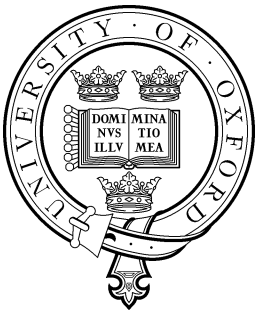
Evaluating selfish spermatogonial selection and its role in human disease



Hannah Kate Ralph

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Abstract

Evaluating selfish spermatogonial selection and its role in human disease

Hannah Kate Ralph, Jesus College.
Submitted for the degree of Doctor of Philosophy, Trinity 2018

The majority of *de novo* mutations originate in the paternal germline. Prior evidence has shown that a subset of activating mutations arise spontaneously within the spermatogonial stem cell (SSC) and can expand clonally along the length of a seminiferous tubule. This process, Selfish Spermatogonial Selection (SSS), explains the high birth prevalence, and increase in frequency with paternal age, of several rare congenital disorders (e.g. Apert syndrome, achondroplasia).

Quantification of these low-level mutations represents a significant technical challenge. Mutation-specific assays have been developed to assess mutation levels directly in sperm and testis samples; to date, however, only 16 loci in six genes, all of which encode members of the RAS-Mitogen-Activated Protein Kinase (RAS-MAPK) pathway, have been assessed.

The aim of this thesis is to evaluate whether other mutations, genes and pathways are subject to SSS by developing scalable screening methods that allow distinction of true variation from confounding errors.

My initial work focused on an existing discovery screen which utilised massively parallel droplet PCR combined with ultra-deep sequencing to assess a candidate gene panel. I undertook manual curation and targeted high-fidelity PCR to resolve a long-list of candidate variants. The validated variants display a striking overlap with clinically relevant mutations and double the number of SSS implicated genes in the RAS-MAPK pathway.

Distinguishing signal from noise remains a significant challenge with PCR-based methods. To improve sensitivity, two different capture-based technologies incorporating unique molecular identifiers (UMIs) were investigated. Target panels were developed to assess hotspots in candidate genes, including genes involved in cancer, intellectual disability and clonal haematopoiesis. The first mutations potentially subject to SSS outside the RAS-MAPK pathway is described, with the identification of mutations at hotspots in *SMAD4*, a key mediator of the TGF-β BMP signalling pathway.

In summary, my work has strengthened the association of the RAS-MAPK pathway with SSS. Using refined methods to increase sensitivity and scalability, I also report the first hints that SSS may impact additional signalling processes in the testis.

Acknowledgements

Firstly, I thank my Anne Goriely and Andrew Wilkie for their supervision during the last 4 years. Thanks are also due to my thesis committee members, Jordan Raff and Tom Brown, as well as my college advisor Shankar Srinivas.

Thank you to the Wellcome Trust for funding this DPhil and to Jesus College for providing funds that enabled me to attend and present my work at conferences.

My warmest thanks go to my lab colleagues: Geoff, Nils, Marie, Hana, Akiko, Kerry, Edu, Yan, Steve and Zhihao. Thank you all for the help you have shared, the friendship you provided and the cakes that we enjoyed. In particular, I want to extend my deepest thanks to Geoff and Nils. Thank you for always selflessly helping me during my time in Oxford. Nils, thank you for teaching me R, for always having time for my questions and for everything else that you did for me.

Finally, and most importantly, my love and gratitude goes to my family. Nothing I've achieved would have been possible without you. Mum and Dad, thank you for the years of support (financial and emotional!) that you have always given me. Anthony, thank you for the love and happiness you bring to my life (and for staying with me through nearly 10 years of study!)

Declaration

I declare that this thesis, *Evaluating selfish spermatogonial selection and its role in human disease*, and the work presented in it is my own. This work was done wholly while in candidature for a research degree at the University of Oxford. Where I have consulted or paraphrased the published work of others, this is clearly referenced. I have acknowledged all main sources of help and assistance I have received. I agree to retain an electronic copy of this work until the publication of my final examination result, and to make this available to the examiners if necessary.

Chapter 3 was published in Genome Research as: Maher, Ralph and Ding. *Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes.* The initial experimental work described in this chapter was performed by Professor Anne Goriely and statistical analysis was performed by Dr Zhihao Ding, with my role being prioritisation and validation of the existing data. I carried this out in collaboration with Dr Geoffrey Maher.

Chapter 6 - Dr Marie Bernkopf assisted with experimental work in this chapter by performing smMIP pool balancing and capture of testicular biopsies. Dr Geoffrey Maher and Dr Maria Neofytou performed PCR validation for a subset of variants in section 6.4.5.

0.1 Abbreviations

ASD	Autism spectrum disorder
APA	Advanced paternal age
DDD	Deciphering Developmental Disorders
DNM	*De novo* mutation
FFPE	Formalin-fixed paraffin embedded
FGF	Fibroblast growth factor
GOF	Gain of function
ID	Intellectual disability
IGV	Integrative Genomics Viewer
LOF	Loss of function
NGS	Next generation sequencing
PAE	Paternal age effect
PCR	Polymerase chain reaction
PGC	Primordial germ cell
RED	Restriction enzyme digestion
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSC	Spermatogonial stem cell
SSS	Selfish spermatogonial selection
smMIP	Single molecule molecular inversion probe
SpT	Spermatocytic tumours
UMI	Unique molecular identifier
VAF	Variant allele frequency
WT	Wild type

Contents

0.1 Abbreviations iv

1 Introduction **1**

1.1 Overview 1

1.2 Genetic mutation 2

1.2.1 Mutation timing 2

1.2.2 Types of mutation 2

1.2.3 Consequences of mutation 4

1.2.4 Origin and signatures of mutation 4

1.2.5 *De novo* mutation and disease 6

1.2.6 Parental origin of mutations 6

1.3 Germ cell biology 8

1.4 Epidemiology of paternal age 9

1.5 Identification of a subset of mutations with an exclusive paternal origin . 12

1.6 Direct measurement of mutation levels in sperm 14

1.7 Activating mutations disrupt SSC homoeostasis 18

1.8 Selfish Spermatogonial Selection 23

1.9 Overlap and parallels of SSS with oncogenesis 27

1.10 Clonal expansion in other contexts 29

1.11 Methods to study selfish selection 31

1.11.1 Identification of clones in testicular biopsies 31

1.11.2 Visual identification of clones 34

1.12 Implications of SSS 36

1.13 Developing alternate strategies to study SSS 38

1.13.1 Next Generation Sequencing 38

1.13.2 Library preparation 40

1.13.3 Overcoming sources of error 40

1.14 Aims 41

2 Materials & Methods **43**
2.1 Ethics approval . . . 43
2.2 General laboratory methods . . . 43
2.2.1 Polymerase Chain Reaction . . . 43
2.2.2 Product visualisation . . . 44
2.3 Testis samples . . . 44
2.4 Testis dissection . . . 44
2.4.1 DNA extraction and quantification . . . 45
2.5 Variant validation . . . 45
2.5.1 Library preparation and sequencing . . . 45
2.5.2 Sequence alignment and variant calling . . . 46
2.6 HaloPlex HS Panel . . . 46
2.6.1 Panel design . . . 46
2.6.2 HaloPlex sample preparation and capture . . . 49
2.6.3 HaloPlex sequencing and variant calling . . . 50
2.7 smMIP . . . 51
2.7.1 smMIP oligonucleotides . . . 51
2.7.2 smMIP pooling and phosphorylation . . . 52
2.7.3 smMIP capture of targeted sequences and circularisation . . . 52
2.7.4 smMIP amplification and library preparation . . . 52
2.7.5 Sequencing analysis . . . 53
2.8 R Scripts . . . 54

3 Validation of a massively-parallel PCR-based discovery screen **55**
3.1 Introduction . . . 55
3.2 Experimental background . . . 55
3.3 Results . . . 58
3.3.1 Coverage . . . 58
3.3.2 Manual curation of variant list . . . 59
3.4 Variant validation . . . 60
3.4.1 Development of amplicon pileup script . . . 60
3.4.2 Variants in overlapping amplicons and replicate samples . . . 67
3.4.3 Tier 2 variants . . . 71

3.4.4 Tier 3 variants . . . 71
3.4.5 Tier 4 variants . . . 72
3.4.6 Validated variants . . . 72
3.5 Visualisation of VAF in biopsies . . . 89
3.6 Synonymous mutations . . . 90
3.7 Mosaic mutations . . . 91
3.8 Discussion . . . 91
3.8.1 Validated variants . . . 91
3.8.2 Technology . . . 92

4 Multiplex target discovery of candidate variants with a commercial capture-based UMI technology 94
4.1 Introduction . . . 94
4.2 Gene selection . . . 97
4.3 Results . . . 97
4.3.1 Analysis of panel performance: coverage . . . 98
4.3.2 Variant calling . . . 99
4.3.3 Analysis of panel performance: variant detection . . . 100
4.3.4 Analysis of panel performance: known positive mutations . . . 101
4.3.5 Detection of variants in seminiferous tubules vs biopsies . . . 103
4.3.6 Variant Prioritisation . . . 104
4.3.7 Variant validation: Tier 1 . . . 112
4.3.8 Variant validation: Tier 2 . . . 117
4.3.9 Variant validation: Tier 3 . . . 117
4.3.10 Validated variants . . . 119
4.4 Benefit of UMI technology . . . 121
4.5 Discussion . . . 124
4.5.1 Validated variants . . . 124
4.5.2 Variant detection in seminiferous tubules . . . 126
4.5.3 Technology Performance . . . 127

5 Assessment of single molecule Molecular Inversion Probes for low level variant detection 129

5.1 Introduction . . . 129
5.2 Effect of gDNA input and probe ratio on smMIP performance . . . 131
5.3 Assessment of low-level mutations in a dilution series . . . 133
5.4 Optimisations to smMIP workflow protocol . . . 134
5.4.1 Synthesis quality of the smMIP . . . 135
5.4.2 Multiple barcoding events for capture . . . 138
5.5 Discussion . . . 140

6 Multiplex target discovery of candidate variants associated with age-related clonal haematopoiesis in the human testes 142

6.1 Introduction . . . 142
6.2 Identification of a recurrent mutation in *DNMT3A* . . . 143
6.2.1 Detection of *DNMT3A* mutation using smMIP panel . . . 144
6.3 Multiplex target discovery in testicular samples . . . 144
6.3.1 Coverage and data analysis . . . 145
6.3.2 Variant prioritisation . . . 151
6.3.3 Variant prioritisation: Tier 1 . . . 151
6.3.4 Variant prioritisation: Tier 2 . . . 151
6.3.5 Validated Variants . . . 153
6.3.6 *PTPN11* deletion . . . 157
6.3.7 Conclusions . . . 158
6.4 Improved resolution of multiplex target discovery in testicular samples . 159
6.4.1 Modification of target smMIP panel . . . 159
6.4.2 Capture and sequence analysis of testicular biopsies . . . 160
6.4.3 Statistical prioritisation of variant sites . . . 166
6.4.4 Top variant sites . . . 166
6.4.5 Novel variants validated . . . 171
6.4.6 Distribution of p.C342S mutations in *FGFR2* . . . 176
6.4.7 Distribution of p.L265P mutations in *MYD88* . . . 177
6.4.8 Discussion and Future Work . . . 180
6.5 Recurrent mutations in *SMAD4* are identified in testicular biopsies . . . 181

6.5.1 Distribution of Myhre syndrome associated mutations in *SMAD4* 182
6.5.2 Candidate variants in TGF-β BMP signaling pathway 182
6.6 Discussion 186
6.6.1 Technology performance 186
6.6.2 Identified variants 186
6.6.3 Recurrent mutations in *SMAD4* 187

7 Conclusion and Future **189**
7.1 Technologies to identify variants associated with Selfish Spermatogonial Selection 189
7.2 Variants identified in this thesis 191
7.3 Future Work 197

A Appendix **198**
A.1 Testis Table Layout 199
A.2 smMIP Sequences 200
A.3 Primer Sequences 218

Bibliography **225**

List of Figures

1.1 Overview of gametogenesis 9
1.2 Assay quantifying c.755 substitutions in *FGFR2* in sperm 15
1.3 Seminiferous tubule structure 19
1.4 Overview of the RAS-MAPK signalling pathway 22
1.5 Selfish Spermatogonial Selection model 24
1.6 Strategies to study Selfish Spermatogonial Selection 32
1.7 Paternal Age Data for England and Wales, 1961 to 2016 37

3.1 Experimental Overview 57
3.2 Analysable sequence per testis biopsy 58
3.3 Example plot of non-reference percentage at variant position 59
3.4 Variant call arising from a batch effect 60
3.5 Example output of pileup graph script 62
3.6 Variant called in overlapping amplicons 67
3.7 Validated variants in *AKT3, APC, BRAF, CBL, FGFR2, FGFR3, KRAS, LRP5, MAP2K1, MAP2K2, NF1, PTPN11, RAF1, RET* and *SOS1* . . . 73
3.8 VAF heatmap of RAF1 p.S257L 89
3.9 Mutation distribution of KRAS p.Q61R and LRP5 p.A97A 90
3.10 Mutation Distribution of NF1 p.M760I 91

4.1 Overview of HaloPlex HS method 96
4.2 HaloPlex Panel coverage performance 98
4.3 *FGFR3* coverage generated by HaloPlex HS panel 99
4.4 Total variants called per sample 100
4.5 Presence of known mutation in *FGFR2* 101
4.6 Variants identified in seminiferous tubules 104
4.7 VAF distribution of variants called 106
4.8 Comparison of variants called by SureCall using HaloPlex HS 106
4.9 Annotation example for known positive mutations 109
4.10 Overview of HaloPlex HS variant filtration 111

4.11 Distribution of UMI supporting variant calls 112
4.12 APC p.E1494G variant 115
4.13 Validated variants in *ADNP, CHD8, NF1* and *SMAD4*. 117
4.14 Validated variants screened in surrounding biopsies 121
4.15 Comparison of target region with and without UMI deduplication 122

5.1 Schematic view of smMIP method 130
5.2 Effect of gDNA input and probe ratio on smMIP performance 132
5.3 Variant detection in a dilution series 134
5.4 smMIP capture efficiency 135
5.5 smMIP synthesis and purification 136
5.6 Effect of oligonucleotide purification on variant detection 137
5.7 Effect of oligonucleotide purification on capture efficiency 138
5.8 Effect of multiple independent barcoding events for smMIP capture . . . 139
5.9 Detection of spiked mutations 140

6.1 *DNMT3A* c.2644C>T p.R882C and c.2645G>A p.R882H variant 146
6.2 UMI supporting p.R882C and p.R882H in DNMT3A 147
6.3 smMIP panel coverage performance 148
6.4 smMIP panel script overview 151
6.5 Novel variants validated in *KIF1C* and MYD88 154
6.6 Novel variants validated in *NAA10* and *TP53* 155
6.7 In-frame deletion in *PTPN11* 157
6.8 Mutations identified in *DNMT3A* in clonal haematopoesis surveys 160
6.9 smMIP panel coverage performance 162
6.10 Distribution of non-reference UMIs and nucleotide changes 167
6.11 Novel variants validated in *PPP2R5D, TET2, GNAS, ADNP, SMAD4,* and *DNMT3A* 172
6.12 Novel variant validated in *GNB1* 173
6.13 Distribution of mutations in *FGFR2* across Tes14 176
6.14 Distribution of *MYD88* mutations in Tes10F 179
6.15 Variants validated in Tes14B 181
6.16 *SMAD4* mutation heatmap 183
6.17 *SMAD4* mutation diagram 184

7.1 Geographic distribution of validated variants 194

A.1 Table layouts for testicular slices 199

List of Tables

1.1 Parental origin of mutations in *FGFR2* and *FGFR3* 13
1.2 Overview of published studies assessing SSS loci 25
1.3 Mutation levels of *FGFR2* c.755C>G in testicular biopsies 33

2.1 Summary of testis samples 44
2.2 HaloPlex HS Panel Targets 47
2.3 Summary of samples sequenced with HaloPlex HS panel 50
2.4 Sequences of smMIP targeting *FGFR2* 51

3.1 Table of replicated calls 69
3.2 Tier 1 validated variants 70
3.3 Tier 2 validated variants 71
3.4 Tier 3 validated variants 72

4.1 Detection of known *FGFR2* variants 101
4.2 Detection of previously validated variants 102
4.3 Number of variants called per seminiferous tubule 103
4.4 Tier 1 Variants 113
4.5 Tier 2 Validation Results 117
4.6 Tier 3 Validation Results 119
4.7 Summary of novel validated variants 120
4.8 Summary of all validated variants 124

5.1 Mutations in *FGFR2* dilution series 133

6.1 Tier 1 variants 153
6.2 Tier 2 variants 153
6.3 Summary of novel validated variants 156
6.4 PTPN11 p.D61 deletion in Tes4 158
6.5 Clonal haematopoiesis studies 159
6.6 Table of smMIP covering key clonal haematopoiesis related genes 159

6.7 Top 25 variants 169
6.8 Summary of top variants by p-value 170
6.9 Summary of novel validated variants 174
6.10 Distribution of z-scores and p-values for *FGFR2* c.1024T>A (p.C342S) . 178
6.11 Variants in *SMAD4* at p.I500 and p.R496 codons 185

7.1 Technology Comparison 190
7.2 Summary of validated variants 192

Chapter 1

Introduction

1.1 Overview

De novo mutations (DNMs) are a major contributor to severe developmental disorders. It is estimated that approximately 400,000 children affected by a developmental disorder as a result of DNM are born each year. The majority of DNM originates during the process of spermatogenesis in the paternal germline. Spontaneously arising mutations have been shown to accumulate with paternal age. A subset of activating mutations that arise within spermatogonial stem cells can clonally expand along the length of a seminiferous tubule through a phenomenon termed Selfish Spermatogonial Selection (SSS). This clonal expansion of acquired mutations explains the high apparent birth rate and exclusive paternal origin of a group of rare congenital disorders including achondroplasia and Apert syndrome. To date, evidence has been gathered for 16 loci in six genes, all of which encode proteins that act at different stages of the RAS-mitogen-activated protein kinase (RAS-MAPK) signalling pathway, involved in the regulation of cell proliferation and survival. The lack of a scalable method to study low-level mutation directly in the male germline has hindered efforts to understand SSS on a broader level. The aim of this thesis is to address whether additional mutations, genes and pathways are subject to SSS through the assessment of different scalable technologies for low-level variant detection.

1.2 Genetic mutation

1.2.1 Mutation timing

Mutations are a source of variation within populations but, depending on the functional effect, may also cause disease. The precise timing of a particular mutation in the growth and differentiation of a developing human will dictate the number and distribution of cells that carry the mutation.

Upon fertilisation, a human zygote inherits half of its genome from its mother and half from its father. Aside from inherited variation, we are all born with novel genetic changes that arose either in one of the parental gametes or occurred following fertilisation (Lynch, 2010). These mutations, termed *de novo*, are by definition absent from the parental genome. Sequencing of children and their parents has provided direct estimates of the human germline mutation rate (Roach et al., 2010), with an estimated 30 to 100 *de novo* mutations (DNM) per individual (section 1.2.5). Of these, one to two DNMs are estimated to occur within coding regions (Veltman and Brunner, 2012).

Mutations also continue to occur throughout post-natal development and adult life, affecting both somatic tissue and germ cells. This can give rise to mosaicism, defined as the presence of genetically distinct cells within an organism as a result of post-zygotic mutational events (Biesecker and Spinner, 2013). Such post-zygotic mutations are an underrepresented source of variation in humans and studies have estimated that at least 3.8% of apparently DNMs are present as mosaic events in the parent (Rahbari et al., 2016). Clinically, determining gonadal involvement of mosaicism is important for the purpose of genetic counselling as the risk of recurrence is higher than for true DNM (Cohen et al., 2015).

1.2.2 Types of mutation

Large scale sequencing projects, such as the 1000 Genomes Project, have provided a comprehensive overview of the genetic variation that exists on a population level. In

comparison to the human reference genome, a typical genome varies at 4.1 to 5 million positions, with the majority of observed variation being shared by 0.5% of the population (The 1000 Genomes Project Consortium, 2015).

In this study of 2504 individuals, the most common genetic variation observed was Single Nucleotide Polymorphisms (SNPs), accounting for 84.7 million of the 88 million variants identified. A SNP is a single nucleotide change in the DNA sequence. The second most common type of variation observed was short insertions/deletions (indels), accounting for 3.6 million of the variants identified. An indel refers to a change in the number of nucleotides present in the sequence relative to the reference sequence.

While large-scale structural variation is a less common event in the human genome, the relative size of these genetic changes means that a greater proportion of the genome is affected ($\sim$20 million bases versus $\sim$5 million) (The 1000 Genomes Project Consortium, 2015). Copy Number Variation (CNV) is defined as a duplication or deletion event affecting more than 1000 nucleotides. Structural variants are more challenging to identify than SNPs. *De novo* structural variation has been estimated to affect an average of 4.1 kbp of genomic sequence and 29 coding bases per generation (Kloosterman et al., 2015).

Finally, mutation can also affect whole chromosomes or parts thereof, through deletion, duplication, inversion or translocation. The most common human chromosome abnormality is aneuploidy (Hassold and Hunt, 2001). Aneuploidy, the presence of an abnormal number of chromosomes, arises as a result of non-disjunction - the failure of homologous chromosomes to correctly separate during cell division. Work on aneuploidy has largely focused on the few existing viable autosomal trisomies - Patau (chromosome 13), Edwards (chromosome 18) and Down syndrome (chromosome 21). Collectively, these are the most common genetic causes of developmental disability (Hassold and Hunt, 2001). DNA polymorphisms have been used to establish the origin of aneuploid chromosomes, revealing a strong maternal age effect. Most trisomies arise during meoisis I of oogenesis and the proportion of cases with a maternal origin increases with maternal age (Hassold et al., 1995; Lamb et al., 1996).

1.2.3 Consequences of mutation

The consequence a mutation has on phenotype depends on where it occurs in the genome and the specific type of mutation it is. For example, in the coding region of a gene the substitution of a nucleotide for another can be functionally silent, meaning that it does not change the encoded amino acid. It can also result in a missense mutation (meaning that the substitution results in the sequence encoding a different amino acid) or it can result in a nonsense mutation (meaning that the substitution results in the introduction of a stop codon). Within the coding region, the addition or deletion of a number of nucleotides that is not a multiple of 3 can lead to the premature introduction of a stop codon in a different reading frame (a frameshift mutation).

Further, a mutation may be classified as a loss-of-function (LOF) or gain-of-function (GOF). A LOF mutation results in the loss of normal protein function. In some cases, the level of expression from the remaining wild type allele may be sufficient to rescue the phenotype. The presence of a mutant form of the protein can also disrupt the wild type function leading to a dominant negative mutation (Herskowitz, 1987). Alternatively, the introduction of a mutation may confer a novel function to the protein (GOF).

Larger structural variants may affect the coding and/or regulatory sequence of multiple genes depending on the size of the event. Such events have been described as important contributors to disease, including neurodevelopmental disorders (O'Roak et al., 2012; Weischenfeldt et al., 2013).

