focuses on the 107 positions where 100+ heteroplasmies ≥0.1% MAF were found, **Supplementary figures 5.5.6 S4** focuses on the 88 positions with 10+ heteroplasmies ≥1%MAF. For each position listed in the figures, the observed heteroplasmy mean and median MAF levels as well as the number of samples found above different MAF thresholds are shown.

Supplementary figure 5.5.6 S3 shows that out of the 107 positions with at least 100 heteroplasmies ≥0.1% MAF, high heteroplasmy levels (≥1% MAF) in at least 5 samples are observed at 17 positions. Of these, very high heteroplasmy levels (≥10% MAF) in at least 5 samples are found at 7 positions. Only one site, 16093T, reaches ≥1% median MAF level across all heteroplasmies however, and none reach median heteroplasmy MAF≥10%.

Supplementary figure 5.5.6 S4 shows that that out of the 88 positions with at least 10 heteroplasmies ≥1% MAF, we see 16 sites with 5 or more samples with very high MAF levels ≥10%. In terms of median MAF level, if all identified heteroplasmies ≥0.1% MAF are considered, none of the 88 positions reach very high 10% median MAF, level, although 14 do reach 1% median MAF. Of these, 6 positions (189A, 195T, 204T, 234A, 16093T, 16390G) have both median MAF ≥ 1% and at least 5 samples with MAF ≥ 10%.

For comparison, out of the total 10804 positions with putative heteroplasmies identified at filter 5, based on position median putative heteroplasmy median MAF, 1477 sites or 13.7% meet the minimum 1% median MAF definition for high heteroplasmy level with 191 positions or 1.8% matching the very high heteroplasmy level definition of at least 10% median MAF. In terms of the number of individual samples with high MAF levels identified, 543 sites or 5% meet the high heteroplasmy definition of at least 5 samples found with MAF ≥1%, with 20 positions meeting the very high heteroplasmy level definition of minimum 5 samples with MAF ≥10%.

When we compare this to what fraction of highest heteroplasmy incidence positions that match these same high and very high heteroplasmy site definitions, at filter 5 out of 107 highest putative heteroplasmy incidence sites 15.9% match high heteroplasmy and 6.5% match very high heteroplasmy definitions based on high MAF samples counts. This is substantially above the overall high and very high heteroplasmy site proportions found at this filter level. But in terms of median MAF, no sites within this set match very high heteroplasmy level definition, and only one position, equivalent to 0.9% matches high heteroplasmy definition of at least 1% median MAF, which is a much lower proportion than what we see across all putative heteroplasmic positions.

Altogether this demonstrates a substantial enrichment of high and very high heteroplasmy level positions between highest putative heteroplasmy incidence sites if we look at the number of samples above 1% and 10% MAF thresholds. In contrast when using median MAF to define high and very high heteroplasmy sites, there is no enrichment of high median MAF sites between the highest heteroplasmy prevalence positions at filter 7, and at filter 5 we in fact see a depletion of such high median MAF sites. This is of course consistent with our earlier observation that uniformly high heteroplasmy levels found at a high frequency in population do not exist.

Supplementary figure 5.5.6 S3: MAF levels at 107 positions with highest heteroplasmy incidence at filter 5

MAF levels at 107 positions with highest heteroplasmy incidence at filter 5 by consensus method

Position	rCRS base	MAF level		Samples with MAF above				
		Median	IQR	0.1%	0.5%	1%	5%	10%
439	A	0.25%	0.09%	859	7	0	0	0
440	A	0.17%	0.05%	113	0	0	0	0
443	A	0.18%	0.06%	422	0	0	0	0
446	A	0.17%	0.06%	173	0	0	0	0
448	A	0.21%	0.07%	1,115	1	0	0	0
464	A	0.23%	0.09%	1,346	3	0	0	0
470	A	0.22%	0.09%	1,430	3	0	0	0
471	T	0.17%	0.06%	163	1	0	0	0
503	A	0.22%	0.10%	1,269	20	1	1	0
512	A	0.20%	0.09%	486	0	0	0	0
517	A	0.16%	0.06%	238	0	0	0	0
519	A	0.16%	0.05%	235	0	0	0	0
521	A	0.16%	0.06%	230	0	0	0	0
523	A	0.17%	0.06%	315	2	1	1	0
538	A	0.24%	0.12%	1,896	26	1	0	0
550	A	0.16%	0.06%	313	0	0	0	0
554	A	0.17%	0.06%	828	0	0	0	0
561	A	0.17%	0.06%	761	1	0	0	0
567	A	0.15%	0.05%	135	0	0	0	0
574	A	0.17%	0.06%	603	1	0	0	0
709[2,3]	G	0.23%	0.34%	205	50	34	12	9
2623[2,3]	A	0.32%	0.55%	224	83	48	12	8
2702[2]	G	0.27%	0.44%	109	29	21	4	2
3244[2,3]	G	0.26%	0.54%	199	68	36	7	5
3489	A	0.12%	0.03%	220	0	0	0	0
3501	A	0.13%	0.04%	223	0	0	0	0
3508	A	0.15%	0.07%	1,267	0	0	0	0
3518	A	0.12%	0.03%	236	0	0	0	0
3547	A	0.18%	0.08%	1,528	3	2	0	0
3561	A	0.15%	0.05%	272	0	0	0	0
3577	A	0.25%	0.11%	1,596	47	0	0	0
3578	T	0.34%	0.14%	1,193	153	3	0	0
3583	A	0.21%	0.09%	1,397	11	0	0	0
3584	A	0.13%	0.04%	321	2	1	1	1
3601	A	0.15%	0.06%	839	0	0	0	0
8279	T	0.15%	0.06%	623	4	3	0	0
8426	T	0.15%	0.05%	1,152	3	1	1	0
8459	A	0.12%	0.03%	111	0	0	0	0
8463	A	0.14%	0.05%	213	2	1	1	0
8597	T	0.14%	0.05%	320	1	1	0	0
8603	T	0.14%	0.05%	196	1	0	0	0
8613	A	0.13%	0.05%	155	2	0	0	0
8701[2]	A	0.18%	0.15%	132	12	**6**	1	0
10398[2,3]	A	0.20%	0.18%	139	19	**7**	7	**6**
10873	T	0.19%	0.14%	108	10	3	0	0
10906	A	0.13%	0.05%	723	1	1	1	1
10915	T	0.13%	0.05%	266	3	0	0	0
10921	A	0.15%	0.05%	200	0	0	0	0
10959	T	0.22%	0.07%	176	2	2	1	1
10963	A	0.20%	0.07%	271	4	3	0	0
10967	A	0.16%	0.05%	135	3	2	2	1
10977	T	0.18%	0.06%	334	0	0	0	0
10978	A	0.18%	0.08%	131	4	2	0	0
10987	A	0.14%	0.04%	134	0	0	0	0
10988	A	0.13%	0.03%	215	1	1	0	0

Position (cont.)	rCRS base	MAF level		Samples with MAF above				
		Median	IQR	0.1%	0.5%	1%	5%	10%
11711[2]	G	0.26%	0.56%	174	52	34	10	4
11914[2]	G	0.20%	0.25%	125	24	14	2	1
12684	G	0.15%	0.07%	789	2	0	0	0
12705	C	0.16%	0.09%	574	15	1	0	0
13105[2]	A	0.16%	0.20%	122	18	8	4	1
13276	A	0.14%	0.06%	121	1	1	1	1
13433	T	0.13%	0.05%	157	3	3	1	0
13650	C	0.14%	0.06%	299	7	1	1	0
13681	A	0.13%	0.06%	101	8	5	2	1
13696	A	0.13%	0.05%	179	0	0	0	0
13744	A	0.13%	0.04%	177	0	0	0	0
13747	A	0.13%	0.04%	582	0	0	0	0
13750	A	0.20%	0.10%	2,355	2	0	0	0
13751	T	0.13%	0.04%	556	0	0	0	0
13752	T	0.18%	0.08%	2,482	2	1	0	0
13753	T	0.12%	0.03%	103	2	0	0	0
13776	A	0.20%	0.07%	342	0	0	0	0
13777	A	0.17%	0.06%	299	0	0	0	0
13779	A	0.25%	0.09%	1,515	22	0	0	0
13780	A	0.22%	0.09%	841	3	2	1	0
13781	T	0.22%	0.09%	693	8	4	2	2
13789	T	0.28%	0.13%	2,227	116	3	0	0
13790	A	0.16%	0.06%	534	4	3	1	1
13794	A	0.13%	0.03%	104	1	1	1	0
13803	A	0.14%	0.04%	140	3	0	0	0
13819	T	0.15%	0.05%	693	0	0	0	0
13853	A	0.12%	0.03%	137	0	0	0	0
13857	A	0.13%	0.04%	886	1	1	0	0
13861	A	0.13%	0.03%	670	0	0	0	0
13864	A	0.11%	0.02%	122	0	0	0	0
13868	T	0.13%	0.04%	453	0	0	0	0
13887	A	0.12%	0.03%	120	0	0	0	0
13894	T	0.12%	0.03%	133	2	2	0	0
14046	A	0.15%	0.06%	196	0	0	0	0
14047	A	0.21%	0.09%	289	3	2	0	0
14053[2]	A	0.19%	0.06%	199	6	5	0	0
14059	A	0.15%	0.05%	731	1	0	0	0
14487	T	0.15%	0.06%	1,246	3	2	1	1
14513	A	0.12%	0.02%	109	1	0	0	0
14783	T	0.16%	0.11%	116	6	1	0	0
15043	G	0.17%	0.12%	138	14	4	0	0
15301	G	0.17%	0.10%	162	13	3	2	2
15443	C	0.15%	0.06%	250	2	0	0	0
16093[1,2,3]	T	**2.22%**	6.90%	237	170	149	69	45
16111[2]	C	0.17%	0.12%	106	15	6	0	0
16126[2]	T	0.18%	0.16%	119	16	5	1	1
16129[2,3]	G	0.16%	0.13%	413	48	31	17	14
16171	A	0.30%	0.13%	147	9	0	0	0
16189[2]	T	0.54%	0.26%	304	188	30	8	5
16192[2,3]	C	0.39%	1.82%	128	58	39	21	16
16223	C	0.21%	0.15%	105	10	2	2	1
16293[2]	A	0.32%	0.24%	162	42	20	6	1