1.2.4 Origin and signatures of mutation

The genome of a given cell has been shaped by distinct sources of error. Mistakes can arise due to errors made during DNA replication as a result of misincorporation of nucleotides by the DNA polymerase (Ségurel et al., 2014). Incorporation of nucleotides during DNA replication is a highly selective process with an average of one mismatch per 10^{-4} to 10^{-5} nucleotides (Korona et al., 2011). Inherent instability of the DNA as well as damaging exogenous or endogenous factors can also contribute to mutation. A common source

of exogenous mutation, for example, is ultraviolet (UV) or ionizing radiation. This is associated with a particular pattern of error and can be observed in clonal mutations that arise in UV-exposed skin (section 1.10). Regardless of the source of error, if repair does not occur correctly, this can result in the mutation becoming fixed during cell division.

Mutations do not occur uniformly throughout the genome and arise throughout life as a result of different mutational processes within a cell, tissue or organism. An insight into distinct mutational signatures has been established through analysis of different human cancers (Alexandrov et al., 2018). Over 23,000 samples have now been analysed, the majority of which (80%) were exome-sequenced, meaning that only the protein-coding genes have been sequenced. Mathematical models were applied to the sequencing data to identify signatures and then count the number of mutations that fit each identified signature in a given data set. This comprehensive analysis has identified 49 signatures associated with single-base substitutions, 11 double-base substitution signatures, 17 indel signatures and four clustered base substitution signatures.

Some of these signatures are now well described and the underlying processes that cause them are understood. For example, signature 1 (C>T transitions) is consistent with deamination of 5-methyl cytosine in the trinucleotide XpCpG context. Methylated CpG dinucleotides are known to be unstable and mutable due to deamination of 5-methyl cytosine, resulting in G:T mismatches (Walser and Furano, 2010). Other signatures remain of unknown cause, however, the delimitation of distinct mutation signatures is important in understanding the processes that contribute to somatic mutation acquisition and oncogenic transformation.

The mutational signature of germline DNM has been shown to correlate best with the described signatures 1 and 5 (Rahbari et al., 2016). Signature 1 is characterised by a high proportion of C:G>T:A mutation and accounts for approximately a quarter of DNM in the germline. Signature 5 corresponds to the remaining 75% of DNM and is represented by T:A>C:G mutations. The origin of signature 5 is less well understood but may be associated with transcription coupled repair (Alexandrov et al., 2015; Acuna-Hidalgo et al., 2016).

1.2.5 *De novo* mutation and disease

DNM is an important contributor to disease. Deciphering Developmental Disorders (DDD), a consortium that aims to identify the genetic basis of undiagnosed developmental disorders, has highlighted the contribution of DNM in disease. Children with severe, undiagnosed developmental disorders were recruited to the study and phenotype information recorded. DNA from the affected child ("proband") is exome sequenced along with the DNA of their unaffected parents, in order to identify mutations that are present in the proband but not their parents. In the most recent publication, more than 4200 families were exome-sequenced resulting in a high-confidence set of 8351 candidate DNMs. DNMs with a probable pathogenic impact were identified in 23% of individuals. Paternal age was a strong risk factor influencing whether a likely pathogenic DNM was identified in a proband. Extrapolating from the exome data to a genome wide figure, these findings correspond to an additional paternal contribution of 1.53 DNMs per year. A weaker, independent contribution was observed due to increasing maternal age (Deciphering Developmental Disorders Study, 2017).

Accounting for the age-and sex-specific effect on mutation rate, the consortium estimated that DNMs cause developmental disorders in approximately 1 in 448 births (for parents aged 20 years), rising to 1 in 213 births (for parents aged 45 years). This corresponds to approximately 400,000 children born with a developmental disorder caused by a DNM each year (Deciphering Developmental Disorders Study, 2017). Given this, understanding the origin and rate at which DNM mutation arises is of great importance.

1.2.6 Parental origin of mutations

Historically, estimates of human mutation rate could only be determined by indirect methods. Early estimates were derived from calculations based on the incidence of genetic disorders; for example, Haldane used the occurence of haemophilia in the population to estimate the mutation rate (Haldane, 1935). The advent of next generation sequencing (NGS) approaches have not only enabled estimates of the human mutation rate, it has

also made it possible to determine the parent of origin of DNM.

A landmark study performed whole genome sequencing (WGS) of 78 trios (proband and both parents) and estimated that each child carries an average of 60 (range: 30 to 100) variations that were absent in either parent (Kong et al., 2012). Further, this study reported that the number of DNMs in the child increases with the age of the father, with each additional year of paternal age being estimated to contribute 1 to 2 additional DNMs. Modelling found that paternal age accounted for the vast majority (97.1%) of variation in the mutation rate observed. Subsequent WGS and WES studies have placed the human mutation rate between 1.2×10^{-8} to 2.17×10^{-8} (Neale et al., 2012; Sanders et al., 2012; O'Roak et al., 2012; Iossifov et al., 2014; Fromer et al., 2014; Turner et al., 2017).

A more recent study undertook the analysis of WGS data from more than ten times the number of trios studied by Kong et al. (2012). This resulted in the refinement of the average number of DNMs per child to approximately 45, following the identification of a total of 36,441 high-confidence DNMs across 816 trios (Goldmann et al., 2016). Importantly, the parental origin of the DNMs could be assigned for a higher proportion of identified variants than previous studies. This revealed that approximately 80% of DNMs originate on the paternal allele and must therefore have arisen during spermatogenesis.

While studies have concluded that the contribution of maternal age to DNM is weaker, positive correlation has been identified with the number of clustered DNM on the maternal allele (Goldmann et al., 2018). WGS of 1291 trios identified 1796 clustered *de novo* mutation (cDNM), defined as mutations that occur within a spatial proximity below 20 kb, and a highly significant correlation with maternal age was identified. Maternal cDNM were enriched on specific chromosomes (8, 9 and 16) implying a specific mutational mechanism results in cDNM (Goldmann et al., 2018).

1.3 Germ cell biology

DNMs arise in the parental germline and so their occurrence is linked to the biology of the germ cell itself. Considering the differences in germ cell biology between males and females, the finding that the majority of point mutations arise in the paternal germline (section 1.2.6) and indeed the accumulation of such mutations over time is not a surprising finding.

Primordial Germ Cells (PGC) are progenitor cells that give rise to the gametes (oocytes and spermatozoa). The first stage of gametogenesis occurs during early embryogenesis when PGCs, derived from the inner cell mass, are formed and migrate to the gonadal ridge at gestational week 5 (Culty, 2009). In the second stage, PGCs respond to specific environmental signals to initiate either oogenesis or spermatogenesis (Nikolic et al., 2016).

Prior to specification of the PGC, a similar number of cell divisions occur in both the male and female embryos. At this stage, and following specification, the estimated mutation rate is similar for both males and females (Figure 1.1 stage I and II; Rahbari et al. (2016)). Around gestational weeks 17 to 18, the gonocytes (precursors of spermatogonial stem cells (SSCs)), undergo rapid division. Changes in expression of pluripotency factors coincides with migration of the gonocytes toward the basement membrane (Jørgensen and Rajpert-De Meyts, 2014). From gestational week 20, pre-SSC enter quiescence until birth (Culty, 2009). After puberty, there is significant divergence in the number of cell divisions that occur in the male and female germline. Oocytes are released each menstrual cycle and complete the first meiotic division and then arrest until fertilisation. This means that, following PGC expansion, oocytes undergo a single round of DNA replication only. This prolonged period in meiotic arrest has been proposed as the reason that aneuploidy increases with maternal age (section 1.2.4).

In contrast, SSCs undergo division approximately 23 times per year in order to support sperm production and maintain the SSC pool (Heller and Clermont, 1963). This means that, as men age, the number of times a SSC has completed DNA replication and cell

division increases. The approximate number of replications of the SSC of a man aged 20 is 190, rising to 650 for a man aged 40 (Figure 1.1). Given this, random errors can occur and accumulate through successive division, supporting the molecular observations that the total number of DNMs in the child increases with the age of the father.

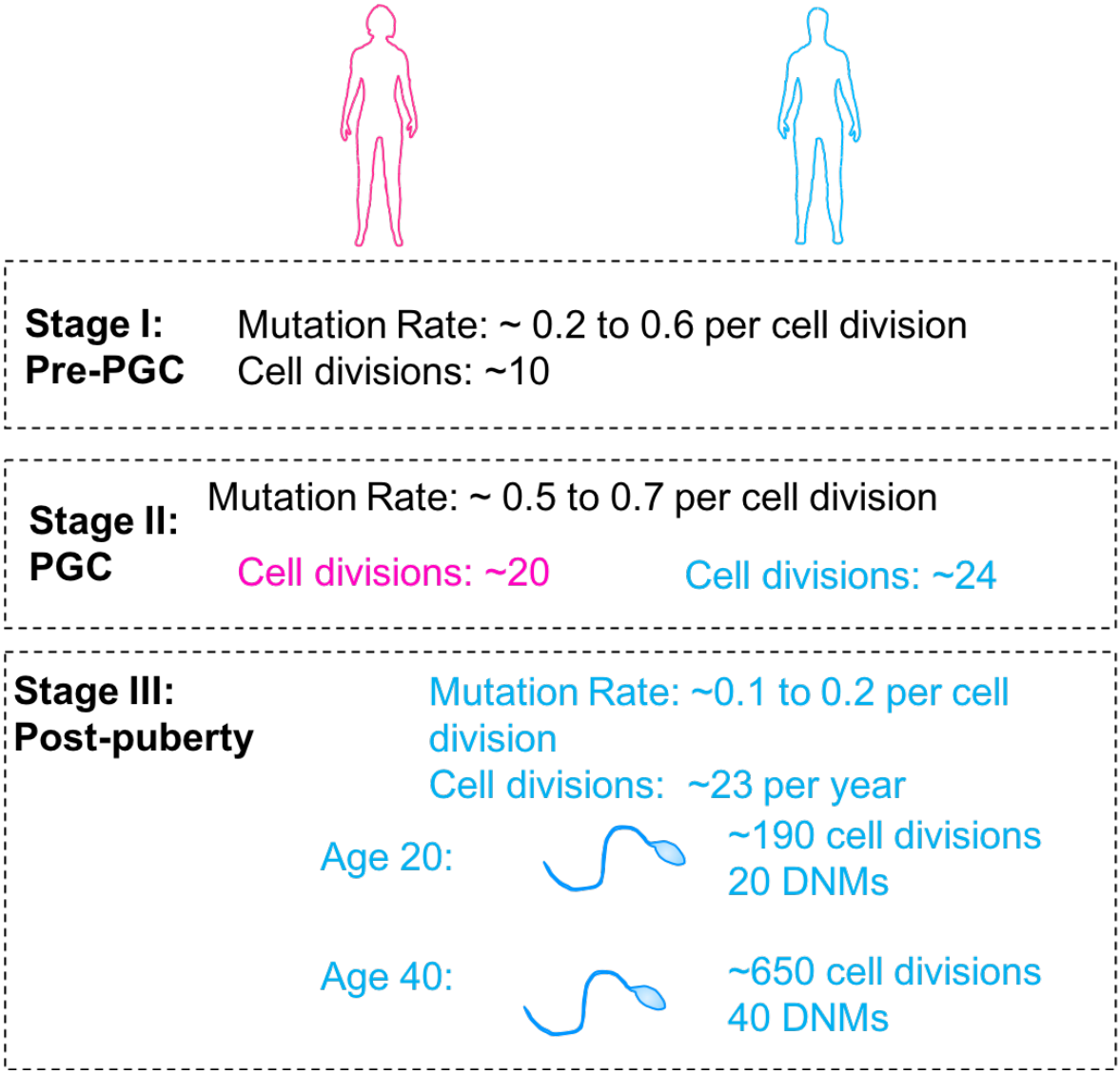


Figure 1.1: Overview of gametogenesis in females (pink) and males (blue) showing the relative number of cell divisions and estimated mutation rate at each stage. Data taken from Rahbari et al. (2016).

1.4 Epidemiology of paternal age

Separately from molecular observations of the parental origin of DNM and the contribution to disease, advanced paternal age (APA) has long been established as a risk factor

for a number of serious conditions through epidemiological observation. The first condition to be linked with a so-called "paternal age effect" was achondroplasia, the most common form of short-limbed dwarfism. More than 100 years ago, Wilhelm Weinberg proposed that birth order was a risk factor for achondroplasia (Weinberg, 1912; Crow, 2003). While it was not yet possible to identify the molecular cause of the disorder, during the decades that followed several publications recorded parental age (Grebe 1944, Krooth 1952, Stevenson 1957). Using the data recorded by Stevenson of 46 cases of sporadic achondroplasia combined with three other studies (n = 222), Penrose calculated the correlation between the father and mother's ages and, separately, the correlation between each of the parent's ages with the incidence of achondroplasia (Penrose, 1957). Maternal age is a confounding factor in such analysis. Penrose's analysis indicated that the increasing age of the father was the main risk factor in occurence of achondroplasia. The average paternal age recorded was 4.3 to 7.2 years higher than the population controls.

Subsequently, epidemiological observation of other monogenic disorders, congenital abnormalities (Green et al., 2010) and complex traits have been associated with increased risk due to APA. The first epidemiological report of the association between APA and a complex neurocognitive trait came from the analysis of a large Israeli cohort (Malaspina et al., 2001). The authors reported that APA was a risk factor for schizophrenia, with the relative risk of conceiving a child with schizophrenia increasing rapidly with paternal age. Autism Spectrum Disorder (ASD) has a number of genetic and phenotypic overlaps with schizophrenia and, intriguingly, cohort studies have also associated ASD with APA (Grether et al., 2009; Reichenberg et al., 2006). Children of fathers above 40 have a five-fold increase risk of ASD, versus fathers under the age of 30 years (Reichenberg et al., 2006). Other complex neurocognitive outcomes have also been linked with APA including bipolar disorder (Frans et al., 2008). The association between APA and risk for several types of cancer has been inconsistent. Some cancer types have been robustly linked to APA, including childhood cancers (Johnson et al., 2009). The daughters of men aged 40 years and older at conception are reported to have a 1.6 fold increased risk of breast cancer (versus a father <30 years) (Choi et al., 2005). A retrospective study of 434,933

births in Northern Ireland revealed an increased risk of acute lymphoblastic leukaemia (ALL) with paternal age above 35 years (Murray et al., 2002). Thus, the abundance of detrimental outcomes associated with APA warrants concern and further study of the mechanisms underlying the so-called "Paternal Age Effect".

The observed epidemiological association of certain disorders with paternal age has been proposed to be due to accumulation of errors during DNA replication in stem cells. This was first proposed by Penrose in 1955 and has also been described by Vogel and Rathenberg (1975). For example, an excess of rare DNM has been reported in individuals with schizophrenia (e.g. Gulsuner et al. (2013)), in line with the findings of the previously discussed sequencing studies that DNM burden increases with paternal age (section 1.2.6). Such a model is consistent with the differences in the biology of the male and female germline (section 1.3).

However, indications that the underlying biology explaining the observed epidemiological paternal age effect was not so simple came from a landmark study on the statistical analysis of spontaneous autosomal dominant disorders with parental age (Risch et al., 1987). This study presented 68 new cases of craniosynostosis-related conditions (Apert, Crouzon and Pfeiffer) for which parental age was known, as well as a review of selected case-control studies from the literature. The age distribution of parents with affected offspring were compared to the general population to calculate an observed/expected ratio for each disorder. If a model of "copy-error" leading to accumulation of mutations over time was true, then syndromes would be expected to increase with paternal age in a linear fashion. To test this, Risch compared the recorded parental ages with the mean paternal and maternal age predicted by a linear model and, separately, with an exponential model. The results indicated that disorders could be split into two groups based on mutation rate. One group was characterised by a low rate of increase in mutation rate, consistent with a linear model. The second group, however, showed a non-linear increase with paternal age and, as such, a model of "copy-error accumulation" was not sufficient to explain the observed birth prevalence of the associated disorders. This second group of syndromes include Apert, Pfeiffer and Crouzon syndromes and achondroplasia, all of

which show a sharp rate of increase with paternal age, even after correction for maternal age. This pattern cannot be explained by a simple model of mutation accumulation in the male germline over time.

1.5 Identification of a subset of mutations with an exclusive paternal origin

Due to the lack of knowledge of the molecular basis of these disorders at the time of publication of the Risch study, it was not possible to fully exclude a maternal component as a contributory factor to the risk of disease outcome with parental age. This changed with the identification of the underlying causative mutations which revealed two important findings.

Firstly, the causative mutations were highly specific. In 1994, two groups independently reported that achondroplasia was caused by mutation affecting a single codon in Fibroblast growth factor receptor 3 *(FGFR3)* (c.1138G>A; G>C), both of which encode the same amino acid substitution in the protein (p.G380R). That all patients - 16 in Shiang et al. (1994) and 23 in Rousseau et al. (1994) - demonstrated such a limited repertoire of mutation was a striking finding. It has subsequently been established that 97% of sporadic achondroplasia cases are caused by mutation at this codon. The birth prevalence associated with achondroplasia, 1 in 15,000, implies that this position is one of the most frequently mutating in the genome (Bellus et al., 1995). It should be noted that this figure includes inherited achondroplasia cases as well as sporadic. This substitution is occurring up to 1000-fold more commonly than expected based on the estimated human germline mutation rate (section 1.2.6 and Rahbari et al. (2016)).

The same pattern was observed for Apert syndrome when the causative mutation in 40 unrelated patients was found to be one of two base substitutions in *FGFR2*. The two substitutions affect adjacent amino acids with c.755C>G (p.S252W) being reported approximately twice as often as mutation in the neighbouring codon c.758C>G (p.P253R) (Wilkie et al., 1995). Of these two substitutions, c.755C>G occurs at a CpG while

c.758C>G does not. Apert syndrome has an estimated birth prevalence of 1 in 70,000 (Cohen et al., 1992). Mutations therefore appear to be elevated 200 to 800-fold above the expected background genomic rate for this type of substitution (section 1.2.6 and Rahbari et al. (2016)). Other mutations in *FGFR2* and *FGFR3* are associated with different phenotypic outcomes. These include reports of milder craniosynostosis related conditions in *FGFR2* such as Crouzon (Reardon et al., 1994) and Jackson-Weiss syndrome (Jabs et al., 1994). This allelic series and the different severity of phenotypes implies that distinct GOF mechanisms are taking place.

The second important finding that came from the identification of the underlying causative mutations in achondroplasia and Apert syndrome was the ability to determine the parent of origin. Natural sequence variation flanking the site of the mutation can be used to determine on which parental allele the mutation arose. This was carried out on informative family groups in multiple studies looking at causative mutations in *FGFR2* and *FGFR3* (Table 1.1). In all cases studied, the mutation was paternal in origin. In line with previous epidemiological findings, the average paternal age in these studies was also above the population average.

Table 1.1: Parent of origin for mutations in *FGFR2* and *FGFR3* associated with known syndromes. The number of cases with an established paternal origin is shown, with the number of informative families in each study indicated in brackets.

Gene	**Codon**	**Syndrome**	**Number of Paternal Origin**	**Average paternal age at birth (years)**	**Reference**
FGFR2	c.755C>G (p.S252W); c.758C>G (p.P253R)	Apert	57 (57)	33.3	Moloney et al. (1996)
FGFR2	multiple	Crouzon/ Pfeiffer	22 (22)	34.69	Glaser et al. (2000)
FGFR3	c.1138G>A (p.G380R)	Achondroplasia	40 (40)	25.86	Wilkin et al. (1998)
FGFR3	c.749C>G (p.P253R)	Muenke	10 (22)	34.5	Rannan-Eliya et al. (2004)

1.6 Direct measurement of mutation levels in sperm

Further study of the described highly localised mutations in *FGFR2* and *FGFR3* provides the opportunity to understand the unusual presentation of the associated syndromes. Given that the causative mutations were highly localised, sensitive mutation-specific assays were developed to directly estimate the mutation levels in sperm.

An initial study focused on codon c.755 in *FGFR2* (Goriely et al., 2003). In addition to the C>G mutation at this codon associated with Apert syndrome, a C>T mutation (encoding p.S252L) was identified associated with the milder phenotypic outcome of Crouzon syndrome (Oldridge et al., 1997). The other possible nucleotide change at this position (C>A) has not been described in humans but would result in a truncated protein. The sequence context, possible nucleotide substitutions at c.755 and the phenotypic effect are shown in Figure 1.2 (A). To circumvent low mutation levels, Restriction Enzyme Digestion (RED) was employed to enrich mutant copies of DNA. An initial *MboI* digestion was used to reduce the wild type background and the undigested fraction was then PCR-amplified and subjected to a second round of *MboI* digestion. Pyrosequencing was then used to measure the relative levels of different mutation species. Quantification of absolute mutation levels was enabled by spiking a known concentration of a sample containing a rare triple mutation at the site of interest (Figure 1.2). Initial validation of this method using a serial dilution indicated that measurements were accurate within twofold above concentrations of 10^{-5}. Mutation levels were then measured in 11 blood samples and 105 sperm samples from healthy individuals. Of these sperm samples, six were from unaffected men who had fathered a child with Apert syndrome caused by mutation at c.755. The sampled level of all possible mutations at this codon was extremely low in all blood samples ($<10^{-6}$) excluding the possibility that levels recorded in the sperm were due to technical artefact and providing an indication that the mutation rate at this site is not uniformly high as initially predicted (shown in grey squares in panel B of Figure 1.2). Similarly low levels ($<6.3 \times 10^{-6}$) were also observed for the C>A mutation and no change with age of the donor was observed (middle panel of B, Figure 1.2). In sperm,

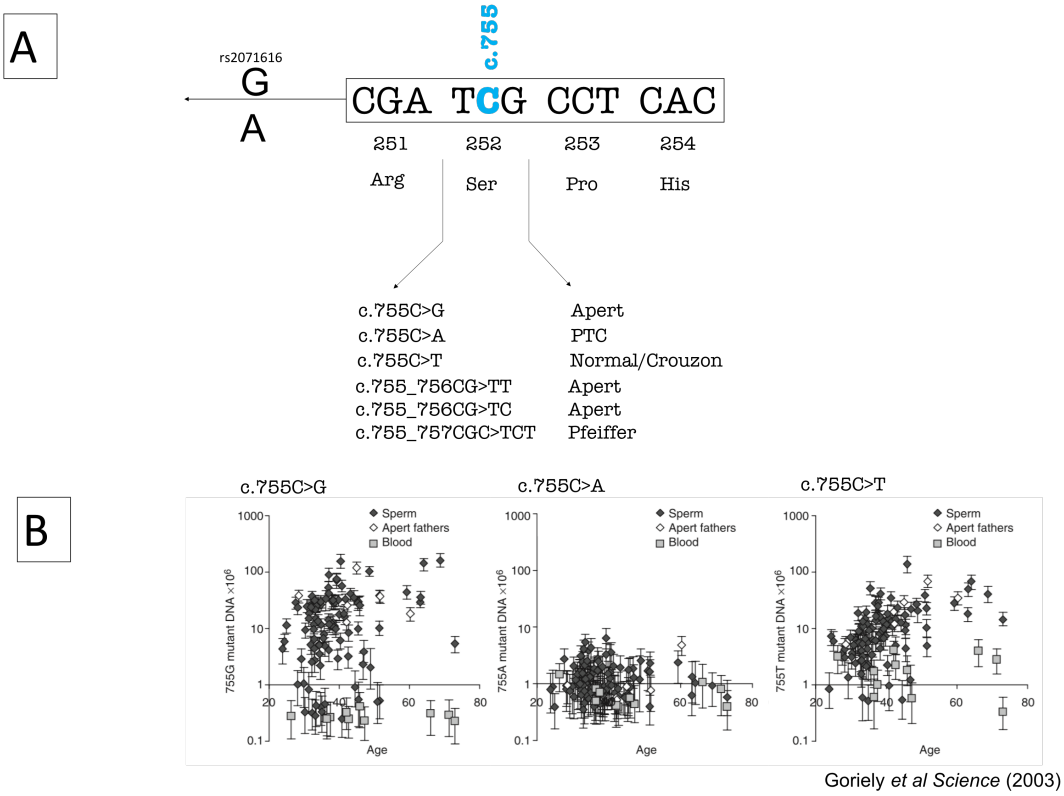


Figure 1.2: Overview of mutation-specific assay designed to quantify substitutions at c.755 of *FGFR2* in sperm. (A) Sequence context and possible nucleotide substitutions affecting c.755 (shown in blue) of *FGFR2*. A nearby SNP (rs2071616) is used for phasing. (B) Mutation levels observed at c.755 of *FGFR2* (y-axis) against age of the donor (x-axis). c.755C>G (left panel), c.755C>A (middle panel) and c.755C>T (right panel). Figure adapted from Goriely et al. (2003).

the most pronounced elevated mutation levels (maximum: 1.6 x 10^{-4}) were observed for the C>G mutation, which was on average 1.66-fold higher than C>T (maximum: 1.4 x 10^{-4}) (Figure 1.2). Levels of both mutations positively correlated with the age of the donor. Importantly, the men who had fathered a child with Apert syndrome had sperm levels within the range measured for the other men sampled in this study. The study concluded a strong correspondence between levels of mutation observed in sperm and the epidemiological birth data (section 1.4 and (Risch et al., 1987)).