Positions highlighted in blue fit one or more high or very high heteroplasmy site definitions:
1 – Putative heteroplasmy median MAF ≥1%;
2 – At least 3 high heteroplasmy samples with MAF ≥1%;
3 – At least 5 very high heteroplasmy samples with MAF ≥10%;

Supplementary figure 5.5.6 S4: MAF levels at 88 positions with highest heteroplasmy incidence at filter 7

MAF levels at 88 positions with highest heteroplasmy incidence at filter 7 by consensus method

Position	rCRS base	MAF at filter 5 Median	MAF at filter 5 IQR	MAF at filter 7 Median	MAF at filter 7 IQR	Samples with MAF above 0.1%	0.5%	1%	5%	10%
152[1]	T	4.02%	10.59%	5.24%	15.08%	13	13	10	5	3
189[1,2]	A	1.47%	2.14%	2.01%	6.43%	64	61	42	15	9
195[1,2]	T	1.86%	7.11%	4.43%	22.06%	53	50	35	17	11
200[1]	A	1.45%	1.61%	1.80%	1.86%	19	19	12	2	1
204[1,2]	T	1.29%	3.33%	3.31%	5.49%	98	88	60	22	13
214[1]	A	1.65%	2.36%	2.66%	6.69%	17	15	11	3	3
215	A	0.99%	2.28%	3.08%	5.03%	33	26	16	5	3
234[1,2]	A	1.50%	7.39%	5.32%	15.93%	41	33	24	13	8
251[1]	G	1.13%	3.37%	3.42%	5.58%	25	19	14	5	2
279[1]	T	1.11%	2.46%	3.00%	12.06%	26	20	13	5	4
385	A	0.57%	1.90%	3.52%	4.47%	40	23	15	5	2
709[2]	G	0.23%	0.34%	2.39%	8.84%	205	50	34	12	9
1082	A	0.27%	0.39%	2.36%	3.95%	67	18	15	4	1
1169	G	0.39%	1.11%	3.67%	5.70%	43	18	12	4	1
1201	A	0.51%	1.54%	3.66%	6.39%	53	28	18	8	4
1435[1]	A	1.04%	1.72%	1.81%	1.08%	19	13	10	1	1
1598	G	0.39%	0.82%	1.74%	1.48%	46	19	11	2	2
2075[1]	T	1.20%	2.39%	1.73%	3.09%	19	11	11	1	1
2174	G	0.75%	1.50%	2.35%	6.72%	23	16	10	3	2
2275	T	0.71%	1.67%	2.03%	4.46%	28	15	12	3	2
2333	G	0.28%	0.65%	1.84%	2.29%	74	21	14	3	0
2492	G	0.26%	0.75%	3.37%	6.55%	51	19	11	4	1
2623[2]	A	0.32%	0.55%	2.27%	4.88%	224	83	48	12	8
2702	G	0.27%	0.44%	1.72%	1.78%	109	29	21	4	2
2778	T	0.31%	1.45%	1.82%	1.90%	42	17	15	3	2
2805	A	0.58%	1.39%	2.59%	2.75%	42	23	15	1	0
2833	A	0.33%	0.82%	2.29%	2.40%	84	40	21	4	3
2960	T	0.37%	0.98%	1.82%	3.16%	48	20	14	2	1
3022	G	0.21%	0.74%	2.19%	4.24%	62	21	12	4	2
3242	G	0.30%	0.68%	1.79%	1.33%	48	18	11	1	0
3244[2]	G	0.26%	0.54%	1.81%	2.20%	199	68	36	7	5
3421	G	0.64%	1.42%	2.10%	2.10%	25	13	10	1	0
4665	G	0.31%	1.39%	2.40%	1.90%	47	21	15	3	2
5262	G	0.25%	0.55%	5.59%	13.85%	54	18	12	6	4
5585	G	0.34%	1.35%	2.87%	5.27%	33	12	11	4	2
5894	A	0.35%	0.81%	1.90%	5.28%	48	19	12	4	1
5970[2]	G	0.39%	0.59%	5.27%	11.29%	63	23	14	7	6
6260	G	0.30%	1.05%	3.69%	3.25%	41	15	11	3	2
7598	G	0.49%	1.86%	2.59%	4.28%	35	17	15	4	1
7609	T	0.29%	1.40%	5.86%	19.52%	39	15	11	6	4
8251	G	0.30%	0.91%	3.12%	6.63%	67	24	18	7	3
9055	G	0.30%	0.76%	2.31%	2.25%	42	16	10	2	0
9716	T	0.53%	1.64%	2.19%	5.87%	34	17	14	5	1
9868	G	0.25%	0.70%	2.52%	2.06%	46	14	10	2	1

Position (cont.)	rCRS base	MAF at filter 5 Median	MAF at filter 5 IQR	MAF at filter 7 Median	MAF at filter 7 IQR	Samples with MAF above 0.1%	0.5%	1%	5%	10%
9947	G	0.43%	1.80%	2.69%	3.41%	37	15	11	3	2
10197	G	0.39%	0.84%	2.86%	8.16%	58	25	15	5	4
11225	G	0.51%	2.89%	3.56%	2.35%	24	12	10	3	2
11456	G	0.25%	0.82%	2.69%	4.72%	42	14	10	4	2
11711	G	0.26%	0.56%	2.59%	4.89%	174	52	34	10	4
11914	G	0.20%	0.25%	2.15%	2.22%	125	24	14	2	1
12172	A	0.33%	0.70%	3.73%	4.08%	40	13	10	4	0
12457	G	0.57%	1.36%	1.68%	2.50%	31	16	13	2	1
12771	G	0.28%	0.69%	5.15%	24.13%	48	15	12	6	4
13063	G	0.28%	0.65%	1.89%	2.53%	54	22	13	2	1
13369	T	0.24%	0.69%	6.06%	13.51%	57	22	13	7	4
13676	A	0.30%	0.49%	1.61%	1.23%	74	22	14	1	1
13722	A	0.28%	0.76%	1.92%	4.59%	44	15	11	4	2
13879	T	0.35%	0.67%	4.60%	5.48%	46	19	10	5	2
14831	G	0.33%	1.44%	2.23%	3.64%	58	22	17	5	3
14905	G	0.20%	0.47%	3.11%	7.69%	52	14	10	4	2
14985	G	0.29%	0.86%	1.96%	4.26%	70	27	18	5	2
15200	G	0.35%	1.69%	4.13%	7.35%	36	16	11	4	3
15884	G	0.25%	0.60%	1.61%	2.79%	55	16	12	3	1
15924	A	0.24%	0.52%	1.99%	1.41%	93	30	18	1	0
16065	G	0.29%	1.51%	6.34%	10.90%	42	15	12	6	4
16086	T	0.39%	0.97%	5.34%	18.04%	56	25	15	8	4
16092[2]	T	0.31%	1.31%	6.89%	23.36%	53	21	15	8	6
16093[1,2]	T	2.22%	6.90%	4.02%	10.87%	237	170	149	69	45
16124	T	0.24%	0.36%	8.95%	19.35%	64	17	11	7	4
16129[2]	G	0.16%	0.13%	8.65%	18.77%	413	48	31	17	14
16172	T	0.25%	0.48%	3.98%	11.90%	57	17	11	4	4
16183[1]	A	1.57%	4.97%	5.29%	7.35%	24	17	14	8	2
16187	C	0.24%	0.57%	1.60%	1.35%	49	18	11	0	0
16189[2]	T	0.54%	0.26%	3.59%	3.67%	304	188	30	8	5
16192[2]	C	0.39%	1.82%	5.50%	15.14%	128	58	39	21	16
16256	C	0.39%	0.84%	3.87%	5.27%	51	22	13	6	2
16261	C	0.30%	1.16%	1.74%	2.72%	43	17	13	2	1
16266	C	0.41%	0.47%	4.46%	15.00%	49	19	10	5	4
16291	C	0.48%	1.11%	1.95%	2.54%	52	25	20	4	4
16292	C	0.41%	0.58%	4.17%	4.86%	54	25	12	4	1
16293	A	0.32%	0.24%	2.69%	5.07%	162	42	20	6	1
16298	T	0.63%	1.03%	1.87%	2.70%	33	18	12	3	1
16301	C	0.64%	1.48%	1.84%	2.04%	22	14	10	1	1
16309	A	0.56%	1.68%	2.99%	3.22%	46	27	16	4	1
16311[2]	T	0.46%	1.24%	5.73%	14.81%	77	35	23	12	8
16327	C	0.66%	1.20%	2.76%	6.67%	31	18	12	5	2
16362[2]	T	0.83%	0.87%	14.39%	33.25%	29	23	10	5	5
16390[1,2]	G	1.05%	1.88%	2.43%	2.70%	57	53	31	7	6

Samples with MAF ≥0.1% & ≥0.5% based on sites passing filter 5; samples with MAF ≥1%, ≥5% & ≥10% based on passing filter 7
All positions pass high heteroplasmy site definition of at least 3 high heteroplasmy samples with MAF ≥ 1%;
Positions highlighted in blue fit one or two other high or very high heteroplasmy site definitions:
1 – Median putative heteroplasmy MAF ≥1% at filter 5;
2 – At least 5 very high heteroplasmy samples with MAF ≥10%;

A3.5 Relationship between mitochondrial heteroplasmy and the genomic context

A3.5.1 Heteroplasmy differences between mitochondrial genome functional elements

Supplementary figure 5.6.1 S1 shows the average number of heteroplasmies per position found with consensus method with MAF above different thresholds across different mitochondrial genome functional element groups.

Supplementary figure 5.6.1 S2 shows the proportions of heteroplasmic sites at which different number of heteroplasmies are detected by consensus method in different genomic contexts.

Supplementary figure 5.6.1 S1: MAF observed at the heteroplasmic sites found in different genomic contexts

MAF levels observed at the heteroplasmic sites found in different genomic contexts

Genomic context	Average MAF central tendency[1]		Average heteroplasmies above different MAF thresholds[2]				
	mean	median	0.1% MAF	0.5% MAF	1% MAF	5% MAF	10% MAF
Control region	2.2% (3.5%)	1.3% (3.1%)	40.1 (164.5)	5.5 (14.4)	3.1 (8.8)	1.2 (3.9)	0.7 (2.6)
Non-coding	1.9% (4.6%)	1.6% (4.6%)	19.9 (93.6)	1.7 (3.4)	1.2 (2.5)	0.4 (0.9)	0.1 (0.4)
tRNA	1.8% (4.2%)	1.3% (4.0%)	4.9 (6.3)	1.6 (2.1)	1.0 (1.5)	0.3 (0.6)	0.2 (0.5)
rRNA	1.6% (3.2%)	0.9% (2.9%)	7.4 (11.6)	2.4 (3.9)	1.5 (2.5)	0.4 (0.9)	0.2 (0.6)
Protein coding	1.6% (3.3%)	0.9% (3.0%)	10.8 (72.4)	1.9 (3.4)	1.1 (1.7)	0.3 (0.7)	0.2 (0.5)
Codon 1st position	*1.6% (3.5%)*	*0.9% (3.2%)*	*14.5 (99.2)*	*2.3 (3.9)*	*1.3 (2.0)*	*0.4 (0.8)*	*0.2 (0.6)*
Codon 2nd position	*1.6% (2.9%)*	*0.9% (2.6%)*	*7.8 (35.1)*	*1.9 (3.7)*	*1.1 (1.5)*	*0.4 (0.7)*	*0.2 (0.5)*
Codon 3rd position	*1.5% (3.6%)*	*1.0% (3.3%)*	*10.3 (68.6)*	*1.4 (2.2)*	*0.8 (1.4)*	*0.3 (0.6)*	*0.2 (0.4)*
Overall	1.6% (3.4%)	1.0% (3.1%)	11.2 (71.2)	2.1 (4.6)	1.2 (2.6)	0.4 (1.1)	0.2 (0.8)

Values represent averages calculated across all identified heteroplasmic positions within each genomic context: (1) Mean and median MAF levels of heteroplasmies observed at the same site, averaged across all heteroplasmic sites; (2) Average number of heteroplasmies above each MAF threshold found per heteroplasmic site in each genomic context. Standard deviation shown in brackets.

Supplementary figure 5.6.1 S2: Heteroplasmic site proportions with different levels of heteroplasmy prevalence

Proportions of heteroplasmic positions with different number of heteroplasmies per site

(a) Ratios of our identified heteroplasmic sites with different identified heteroplasmy counts

Heteroplasmies MAF ≥ 0.1%	Overall	MT genome functional elements					Protein codon positions		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
1	36.5%	19.8%	27.3%	28.2%	22.5%	27.0%	25.4%	23.9%	32.0%
2 to 9	53.9%	46.3%	61.4%	57.9%	53.4%	53.9%	51.4%	59.2%	50.9%
10 to 99	19.0%	27.9%	9.1%	13.8%	23.9%	18.1%	22.0%	16.3%	16.1%
100+	1.0%	6.1%	2.3%	0.1%	0.2%	1.0%	1.3%	0.6%	1.0%
Total sites	*11,927*	*495*	*44*	*942*	*1,662*	*7,682*	*2,582*	*2,686*	*2,454*

(b) Ratios of previously reported heteroplasmic sites with different reported heteroplasmy counts

Reported heteroplasmies	Overall	MT genome functional elements					Protein codon positions		
		CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
1	24.5%	15.7%	9.8%	31.5%	29.2%	23.6%	26.1%	33.1%	16.0%
2 to 9	54.0%	32.8%	52.9%	51.9%	55.7%	55.7%	53.0%	53.4%	58.9%
10 to 99	19.7%	36.7%	33.3%	16.1%	14.0%	19.8%	20.0%	13.1%	23.8%
100+	1.8%	14.8%	3.9%	0.5%	0.8%	0.8%	1.0%	0.4%	1.3%
Total sites	*11,098*	*682*	*51*	*915*	*1,581*	*7,880*	*2,410*	*2,094*	*3,415*

*(**a**) Positions where heteroplasmies ≥0.1% MAF are identified using consensus read frequency based method; (**a**) Heteroplasmic positions reported across 20 previous studies reviewed earlier in section 5.3; Values in tables represent relative proportion of heteroplasmic positions with the specified number of total heteroplasmies in the specific genomic context, total heteroplasmic sites in each context listed at the bottom of each table; MAF – minor allele frequency; CR – control region, NC – non-coding nucleotides*

Supplementary figure 5.6.1 S3 provides a comparison of this to the heteroplasmy discovery rates of different genomic regions reported in 20 reviewed previous studies.

Supplementary figure 5.6.1 S3: Heteroplasmy discovery rates at different genomic contexts in the literature

Heteroplasmy discovery rates found in previous literature at different genomic contexts

(a) Heteroplasmy discovery rates at different genomic contexts reported in 20 reviewed studies

Study	MAF threshold	Overall HDR	MT genome functional elements					Codon positions		
			CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
Bolze 2020	1%	0.024	0.168	0.034	0.014	0.009	0.014	0.013	0.006	0.023
Calabrese 2022	5%	0.184	0.585	0.069	0.093	0.094	0.176	0.127	0.058	0.342
Laricchia 2022	10%	0.024	0.188	0.117	0.009	0.009	0.012	0.012	0.007	0.018
Calabrese 2014	1%	0.386	2.066	0.321	0.133	0.281	0.273	0.231	0.133	0.452
Liu 2018	5%	1.693	8.110	2.276	0.867	1.175	1.277	1.124	1.395	1.310
Diroma 2014	1%	0.389	1.675	2.166	0.111	0.261	0.315	0.313	0.124	0.508
Ye 2014	1%	0.242	0.655	0.079	0.127	0.231	0.220	0.248	0.193	0.219
Wei 2019	1%	0.038	0.155	0.164	0.028	0.027	0.030	0.032	0.020	0.037
Ding 2015	4%	0.043	0.210	0.125	0.013	0.023	0.034	0.024	0.017	0.060
Abecasis 1012	10%	0.022	0.068	0.016	0.014	0.017	0.019	0.025	0.017	0.016
Sondheimer 2011	5.4%	1.293	1.811	0.800	0.945	1.257	1.243	1.878	0	1.130
Duan 2019	0.5%	7.243	23.458	8.621	5.319	10.549	5.114	4.340	1.053	10.142
Abecasis 2010	10%	0.040	0.048	0	0.029	0.037	0.042	0.035	0.047	0.042
Rebolledo 2014	1%	0.086	0.446	0	0.068	0.071	0.059	0.047	0.047	0.081
Sosa 2012	5%	0.107	0.282	0	0.033	0.159	0.088	0.053	0.053	0.158
He 2010	2%	0.311	2.433	0.862	0.100	0.100	0.159	0.237	0.171	0.066
Li 2010	10%	0.017	0.086	0	0.005	0.006	0.014	0.012	0.006	0.026
Goto 2011	2%	0.080	0	0	0	0	0.118	0.205	0	0.146
Skonieczna 2015	2%	4.5×10^{-4}	3.7×10^{-3}	0	0	2.6×10^{-4}	2.3×10^{-4}	1.7×10^{-4}	2.6×10^{-4}	2.6×10^{-4}
Stoneking 2000	10%	0.051	0.052	n/a	0	n/a	n/a	n/a	n/a	n/a
Literature aggregate		*0.002*	*0.011*	*0.004*	*0.001*	*0.001*	*0.001*	*0.001*	*0.001*	*0.002*

(b) Genomic region-specific heteroplasmy discovery rates scaled by study overall discovery rates

Study	MAF threshold	Overall HDR	MT genome functional elements [1]					Codon positions [2]		
			CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
Bolze 2020	1%	0.024	6.96	1.41	0.58	0.36	0.59	0.92	0.45	1.63
Calabrese 2022	5%	0.184	3.18	0.38	0.50	0.51	0.96	0.72	0.33	1.94
Laricchia 2022	10%	0.024	7.80	4.86	0.38	0.38	0.51	0.98	0.57	1.45
Calabrese 2014	1%	0.386	5.35	0.83	0.35	0.73	0.71	0.85	0.49	1.66
Liu 2018	5%	1.693	4.79	1.34	0.51	0.69	0.75	0.88	1.09	1.03
Diroma 2014	1%	0.389	4.30	5.56	0.28	0.67	0.81	0.99	0.39	1.61
Ye 2014	1%	0.242	2.71	0.33	0.53	0.96	0.91	1.13	0.88	0.99
Wei 2019	1%	0.038	4.04	4.28	0.74	0.71	0.78	1.08	0.68	1.24
Ding 2015	4%	0.043	4.89	2.90	0.31	0.55	0.79	0.72	0.51	1.76
Abecasis 1012	10%	0.022	3.12	0.74	0.62	0.76	0.89	1.28	0.89	0.82
Sondheimer 2011	5.4%	1.293	1.40	0.62	0.73	0.97	0.96	1.51	0	0.91
Duan 2019	0.5%	7.243	3.24	1.19	0.73	1.46	0.71	0.85	0.21	1.98
Abecasis 2010	10%	0.040	1.20	0	0.71	0.92	1.04	0.85	1.12	1.01
Rebolledo 2014	1%	0.086	5.19	0	0.79	0.83	0.68	0.80	0.80	1.38
Sosa 2012	5%	0.107	2.64	0	0.31	1.49	0.82	0.60	0.60	1.79
He 2010	2%	0.311	7.83	2.77	0.32	0.32	0.51	1.49	1.08	0.41
Li 2010	10%	0.017	5.06	0	0.30	0.36	0.83	0.85	0.43	1.85
Goto 2011	2%	0.080	0	0	0	0	1.46	1.74	0	1.24
Skonieczna 2015	2%	4.5×10^{-4}	8.14	0	0	0.57	0.51	0.75	1.12	1.12
Stoneking 2000	10%	0.051	1.02	n/a	0	n/a	n/a	n/a	n/a	n/a
Literature aggregate		0.002	5.78	1.93	0.50	0.48	0.67	0.90	0.51	1.59

*HDR – heteroplasmy discovery rate (HDR) expressed as heteroplasmies reported per sample per kilobase analysed sequence; CR – Control region, NC – non-coding nucleotides; Colour scale: **(a)** highest (red) to lowest (white) genomic context specific HDR values in each study, **(b)** genomic context specific HDR values above (red) and below (blue) the overall HDR. (1) CR/NC/tRNA/rRNA/Protein specific HDR scaled relative to study overall HDR; (2) codon position specific HDR scaled relative to protein-coding gene overall HDR*

A3.5.2 Heteroplasmy patterns in protein coding genes

Supplementary figure 5.6.2 S1 compares the reported protein coding gene heteroplasmy distribution between the three codon positions in 20 previously published studies.