The authors of this study then proceeded to investigate how mutational events were distributed across the two alleles of the *FGFR2* gene. If mutations at this codon truly occurred at random frequently enough to reach the elevated mutation levels observed, it would be expected that the mutation would be observed approximately equally with

the two possible alleles of a proximal SNP (rs2071616). Allele-specific PCR was used to detect the occurrence of the C>G and C>T mutation in 46 sperm samples heterozygous for the upstream G/A SNP. Most individuals studied demonstrated a tendency for the mutation to be present with one or other of the SNP alleles, with this effect seen more strongly for the C>G mutation. This indicates that the mutation does not occur as a hotspot. Rather, the combination of the higher absolute levels of the C>G mutation at c.755, along with the skewing towards one allele over the other in an individual, are compatible with the interpretation that the c.755C>G variant confers a positive selective effect. This indicates that, while the C>G mutation does not arise as frequently as the C>T mutation, it is selected for more strongly when it does arise.

A follow up study by the same group extended analysis to the three nucleotides preceding position 755 (752-754, Figure 1.2), again employing RED to reduce the wild type background and measure mutation levels in blood and sperm (Goriely et al., 2005). The data again showed that only 755 C>G and C>T mutation levels were elevated in sperm. The other positions surveyed included another position in a CpG dinucleotide (752G) which would be expected to have a comparable mutation level to 755C if the levels observed were due to the CpG context alone. Instead, levels of 752G>A mutation in the blood and sperm were in line with levels of 755C mutation observed in the blood, a tissue in which there is not proposed to be selection. Together, the mutation levels surveyed for the other position/mutation combinations supported the conclusion that this region of *FGFR2* is not generally highly mutable. A further finding of this study was the occurrences of double base substitutions involving position 755C (755_756CG>TC (p.S252F), 755_756CG>AC (p.S252Y). Tandem base substitutions are anticipated to arise extremely rarely in the genome by chance (Kondrashov, 2003). However, the occurrence of a mutation with relatively weak proliferative advantage (e.g. the c.755C>T mutation, associated with a milder phenotypic outcome (Oldridge et al., 1997)) could be enhanced by the occurrence of a second mutation on the same allele. The proband carrying the c.755_756CG>TC (p.S252F) mutation presented with classical Apert syndrome (Goriely et al., 2005). Her clinically unaffected father was heterozygous for c.755C>T

and her mother was wild type. This same tandem base substitution had previously been identified in the sperm sample of an unrelated individual (Goriely et al., 2003). Furthermore, a subsequent study of mutation levels at the p.G12 codon of HRAS in sperm identified 31 independent events of tandem base substitution (Giannoulatou et al., 2013).

These two studies of c.755 in *FGFR2* (Goriely et al., 2003, 2005) concluded that the observed mutation level in sperm correlated well with the observed birth prevalence of Apert syndrome and that the exponential age-related increase in prevalence could be accounted for by positive selection of the mutations. However, two other studies published around the same time came to conflicting conclusions (Tiemann-Boege et al., 2002; Glaser et al., 2003).

The first study (Tiemann-Boege et al., 2002) reported sperm mutation levels of the achondroplasia associated mutation (c.1138G>A, p.G380R) in *FGFR3* in 118 sperm donors. The authors reported no change in mutation level between the ages of 18 to 40 years and 55 to 80 years, but did observe a 2-fold increase in mutation level between the ages of 40 and 55 years. Using the mutation levels they collected only for the men aged between 18 and 59 years, the authors compared their observed age/mutation data with the reported age distribution of fathers of a child with achondroplasia from Risch et al. (1987). They reported that the number of actual achondroplasia births to fathers aged <30 years is smaller than would be expected from the mutation frequency data they observe. Conversely, the number of births to fathers aged >35 years was higher than predicted.

Similarly, a study by Glaser et al. (2003) came to the conclusion that sperm levels of the Apert syndrome mutations (c.755C>G and c.758C>G) in *FGFR2* did not match the birth prevalence. Mutation levels in 15 fathers of children with Apert syndrome were assayed separately and the authors concluded that these men had significantly higher mutation frequencies at an earlier age than men who had not fathered a child with Apert syndrome, although in fact only two of the fathers had mutation levels outside the range of the rest of the men.

The mutation frequencies reported in these two studies (Tiemann-Boege et al. (2002);

Glaser et al. (2003)) did not explain the epidemiological observations, but they did, however, highlight the extreme difficulty in quantifying low-level mutations. For example, the Tiemann-Boege et al. (2002) study showed a tendency to overestimate the true copy number of the c.1138G>A mutation at the lower level of the spectrum. If the mutation level was below the level the authors could detect (1 in 26,500), an individuals mutation level would be underreported, potentially explaining why the observed mutation level did not increase as expected with age in line with the epidemiological data. The Glaser et al. (2003) study performed mutation-detection assays using only 150 ng (50,000 sperm) per reaction. While the authors claim that they could detect a single copy of the mutation in this wild type background, such a low level of input DNA means that their ability to robustly detect low-level mutation is seriously impaired resulting in underquantification. Combined, these technical problems point to an inability to resolve low-level mutations accurately and sensitively. A summary of other mutation-specific assays performed directly in sperm is provided in Table 1.2.

1.7 Activating mutations disrupt SSC homoeostasis

A proposed explanation for the observation of elevated levels of specific activating mutations in *FGFR2* in the sperm of healthy men was positive selection of SSCs bearing the mutation (Goriely et al., 2003). Early evidence indicated the activating nature of the p.S252W substitution in FGFR2. Substitution of the amino acid tryptophan at p.S252 reduces the rate of ligand dissociation (Anderson et al., 1998) and crystal structures have shown that additional contacts are formed between the FGF2 ligand and the tryptophan side chain (Ibrahimi et al., 2004). However, at the time, the functional role of FGFR2 in SSCs was not yet clear. It was later demonstrated that the b and c isoforms of *FGFR2* are expressed in rat SSC lines (Goriely et al., 2005). More recently, expression of FGFR2 overlapping the known location of SSC in the human testis was demonstrated by immunohistochemistry (Martin et al., 2014).

As discussed in section 1.3, SSC divide asymmetrically to support self-renewal and

sperm production throughout the male reproductive life. It is important to note that much of the understanding of spermatogenesis and SSC maintenance in mammals comes from studies in mice rather than humans, although the basic elements of spermatogenesis are conserved from mice to humans (Fayomi and Orwig, 2018).

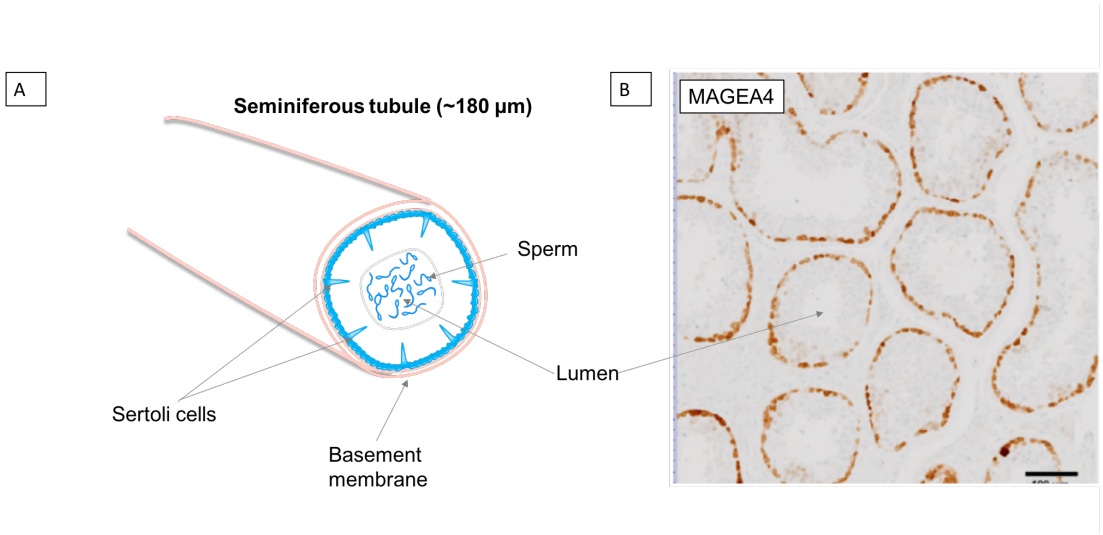


Figure 1.3: (A) Schematic view of a seminiferous tubule cross section showing the location of sertoli cells, basement membrane and the lumen. SSC are located at the periphery of the seminiferous tubule and divide to produce sperm which collects in the lumen. (B) MAGEA4 stained section of FFPE testis showing seminiferous tubule cross-sections. Image provided by Dr Geoffrey Maher.

SSCs occupy the periphery of seminiferous tubules (see tubule cross section in Figure 1.3). These cells account for approximately 0.02 to 0.03% of the total testis cell population (Tagelenbosch and de Rooij, 1993). Secretion of Glial cell-line derived neurotrophic factor (GDNF) by sertoli cells has been demonstrated to be essential for self-renewal of SSCs (Meng et al., 2000). GDNF is the ligand for RET, a receptor expressed on the surface of SSCs (Naughton et al., 2006). Expression of GDNF induces transcriptional regulators that have been associated with SSC self-renewal, including *Etv5, Bclb6, Egr2, Egr3, Lhx1* and *Tspan8* (Oatley et al., 2006). In turn, Etv5 instigates a positive feedback loop for GDNF signalling by inducing *Ret* expression. Other growth factors, including fibroblast growth factor (FGF) and epidermal growth factor (EGF), are known to be required for *in vitro* maintenance of SSC (Kubota et al., 2004). Levels of *Etv5* and *Bclb6* expression can also be increased through MAPK pathway activation by the ligand FGF2, which signals

through the MAPK or AKT pathway upon association with the receptor tyrosine kinase, FGFR2 (Ishii et al., 2012). Activation of the PI3K-Akt and MAPK pathways have also been shown to play a key role in self-renewal of SSCs (Lee et al., 2007).

Positive selection can be envisioned to act upon a mutation in a stem cell due to increased survival or altered proliferative capacity. To test the mechanism of selection associated with FGFR2 p.S252W, Martin et al. (2014) used a mouse model to compare the fitness of WT SSCs with SSCs carrying the FGFR2 p.S252W mutation.

Lentiviral transduction was used to create mouse SSCs that stably expressed either wild type human FGFR2 or p.S252W. To address whether there were changes to asymmetric cell division rate, levels of expression of self-renewal and differentiation-associated genes were measured. Higher expression of a transcription factor involved in SSC maintenance (Etv5) was observed in S252W SSCs, suggesting an increased capacity for self-renewal. The authors measured levels of pMAPK and pAKT at different concentrations of FGF2 in the two SSC subpopulations. They found that FGFR2 S252W SSCs produced greater pMAPK and pAKT levels than WT SSCs at low concentrations of the ligand. No difference was observed in the proliferative ability of either WT or mutant SSCs at reduced ligand concentration. This suggests that SSCs bearing the FGFR2 S252W mutation have an increased capacity for self-renewal in growth factor limited conditions. Competition between WT and FGFR2 S252W SSCs was tested by fluorescently labeling the respective SSCs and measuring their ratio by flow cytometry following coculture for up to 100 days. Regardless of the initial ratio of WT to S252W SSCs, S252W SSCs always outcompeted the WT SSCs in the presence of reduced FGF ligand. Finally, stem cell activity was assessed by transplanting the cultured WT and mutant SSCs into mice. After 2 months, significantly more colonies arising from the FGFR2 S252W SSCs were observed ($p < 0.005$). This work provides direct evidence that a mutation proposed to enter the population as a result of positive selection in the male germline is indeed associated with a functional effect that confers a selective advantage to the SSC in which it arises.

In addition to the specific activating mutations in *FGFR2* and *FGFR3* described

above (sections 1.5 and 1.6), mutations associated with other disorders are also proposed to be subject to positive selection in the male germline. A small class of disorders were identified that shared the common features of (1) an extreme bias in paternal origin of the mutations, (2) a paternal age effect, whereby fathers were on average 2 years older than the general population at the time of conception and (3) a high apparent germline mutation rate based on birth prevalence. These disorders include severe skeletal dysplasias (FGFR3: achondroplasia and thanatophoric dysplasias), craniosynostosis associated conditions (FGFR2: Apert, Crouzon and Pfeiffer), cancer predisposition syndromes such as multiple endocrine neoplasia (RET: MEN2A/2B) and neurocardiofaciocutaneous syndromes (HRAS and KRAS: Costello syndrome) and (PTPN11: Noonan syndrome). The corresponding mutations in these genes and the studies which have provided direct evidence of mutation levels in the male germline are listed in Table 1.2.

Strikingly, the currently identified mutations all affect genes which encode proteins acting within the growth factor RAS/MAPK signalling-pathway (Figure 1.4). As described above, this pathway plays an important role in SSC self-renewal and dysregulation has the potential to alter the growth and proliferative capacity of SSC.

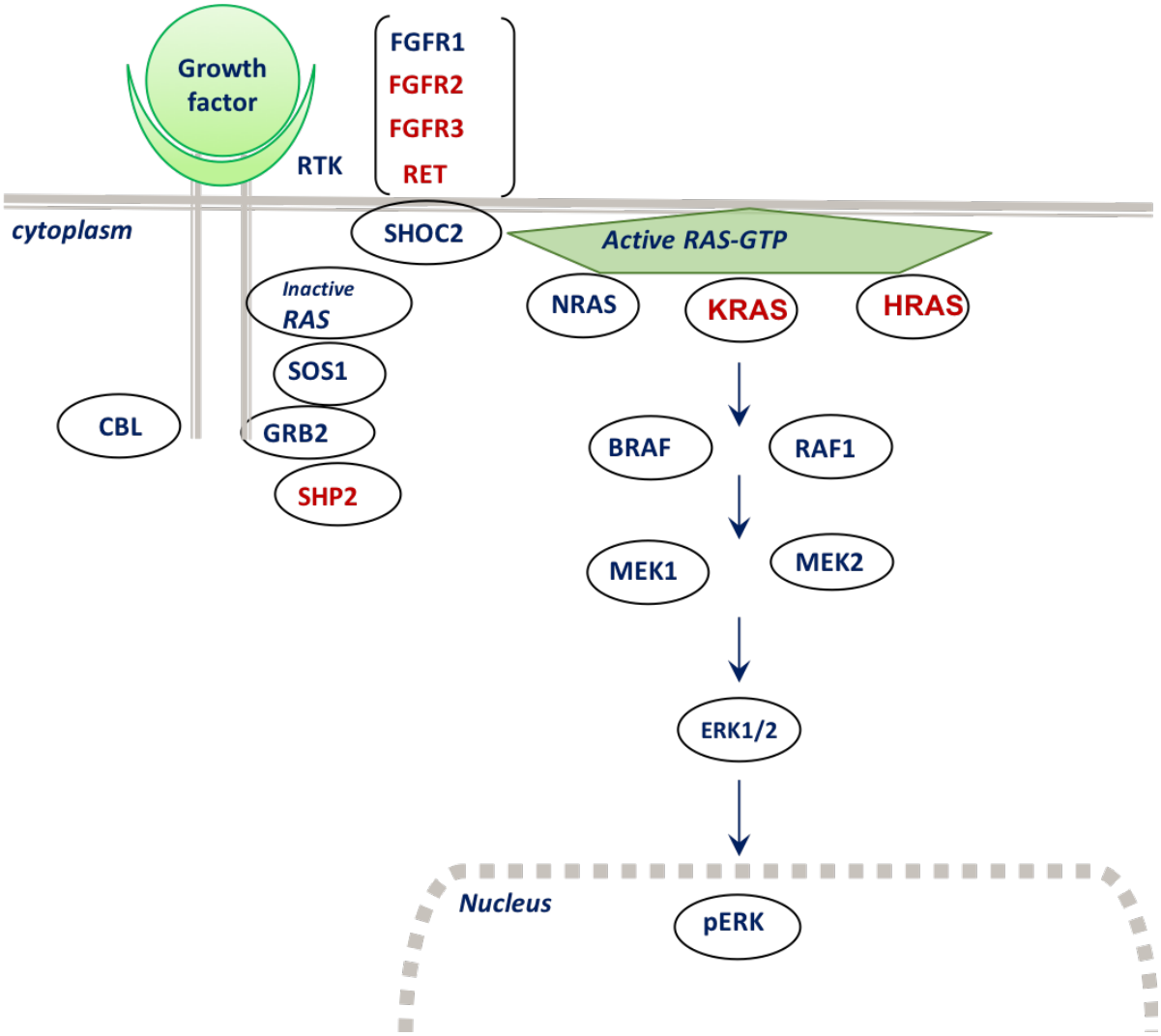


Figure 1.4: Overview of the RAS-Mitogen-Activated Protein Kinase (RAS-MAPK) signalling pathway. Proteins encoded by genes in which SSS associated mutations have been identified (Table 1.2) are coloured in red.

1.8 Selfish Spermatogonial Selection

The study of a subset of activating mutations associated with a high apparent mutation rate based on birth prevalence of the associated disorders, led to the proposal that specific activating mutations are positively selected in the male germline. This phenomenon was termed 'Selfish Spermatogonial Selection' to account for the apparent paradox of mutations with a detrimental outcome for the offspring being positively selected in the male germline (Goriely and Wilkie, 2012). Mutations occur at random in SSC but their functional effect may result in positive selection and expansion along the length of the seminiferous tubule in which the SSC is located (Figure 1.5). This is proposed to occur in the testes of all men as they age and accounts for the unusual features of a subset of PAE disorders (described in section 1.4) that could not be explained by a simple model of copy-error accumulation.

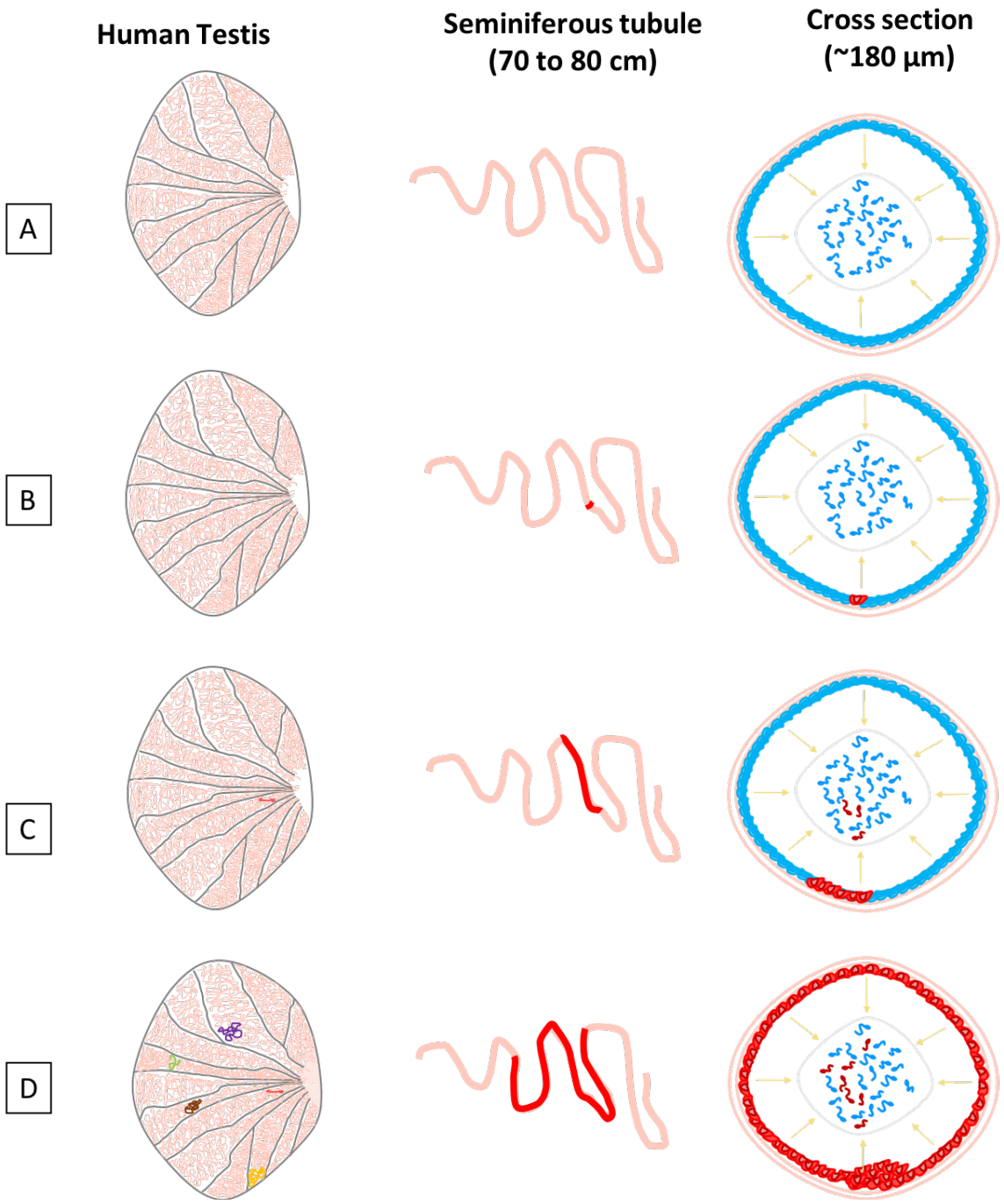


Figure 1.5: Proposed model of Selfish Spermatogonial Selection. (A) Within a human testis (left panel), there are 300 to 500 seminiferous tubules (Glass, 2005). Each seminiferous tubule is 70 to 80 cm in length (middle panel). Around the periphery of the seminiferous tubule, there are spermatogonial stem cells which divide to produce sperm (right panel). (B) An activating mutation may arise at random in a SSC (shown in red). (C) Over time, the SSC bearing the mutation clonally expands along the length of the seminiferous tubule accompainted by an enrichment in mutant sperm production. (D) Clonal expansion of the mutation continues over time. Multiple independent clonal events are anticipated in a single testis.