Supplementary figure 5.6.2 S1: Variability in literature: heteroplasmy distribution between codon positions

Variability in previous literature: reported heteroplasmy distribution between codon positions

Study	Study MAF threshold	Protein-coding gene heteroplasmies		Protein heteroplasmy distribution between codon positions			Heteroplasmy dN/dS ratio	Codon 2nd to 1st position heteroplasmy ratio
		count	% of total	1st	2nd	3rd		
Bolze 2020	1%	31,660	40.5%	30.8%	14.8%	54.4%	0.84	0.48
Calabrese 2022	5%	12,412	65.5%	24.1%	11.1%	64.8%	0.54	0.46
Laricchia 2022	10%	7,889	34.9%	32.6%	19.1%	48.3%	1.08	0.58
Calabrese 2014	1%	7,482	48.3%	28.3%	16.3%	55.4%	0.81	0.57
Liu 2018	5%	5,157	51.6%	29.4%	36.4%	34.2%	1.94	1.24
Diroma 2014	1%	4,443	55.5%	33.2%	13.1%	53.7%	0.87	0.39
Ye 2014	1%	2,706	62.3%	37.7%	29.2%	33.1%	2.04	0.78
Wei 2019	1%	1,029	53.2%	36.1%	22.7%	41.2%	1.44	0.63
Ding 2015	4%	799	54.1%	24.2%	17.0%	58.8%	0.70	0.70
Abecasis 1012	10%	236	60.8%	42.8%	29.7%	27.5%	2.63	0.69
Sondheimer 2011	5.4%	143	64.1%	37.8%	0%	62.2%	0.61	0
Duan 2019	0.5%	116	48.3%	28.0%	6.8%	65.3%	0.55	0.24
Abecasis 2010	10%	77	71.3%	28.6%	37.7%	33.8%	1.96	1.32
Rebolledo 2014	1%	52	46.8%	26.9%	26.9%	46.2%	1.17	1.00
Sosa 2012	5%	40	56.3%	20.0%	20.0%	60.0%	0.67	1.00
He 2010	2%	36	35.0%	50.0%	36.1%	13.9%	6.20	0.72
Li 2010	10%	21	56.8%	27.3%	13.6%	59.1%	0.75	0.50
Goto 2011	2%	12	100%	58.3%	0%	41.7%	1.40	0
Skonieczna 2015	2%	8	34.8%	25.0%	37.5%	37.5%	1.67	1.50
Literature aggregate		74,318	45.8%	30.1%	17.0%	53.4%	0.89	0.56

dN/dS – nonsynonymous-to-synonymous heteroplasmy ratio, expressed as heteroplasmies reported at codon 1st and 2nd positions divided by heteroplasmies reported at codon 3rd 'wobble' positions in each reviewed study. Data by Stoneking 2000 excluded due to study reporting mitochondrial heteroplasmies in the D-loop region only

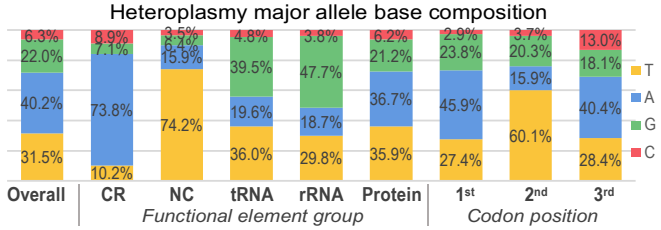
A3.5.3 Relationships between heteroplasmy and the base type in different genomic contexts

Supplementary figure 5.6.3 S1 illustrates how heteroplasmy major allele base composition changes between different genomic contexts depending on whether all heteroplasmies are analysed or whether we look only at subsets of heteroplasmies with MAF levels within some specific range.

Supplementary figure 5.6.3 S1: Base composition of heteroplasmies in different genomic contexts

Base composition of different MAF level heteroplasmies found in different genomic contexts

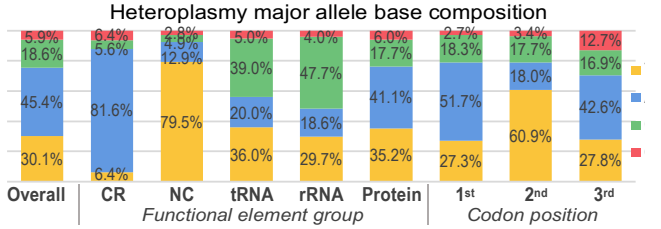
(a) Major allele base composition of all identified heteroplasmies with MAF ≥0.1%



Major allele base composition deviation from expected

Major allele	MT genome functional element group					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	+31%	-29%	-26%	-26%	-2%	+5%	-11%	-7%
C	+2%	-6%	0%	-1%	-1%	-3%	-2%	+4%
G	-12%	-6%	+15%	+17%	+1%	-8%	+1%	+10%
T	-18%	+59%	+2%	+2%	+4%	+1%	+10%	+9%
Total	*19,848*	*875*	*4,876*	*12,224*	*83,334*	*37,469*	*20,899*	*25,222*

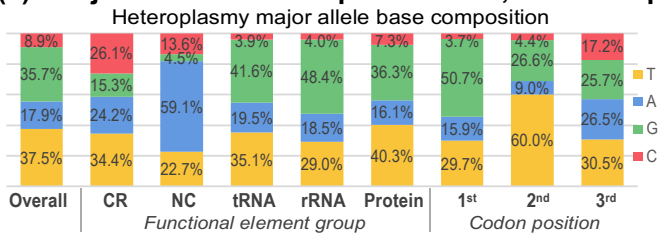
(b) Major allele base composition of 98,002 heteroplasmies with MAF between 0.1% and 0.5%



Major allele base composition deviation from expected

Major allele	MT genome functional elements					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	+34%	-37%	-31%	-32%	-2%	+6%	-12%	-11%
C	0%	-6%	1%	-1%	0%	-3%	-2%	+5%
G	-11%	-5%	+18%	+22%	+1%	-9%	+1%	+10%
T	-21%	+65%	+3%	+3%	+4%	+2%	+13%	+9%
Total	*17,121*	*800*	*3,284*	*8,194*	*68,869*	*31,573*	*15,769*	*21,736*

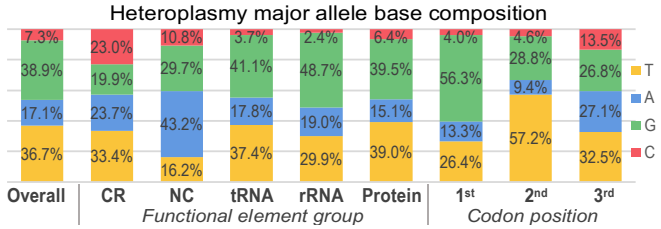
(c) Major allele base composition of 9,344 heteroplasmies with MAF between 0.5% and 1%



Major allele base composition deviation from expected

Major allele	MT genome functional element group					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	+5%	+39%	-1%	-1%	-1%	-2%	-3%	+5%
C	+17%	0%	-3%	-3%	-2%	-4%	-4%	+5%
G	-16%	-15%	+1%	-1%	+4%	-1%	-5%	+12%
T	+1%	+4%	-6%	-4%	+2%	-2%	0%	+7%
Total	*1,171*	*22*	*616*	*1,610*	*5,986*	*2,455*	*2,062*	*1,480*

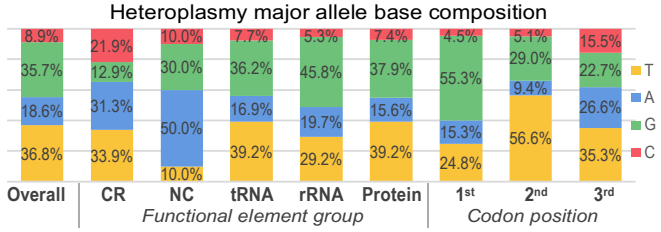
(d) Major allele base composition of 9,157 heteroplasmies with MAF between 1% and 5%



Major allele base composition deviation from expected

Major allele	MT genome functional element group					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	+6%	+24%	-1%	0%	-1%	-4%	-2%	+7%
C	1+5%	0%	-2%	-4%	-1%	-3%	-2%	+4%
G	-14%	+8%	-3%	-5%	+4%	0%	-6%	+12%
T	0%	-2%	-3%	-3%	+1%	-5%	-2%	+9%
Total	*975*	*37*	*679*	*1,711*	*5,811*	*2,364*	*2,125*	*1,346*

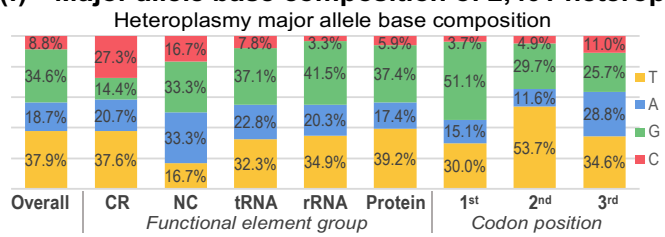
(e) Major allele base composition of 1,850 heteroplasmies with MAF between 5% and 10%



Major allele base composition deviation from expected

Major allele	MT genome functional element group					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	+12%	+29%	-4%	-1%	-2%	-4%	-3%	+5%
C	+12%	-3%	+1%	-2%	-2%	-4%	-3%	+3%
G	-19%	+10%	-4%	-4%	+5%	+3%	-3%	+9%
T	+1%	-8%	-1%	-3%	+1%	-6%	-2%	+12%
Total	*233*	*10*	*130*	*319*	*1,169*	*463*	*435*	*278*

(f) Major allele base composition of 2,401 heteroplasmies with MAF ≥10%



Major allele base composition deviation from expected

Major allele	MT genome functional element group					Codon position		
	CR	NC	tRNA	rRNA	Protein	1st	2nd	3rd
A	1%	+13%	+2%	0%	0%	-4%	-1%	+7%
C	+18%	+4%	+1%	-4%	-3%	-4%	-3%	-1%
G	-16%	+14%	-2%	-7%	+6%	+1%	-1%	+12%
T	+3%	-2%	-9%	+1%	0%	-2%	-7%	+11%
Total	*348*	*6*	*167*	*390*	*14,99*	*614*	*508*	*382*

*(**a**) All 120,754 heteroplasmies ≥0.1% MAF identified by consensus method; (**b-f**) Heteroplasmy subsets with MAF levels within a specified range; Graph columns (left) visualise major allele relative base composition of all heteroplasmies identified within a specific genomic context; Table values (right) represent the difference between the observed heteroplasmy major allele base composition across each mitochondrial genome functional element group from the expected base composition if base type specific heteroplasmy discovery rates were independent of genomic context; MAF – minor allele frequency; CR – control region, NC – non-coding nucleotides*

A3.6 Comparison between novel versus previously known heteroplasmies

In this supplementary section we compare sets of putative heteroplasmies identified at MT genome positions which have already been previously reported as heteroplasmic in literature versus heteroplasmies we identified at novel sites without any previous literature reports of heteroplasmies found. Question explored here is whether there are differences between these sets of known versus novel positions themselves or between the detected heteroplasmic variants at them, and if so, why that might be.

A3.6.1 Distributions of the known versus novel heteroplasmies across the mitochondrial genome

Reviewing previous literature, we found reports of heteroplasmies at 11,098 mtDNA positions. In our dataset, using the allele consensus frequency based method, we found a total of 120,681 putative heteroplasmies with MAF ≥0.1% across 10,804 mtDNA positions. Of these, 13,403 putative heteroplasmies distributed across 5,842 mtDNA positions were found to have MAF ≥1%.

Supplementary figures 5.7.1 S1 and S2 summarise how these putative heteroplasmies are distributed between known heteroplasmic positions previously reported in literature versus novel positions for which we did not find any previous heteroplasmy reports. **Supplementary figure 5.7.1 S1** shows the data for all putative heteroplasmies with MAF≥ 0.1% identified by consensus method, whereas **Supplementary figure 5.7.1 S2** focuses on heteroplasmies with MAF ≥1% only. Known versus novel putative heteroplasmic position and identified heteroplasmy counts are summarised in **panel A**, with **panel B** showing at how many of the sites within each group different numbers of putative heteroplasmic samples were found. **Panel C** provides a side by side comparison of mitochondrial genome locations of known versus novel putative heteroplasmic sites, and **panel D** shows landscapes of putative heteroplasmy counts found at these known and novel positions. In the landscapes previously reported heteroplasmic sites are shown in blue, with novel positions shown in red.

Supplementary figure 5.7.1 S1: Heteroplasmies with MAF ≥0.1% found at previously known versus novel positions

Putative heteroplasmies with MAF ≥0.1% found at previously known versus novel positions

(a) Known versus novel putative heteroplasmic positions and heteroplasmies found at them

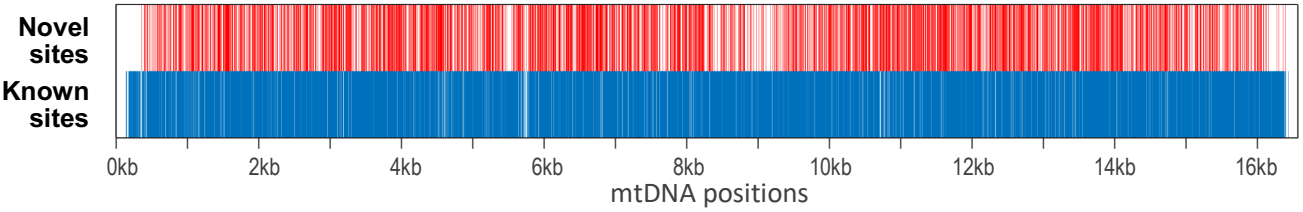
Samples with MAF ≥0.1%	All found	At previously known sites	At novel sites
Total heteroplasmies	120,681	104,030 *(86.2% of total)*	16,651 *(13.8% of total)*
Positions found at	10,804	8,367 *(77.4% of total)*	2,437 *(22.6% of total)*

(b) Positions with different number of putative heteroplasmies with MAF ≥0.1% found

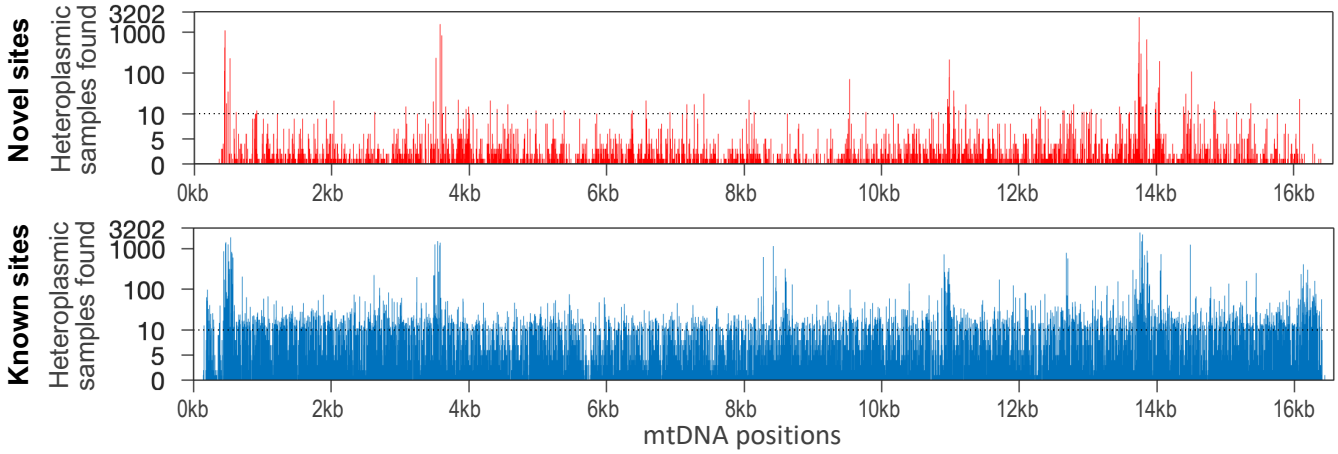
Samples with MAF ≥0.1%	All positions	Previously known	Novel
1	2821 *(26.1%)*	1746 *(20.9%)*	1075 *(44.1%)*
2 to 9	5820 *(53.9%)*	4567 *(54.6%)*	1253 *(51.4%)*
10 to 99	2056 *(19.0%)*	1963 *(23.5%)*	93 *(3.8%)*
100+	107 *(1.0%)*	91 *(1.1%)*	16 *(0.7%)*

Percentages indicate fractions of total positions within each site set

(c) Locations of 2437 novel vs 8367 known sites with putative heteroplasmies with MAF ≥0.1%



(d) Distribution of putative heteroplasmies with MAF ≥0.1% across 2437 novel vs 8367 known sites



Of the of 120,681 putative heteroplasmies with MAF ≥0.1% found across 10,804 mtDNA positions, summarised in **Supplementary figures 5.7.1 S1**, 86.2% or 104,030 heteroplasmies were identified at 8,367 previously reported heteroplasmic positions. The remaining 13.8% or 16,651 putative heteroplasmies were found at 1,437 novel positions. The 8,367 previously reported positions at which we found putative heteroplasmies MAF≥0.1% represent 77.4% of our identified putative heteroplasmic positions and 50.5% of all previously reported heteroplasmic positions in literature. Novel putative heteroplasmic positions we found amount to 22.6% of all of our identified putative heteroplasmic sites, with heteroplasmic variants found at them representing 13.8% of all identified putative heteroplasmies with MAF ≥0.1%.

Supplementary figure 5.7.1 S2: Heteroplasmies with MAF ≥1% found at previously known versus novel positions

Putative heteroplasmies with MAF ≥1% found at previously known versus novel positions

(a) Number of known vs novel putative heteroplasmic positions and heteroplasmies found at them

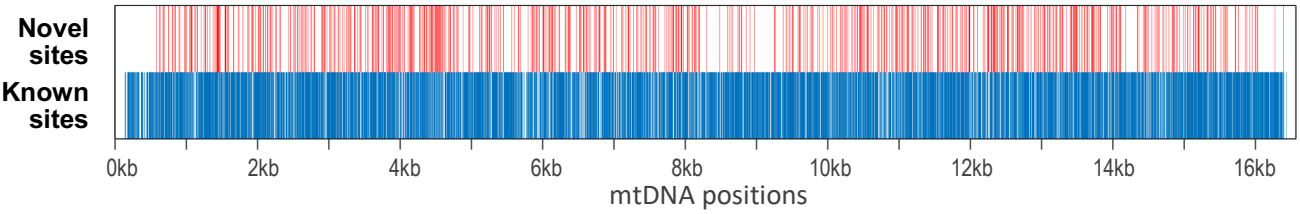
Samples with MAF ≥1%	All found	At previously known sites	At novel sites
Total heteroplasmies	13,403	12,345 *(92.1% of total)*	1,058 *(7.1% of total)*
Positions found at	5,842	5,066 *(86.7% of total)*	776 *(13.3% of total)*

(b) Positions with different number of putative heteroplasmies with MAF ≥1% found (filter 7)