Table 1.2: Overview of loci where evidence for SSS has been established. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical DNA substitutions.

Gene	**Codon**	**Selfish Mutation**	**Germline disorder**	**COSMIC v86**	**Strategy A (Sperm)**	**Strategy B (Biopsy)**	**Strategy C (IHC)**
FGFR2	p.S252	c.755C>G (p.S252W)	Apert syndrome	66	Goriely et al. (2003); Goriely et al. (2005); Qin et al. (2007); Choi et al. (2008); Yoon et al. (2009)	Qin et al. (2007); Choi et al. (2008); Choi et al. (2012)	-
FGFR2	p.S252	c.755C>T (p.S252L)	Crouzon/ Pfeiffer syndrome	3	Goriely et al. (2003); Goriely et al. (2005)	-	-
FGFR2	p.P253	c.758C>G (p.P253R)	Apert syndrome	11	Yoon et al., 2007	Choi et al. (2008)	Maher et al. (2016a)
FGFR2	p.W290	c.870G>T (p.W290C)	Pfeiffer syndrome	9	-	-	Maher et al. (2016a)
FGFR2	p.Y340	c.1019A>G (p.Y340C)	Pfeiffer syndrome	-	-	-	Maher et al. (2016a)
FGFR2	p.C342	c.1024T>A (p.C342S)	Crouzon/ Pfeiffer syndrome	-	-	-	Maher et al. (2016a)
FGFR3	p.R248	c.742C>T (p.R248C)	Thanatophoric dysplasia I	296	-	-	Maher et al. (2016a)

FGFR3	p.Y373	c.1118A>G (p.Y373C)	Thanatophoric dysplasia I	469	-	-	Maher et al. (2016a)
FGFR3	p.G380	c.1138G>A (p.G380R)	Achondroplasia	25	Tiemann-Boege et al. (2002)	Dakouane Giudicelli et al. (2008); Shinde et al. (2013)	-
FGFR3	p.K650	multiple variants (c.1948A, c.1949A, c.1950G), including p.K650E; p.K650T; p.K650M; p.K650N; p.K650Q	Thanatophoric dysplasia II; Familial acanthosis nigricans; Severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN); hypochondroplasia	105	Goriely et al. (2005)9	-	Maher et al. (2016a) (c.1948A>G; p.K650E only)
HRAS	p.G12	multiple variants (c.34G, c.35G), including p.G12S; p.G12D; p.G12C; p.G12V	Costello syndrome	509	Giannoulatou et al. (2013)	-	-

HRAS	p.G13	c.37G>C (p.G13R)	-	482	-	-	Maher et al. (2016a)
KRAS	p.Q61	c.182A>G (p.Q61R)	-	143	-	-	Maher et al. (2016a)
PTPN11	p.D61	c.181G>T (p.D61Y)	-	53	-	-	Maher et al. (2016a)
PTPN11	p.A72	c.215C>T (p.A72V)	-	78	-	-	Maher et al. (2016a)
PTPN11	p.N308	c.922A>G (p.N308D)	Noonan syndrome	5	-	Yoon et al. (2013); Eboreime et al. (2016)	-
RET	p.M918	c.2753T>C (p.M918Y)	Multiple endocrine neoplasia 2B (MEN2B)	417	-	Choi et al. (2012)	-

1.9 Overlap and parallels of SSS with oncogenesis

The proposed expansion of advantageous mutations in the germline has parallels with the process of oncogenesis. Of the 16 codons thus far studied, in 14 of these identical or allelic amino acid substitutions have been been described in different human cancers, further demonstrating the activating nature of these mutations (Table 1.2). For example, mutations in *HRAS* at c.34 and c.35 (p.G12) are associated with the germline disorder Costello syndrome (Aoki et al., 2005) and also represent one of three reported hotspots in cancer (p.G12, p.G13 and Q61). Missense substitutions at these positions

cause constitutional signalling activation as a result of the RAS molecule being locked in its GTP-bound conformation (Scheffzek et al., 1997). Levels of mutation at the p.G12 site were quantified in blood and sperm samples and compared to an adjacent codon (p.A11) where activating mutations do not occur (Giannoulatou et al., 2013). Using the previously described RED-enrichment approach, the authors found the presence of known oncogenic mutations at p.G12 at elevated levels in sperm samples. The most abundant mutation observed at this site in sperm was p.G12S which was detected at levels above 10^{-5} (Giannoulatou et al., 2013).

Given the parallels that this process has with oncogenesis, and the identification of mutations in the testis that are also reported in cancer, the question naturally arises as to whether expansion of mutations in SSCs result in germline cancer. Spermatogonia very rarely give rise to tumours. Termed Spermatocytic Tumours (SpT), this type of tumour has an incidence in the USA of 0.4 to 2 cases per 1,000,000 (Ghazarian et al., 2015). This type of cancer has a distinct developmental origin, epidemiology and pathology from classical testicular cancer which develops from gonocytes and has a mean age of diagnosis in the early 20s (Moch et al., 2016). SpT originates from adult progenitors that are more differentiated (Giannoulatou et al., 2017). In contrast, SpT are well defined, slow growing tumours with a generally good clinically progression. Typically, they are more prevalent in older men with a mean age of diagnosis of 54 years (range 19 to 92 years) (Carrière et al., 2007).

An initial study screened 30 SpT samples for mutation in genes in the MAPK or PI3K signalling pathway that were either associated with a PAE syndrome or in which activating mutations had been described (Goriely et al., 2009). Activating mutations were identified in 20% of tumours, with two mutations in *FGFR3* (c.1948A>G p.K650E) and five in *HRAS* (3 x c. c.182A>G p.Q61R, 2 x c.181C>A p.Q61K) (Goriely et al., 2009). A subsequent study of 29 SpT samples identified two identical oncogenic mutations in NRAS (c.182A>G, p.Q61R) (Giannoulatou et al., 2017). Whole genome sequencing of five SpT and matched normal tissue was then performed. This revealed that the acquired SNV burden is low (approximately 0.2 per megabase (Mb), with an average of 6 (range: 2 to 9)

non-synonymous substitutions per tumour. Compared to the average burden of mutations in other tumours, this rate is exceedingly low. An analysis of point mutation burden through exome sequencing of 21 cancer types, totalling 3083 tumour-matched normal samples, revealed huge variation in the mutation frequency across different subtypes of cancer and individuals (Lawrence et al., 2013). Variation between the highest and lowest mutation rates across tumour types varied by more than 1000-fold. Some melanoma and lung cancer patients, for example, exhibited mutation frequencies in excess of 100 per Mb. This places SpT at the lower end of this spectrum (0.2 per Mb), which is not surprising given that the germline is under evolutionary pressure to limit mutations and preserve genome integrity. The signature of these mutations was similar to the previously described signature of DNM in the germline (section 1.2.4 and Rahbari et al. (2016)). While the SNV burden of SpT samples was low, extensive whole chromosome aneuploidy was observed, with 50 to 99 autosomes per tumour. Gene imbalance is proposed to be the initiating oncogenic event in SpT formation (Giannoulatou et al., 2017).

1.10 Clonal expansion in other contexts

Positive selection and subsequent clonal expansion of mutations as a result of their functional effect has precedent in other biological contexts, including oncogenesis and somatic mosaicism. Low level somatic mutations have been described in tissues with cycling stem cell populations, including the skin (Martincorena et al., 2015), blood (Genovese et al., 2014; Young et al., 2016; Acuna-Hidalgo et al., 2017) and colonic epithelium (Nicholson et al., 2018). Clonal haematopoiesis, the expansion of haematopoietic stem cells into distinct genetic subpopulations, is described in more detail in chapter 6.

Seborrheic keratosis is a common benign skin lesion and a frequent observation in older patients. An Australian study estimated that 100% of individuals over the age of 50 years exhibit seborrheic keratosis lesions, with an average individual burden of 69 lesions in individuals over the age of 75 (Yeatman et al., 2008). A lower prevalence was noted in a British study (Memon et al., 2001), although this is not surprising given that

these lesions are more common in parts of the skin frequently exposed to the sun (Kwon et al., 2003).

Strikingly, mutations in genes known to be subject to SSS have also been implicated in seborrheic keratosis. *FGFR3* mutations have been shown to have an important role in the development of these benign lesions. In a screen of 62 cases of seborrheic keratosis, activating mutations in *FGFR3* were identified in 39% of cases (Logié et al., 2005). These mutations were identical to those described in the germline in association with thanatophoric dysplasia (Tavormina et al., 1995) and SADDAN (Severe Achondroplasia with Developmental Delay and Acanthosis Nigricans). With one exception, the identified mutations have also been described in cancer. Activating RAS mutations (in *HRAS, KRAS, NRAS*) at known hotspots were found in 39% of seborrheic keratosis lesions (Hafner et al., 2012). Additionally, mutations in *PIK3CA* have been described (Hafner et al., 2007).

Mutations in genes with known roles in cancer have also been identified in biopsies from normal skin (Martincorena et al., 2015). Taking small biopsies from eyelids of donors aged 55 to 73 years, targeted ultradeep sequencing identified an average burden of 2 to 6 mutations per Mb per cell. These mutations were in known cancer driver genes including previously described hotspot mutations in *KRAS*, *NRAS* and *HRAS*. The NOTCH receptors, frequently mutated in the context of cancer, also had a high burden of mutation with *NOTCH1* being the most frequently mutated gene in this study. They found that the average clone size for neutral mutations ($0.15mm^2$) was significantly smaller than clones containing known driver mutations, with, for example, the identification of an activating *FGFR3* mutation in a clone spanning six adjacent biopsies ($0.69mm^2$). As with seborrheic keratosis, the epidermis of the eyelid is exposed to UV light and a mutational signature consistent with this exposure was observed.

In the colonic epithelium of apparently healthy individuals, mutations at a hotspot in *KRAS* have been described (Parsons et al., 2010). Targeted sequencing of bulk colonic tissue established the presence of the KRAS p.G12D mutation in the range of 0.2% to 1.8% (Nicholson et al., 2018). Modelling established that the only way for this activating

point mutation to reach the observed variant allele frequencies is through altered stem cell dynamics, specifically a 10-fold increase in lateral expansion of crypts carrying the mutation (Nicholson et al., 2018). This provides further evidence that pro-oncogenic mutations can alter the dynamics of stem cell replacement and drive accelerated fixation of the mutation and subsequent clonal expansion.

1.11 Methods to study selfish selection

As discussed in section 1.6, the detection and accurate quantification of low-level mutations in sperm represents a significant technical challenge. Studies were limited to assessing single codons in genes of interest (Table 1.2). To circumvent this, two alternative approaches have been utilised based on the expectation that SSS would lead to localised clonal enrichment within the testes (Figure 1.5). Within a single testis, there are 300 to 500 seminiferous tubules (Glass, 2005). Consequently, a biopsy of testis will contain distinct cross sections of multiple seminiferous tubules (Figure 1.6). A clonally expanding SSC is physically restricted to the seminiferous tubule in which it arises and mutations may reach elevated, more readily detectable levels.

1.11.1 Identification of clones in testicular biopsies

Systematic biopsying of the testes has been employed by Norman Arnheim's group and has been used to quantify and examine the spatial distribution and mutation frequency of five loci in *FGFR2, FGFR3, PTPN11* and *RET* (Table 1.2).

The first study examined the distribution and level of *FGFR2* c.755C>G (p.S252W) mutation (Qin et al., 2007). Four testes (from 2 individuals) were dissected into 192 biopsies each and a specialised PCR assay was used to measure mutation levels. Consistent with the expectation that the mutation arises and expands along the length of a tubule, the majority of mutations were contained within a small number of biopsies. The biopsies in which elevated mutation levels were detected clustered together in a non-random distribution in adjacent biopsies in and across testicular slices (Figure 1.6).

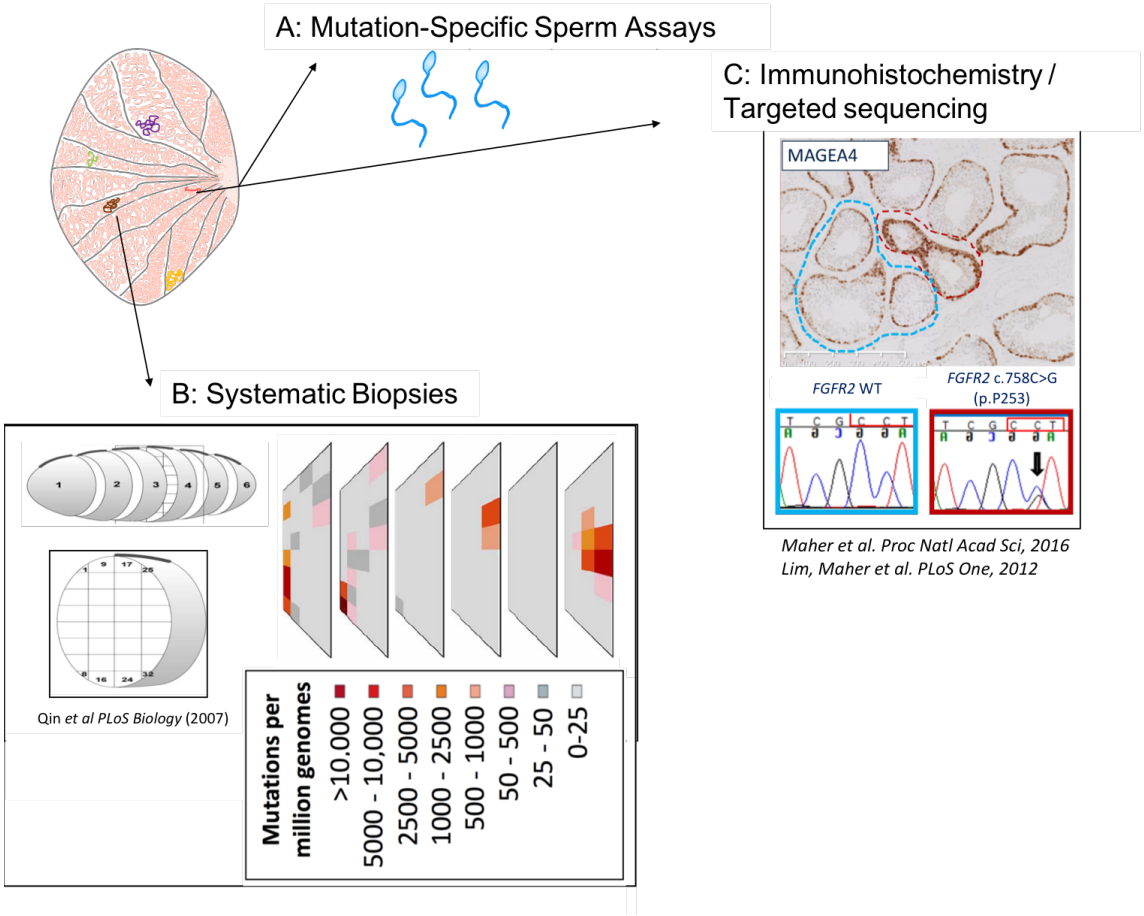


Figure 1.6: Strategies to study selfish spermatogonial selection. Details of the currently published studies utilising these three strategies are provided in Table 1.2. (A) Mutation specific assays have been developed to assess the levels of SSS associated mutations directly in sperm. (B) Mutation levels have been assessed by combining mutation specific assays with systematic biopsying of testicular slices. Image from Qin et al. (2007) shows the systematic dissection of a testis first into slices (top left) and then into biopsies (bottom left). The image on the right shows six slices of a single testis. The distribution of c.755C>G mutation is shown as the number of mutant molecules detected per million genomes. This reveals the clustering of biopsies with elevated mutation levels in and between testicular slices. (C) Identification of mutations through visualisation of immunopositive tubules followed by microdissection and sequencing. Image from Maher et al. (2016a) shows seminiferous tubule cross sections in a MAGEA4-stained section of FFPE human testis (top panel). Scale bar: 500µm. A tubule with a consistent pattern of increased immunopositivity around the periphery is shown outlined in red with surrounding tubules (e.g. the one outlined in blue) showing a normal appearance. Sanger sequencing of the immunopositive tubule (red) confirmed the presence of an Apert syndrome associated mutation in *FGFR2*. This mutation was not identified in the adjacent visually normal tubule.

Positive biopsies had mutation levels 10^3 to 10^4-fold higher than the rest of the testis (Table 1.3). The average mutation level observed across the biopsies as a whole closely matched the mutation levels in sperm extracted from the epididymis of the corresponding testis.

Table 1.3: Summary of mutation levels of *FGFR2* c.755C>G detected in biopsies from 4 testes of 2 men. Adapted from Qin et al. (2007)

Testis ID	**Age (years)**	**Mutation Frequency (Range)**	**Mutation Frequency (Average)**	**Mutation Frequency (Sperm)**	**Percentage of testicular biopsies (n=192) containing 95% of mutant genomes**
374-1	62	$<10^{-6}$ to 0.027	3.8×10^{-4}	4.5×10^{-4}	6.25
374-2	62	$<10^{-6}$ to 0.007	6.7×10^{-5}	3.9×10^{-5}	2.60
854-1	54	$<2 \times 10^{-6}$ to 0.032	5.9×10^{-4}	2.6 x 10-4	3.13
854-2	54	<2 x10^{-6} to 0.047	6.8×10^{-4}	1.1×10^{-4}	3.65

The Arnheim group have also developed and employed a computational model to test whether the localised distribution and mutation levels observed could be explained by the position of interest being a mutational hotspot (Qin et al., 2007). There are two phases of this model. In the first phase ('growth', from zygote formation to puberty), germline cells undergo symmetric division and increase in number exponentially. A mutation could arise during any division and has the potential to reach high levels across a region of the testis due to the exponential increase in cell number that is occurring. In the second phase of the model ('adult'), germline cells undergo asymmetric division to support self-renewal and sperm production. Cell numbers during this stage remain largely constant. Consequently, a mutation occurring during this phase would exhibit a more restricted distribution in the testis. The nucleotide substitution rate per cell division is inferred from simulations to find the rate that would match the observed mutation data. By varying this parameter during the two different phases of the model, the authors were not able to find conditions that would match the biologically observed data. The occurrence of mutations early in the growth phase (which have the potential to expand through the

exponential increase in cell number) were found to be unlikely. The majority of mutations occurred in the adult phase, resulting in little variation in the mutation level across the testis. This is inconsistent with the biological observation of mutation distribution in localised biopsies (Figure 1.6). Increasing the mutation rate per cell division to produce more mutations in the growth phase resulted in the mutation level being much higher than was actually observed (Table 1.3). The results of these computer simulations resulted in the authors rejecting the hypothesis ($p < 10^{-6}$) that the genomic position was a hotspot for mutation. If, however, the presence of the mutation in the SSC promotes rare symmetric divisions, this could allow the mutant SSC to expand locally over time. The authors therefore added a selection parameter such that a SSC has a probability of undergoing asymmetric or symmetric division. They used the experimentally observed mutation levels and distribution within the testis in their model and inferred that the probability of a symmetric division is approximately 1 in every 100 divisions. Simulations with these model parameters reveal clustered biopsies with elevated mutation levels, consistent with the biological observations. Thus, computer simulations support the rejection of the hypothesis that SSS associated sites of mutation appear to have a mutation rate above the background level due to these sites being hotspots. Furthermore, incorporating a selection parameter into the model meant that simulations matched observed biological distributions and mutation levels (Qin et al., 2007). As was described in section 1.7, positive selection of SSCs bearing this mutation in *FGFR2* (c.755C>G, p.PS252W) has been demonstrated (Martin et al., 2014).

1.11.2 Visual identification of clones

The clustered mutation foci identified by the above strategy represent putative clones occurring and expanding along the length of a seminiferous tubule (Figure 1.5). These events have been visually detected in the testis using a strategy based on immunohistochemistry (Lim et al., 2012; Maher et al., 2014, 2016a). In serial sections of FFPE (formalin-fixed, paraffin embedded) tissue from human testes, tubules were identified with abnormal expression of spermatogonial markers (MAGEA4, FGFR3, phospho-AKT). An

example is shown in Figure 1.6. This approach revealed clustered regions of seminiferous tubules that exhibited reproducibly enhanced staining at the periphery, a pattern distinct from the appearance of other neighbouring tubules. Following laser capture microdissection, whole genome amplification and targeted sequencing of the coding regions of 135 candidate genes, GOF mutations were identified in these visually abnormal tubules. In the example tissue section shown in Figure 1.6, the presence of a mutation associated with Apert syndrome (*FGFR2* c.758C>G p.P253R) was confirmed by Sanger sequencing. This mutation was not detected in an adjacent tubule with normal immunostaining. FFPE sections from 15 men, ranging in age from 39 to 96 years old, were analysed leading to the identification of 24 immunopositive tubules in 14 testes. In 16 of these immunopositive tubules, 11 distinct GOF mutations in five genes (*FGFR2, FGFR3, PTPN11, HRAS* and *KRAS*) were identified (Maher et al., 2016a). Details of these mutations, all of which have been previously associated with disease or cancer, are listed in Table 1.2.

This was the first study of SSS that moved beyond mutation specific assays, significantly widening the number of nucleotides assessed and providing the first visual representations of the geographic extent of clones. It was successful in increasing the range of mutations detected and provided evidence for the involvement of an additional gene (*KRAS*), also encoding a member of the RAS-MAPK pathway (Figure 1.4). Of the 11 GOF mutations described in this study, only 3 are associated with long term survival of the affected offspring. This may reflect a limitation of the experimental approach as tubules with a strongly immunopositive appearance are more likely to be identified. Thus, with current spermatogonial markers, it is not possible to identify tubules that may carry a mutation which is less strongly activating and leads to milder pathway disruption. The approach also relies on the availability of good quality FFPE tissue. Genetic analysis of this material is challenging due to the poor quality of DNA extracted from FFPE tissue (Maher et al., 2016b). This limits the application of this strategy in identifying additional sites in the genome subject to SSS.

1.12 Implications of SSS

Preferential expansion of disease-associated and oncogenic mutations in the male germline, where the mutations may be inherited by the next generation, demonstrates the biological relevance of the phenomenon of SSS. Although the disorders resulting from mutations that have been associated with SSS (Table 1.2) are individually rare, collectively they are estimated to account for between 1 in 2000 and 1 in 5000 live births (Sharland et al., 1992), representing a considerable disease burden. Direct assessment of mutation levels in sperm has established that men who have fathered an affected child have mutation levels within the normal range for their age group (Goriely et al., 2003, 2005). Given the demonstrated bias in paternal origin of DNM (section 1.2.6) and the significant contribution DNMs have to developmental disorders (section 1.2.5), an understanding of the wider implications of SSS on a population level is essential.