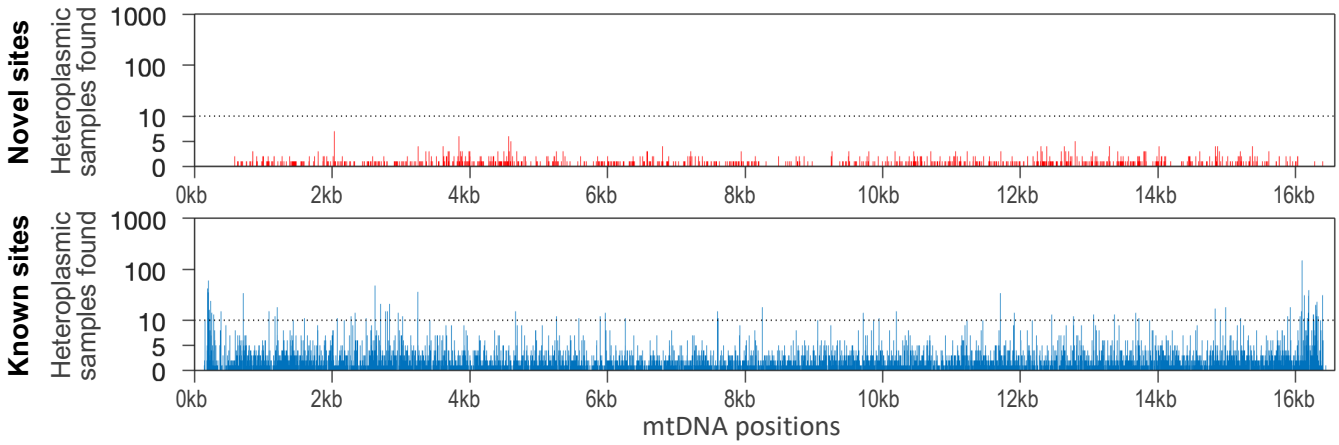
Samples with MAF ≥1%	All positions	Previously known	Novel
1	2,964 *(50.7%)*	2,385 *(47.1%)*	579 *(74.6%)*
2 to 4	2,335 *(40.0%)*	2,143 *(42.3%)*	192 *(24.7%)*
5 to 9	455 *(7.79%)*	450 *(8.88%)*	5 *(0.64%)*
10+	88 *(1.51%)*	88 *(1.74%)*	0 *(0%)*

Percentages indicate fractions of total positions within each site set

(c) Locations of 5066 known and 776 novel positions with putative heteroplasmies with MAF ≥1%



(d) Distribution of putative heteroplasmies with MAF ≥1% across 5066 novel vs 776 known sites



Looking at putative heteroplasmies with a higher minimum MAF threshold of 1%, summarised in **Supplementary figure 5.7.1 S2**, we identified a total of 13,403 such heteroplasmies across 5,842 mtDNA positions, with 92.1% or 12,345 heteroplasmies found at 5,066 previously reported heteroplasmic positions, and 7.1% or 1,058 heteroplasmies found at 776 novel sites. Novel positions in this case represent 13.3% of all sites at which any heteroplasmies with MAF≥1% were found. Whether we look at all putative heteroplasmies with MAF≥0.1% or only those with MAF≥1%, fewer putative heteroplasmies are identified per site at novel positions compared to previously reported sites, as shown by novel site proportion of all putative heteroplasmic positions being larger than the corresponding fraction of all putative heteroplasmies that is found at these novel sites.

In terms of the number of samples with MAF ≥0.1% identified, singletons with only one putative heteroplasmy found make up 20.9% of previously reported heteroplasmic positions compared to 44.1% of novel sites. Whereas 10 or more putative heteroplasmies were found at 24.6% of previously known sites compared to only 4.5% of novel sites. Looking at putative heteroplasmies with MAF ≥1% only, of the positions with such putative heteroplasmies found singletons make up 47.1% of known sites compared to 74.6% of novel positions. Furthermore, none of the 88 positions where 10 or more putative heteroplasmic samples with MAF≥1% were found are novel sites.

The previously reported versus novel putative heteroplasmic position distribution along the genome look fairly evenly spread out. The one exception to this are areas in the D-Loop region where almost no novel heteroplasmic positions are found, especially so when looking at putative heteroplasmies with MAF ≥1%. But as heteroplasmies in these regions seem to be generally more frequent in the population compared to elsewhere as shown in the earlier results, it is reasonable that these positions would be more likely to have already been reported in literature.

In terms of the landscapes of putative heteroplasmic sample frequencies along the genome, it is interesting that most of the distinct low level putative heteroplasmy hotspot peaks observed earlier clearly contain both previously known and novel positions, rather than being present in only one or the other of the two landscapes. This points to low level putative heteroplasmy hotspots we see being unrelated to actual heteroplasmy frequency in the population, unless heteroplasmy levels at these spots are always too low to pass the minimum threshold used by most studies. If these peaks were predominantly found to consist of known positions, it would strengthen the case for this being a real biological phenomenon. This not being the case certainly does not disprove the cause being biological, but the hotspots being made up of both novel and known sites certainly could be consistent with these high heteroplasmy prevalence regions being an artefact of the data generation or handling.

A3.6.2 Base composition comparison

The question of whether base composition differences exist between putative heteroplasmies found at positions with and without previous heteroplasmy reports in literature is addressed in **Supplementary figures 5.7.2 S1** below. **Panel A** shows base composition of the previously known versus novel putative heteroplasmic position sets. **Panel B** compares the major allele base composition of putative heteroplasmies found at the known versus novel sites. Colour scale in panels A and B visualise the extent to which each base type frequency deviated from MT genome baseline frequency of that base. **Panels C and D** show mean and median number of putative heteroplasmies found of each major allele base type are shown. For this, for each base type only the positions where some heteroplasmies with that major allele base were found are included in the calculation of these averages.

Supplementary figure 5.7.2 S1: Base composition of previously known versus novel heteroplasmic positions

Base composition of previously known versus novel heteroplasmic positions with putative heteroplasmies with MAF ≥0.1% identified by consensus method at filter level 5

(a) Base composition of 8,367 known versus 2,437 novel putative heteroplasmic positions

Position rCRS base	All found (10,804)	Previously known (8,367)	Novel sites (2,437)
A	25.8% *(-5.1%)*	25.6% *(-5.3%)*	26.6% *(-4.3%)*
C	21.3% *(-9.9%)*	19.9% *(-11.4%)*	26.2% *(-5.1%)*
G	18.7% *(+5.6%)*	20.4% *(+7.3%)*	12.7% *(-0.4%)*
T	34.2% *(+9.5%)*	34.1% *(+9.4%)*	34.6% *(+9.8%)*

Values in brackets show position subset base frequency deviation from mtDNA overall base composition

(b) Major allele base composition of putative heteroplasmies found at known vs novel positions

Major allele base	All found (120,681)	At known sites (104,030)	At novel sites (16,651)
A	40.4% *(+9.5%)*	36.2% *(+5.3%)*	66.7% *(+35.7%)*
C	6.9% *(-24.4%)*	7.2% *(-24.1%)*	5.1% *(-26.2%)*
G	22.0% *(+8.9%)*	23.9% *(+10.8%)*	10.2% *(-2.9%)*
T	30.7% *(+6.0%)*	32.8% *(+8.1%)*	18.1% *(-6.6%)*

Values in brackets show putative heteroplasmy major allele base frequency deviation from mtDNA base composition

(c) Mean putative heteroplasmies found across sites with that major allele by base type

Major allele base	All sites	At known	At novel sites
A	15.6 *SD: 105.8*	15.2 *SD: 97.9*	17.2 *SD: 131.8*
C	3.0 *SD: 13.7*	3.4 *SD: 15.6*	1.3 *SD: 0.7*
G	10.2 *SD: 19.9*	10.9 *SD: 21.1*	5.4 *SD: 3.9*
T	8.9 *SD: 65.1*	10.3 *SD: 73.1*	3.5 *SD: 4.1*

Base type mean calculated excluding any positions where no heteroplasmies with that major allele were found

(d) Median putative heteroplasmies found across sites with that major allele by base type

Major allele base	All sites	At known sites	At novel sites
A	2 *IQR: 4*	2 *IQR: 4*	1 *IQR: 1*
C	1 *IQR: 1*	1 *IQR: 1*	1 *IQR: 0*
G	7 *IQR: 10*	8 *IQR: 11*	4 *IQR: 4*
T	4 *IQR: 5*	5 *IQR: 6*	3 *IQR: 3*

Base type median calculated excluding any positions where no heteroplasmies with that major allele were found

Supplementary figure 5.7.2 S1 panel A shows that the previously reported and novel putative heteroplasmic positions differ substantially in their base composition. Bases with biggest frequency differences between the two sets of positions are Cs (+6.3% for novel positions) and Gs (-7.7% for novel positions). Relative to MT genome overall base compositions this represents an excess of Gs between known heteroplasmic sites only but not novel sites, and more than twice as big a depletion of Cs between known sites compared to novel sites. Both sets however show similarly reduced T frequency and increased A frequency compared to mtDNA as a whole.

If we look at the detected putative heteroplasmy major allele base composition across the same sets of positions, shown in **panel B**, the most obvious is the overwhelming excess of heteroplasmies with a major allele A between

novel sites 66.7% compared to only 36.2% at previously reported positions. Heteroplasmies with C as the major allele are nearly equally rare across both sets of sites at only 7.2% frequency across known positions and 5.1% frequency across novel sites. This is a significant depletion relative to mtDNA overall C frequency of 31.3%. Putative heteroplasmies with major alleles G and T also show substantial differences between the sets of known and novel positions, with both bases found across previously reported heteroplasmic positions at frequencies above their genome baseline levels by 10.8% for G, and 8.1% for T. In comparison frequencies of putative heteroplasmies with G and T major alleles found across novel sites are 2.9% and 6.6% below the respective mtDNA baseline frequencies for these two base types.

The deviation between heteroplasmic position base composition and the total detected putative heteroplasmy major allele composition points to differences in how common heteroplasmies are for different bases. While frequencies of previously reported putative heteroplasmic positions with rCRS base G and T and fairly similar to the corresponding major allele G and T frequencies amid the detected heteroplasmies across the known sites, this is not the case though for positions with reference base A or C. Positions with rCRS base A are slightly less common between putative heteroplasmic sites compared to mtDNA overall, but heteroplasmies with the major allele A are more common than that. In comparison, positions with reference base C are only slightly less frequent between putative heteroplasmic sites compared to MT genome overall, but putative heteroplasmies with C major allele are much less common than that, indicating that even where found, heteroplasmies with major allele C are not present in large numbers. Whereas where heteroplasmies with major allele A are found, it is in much higher numbers.