This is particularly relevant given that paternity is increasingly being delayed. The global trend of delaying parenthood - attributable to social, economic, education and medical factors - shows no signs of abating (Mills et al., 2011). Examining the publicly available parental age data for England and Wales demonstrates a very clear trend towards delaying parenthood (Figure 1.7). Over the last five decades, the proportion of births to men aged 25 to 29 years has fallen by approximately half (from 31.9% in 1961 to 15.9% in 2016). Conversely, the births to men aged 35 to 39 has doubled (from 14.0% to 28.7% (Figure 1.7 panel B). Additionally, the average paternal age at conception of the first child has increased by almost four years between the 1970s and 2011, rising from 29.06 to 32.75 years (Figure 1.7). A similar pattern can be observed in other Western countries, including Germany, Spain and Estonia (Willführ and Klüsener, 2016). The same trend is also observed in the USA: a recent retrospective analysis of 168,867,480 recorded births revealed that births to men aged 40 had more than doubled, accounting for 8.9% of births in the period 2011-2015 (up from 4.1% between 1972 and 1975) (Khandwala et al., 2017).

Published work (Table 1.2) has demonstrated that members of the RAS-MAPK sig-

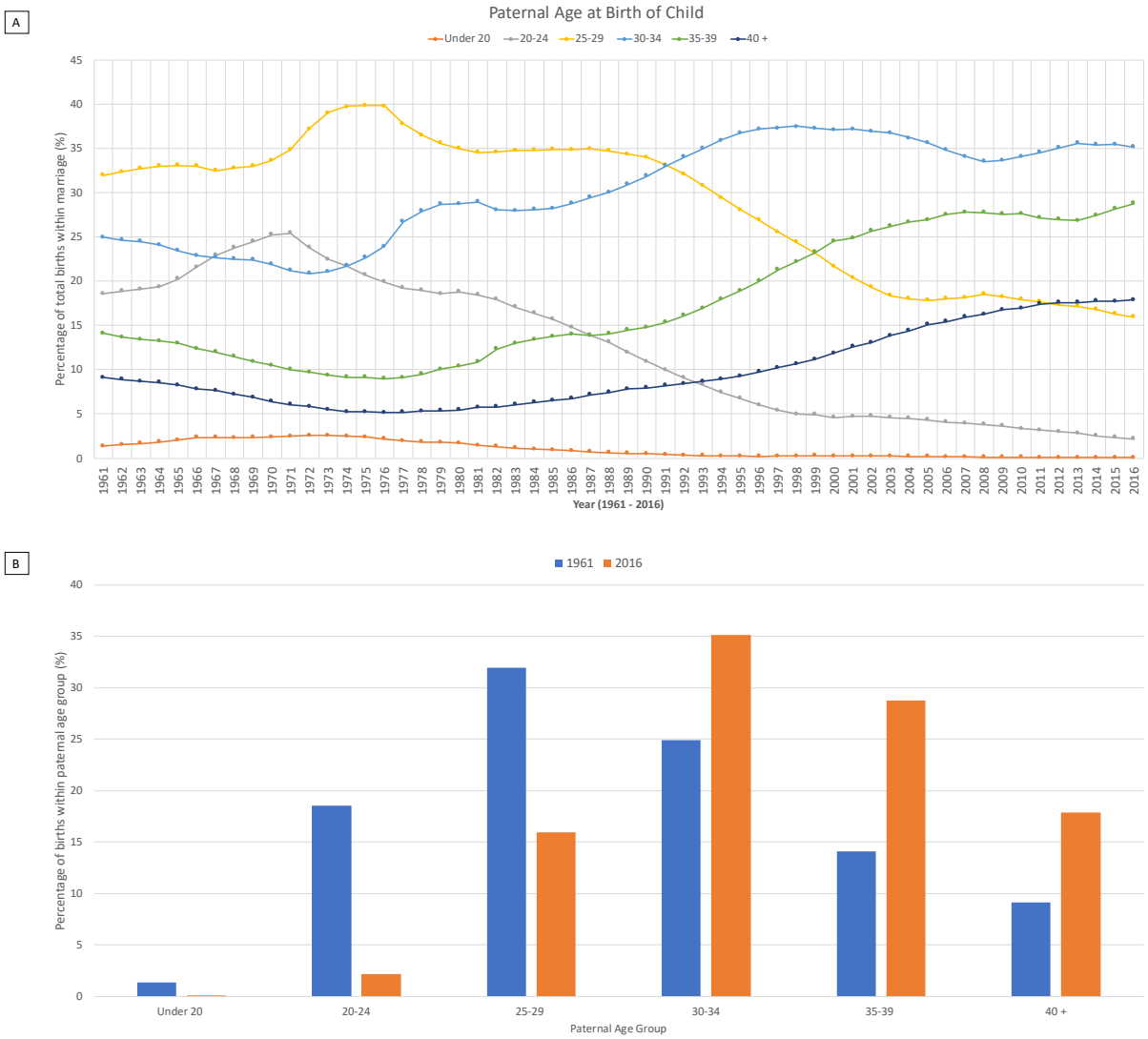


Figure 1.7: (A) Percentage of total births within marriage to each paternal age group (<20 years (orange), 20 to 24 (grey), 25 to 29 (yellow), 30 to 34 (light blue), 35 to 39 (green) and >40 (dark blue) (y-axis) between 1961 and 2016 (x-axis). Paternal age information was taken from the Office for National Statistics for England and Wales. (B) Comparison of the proportion of births within each paternal age group in 1961 (blue) and 2016 (orange).

nalling pathway are vulnerable to SSS. This pathway is important for SSC maintenance in the testis and also in other cellular contexts, for example in brain development (Kim et al., 2003), memory (David Sweatt, 2001) and synaptic plasticity (Brambilla et al., 1997). Other signalling pathways are also active in the testis (e.g PI3K-AKT, Lee et al. (2007)) and members of these pathways also represent candidates for SSS.

Thus far, only strongly activating mutations have been described in association with SSS. In the germline, these cause deleterious disorders that will be eliminated from the

population due to low reproductive fitness. For example, mutation levels in sperm of the strongly activating FGFR3 p.K650E mutation have been shown to increase with age (Goriely et al., 2009). In the germline, this mutation causes the embryonic lethal condition thanatophoric dysplasia type II (Tavormina et al., 1995). Mutations which are less strongly activating, though still deleterious, may be maintained in the population as a result of their positive selection in the male germline (Goriely and Wilkie, 2012). However, assessment of mutations with a milder phenotypic outcome (and therefore less relative enrichment in sperm), has been limited due to the technical difficulties associated with studying low-level mutation.

1.13 Developing alternate strategies to study SSS

The technical challenge in studying SSS is to find a balance between surveying genomic sites at scale while achieving the sensitivity required to distinguish true variation from inherent error associated with the technology used.

1.13.1 Next Generation Sequencing

Next Generation Sequencing (NGS) techniques offer affordable scalability but are confounded by the associated error profile which is both context-and platform-specific. Currently, Illumina sequencing platforms dominate the market. Illumina uses sequencing-by-synthesis technology whereby modified dNTPs are cleaved to allow addition of the next base in the template. Cleavage removes a fluorophore that emits a signal identifying the incorporated nucleotide as A, C, G or T. To generate enough signal for the camera to detect incorporation, template molecules are first amplified on a solid surface. Multiple studies have demonstrated that errors associated with Illumina sequencing are non-random and specific biases can be related to sequence context (Schirmer et al., 2016; Nakamura et al., 2011; Manley et al., 2016).

A systematic comparison of three Illumina platforms (HiSeq2000, MiSeq, and now the superseded Genome Analyzer II) was conducted using *in vitro* metagenomic data

to determine biases associated with sequencing (Schirmer et al., 2016). To examine the context in which an error occurs, the data were divided into 4-mer motifs (corresponding to 64 possible sequences) and it was found that the most common sequence motifs occurring in the three nucleotides prior to the position with an error were "GGG", "CGG" and "AGG". Higher error rates were observed in the context of multiple ddGTPs with a bias in substitution errors following motifs that end "GG". This has also been reported by other studies (Nakamura et al., 2011; Minoche et al., 2011; Dohm et al., 2008). The authors also observed a bias of incorporation of ddGTPs, which they attributed to the preferential interaction of the guanine base with a side chain of the *Taq* DNA polymerase employed in reversible dye-terminator sequencing (Schirmer et al., 2016).

Error profiles have also established that error rates increase towards the end of the read. This is attributed to problems with phasing which accumulate as the run progresses. With each sequencing cycle, individual molecules in a cluster, which should be the same length, fall out of synchronisation with each other. An individual molecule may fall behind ("phasing") or move ahead ("pre-phasing") of the rest. Error rates are approximately double for read 2 compared to read 1. This has been recorded for error rates associated with substitutions (read 1: 0.0021 errors per base; read 2 0.0042 errors per base) and indels, which are a much rarer occurence (read 1: $<5.1 \times 10^{-6}$; read 2: $<4.9 \times 10^{-6}$) (Schirmer et al., 2016). Base substitution does not occur randomly, with 'G' being more likely to be incorporated in place of 'A', 'C' or 'T' and a 'T' being more likely to be incorporated in place of 'G'. Indel errors are not as sequence specific as substitution errors and were observed across the length of the read, although repeat motifs (e.g. "AAA" and "TTT") were seen in association with deletion.

Point mutations are the predominant source of error in Illumina sequencing and, understanding this, and the patterns that lead to them, is important when using NGS technology to identify variants. Phred quality scores are a measure of the confidence that a given nucleotide has been correctly called. For example, a quality score of 30 refers to a probability that a base call is incorrect of 1 in 10^3. Quality scores are obtained by matching properties of the sequencing clusters (e.g. intensity and signal to noise ratios)

to empirical metrics calculated by Illumina platforms (Zhang et al., 2017). Maintaining an optimal cluster density is important to obtain a high proportion of reads passing the quality filter. Increasing the cluster density results in a higher number of reads, only to a certain level, after which a lower number of reads will pass the quality filter (Manley et al., 2016).

1.13.2 Library preparation

Additional sources of error can arise during preparation of a sequencing library. Errors can be introduced during PCR by DNA polymerase incorporating the incorrect nucleotide. As well as nucleotide misincorporation, other types of error that can be introduced during PCR include template-switching, DNA damage and PCR-mediated recombination (Potapov and Ong, 2017). The errors are both DNA sequence context and polymerase specific (Hestand et al., 2016). An error that arises during the first cycle of PCR amplification would affect 100% of the product, however, a mutation occurring later affects a smaller proportion; understanding this pattern is necessary when trying to distinguish true sequence variants from errors introduced during PCR amplification.

Canonical *Taq* polymerase is estimated to have a total error rate of 1.8×10^{-4} errors per base per doubling (Lee et al., 2016). High-fidelity polymerases can have fidelity rates 1 to 2 orders of magnitude higher than canonical polymerases. In a single-molecule sequencing assay to study errors generated during PCR, base substitutions introduced by high-fidelity polymerases such as Q5 were found to occur at rates two-fold lower than substitutions introduced by thermocycling induced DNA damage (Potapov and Ong, 2017). The authors of this study reported that DNA damage introduced by temperature cycling was the major source of error for high-fidelity polymerases.

1.13.3 Overcoming sources of error

The problem of distinguishing true variation from introduced errors can be aided by technology that uniquely tags the progenitor molecules with a barcode sequence prior

to PCR amplification and sequencing. Termed unique molecular identifiers (UMI), their incorporation facilitates downstream consensus calling and PCR de-duplication using bioinformatic approaches (Salk et al., 2018).

An example of technology that uses UMIs are single molecule Molecular Inversion Probes (smMIPs) (method details are in Figure 5.1 on page 130). smMIPs are designed to flank a genomic region of interest and allow the independent capture of a single DNA strand. The probes are comprised of two hybridisation arms and a common backbone that allows multiplex amplification. A 10 nucleotide UMI provides the theoretical ability to tag more than 1 million progenitor molecules. smMIPs have previously been used in high-throughput resequencing projects of candidate autism genes and clinical cancer samples (Roak et al., 2012; Hiatt et al., 2013; O'Roak et al., 2014). Additionally, commercially produced capture-based solutions that make use of UMIs are also available, including the HaloPlex High Sensitivity panel (Agilent Technologies). The ability to perform high-throughput screens and generate consensus calls may enable the sensitive detection of low-level variants in testis samples. This will be explored in this thesis.

1.14 Aims

Therefore, the aims of this thesis are to:

- Survey sites in the genome that are potentially subject to SSS. By examining genes that encode members of the RAS-MAPK pathway, members of other signalling pathways in the testis and candidate genes involved in oncogensis, developmental disorders, overgrowth conditions and intellectual disability, I am to address whether SSS has broader implications than previously known (chapters 3, 4, 5 and 6).

- Examine technology and approaches that provide the scalability required to survey multiple sites in the genome at once (chapters 3, 4, 5 and 6).

- Assess whether technology that incorporates UMIs provide the sensitivity and confidence required to distinguish true low-level variation from confounding errors (chap-

ters 4, 5 and 6).

Chapter 2

Materials & Methods

2.1 Ethics approval

Ethical approval was given for the use of human testicular tissue by the Oxfordshire Research Ethics Committee A (C03.076: Receptor tyrosine kinases and germ cell development: detection of mutations in normal testis, testicular tumours and sperm).

2.2 General laboratory methods

2.2.1 Polymerase Chain Reaction

PCR amplification was performed in a volume of 25 µl using 30 ng genomic DNA, with 200 µm dNTPs, 0.5 µm primers and 0.5 units of Q5 High-Fidelity DNA Polymerase (NEB). Cycling conditions consisted of a 30 second initial denaturation step at 98°C, followed by 30 cycles of 98°C for 10 seconds, 30 seconds annealing (temperature in Table A.2 in Appendix A.3) and 30 second extension at 72°C, with a final extension at 72°C for 2 minutes. All oligonucleotides were ordered from IDT and were prepended with a universal adaptor sequence as follows ACACTGACGACATGGTTCTACA (forward primer) and TACGGTAGCAGAGACTTGGTCT (reverse primer). The sequences of oligonucleotides used in this thesis are listed in appendix A.3.

2.2.2 Product visualisation

PCR products were visualised by agarose gel electrophoresis, using 2% gels in 1x TBE buffer (8.9 mM Tris-base, 8.9 mM boric acid, 2.5 mM EDTA). Ethidium bromide (1 µg/ml was added to the gels prior to pouring to allow visualisation of the DNA by transillumination with UV light. Samples were loaded in buffer (6x: 0.25% bromophenol blue, 30% w/v sucrose) and run simultaneously with the appropriate ladder (NEB).

2.3 Testis samples

Testes from seven men were either commercially sourced or obtained locally from research banks post-mortem (Table 2.1). Biopsies from Tes3 (slice D, 32 biopsies), Tes4 (slices B-G, 153 biopsies), Tes5 (slice J, 27 biopsies), Tes14 (slice D, 34 biopsies) and Tes10 (slice F, 30 biopsies) had previously been extracted and quantified (Indira Taylor and Professor Anne Goriely).

Table 2.1: Summary of testis samples

Testis ID	**Source**	**Age (years)**	**Cause of Death/Surgery**
3	Medical Solutions (Peterborough, UK)	87	Pneumonia
4	Medical Solutions (Peterborough, UK)	90	Heart failure
5	Medical Solutions (Peterborough, UK)	34	Lung cancer
10	John Radcliffe Hospital (Oxford, UK)	83	Hernia
14	John Radcliffe Hospital (Oxford, UK)	71	Prostate cancer
18*	John Radcliffe Hospital (Oxford, UK)	58	Cardiac arrest
19*	John Radcliffe Hospital (Oxford, UK)	58	Cardiac arrest
24	John Radcliffe Hospital (Oxford, UK)	50	Cancer

* indicates that the testes were from the same individual.

2.4 Testis dissection

Testes were collected and stored in saline solution within 48-hours post-mortem. During dissection, each testis was placed in PBS and kept on ice. Extraneous tissue (epididymis,

tunica albuginea) was removed and individual seminiferous tubules were dissected out using forceps. Seminiferous tubules were measured and placed on dry ice prior to long term storage at -80°C.

Tes14 slice B was removed from -80°C storage and defrosted on ice in PBS. The tissue was dissected into 102 biopsies.

2.4.1 DNA extraction and quantification

DNA extraction from biopsies and seminiferous tubules was performed using the Gentra Puregene Tissue Kit (Qiagen) following the manufacturer's protocol. Sample concentration and purity was measured using the Qubit fluorometer (ThermoFisher Scientific) and NanoDrop spectrophotometer (ThermoFisher Scientific).

2.5 Variant validation

DNA from at least one putative-positive biopsy sample and control samples (unrelated blood genomic DNA and DNA from other testicular biopsies) were screened by PCR as described in section 2.2.1.

2.5.1 Library preparation and sequencing

Illumina-specific adaptors and 10 bp sample indexes were attached to amplified products using the Access Array Barcode Library for Illumina Sequencers-384, Single Direction (Fluidigm) at a concentration of 0.4 µM per primer, with iProof (Bio-Rad) for 7 cycles.

Indexed products were diluted, pooled and purified with Agencourt AmPure XP Beads (Beckman Coulter). Quantification was performed using the 2200 TapeStation (Agilent Technologies) with a High Sensitivity D1000 ScreenTape and using the Qubit Fluorometer (ThermoFisher Scientific).

Libraries were diluted to a final concentration of 9 pM and sequenced using the Illumina MiSeq platform (300v2 reagent kit), following the manufacturer's instructions for custom indexing. Custom primers added to the sequencing cartridge were as follows:

CS1-Seq for read 1 (5′-A+CA+CTG+ACGACATGGTTCTACA-3′)

CS2-Seq for read 2 (5′-T+AC+GGT+AGCAGAGACTTGGTCT-3′)

RC-CS2 for the indexing read (5′A+GAC+CA+AGTCTCTGCTACCGTA-3′).

+ indicates a Locked Nucleic Acid (LNA) modified nucleotide.

2.5.2 Sequence alignment and variant calling

Demultiplexed sequencing reads were aligned to the human genome (hg19) using BWA-MEM version 0.7.12 (Li, 2014). Summary tables of the calls across the aligned target region were generated using a custom script based on SAMtools mpileup (described in section 3.4.1). A base call was only considered if the mapping quality was Q20 and phred score Q30. Nucleotide changes within the target regions were visualised in R using ggplot2.

All genomic coordinates refer to GRCh37 (hg19).

Variants were annotated according to the following transcripts - *ADNP*: NM_001282531, *APC*: NM_001127510, *AKT1*: NM_001014431, *AKT3*: NM_005465, *BRAF*: NM_004333, *CHD8*: NM_020920, *CBL*: NM_005188, *CTNNB1*: NM_001098209, *DNMT3A*: NM_175629, *DYRK1A*: NM_001396, *EZH2*: NM_004456, *FGFR2*: NM_000141, *FGFR3*: NM_000142, *HRAS*: NM_001130442, *KRAS*: NM_033360, *LRP5*: NM_002335, *MAP2K1*: NM_002755, *MAP2K2*: NM_030662, *MECP2*: NM_004992, *MYD88*: NM_002468, *NAA10*: NM_003491, *NF1*: NM_001042492, *NRAS*: NM_002524, *PPP2R5D*: NM_006245, *PTEN*: NM_000314, *PTPN11*: NM_002834, *RAF1*: NM_002880, *RET*: NM_020975, *SOS1*: NM_005633, *TP53*: NM_000546, *UBE3A*: NM_130838

2.6 HaloPlex HS Panel

2.6.1 Panel design

Candidate genes were prioritised and selected for inclusion in the panel based on a number of criteria including the size of the gene, G/C content and testicular expression (section

4.2).

The selected list of genes and target regions was uploaded to SureDesign, an online panel design program from Agilent Technologies (https://earray.chem.agilent.com/suredesign/). Probes were designed to cover the desired targets using the default parameters of the program with both strands and 5 bp either side of the exon-intron boundary of the default transcript covered. A final panel was chosen consisting of 4929 amplicons with a total analysable target size of 63.18 kbp. The panel was designed for sequencing with the Illumina platform (100 bp reads).

Table 2.2: HaloPlex HS Panel Targets

Gene	**Category**	**Target (hg19)**	**Regions**	**Predicted coverage**
ADNP	ASD candidate	chr20:49507930-49511049	1	98.94%
AKT1	Overgrowth	chr14:105246521-105246591	1	100%
APC	Signalling	chr5:112175742-112175782	1	100%
BRAF	RAS-MAPK	chr7:140449176-140449216	1	100%
BRAF	RAS-MAPK	chr7:140453112-140453152	1	100%
BRAF	RAS-MAPK	chr7:140477793-140477833	1	100%
BRAF	RAS-MAPK	chr7:140481382-140481422	1	100%
CBL	RAS-MAPK	chr11:119148971-119149011	1	100%
CHD8	ASD candidate	whole gene	39	99.64%

COL43ABP	Recurrent hotspot	chr5:74722212-74722302	1	100%
CTNNB1	Signalling	whole gene	14	98.67%
DNMT3A	Recurrent hotspot	chr2:25457223-25457263	1	100%
DYRK1A	ASD candidate	whole gene	11	99.66%
EZH2	Overgrowth	whole gene	21	95.51%
FGFR2	SSS	whole gene	24	95.89%
FGFR3	SSS	whole gene	18	100%
FMR1	Syndromic autism	whole gene	18	98.11%
HRAS	SSS	whole gene	5	98.77%
IDH1	Recurrent hotspot	chr2:209113090-209113130	1	100%
KRAS	SSS	whole gene	6	96.14%
MAP2K1	RAS-MAPK	chr15:66727435-66727475	1	100%
MECP2	Syndromic autism	whole gene	5	87.27%
MTOR	Recurrent hotspot	chr1:11190759-11190849	1	100%
NF1	RAS-MAPK	whole gene	3	95.76%
NRAS	RAS-MAPK	whole gene	4	100%
PACS1	Recurrent hotspot	chr11:65978632-65978722	1	100%
PIK3CA	Signalling	chr3:178928092-178928132	1	100%
PPP2R1A	Recurrent hotspot	chr19:52715934-52716024	1	100%

PPP2R5D	Recurrent hotspot	chr6:42974958-42975048	1	100%
PTEN	ASD candidate	whole gene	9	85.98%
PTPN11	SSS	whole gene	16	93.84%
RAF1	RAS-MAPK	chr7:12645679-12645719	1	100%
RET	SSS	whole gene	20	98.33%
SETBP1	Recurrent hotspot	chr18:42529846-42533305	1	99.65%
SKI	Overgrowth	chr1:2160199-2160399	1	100%
SMAD4	Recurrent hotspot	chr18:48604632-48604722	1	100%
SOS1	RAS-MAPK	chr2:39250272-39250312	1	100%
SYNGAP1	ID	whole gene	21	100%
TGFBR2	Signalling	chr3:30729964-30730004	1	100%
TP53	Cancer	chr19:7576877-7576917	1	100%
UBE3A	Syndromic autism	whole gene	12	91.29%
ZBTB20	ID	chr3:114058157-114058247	1	100%

2.6.2 HaloPlex sample preparation and capture

25 testicular biopsies from five individuals and 19 seminiferous tubules from two individuals were chosen for inclusion (Table 2.3). Additionally, an Agilent provided sample (Enrichment Control DNA, ECD) and three blood samples were simultaneously processed

and sequenced as controls. Samples were diluted to 9 ng/µl and verified by Qubit fluorometer (ThermoFisher Scientific). A total of 250 ng of each DNA sample was processed.

Samples were processed in four batches following the HaloPlex HS Illumina Protocol, Version C0 (December 2015). The first 8 samples were captured using a 4-hour hybridisation and all remaining samples used a 16 hour-hybridisation.

Table 2.3: Summary of samples sequenced with HaloPlex HS panel

Testis	**Number of Biopsies**	**Number of Tubules**
3	1	-
4	8	-
10	2	-
14	6	-
18	7	-
19	1	12
24	-	7

2.6.3 HaloPlex sequencing and variant calling

Libraries were verified using the 2200 TapeStation (Agilent Technologies). Samples were sequenced in three libraries (16 samples in library 1, 16 samples in library 2 and 18 samples in library 3). The sequencing was performed by the Oxford Genomics Centre at the Wellcome Trust Centre for Human Genetics using the Illumina HiSeq2500 platform.

Raw sequencing fastQ files were processed using Agilent SureCall software (version 3.5.1.46) to produce trimmed aligned files. Variant calls were produced using the following parameters: variant score threshold 0.01, minimum base quality score 20, minimum variant call quality score 5, minimum allele frequency 0.001, minimum of 3 reads supporting the variant allele and SNP filter off.

Annotations were performed using version 3 of the built-in databases (ClinVar, Cosmic 10-07-15 and GWAS V3 UCSC 28-12-15).

In order to prioritise variants, a custom script was developed to automatically extract sequence context information regarding variant positions using the aligned sequencing files (section 4.3.6).

2.7 smMIP

2.7.1 smMIP oligonucleotides

smMIPs were designed using MIPGEN (Boyle et al., 2014) and double-tiled across the target region of interest. Probes were chosen from a long list of possible smMIPs based on a low predicted copy number of the extension and ligation arms, visual inspection and, where possible, avoidance of underlying SNP variation. A panel of smMIP were received from collaborator Dr Alexander Hoischen. All other smMIPs were ordered from IDT and details are listed in appendix A.2.

For the proof of principle work described in Chapter 5, two manually designed probes (Probe 1 and 2, Professor Anne Goriely) and two MIPGEN designed probes (Probe 3 and 4) targeting *FGFR2* were used (Table 2.4).

Table 2.4: Sequences of smMIP targeting *FGFR2*

Probe	**Gene**	**Target (hg19)**	**Strand**	**Sequence**
1	*FGFR2*	chr10:123279597-123279820	-	TCAAAGGTGTCAGCCAGCANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGAGACGTAGAGTTTGTCTGC
2	*FGFR2*	chr10:123279655-123279843	+	TGGAGGATGGGCCGGTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNACTGTGGCCCCATGAATAACAG
3	*FGFR2*	chr10:123279583-1232796947	-	GCCCAGCCCCACATCCAGTGGANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNACTCCTCCTTTCTTCCCT
4	*FGFR2*	chr10:123279607-123279718	+	GGGAGAATGAGAGACCAATAAANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGTAAACCTTGCAGACAAA

2.7.2 smMIP pooling and phosphorylation

For initial testing, smMIP were pooled at equimolar concentrations (100 µm). Phosphorylation of the 5′ end of the smMIP was performed in 10x T4 DNA ligase buffer with 10 mM ATP with T4 polynucleotide kinase (PNK). The reaction was incubated at 37°C for 45 minutes and inactivated at 65°C for 20 minutes. smMIP were diluted and added at either 800 or 4000 smMIP copies per haploid genome copy (section 5.2). All subsequent experiments used a ratio of 1:800.

2.7.3 smMIP capture of targeted sequences and circularisation

Hybridization of smMIP to genomic DNA, gap filling, and ligation were performed in one 25 µl reaction of 10x Ampligase buffer, 0.025 mM dNTPs, 10 units of Hemo Klentaq and 100 units of Ampligase. Following protocol optimisation, Hemo Klentaq enzyme was substituted with 5 units of FastStart HiFi enzyme (Roche). This modified protocol was used for smMIP experiments described in chapter 6.

For section 5.2, smMIP were captured on 100, 250, 500 or 1000 ng of human blood DNA. All subsequent experiments used 250 ng of genomic DNA. Reactions were incubated at 95°C for 10 minutes and then 60°C for 22 to 24 hours.

To remove non-circularised DNA at the end of the reaction, 2 µl of exonuclease mix (containing 10 units of exonuclease I (NEB) and 50 units of exonuclease III (NEB) in 10x Ampligase buffer) was added to each sample and the reaction was incubated at 37°C for 45 minutes then 95°C for 2 minutes.

2.7.4 smMIP amplification and library preparation

Post-capture PCR amplification was carried out using generic primers corresponding to the common smMIP backbone to append the Illumina MiSeq adapter sequences and sample specific barcodes.

Captured DNA was amplified by performing PCR in a 25 µl reaction using 3 µl of exonuclease-treated capture reaction, 100 µM read 1 primer (Illumina), 10 µM read 2

barcoded primer (different for each sample), and 2x iProof HF Master Mix (Bio-Rad) at 98°C for 30 seconds, X cycles of 98 °C for 10 seconds, 60 °C for 30 seconds, 72 °C for 30 seconds, and finally 72 °C for 2 minutes. The number of cycles (X) was defined as the beginning of the exponential phase and was determined by an initial quantitative real-time PCR using 1:100 SYBR green. Following the experiment described in section 5.4.2, 1.5 µl of exonuclease-treated captured product was split into 12 wells (with two independent sample barcodes) and amplified in a 7.5 µl reaction.

If samples were amplified using the same number of cycles during real-time PCR, equal volumes of each PCR sample were pooled together. Otherwise, the volume was adjusted to account for the difference in amplification efficiency between samples. Pooled samples were purified using 0.8x AMPure XP beads following the standard protocol. Libraries were resuspended and then quantified using TapeStation High Sensitivity D1000 ScreenTape an Qubit Fluorometer (ThermoFisher Scientific) to confirm library size and determine the concentration.

Libraries were diluted to a final concentration of 9 pM and sequenced using the Illumina MiSeq platform (150v3 reagent kit), following the manufacturer's instructions for custom indexing. Custom primers added to the sequencing cartridge were as follows:
Forward: CATACGAGATCCGTAATCGGGAAGCTGAAG
Reverse: ACACGCACGATCCGACGGTAGTGT
Index: ACACTACCGTCGGATCGTGCGTGT

smMIP libraries in section 6.4.2 and 6.3 were sequenced using the Illumina NextSeq platform.

2.7.5 Sequencing analysis

Sequencing data was processed to produce demultiplexed alignment files, coverage and allele counts using amplimap (https://github.com/koelling/amplimap), a pipeline built using Snakemake version 3.11.2 (Dr Nils Koelling). A requirement for 2 supporting reads per UMI was set.

2.8 R Scripts

Relevant R scripts are available for download on GitHub (https://github.com/hkralph/analysis_scripts). Library dependencies are listed at the top of each script.

Chapter 3

Validation of a massively-parallel PCR-based discovery screen

3.1 Introduction

As discussed in chapter 1, direct experimental evidence of selfish spermatogonial selection (SSS) has thus far been established for specific activating mutations at 16 codons in six genes (Table 1.2). In order to establish further evidence of SSS and investigate other candidates, a discovery screen was performed using testicular biopsies from five men (strategy B, Figure 1.6). Massively parallel simplex PCR of a panel of candidate hotspots was combined with ultra-deep sequencing. This technology theoretically allows the increased sensitivity required for detection of low-level variants by partitioning sample DNA and primers into separate droplets. However, as this method is based on PCR, interpretation of variants is confounded by amplification errors (Minoche et al., 2011; Hestand et al., 2016). Extensive manual curation is required in order to distinguish variant calls from artefacts.

3.2 Experimental background

A discovery screen using massively parallel simplex PCR (RainDance technology) to interrogate mutational hotspots in 67 genes was performed (Professor Professor Anne Goriely). The target genes encompassed 17 members of the RTK-RAS-MAPK pathway as well as hotspots in putative candidates in other signalling pathways (PI3K-AKT,

TGF-β-BMP, Notch, WNT, EGF), candidates recurrently mutated in cancer and control genes. A total of 276 biopsies from testes of five men (median age: 83 years, range 34 to 90 years) were analysed, with 12 of the biopsies sequenced in duplicate across separate libraries. 16 libraries of 18 samples were constructed and sequenced to an ultra-deep level (∼16,000x) across 8 lanes on two flow-cells of the Illumina Hi-Seq platform.

Following sequencing, quality filtering and data normalisation was performed to account for differences in the flow cell, sequencing lane and average base quality at each position (Dr Zhihao Ding). Primer sequences were excluded from the reads as well as read pairs with >10 non-reference bases. Following alignment, nine amplicons were removed from further analysis due to a failure to map to the predicted sequence. A custom statistical prioritisation pipeline was developed to independently test each possible nucleotide change at every position in all remaining amplicons. Filtration criteria of a minimum of 10 variant reads was applied. This statistical model resulted in a list of 19,625 genomic positions with at least one non-reference call with a P-value -log10P <20. Identification and removal of seven amplicons with excessive variants and the blacklisting of 185 sample-nucleotide change combinations which displayed an elevated number of variant calls reduced this list to 5980 variant positions.

Further refinement of the variant list included exclusion of positions within 1 bp of the primer and variants with a VAF >3%. Variants present above >3% were manually inspected and determined to be artefacts. Only variants with at least 2 independent calls (either in a sample or in an overlapping amplicon) were retained. This yielded a variant list of 374 variant calls at 361 distinct genomic locations which then required manual inspection and validation.

An overview of the experimental, analysis and validation steps performed is shown in figure 3.1.

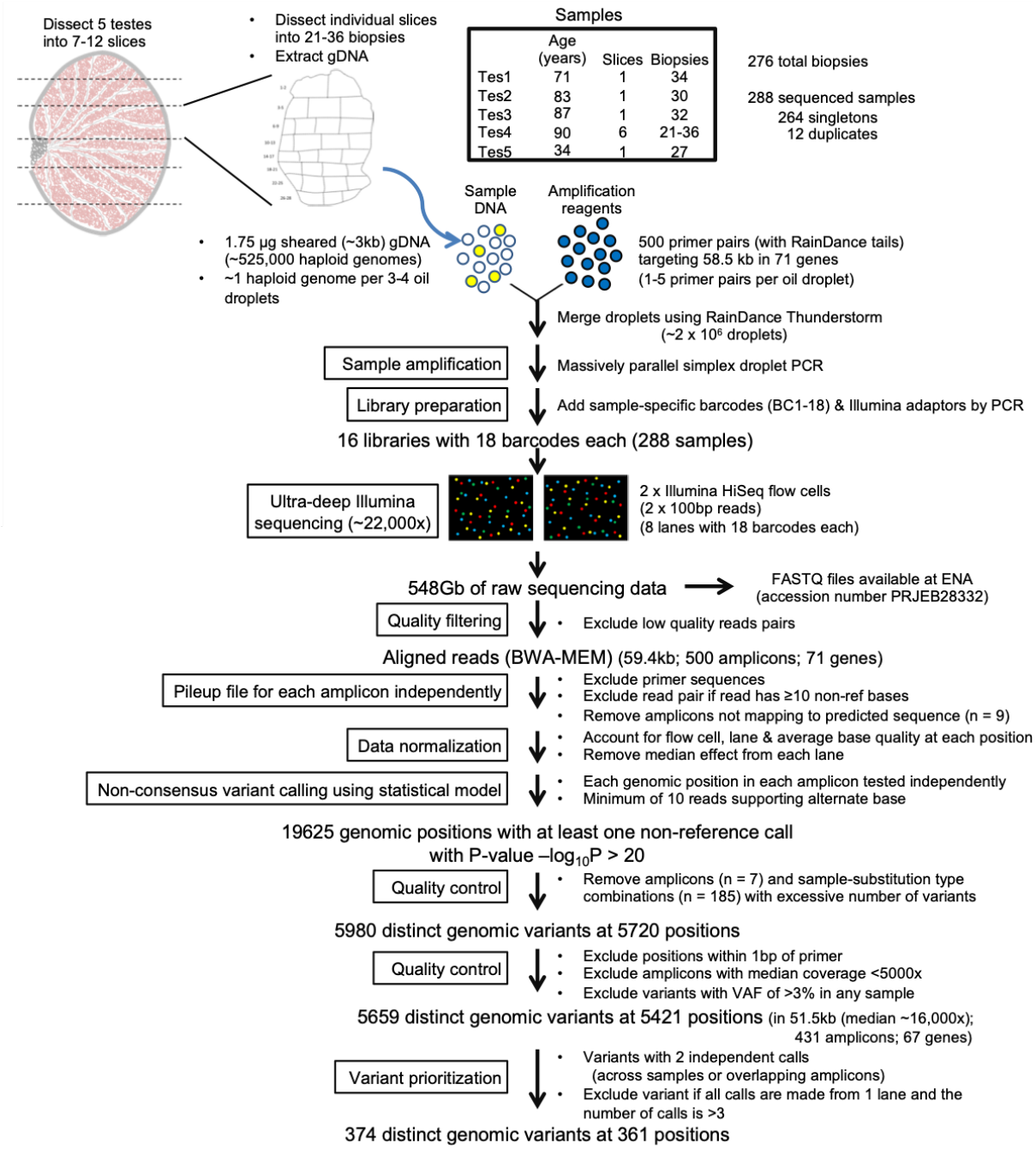


Figure 3.1: A schematic overview of the experimental design. Modified from Maher et al. (2018).

3.3 Results

3.3.1 Coverage

The panel was initially designed to cover 66.7 kb of analyzable sequence. To calculate the callable sequence, the median coverage per amplicon was extracted for all 288 samples. This value was subsequently used to plot coverage in figure 3.7.

Of the 484 amplicons that remained following the previously described quality control, a further 53 were excluded due to insufficient coverage (median coverage <5000). After removing amplicons where variants could not be called, bases that were within 1 bp of a primer and bases that were overlapped by another amplicon, a total of 51.5 kb unique callable bases remained. Of this, 21.5 kb (41.9%) belonged to genes in the RTK-RAS-MAPK pathway, 24.7 kb (47.9%) belonged to other candidate genes and 0.5 kb (10.2%) belonged to control genes. The percentage of the target panel that was analysable above 5000 reads in each of the 288 testicular biopsies is shown in Figure 3.2.

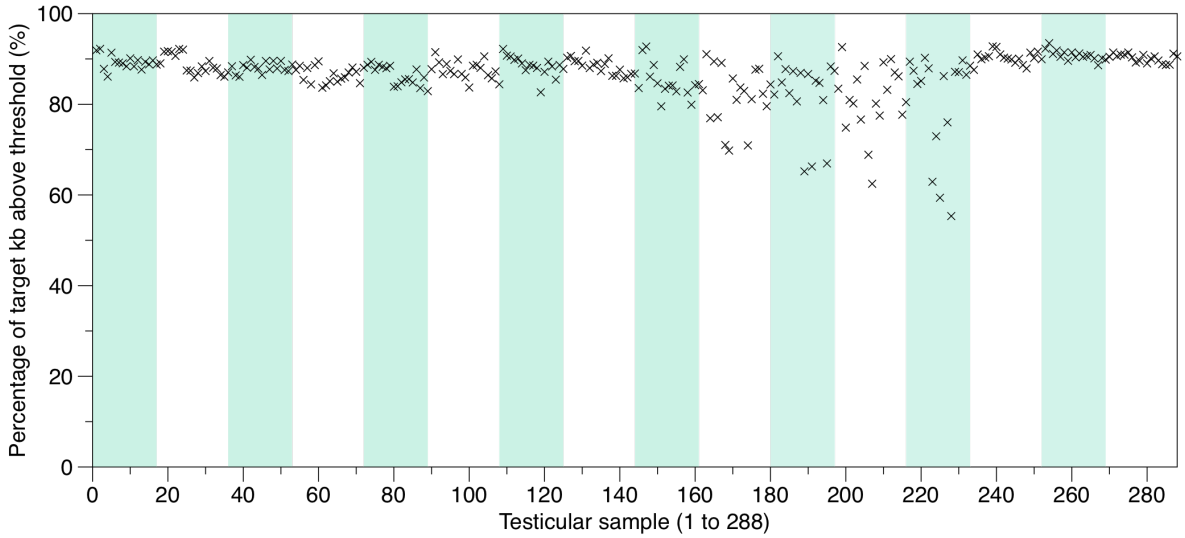


Figure 3.2: Percentage of target panel (y-axis) analysable above a threshold of 5000 reads for each of the 288 testicular biopsies (x-axis). The coloured bars partition the 16 library groups that the testicular biopsies were sequenced in section (3.2)

3.3.2 Manual curation of variant list

In order to visualise all candidate variants, a script was written to automate the production of individual plots to be inspected for each variant. At every position of interest, the total number of reads and non-reference reads were extracted for all testicular biopsies. This was used to calculate the percentage non-reference for each possible nucleotide change at a position and this was then plotted using ggpplot2. An example plot is shown in Figure 3.3 for a candidate variant called in two biopsies with a low background in other samples.

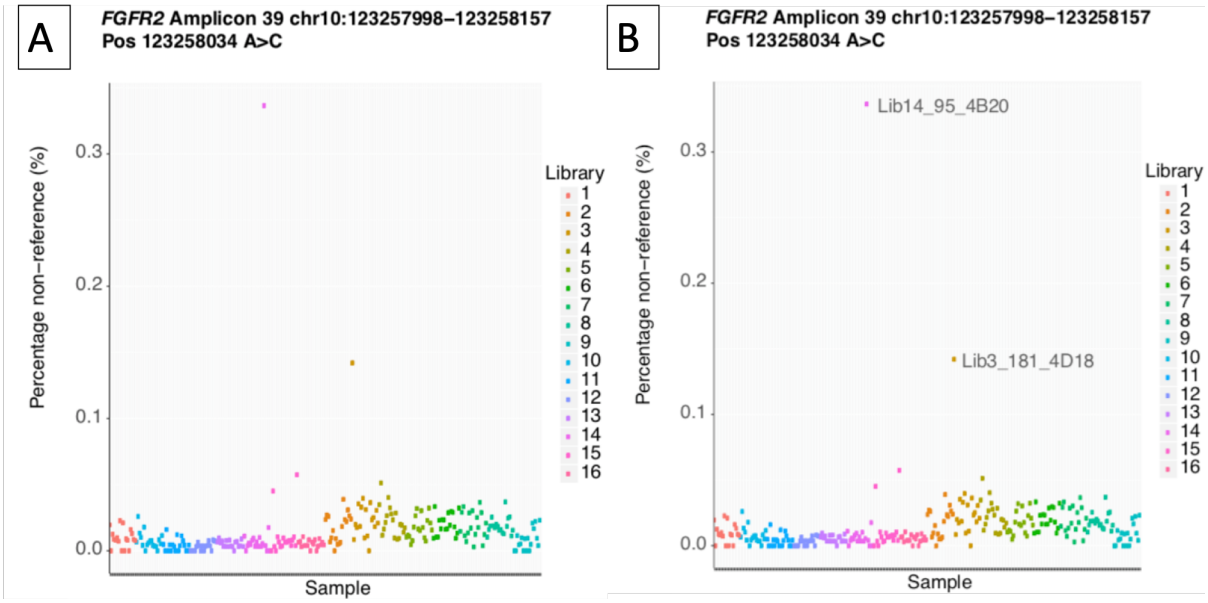


Figure 3.3: Percentage non-reference nucleotide at a candidate variant position. A variant in *FGFR2* was called in two samples from Tes4 (4B20 and 4D18). (A) Non-reference percentage levels (y-axis) for A>C nucleotide change at chr10:123258034. Samples 1 to 288 are plotted along the x-axis and are coloured separately by each of the 16 libraries. (B) As (A) with label indicating the library number, sample number and sample name.

All variant plots were inspected and this was used to categorise variant calls. A distinctive pattern of error was observed. A subset of variants were called in a few samples but visual inspection revealed that all samples from the same library were in-fact elevated above background (termed a 'batch effect') (Figure 3.4). Re-sequencing of five of these

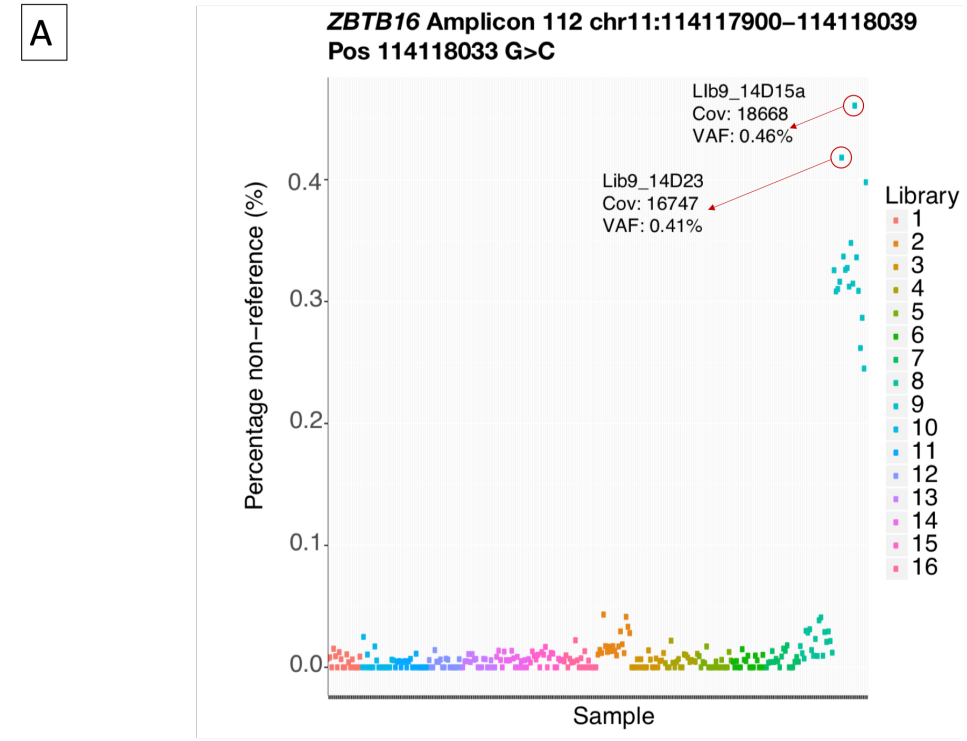
variants was performed and in all cases the result was negative. Variant calls of indeterminate appearance were further inspected using IGV and BLAT. Exclusion of variants that were determined to be batch effects or alignment errors resulted in the identification of 115 high confidence variant calls (at 105 genomic positions) in 165 testicular biopsies.