What is interesting though is that this changes disproportionately between putative heteroplasmies detected at previously reported heteroplasmic positions and novel positions. For example, positions with reference base A frequency increases by 1% for novel positions compared to previously reported ones, while at the same time putative heteroplasmies with major allele A nearly double in frequency in the novel position set. Frequencies of position with reference base C and the detected putative heteroplasmies with major allele C change between previously known versus novel heteroplasmic sites in opposite directions, with rCRS base C being more common between novel positions while the actual detected heteroplasmies with major allele C are less common. The opposite is true for T, whereby rCRS base T frequency is higher amid novel positions, yet putative heteroplasmies with major allele T are actually more common at previously reported sites.

We have looked at how many putative heteroplasmies of each major allele base type were identified at any position where heteroplasmy with that major allele exists. Mean and median number of heteroplasmies with each major allele base found across all positions where that major allele is present in a heteroplasmic form are shown in **Supplementary figure 5.7.2 S1 panels C and D**. For each of the four bases, for the calculation of the mean and median number of heteroplasmies with that major allele, any positions with only putative heteroplasmies with other major alleles found were excluded.

If we look at the average number of putative heteroplasmic samples with each major allele, across known heteroplasmic positions with any heteroplasmies with major allele A found there are on average 15.2 such heteroplasmies, whereas at novel positions this average is slightly higher at 17.2. For the three other bases though the average number of putative heteroplasmies with that major allele found at novel sites (1.3 C, 5.4 G, 3.5 T) is at least two times lower than that found at previously reported sites (3.4 C, 10.9 G, 10.3 T).

Comparing this to the median number of heteroplasmic samples with different major alleles found, A is the base type with the biggest discrepancy from the mean, with a median of only 2 heteroplasmies found at known sites and 1 at novel sites. Major allele with highest median number of heteroplasmies is G at 8 for known sites and 4 for novel sites where this major allele is found. The second highest median heteroplasmy count is for major allele T with 5 at

known sites and 3 at novel sites, and the lowest median number of heteroplasmies found are with major allele C at just 1 for both known and novel positions. This shows us that even though heteroplasmies with major allele A are on average more common, for this major allele base the heteroplasmy count per position is clearly very skewed, hence the extreme discrepancy between the two central tendency measures and the very large standard deviation values.

Across novel positions, most putative heteroplasmies with major allele A found at the same site is 2355 at position 13750 A. For all other major alleles, the maximum number is putative heteroplasmies found at the same site is much lower. For major allele C most putative heteroplasmies found per site is 12 at position 11125 C, for G the most is 21 heteroplasmies found at position 3842 G, and for T this is 97 at position 13739 T. In fact of the top 30 novel positions ranked by number of putative heteroplasmies, rCRS base is A at 29. Consequently, of the 9939 putative heteroplasmies found across these top 30 sites, major allele is A for 9842. For comparison across the previously reported heteroplasmic positions most putative heteroplasmies with major allele A found per position is 1896 at position 538 A. With major allele C the most heteroplasmies found per site is 510 at position 12705 C. For G the most heteroplasmies found per site is 789 at position 12684 G and for T this is 2482 found at position 13752 T (Data not shown).

A3.6.3 Heteroplasmy level comparison

Supplementary figure 5.7.3 S1 addresses the differences between putative heteroplasmies found at previously reported versus novel positions in terms of the measured heteroplasmy levels. **Panel A** heteroplasmy median MAF central tendency across previously reported versus novel putative heteroplasmic poisons. **Panel B** shows how many putative heteroplasmies above MAF thresholds of 0.1%, 1% and 10% were identified on average per site across known versus novel position. **Panel C** shows how many positions in total within the novel or previously reported site groups were found to match high or very high heteroplasmy site definitions used in the previous sections (i.e. positions with median putative heteroplasmy MAF ≥1% and ≥10%, or positions with at least 5 samples with MAF ≥ 1% and 10%)

Supplementary figure 5.7.3 S1: MAF and heteroplasmy prevalence at previously known versus novel sites

Comparison between frequency and MAF levels of putative heteroplasmies identified at novel versus previously known heteroplasmic positions by consensus method at filter 5

(a) Average position-specific median putative heteroplasmy MAF levels detected at filter 5

Median MAF level per site	All positions	Previously known	Novel
Mean across sites	0.97% *SD: 3.13%*	1.02% *SD: 3.32%*	0.78% *SD: 2.34%*
Median across sites	0.30% *IQR: 0.35%*	0.30% *IQR: 0.35%*	0.28% *IQR: 0.34%*

SD – Standard deviation; IQR – interquartile range

(b) Average number of putative heteroplasmies above different MAF thresholds found per site

Heteroplasmies per site		All positions	Previously known	Novel
Mean	with MAF ≥0.1%	11.18 *SD: 71.21*	12.44 *SD: 72.03*	6.83 *SD: 68.18*
	with MAF ≥1%	1.24 *SD: 2.63*	1.48 *SD: 2.92*	0.43 *SD: 0.76*
	with MAF ≥10%	0.22 *SD: 0.76*	0.27 *SD: 0.84*	0.05 *SD: 0.22*
Median	with MAF ≥0.1%	3 *IQR: 7*	4 *IQR: 7*	2 *IQR: 3*
	with MAF ≥1%	1 *IQR: 2*	1 *IQR: 2*	0 *IQR: 1*
	with MAF ≥10%	0 *IQR: 0*	0 *IQR: 0*	0 *IQR: 0*

SD – Standard deviation; IQR – interquartile range

(c) High and very high heteroplasmy site distribution between known and novel positions

High heteroplasmy sites	All positions	Previously known	Novel
High median MAF ≥1%	1,477	1,172 *(79.4%)*	305 *(20.6%)*
Very high median MAF ≥10%	191	163 *(85.3%)*	28 *(14.7%)*
High MAF ≥1% in 5+ samples	543	538 *(99.1%)*	5 *(0.9%)*
Very high MAF ≥10% in 5+ samples	20	20 *(100%)*	0 *(0%)*

Percentages indicate fraction of all high heteroplasmy sites found at known vs novel putative heteroplasmic positions

At the first glance position-specific median putative heteroplasmy MAF level tends to be on average slightly higher at previously reported heteroplasmic positions compared to novel sites, with the mean for known sites being 1.02% compared to 0.78% for novel sites. The medians for both sets are lower though and much more similar at 0.30% for novel sites and 0.28% for novel sites, indicating that the differences in between the means are due to there being more very high median MAF outlier sites between the known positions compared to novel sites. Looking at the actual position in these groups this is indeed true. Within the previously reported site set we found median putative heteroplasmy MAF above 30% at 31 positions compared to only 2 such positions in the novel position set.

We showed earlier that MAF levels tend to be higher for singleton putative heteroplasmies compared to those found in several individuals. But in terms of MAF levels at the singleton sites, we find that even though novel positions have a higher overall proportion of singletons at 44.1% of positions compared to known sites where singletons are 22.9% of sites, only 2.1% of the novel singleton positions (23 positions) have MAF ≥10% compared to 5.2% of all singletons found to have MAF ≥10% at previously reported sites (91 positions) (data not shown).

Previously reported positions make up 77.4% of all putative heteroplasmic sites we found. Of the total 1477 high median MAF (≥1%) sites 79.4% are found at these previously reported positions, and for very high median MAF (≥10%) sites the proportion found at previously reported sites increase to 85.3%. The bias towards having been previously reported is even more extreme for positions with 5 or more high or very MAF samples, with 99.1% of positions of 5+ samples with MAF≥1% and 100% of positions with 5+ samples with MAF≥10% found at previously reported heteroplasmic sites.

A3.6.4 Heteroplasmic variants

Figure below address how many different heteroplasmic variants were found per position across the previously reported sites versus novel putative heteroplasmic positions. At positions where more than one minor allele was identified, we defined the variant based on the highest frequency minor allele only. Thus the total number of possible variants found in the population at a position is 12 (4 major alleles, with three possible minor alleles each)

Supplementary figure 5.7.4 S1 panel A shows the mean and the median number of different variants found per position. **Panels B and C** visualise how the known and novel positions with different number of identified putative heteroplasmic variant are distributed along the mitochondrial genome. **Panel B** shows all putative heteroplasmic variants detected by consensus method at filter 5, whereas **panel C** only displays distribution of variants with MAF ≥1%.

Supplementary figure 5.7.4 S1: Heteroplasmic variants found at previously known vs novel heteroplasmic sites

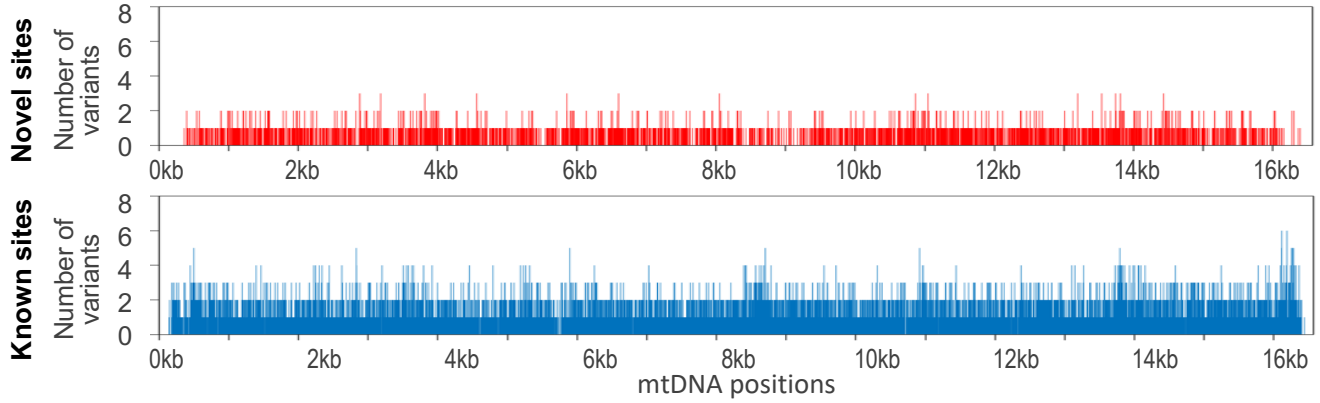
Putative heteroplasmic variants found at previously known vs novel heteroplasmic sites