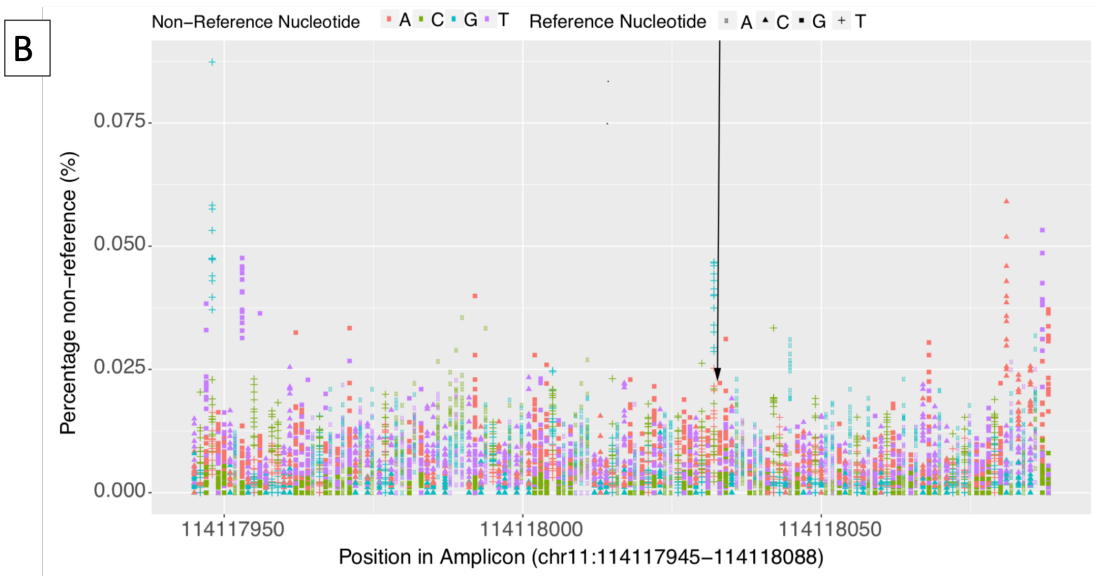
3.4 Variant validation

3.4.1 Development of amplicon pileup script

Variants selected for further validation were amplified and sequenced as described in the Materials and Methods (section 2.5). In order to produce nucleotide counts across a target region of interest, I wrote a script in R. The script utilises the packages Rsamtools (to provide an interface to open and read sequence alignment files), Bsgenome (which provides an indexed copy of the human reference genome sequence hg19) and ggplot2 (to provide the functionality to create graphics). Given an input sequence alignment file ('BAM') and a target chromosome position, a table of the nucleotide counts and coverage across a region of interest is produced ('pileup'). Graphs are also made automatically. The graphical output includes all nucleotide changes across the target region and coverage graphs (Figure 3.5).

Figure 3.4 *(following page)*: Percentage non-reference for a variant in *ZBTB16* at chr11:114118033 (A) Plot showing G>C non-reference percentage (y-axis) for all 288 testicular samples (x-axis). The two samples which were called by the statistical prioritisation method are outlined in red. The other samples from this library (shown in light blue) are also elevated above background. (B) The variant fails to validate upon amplification of the target region in the samples of interest. The arrow indicates the target position. (C) Average coverage across the *ZBTB16* amplicon for the samples of interest (outlined in red) and control DNA samples.


A
ZBTB16 Amplicon 112 chr11:114117900–114118039
Pos 114118033 G>C
LIb9_14D15a
Cov: 18668
VAF: 0.46%
Lib9_14D23
Cov: 16747
VAF: 0.41%
Percentage non–reference (%)
0.4
0.3
0.2
0.1
0.0
Sample
Library
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


B
Non–Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non–reference (%)
0.075
0.050
0.025
0.000
114117950
114118000
114118050
Position in Amplicon (chr11:114117945–114118088)

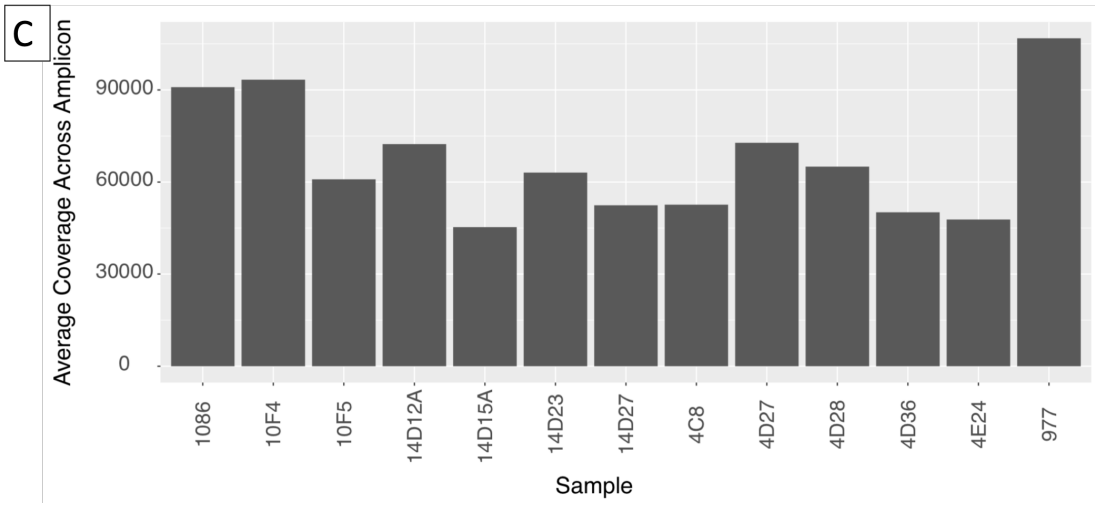
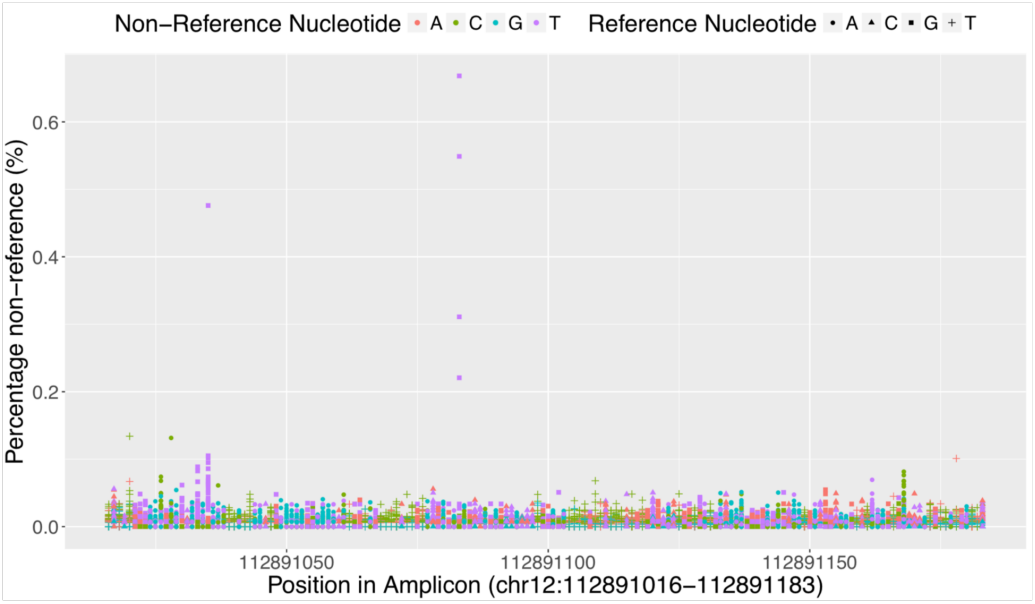
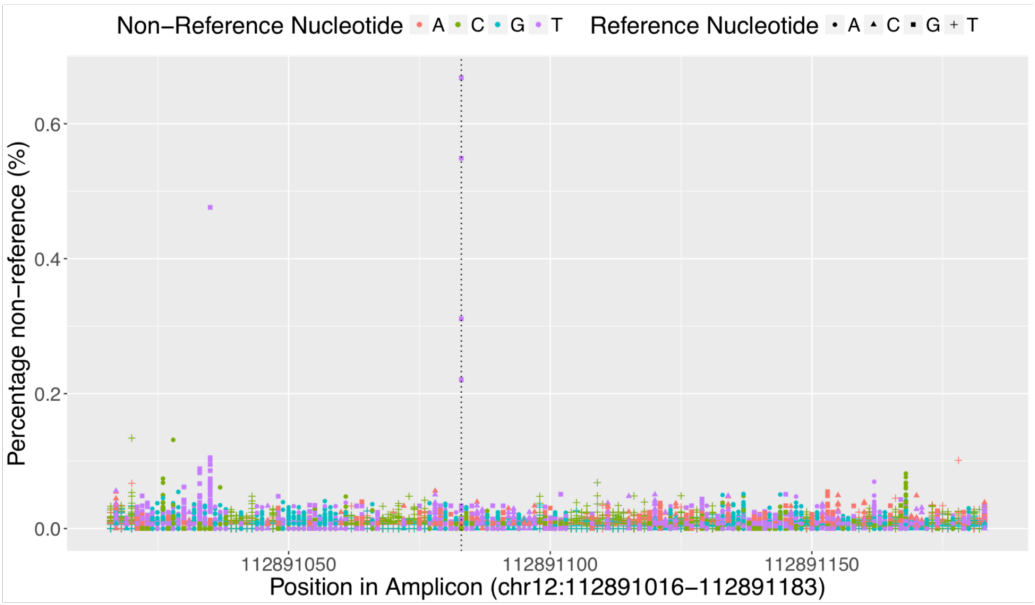

C
Average Coverage Across Amplicon
90000
60000
30000
0
1086
10F4
10F5
14D12A
14D15A
14D23
14D27
4C8
4D27
4D28
4D36
4E24
977
Sample

Figure 3.5 *(following page)*: Output created by a custom script to produce amplicon pileup. Twelve graphs are automatically produced for the target region of interest, which can be specified in the input file. The example shown is *PTPN11* chr12:112891083 (c.417G>T p.E139D). (A) Plot showing the percentage non-reference nucleotide levels (y-axis) at each position in the target amplicon (x-axis). The legend shows the non-reference nucleotide (red = A, green = C, blue = G, purple = T) with reference nucleotide indicated by shape (circle = A, triangle = C, square = G, plus = T). Thus, a purple square indicates a change from the reference base 'G' to the non-reference base 'T'. (B) As (A) with a dotted line to indicate the target variant position (C) As (A) plotted on a log-scale (D) As (C) with a dotted line to indicate the target variant position (E) Plot as (A) with each point coloured by DNA sample rather than nucleotide change (F) As (E) with a dotted line to indicate variant position (G) Plot showing 5 bp either side of the target region with each point coloured by nucleotide change (H) As (G) with each point coloured by DNA sample (I) Percentage G>T change at the position of interest in all samples. Crosses are coloured by replicate (Replicate 1 = green, Replicate 2 = purple) (J) As (I) for G>A nucleotide change (K) As (I) for G>C nucleotide change (L) Average coverage across the target amplicon for each sample.

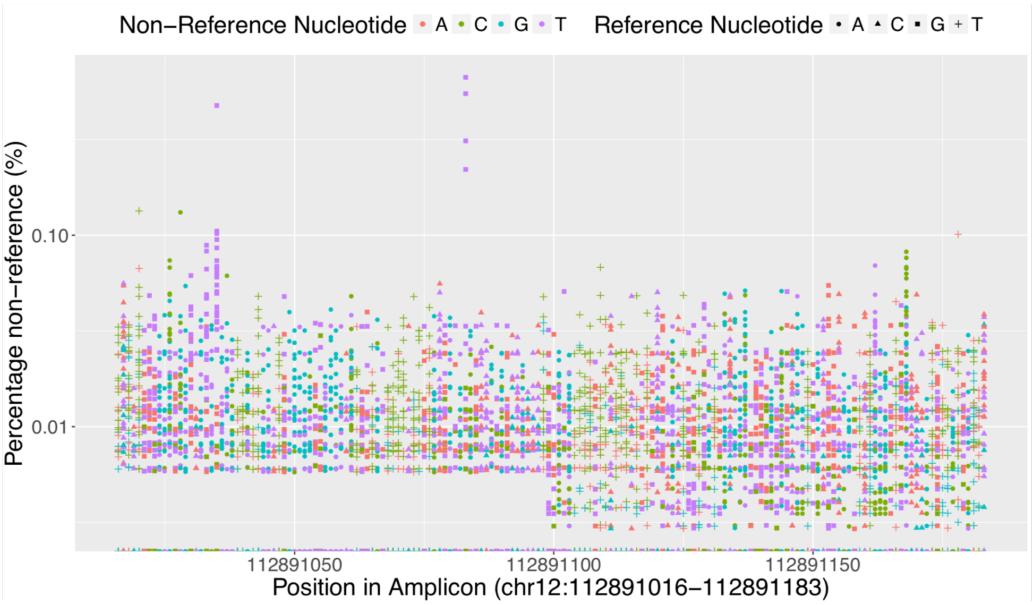
A


Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

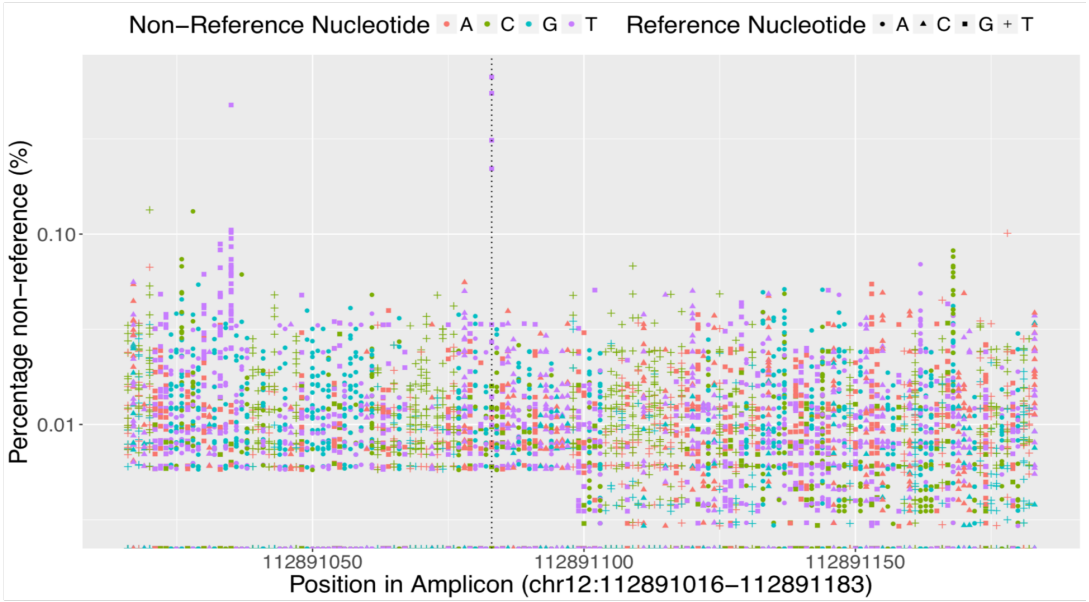
B


Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

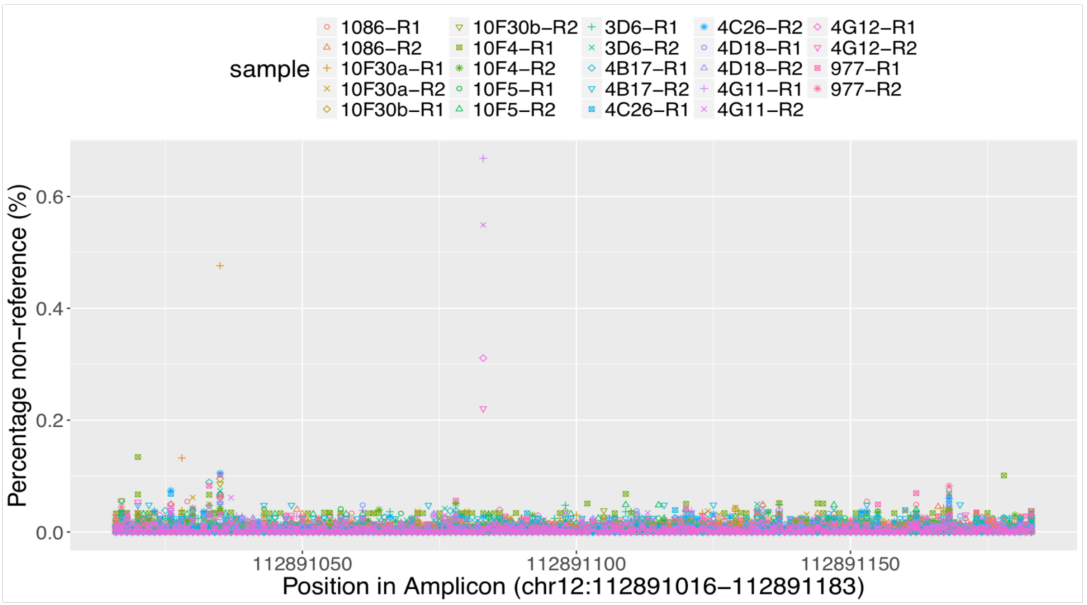
C


Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.10
0.01
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

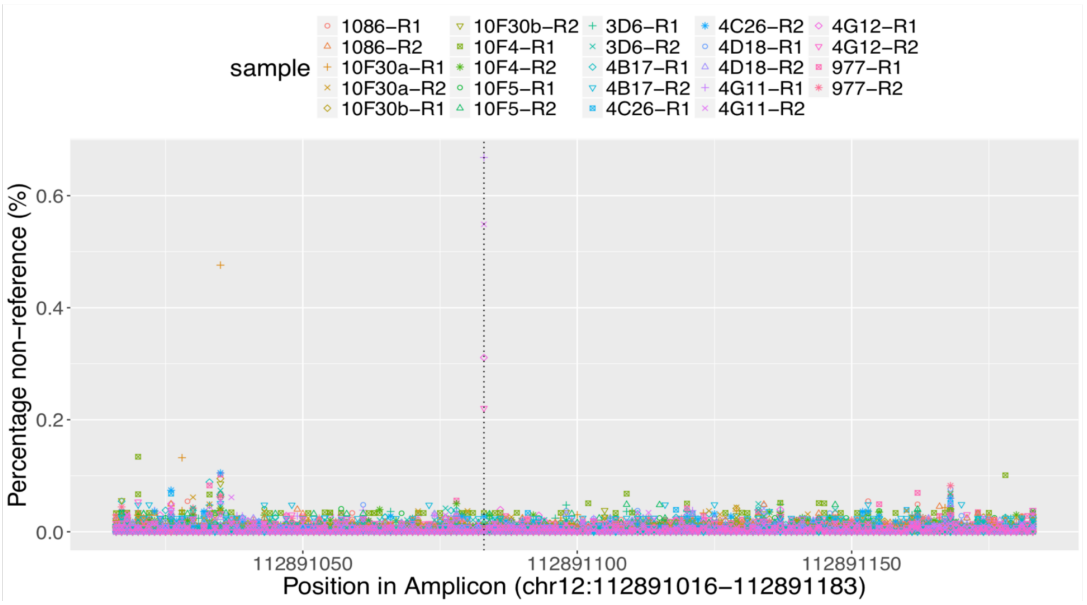
D


Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.10
0.01
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

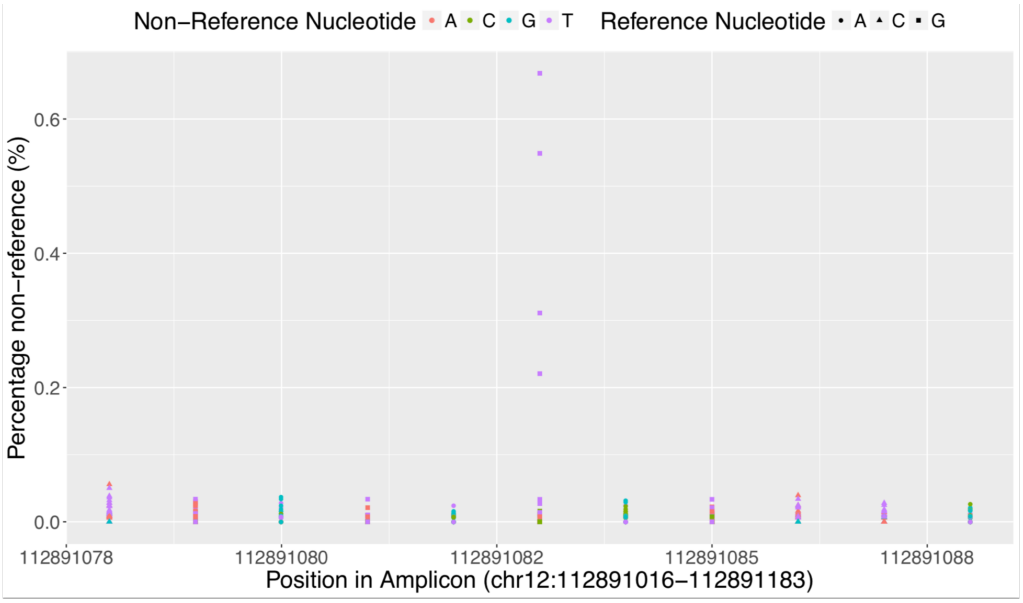
E


sample
1086−R1
1086−R2
10F30a−R1
10F30a−R2
10F30b−R1
10F30b−R2
10F4−R1
10F4−R2
10F5−R1
10F5−R2
3D6−R1
3D6−R2
4B17−R1
4B17−R2
4C26−R1
4C26−R2
4D18−R1
4D18−R2
4G11−R1
4G11−R2
4G12−R1
4G12−R2
977−R1
977−R2
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

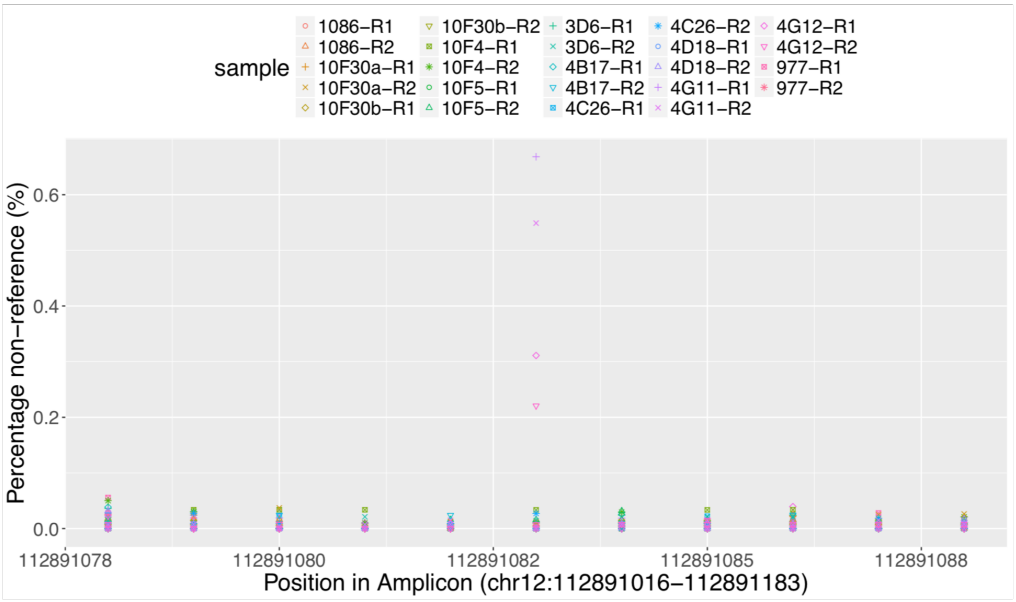
F


sample
1086−R1
1086−R2
10F30a−R1
10F30a−R2
10F30b−R1
10F30b−R2
10F4−R1
10F4−R2
10F5−R1
10F5−R2
3D6−R1
3D6−R2
4B17−R1
4B17−R2
4C26−R1
4C26−R2
4D18−R1
4D18−R2
4G11−R1
4G11−R2
4G12−R1
4G12−R2
977−R1
977−R2
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891050
112891100
112891150
Position in Amplicon (chr12:112891016−112891183)