(a) Average number of different heteroplasmic variants found per position at known vs novel sites

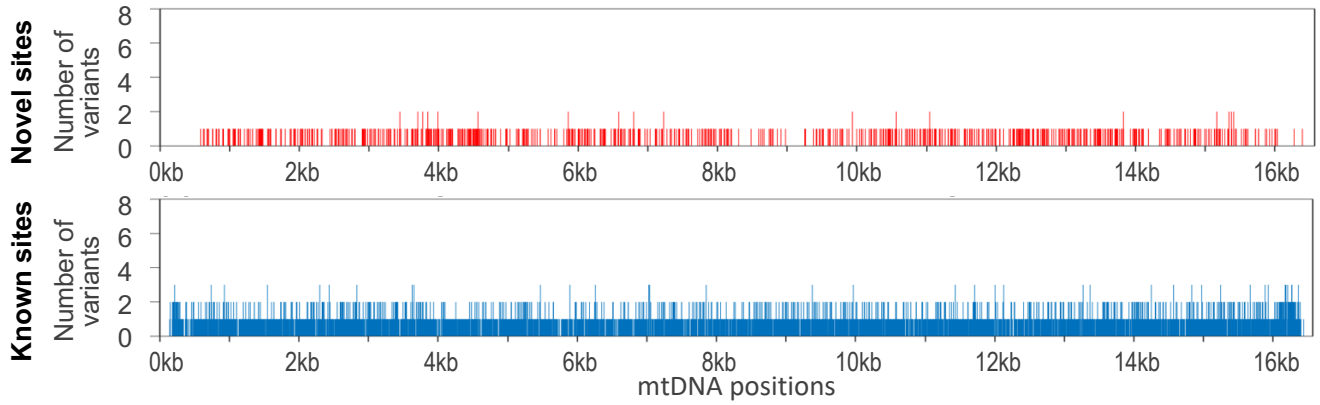
Variants per position	All positions	Previously known	Novel
Mean	1.34 *SD: 0.60*	1.41 *SD: 0.64*	1.10 *SD: 0.32*
Median	1 *IQR: 1*	1 *IQR: 1*	1 *IQR: 0*

SD – standard deviation; IQR – interquartile range

(b) Landscapes of number of putative heteroplasmic variants with MAF ≥0.1%



(c) Landscapes of number of putative heteroplasmic variants with MAF ≥1%



Supplementary figure 5.7.4 S1 shows that that on average fewer putative heteroplasmic variants were found per site across the novel positions compared to previously known sites. At no novel position were more than 3 variants detected across the population, whereas the maximum number of variants found at the previously reported position was 6. Similarly if we look only at the variants found in samples with MAF≥1%, the maximum is 2 per site across novel position compared to 3 per site for previously reported positions.

Positions with more than one heteroplasmic variant found in population are generally distributed fairly evenly across the mitochondrial genome. The one possible exception may be regions where sites with 3 or more variants appear to be clustered more densely. These regions correspond roughly to the same areas where the previously discussed low level putative heteroplasmy hotspot regions were identified at filter 5. This is more obvious at known sites, but also to some extent is present in the novel site set.

We also looked at whether there is a discernible difference between putative heteroplasmies identified at previously known versus novel position in terms of the ratios of putative heteroplasmic transition versus transversion variants. This is addressed in **Supplementary figure 5.7.3 S2**, where **panel A** lists the total number of putative heteroplasmies that are transitions and transversions across all heteroplasmic sites and at specifically only the previously reported or novel position sets. For known and novel sites, percentages in brackets indicate what fraction of total detected variants of each type was across these positions. **Panel B** shows how many heteroplasmic transitions and transversions we found per putative heteroplasmic position on average. **Panels C and D** break down what percentage of all transitions and transversions respectively identified across known versus novel positions was found at positions with a specific rCRS base type, with **panel E** listing the ratio between transitions and transversion found across known and novel positions with different reference bases.

Supplementary figure 5.7.4 S2: Transitions and transversions at previously known vs novel heteroplasmic sites

Transitions and transversions found at previously known vs novel heteroplasmic sites

(a) Total number of heteroplasmies of different types found at known vs novel sites

Putative heteroplasmies	Across all positions	At known sites	At novel sites
All found	120,681	104,030 *(86.2% of all)*	16,651 *(13.8% of all)*
Transitions	66,816	60,825 *(91.0% of all)*	5,991 *(9.0% of all)*
Transversions	53,865	43,205 *(80.2% of all)*	10,660 *(19.8% of all)*

Percentage indicates fraction of all heteroplasmies of that type found

(b) Heteroplasmies of different types found on average per position at known vs novel sites

Heteroplasmies per position		Across all positions	At known sites	At novel sites
Mean	All found	11.1 *SD: 71.2*	12.4 *SD: 72.0*	6.8 *SD: 68.2*
	Transitions	6.2 *SD: 14.9*	7.3 *SD: 16.7*	2.5 *SD: 2.6*
	Transversions	5.0 *SD: 69.8*	5.2 *SD: 70.3*	4.4 *SD: 68.3*
Median	All found	3 *IQR: 7*	4 *IQR: 7*	2 *IQR: 3*
	Transitions	3 *IQR: 6*	4 *IQR: 7*	1 *IQR: 2*
	Transversions	0 *IQR: 0*	0 *IQR: 0*	0 *IQR: 0*

SD – standard deviation; IQR – interquartile range

(c) Transitions found at known vs novel sites by position rCRS base type

Position rCRS base	All transitions Total: 66,816	At known sites Total: 60,825	At novel sites Total: 5,991
A	11,104 *(16.6%)*	10,223 *(16.8%)*	881 *(14.7%)*
C	6,928 *(10.4%)*	6,185 *(10.2%)*	743 *(12.4%)*
G	25,755 *(38.5%)*	24,103 *(39.6%)*	1,652 *(27.6%)*
T	23,029 *(34.5%)*	20,314 *(33.4%)*	2,715 *(45.3%)*

Percentage indicates fraction of total transitions found across the respective group of sites

(d) Transversions found at known vs novel sites by position rCRS base type

Position rCRS base	All transversions Total: 53,865	At known sites Total: 43,205	At novel sites Total: 10,660
A	37,436 *(69.5%)*	27,206 *(63.0%)*	10,230 *(96.0%)*
C	685 *(1.3%)*	564 *(1.3%)*	121 *(1.1%)*
G	817 *(1.5%)*	779 *(1.8%)*	38 *(0.4%)*
T	14,927 *(27.7%)*	14,656 *(33.9%)*	271 *(2.5%)*

Percentage indicates fraction of total transversions found across the respective group of sites

(e) Putative heteroplasmy transition to transversion ratios by position rCRS base type

Position rCRS base	Across all positions Overall Ti/Tv ratio: 1.24	At known sites Overall Ti/Tv ratio: 1.41	At novel sites Overall Ti/Tv ratio: 0.56
A	0.30	0.38	0.09
C	10.11	10.97	6.14
G	31.52	30.94	43.47
T	1.54	1.39	10.02

Ti/Tv ratio – Transitions to transversion ratio

Supplementary figure 5.7.4 S shows that across all putative heteroplasmies with MAF ≥0.1% that we detected, transition variants make up a slight majority 55.4%. Comparing heteroplasmies at known versus novel sites however, transitions make up 58.5% of all heteroplasmies found at previously reported sites, but only 36% of those found at novel sites. In fact, out of all identified putative heteroplasmic transitions 91.0% were found at positions with previously reported heteroplasmies, even though such known positions themselves only make up 77.4% of all our identified putative heteroplasmic sites.

When we break this down by position reference base, very few putative heteroplasmic transversions are found at positions where rCRS base is C or G irrespective of whether the sites have been previously reported as heteroplasmic or not. Instead we see that the overwhelming majority of putative heteroplasmic transversions are found at positions where rCRS base is A, with the second most common being T. The big difference in transversion base

composition at novel position compared to known ones comes from sites where reference base is T, as such sites account for 33.9% of transversions found at known heteroplasmic positions but only 2.5% of those found at novel sites. These differences cannot be accounted for by differences in base composition of previously reported versus novel position sets, as both sets of positions have proportions of site with rCRS bases A and T.

Transitions on the other hand are distributed much more evenly across positions with different reference bases, with the biggest differences seen between known versus novel positions with reference bases G and T. Compared to previously known heteroplasmic sites, novel positions with reference base G account for a smaller proportion of putative heteroplasmic transitions, although this is consistent with positions where rCRS base is G making up a smaller proportion if novel sites in general. In contrast, positions with reference base T make up similar proportions of known and novel sites, but account for higher proportion of transition heteroplasmies at the novel positions.

In literature the Ti/Tv ratio for human nuclear DNA SNPs is usually estimated between 2 and 3 depending on the genome region, whereas in the mitochondrial genome mutation Ti/Tv ratio is thought to be a lot higher, with values between 21 and 37 reported in previous literature (Wang *et al.*, 2015). Our data however shows a much lower overall Ti/Tv ratio across our identified putative heteroplasmies at 1.24, and of we look at previously reported versus novel sites separately, the respective Vi/Tv ratios are 1.41 and 0.56. However when we look at positions with different reference bases separately, we see very different Ti/Tv ratios. For positions with rCRS base C and G our data shows Ti/Tv ratios that are reasonably consistent with values reported in literature. But a much lower Ti/Tv ratio is observed at previously reported putative heteroplasmic positions where rCRS base is A or T, and at novel position with rCRS base A.

If all mutations were equally likely, we would expect a Ti/Tv ratio of 0.5 as there are 2 possible transitions and 4 possible transversions. Such Ti/Tv ratio close to 0.5 is therefore sometimes used in literature as an indicator for a higher proportion of sequencing errors (Wang *et al.*, 2015)

Sequencing error patterns specific to Illumina platform are reported to include higher frequencies of transversions, with the most common miscalls being A-C and T-G (Abnizova *et al*., 2012), both of which are transversions affecting A and T bases. This certainly suggests caution when interpreting our observed high A and T transversion proportions in our putative heteroplasmy dataset.