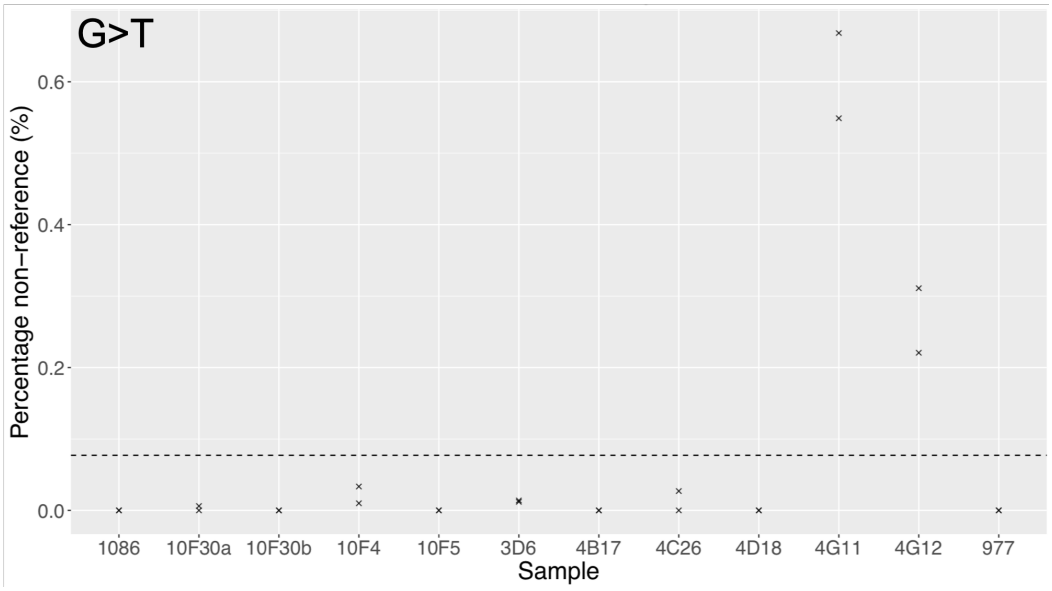
G


Non−Reference Nucleotide A C G T
Reference Nucleotide A C G
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891078
112891080
112891082
112891085
112891088
Position in Amplicon (chr12:112891016−112891183)

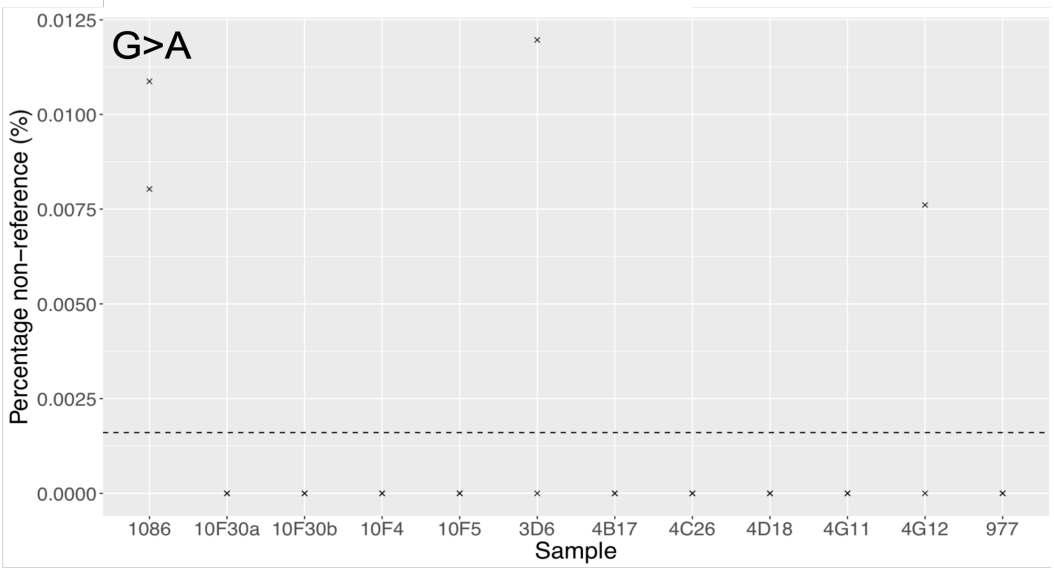
H


sample
1086−R1
1086−R2
10F30a−R1
10F30a−R2
10F30b−R1
10F30b−R2
10F4−R1
10F4−R2
10F5−R1
10F5−R2
3D6−R1
3D6−R2
4B17−R1
4B17−R2
4C26−R1
4C26−R2
4D18−R1
4D18−R2
4G11−R1
4G11−R2
4G12−R1
4G12−R2
977−R1
977−R2
Percentage non−reference (%)
0.6
0.4
0.2
0.0
112891078
112891080
112891082
112891085
112891088
Position in Amplicon (chr12:112891016−112891183)

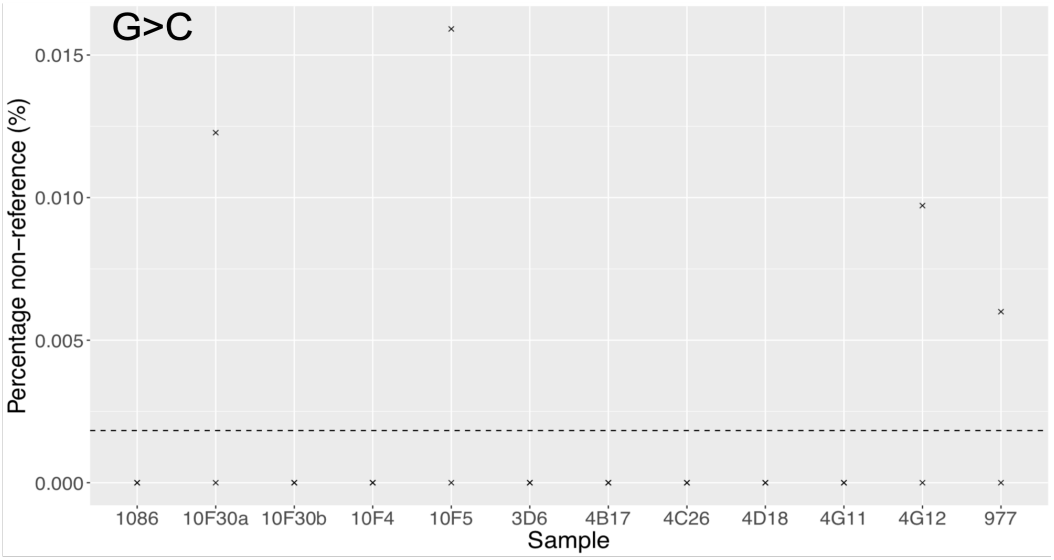
I


G>T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
1086
10F30a
10F30b
10F4
10F5
3D6
4B17
4C26
4D18
4G11
4G12
977
Sample

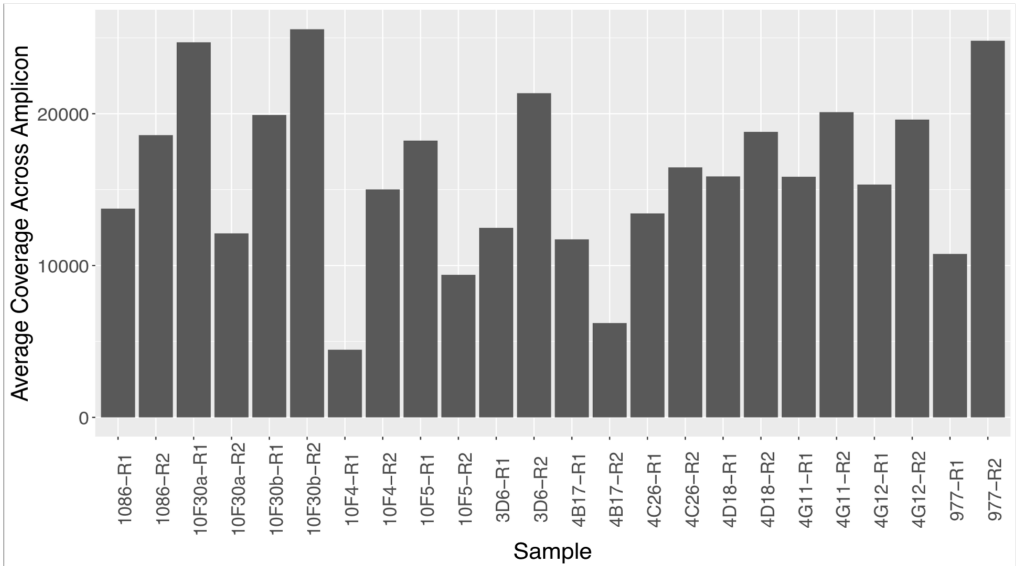
J


G>A
Percentage non−reference (%)
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
1086
10F30a
10F30b
10F4
10F5
3D6
4B17
4C26
4D18
4G11
4G12
977
Sample

K


G>C
Percentage non−reference (%)
0.015
0.010
0.005
0.000
1086
10F30a
10F30b
10F4
10F5
3D6
4B17
4C26
4D18
4G11
4G12
977
Sample

L


Average Coverage Across Amplicon
20000
10000
0
1086−R1
1086−R2
10F30a−R1
10F30a−R2
10F30b−R1
10F30b−R2
10F4−R1
10F4−R2
10F5−R1
10F5−R2
3D6−R1
3D6−R2
4B17−R1
4B17−R2
4C26−R1
4C26−R2
4D18−R1
4D18−R2
4G11−R1
4G11−R2
4G12−R1
4G12−R2
977−R1
977−R2
Sample

3.4.2 Variants in overlapping amplicons and replicate samples

Variants that were called independently in replicate samples or in overlapping amplicons were reasoned to be less likely to have arisen due to error. In the 12 samples that were sequenced in duplicate, 15 unique variant calls were identified (Table 3.1). Of these, 7 of the calls were replicated in the second sample and validation of these variants was successful in all but one case. In the overlapping amplicons, a total of 36 unique variants (ranging in non-reference percentage from 0.1% to 2.63%) were called. An example variant is shown in (Figure 3.6). 18 of the variants identified in overlapping amplicons were selected for further validation, with preference given to variants that had not previously been associated with SSS. The variants were re-screened by PCR or by smMIPs (Dr Geoffrey Maher) and ultra-deep sequencing. A high true positive rate was achieved for these variants (17/18, 94%) suggesting that the majority of variants replicated between samples or amplicons are true positive calls. This give a total of 39 unique validated variants that were classified as Tier 1 and are listed in Table 3.2.

Figure 3.6 *(following page)*: Percentage non-reference nucleotide level for variant in *APC* at chr5:112175762 T>G (encoding p.F1491V) (A) Plot showing T>G non-reference percentage (y-axis) for all 288 samples (x-axis) in amplicon number 371. The sample that was called by the statistical prioritisation method is outlined in red. (B) As (A), for amplicon number 372. (C) PCR validation of the target region in the called sample shows that the variant is present in the sample at 0.55%. (D) Average coverage across the *APC* amplicon amplified for resequencing for the sample of interest (outlined in red) and control DNA samples.

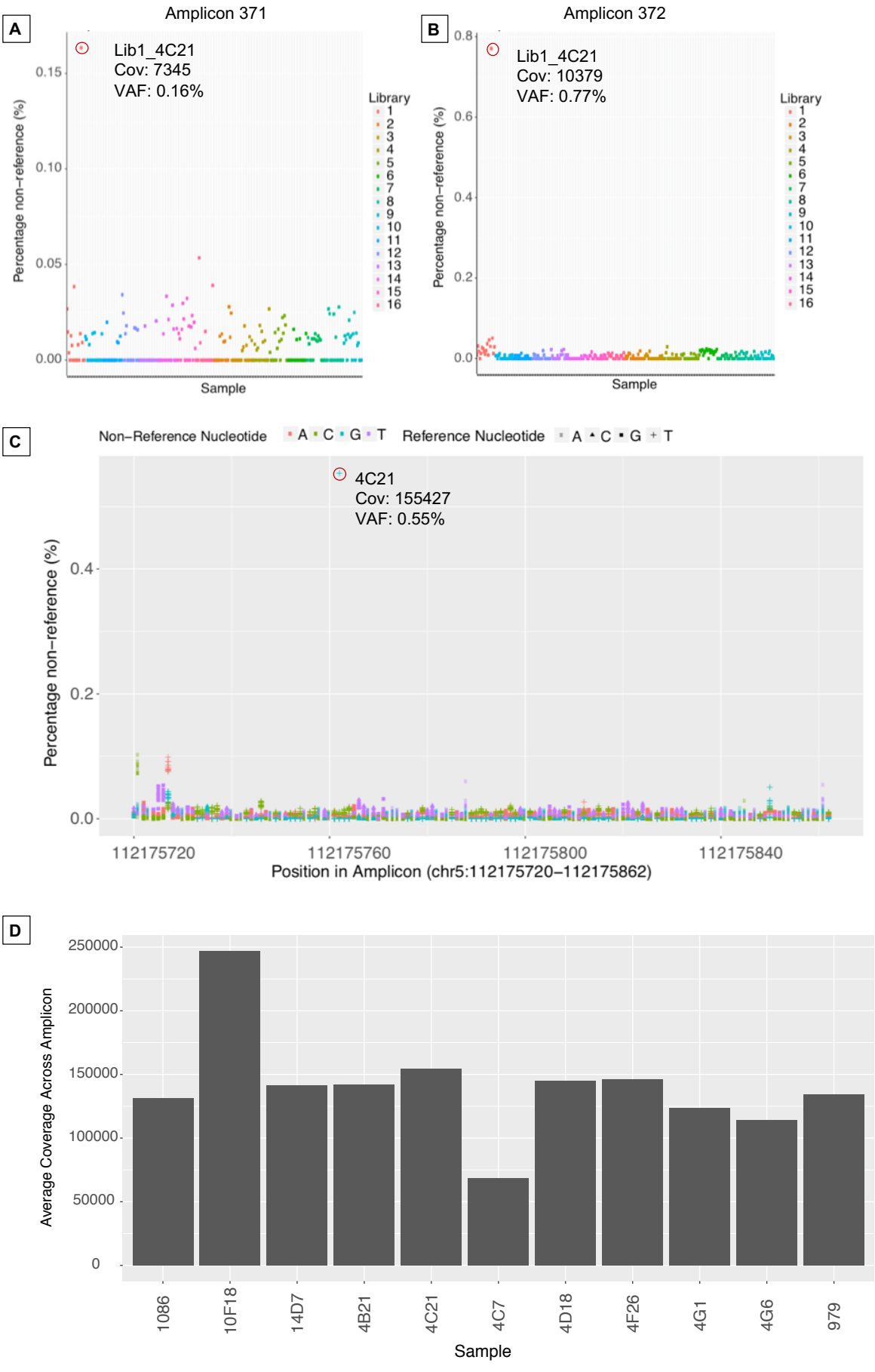

A
Amplicon 371
Lib1_4C21
Cov: 7345
VAF: 0.16%
Percentage non−reference (%)
0.15
0.10
0.05
0.00
Sample
Library
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
B
Amplicon 372
Lib1_4C21
Cov: 10379
VAF: 0.77%
Percentage non−reference (%)
0.8
0.6
0.4
0.2
0.0
Sample
Library
C
Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
4C21
Cov: 155427
VAF: 0.55%
Percentage non−reference (%)
0.4
0.2
0.0
112175720
112175760
112175800
112175840
Position in Amplicon (chr5:112175720−112175862)
D
Average Coverage Across Amplicon
250000
200000
150000
100000
50000
0
1086
10F18
14D7
4B21
4C21
4C7
4D18
4F26
4G1
4G6
979
Sample

Table 3.1: Table of replicated calls

Gene	**Position (hg19)**	**Variant**	**Sample**	**Replicated**	**Coverage (1st amplicon)**	**VAF (%) (1st amplicon)**	**Coverage (2nd amplicon)**	**VAF (%) (2nd amplicon)**	**Validation Result**
AKT3	chr1:243668575	c.1416C>G (p.S472S)	4D28	Y	26319 ; 25165	0.08 ; 0.1	-	-	Positive
BRAF	chr7:140439614	c.2125C>A (p.Q709K)	4D28	Y	16269 ; 13567	0.34 ; 0.18	-	-	Positive
BRAF	chr7:140481402	c.1406G>C (p.G469A)	4D4	Y	828 ; 3221	0.04 ; 0.03	16447 ; 18578	0.15 ; 0.2	Positive
BRAF	chr7:140482928	c.1207C>G (p.P403A)	4D10	Y	14163 ; 6096	0.13 ; 0.05	21485 ; 8868	0.11 ; 0.14	Positive
FGFR2	chr10:123274803	c.1115C>G (p.S372C)	4D13	Y	33864 ; 28557	0.41 ; 0.25	-	-	Positive
PTPN11	chr12:112888197	c.213T>G (p.F71L)	4D28	Y	19464 ; 26058	0.31 ; 0.65	21543 ; 26931	0.24 ; 0.32	Positive
MAP2K2	chr19:4117383	intronic C>G	5J27	Y	18488 ; 12456	0.06 ; 0.13	-	-	Negative
SOS1	chr2:39262566	c.940C>A (p.Q314K)	4D10	N	13880	0.10	-	-	NA - Tier 3
SOS1	chr2:39285879	c.280A>C (p.I94L)	5J27	N	12746	0.13	-	-	Negative
PIK3CA	chr3:178941885	c.2204T>G (p.L735X)	5J27	N	12994	0.09	-	-	NA - Tier 4
BRAF	chr7:140476802	c.1604C>G (p.S535C)	4D4	N	12935	0.09	-	-	NA - Tier 3
FGFR2	chr10:123279677	c.755C>G (p.S252W)	4D4	N	17355	0.10	12002	0.19	Positive
ZBTB16	chr11:113934757	c.735G>C (p.M245I)	5J25	N	17148	0.10	-	-	Negative
CBL	chr11:119148991	c.1211G>A (p.C404Y)	5J17	N	5186	0.52	59	0.00	Positive
TP53	chr17:7577144	c.794T>G (p.L265R)	3D13	N	19024	0.12	-	-	NA - Tier 3

Table 3.2: Tier 1 validated variants. Counts from COSMIC v82 refer to identical DNA substitutions, identical amino acid substitutions and total substitutions at the specific amino acid.

Variant number	**Gene**	**Position (hg19) and amino acid substitution**	**VAF range (%)**	**Number of positive biopsies**	**COSMIC v82**	**Germline disorder**
1	*AKT3*	chr1:243668575 G>C (p.S472S)	0.09	1	0 (1, 1)	-
2	*APC*	chr5:112175762 T>G (p.F1491V)	0.47	1	0 (0, 1)	-
3	*BRAF*	chr7:140482928 G>C (p.P403A)	0.12	1	0 (0, 1)	-
4	*BRAF*	chr7:140481402 C>G (p.G469A)	0.16 - 0.32	4	52 (52, 123)	-
5	*BRAF*	chr7:140449196 T>G (p.Q628P)	0.19	1	-	-
6	*BRAF*	chr7:140439614 G>T (p.Q709K)	0.26	1	-	-
7	*FGFR2*	chr10:123279677 G>C (p.S252W)	0.14 - 1.55	23	54 (54, 57)	Apert syndrome
8	*FGFR2*	chr10:123279674 G>C (p.P253R)	0.06 - 0.56	8	9 (9, 11)	Apert syndrome
9	*FGFR2*	chr10:123279605 A>C (p.F276C)	0.32	1	1 (1, 1)	-
10	*FGFR2*	chr10:123279566 T>G (p.Q289P)	0.12	1	-	Crouzon syndrome
11	*FGFR2*	chr10:123279562 C>A (p.W290C)	0.18 - 0.69	3	0 (7, 7)	Pfeiffer syndrome
12	*FGFR2*	chr10:123279562 C>G (p.W290C)	0.12 - 0.34	3	7 (7, 7)	Pfeiffer syndrome
13	*FGFR2*	chr10:123274803 G>C (p.S372C)	0.11 - 0.33	5	1 (1, 2)	Beare Stevenson syndrome
14	*FGFR3*	chr4:1805519 C>T (p.S344F)	0.21	1	1 (1, 2)	-
15	*FGFR3*	chr4:1807371 C>A (p.N540K)	0.07 - 0.18	2	-	Hypochondroplasia
16	*FGFR3*	chr4:1807371 C>G (p.N540K)	0.07	1	-	Hypochondroplasia
17	*FGFR3*	chr4:1807488 G>A (p.V553M)	0.37	1	0 (0,1)	-
18	*KRAS*	chr12:25398284 C>G (p.G12A)	0.12 - 0.37	3	2255 (2256, 33497)	-
19	*KRAS*	chr12:25398284 C>T (p.G12D)	0.12 - 1.82	6	14126 (14128, 33497)	-
20	*KRAS*	chr12:25380282 G>C (p.A59G)	0.29	1	8 (8, 41)	-
21	*KRAS*	chr12:25380282 G>T (p.A59E)	0.14 - 0.50	2	6 (6, 41)	-
22	*KRAS*	chr12:25380276 T>C (p.Q61R)	0.62 - 2.63	9	115 (116, 601)	-
23	*MAP2K1*	chr15:66727455 G>C (p.K57N)	0.14	1	1 (14, 19)	-
24	*PTPN11*	chr12:112888197 T>G (p.F71L)	0.48	1	1 (22, 22)	Noonan syndrome
25	*PTPN11*	chr12:112888198 G>C (p.A72P)	0.28	1	0 (0, 137)	Noonan syndrome
26	*PTPN11*	chr12:112888199 C>A (p.A72D)	0.13 - 0.25	3	11 (11, 137)	-
27	*PTPN11*	chr12:112888199 C>G (p.A72G)	0.11	1	2 (2, 137)	Noonan syndrome
28	*PTPN11*	chr12:112888199 C>T (p.A72V)	0.88	1	72 (73, 137)	-
29	*PTPN11*	chr12:112888202 C>T (p.T73I)	0.39 - 0.59	2	19 (19, 19)	Noonan syndrome
30	*PTPN11*	chr12:112888210-112888211 GA>CT (p.E76L)	0.19	1	0 (0, 203)	-
31	*PTPN11*	chr12:112888211 A>C (p.E76A)	0.45 - 0.69	2	20 (20, 203)	-
32	*PTPN11*	chr12:112888211 A>T (p.E76V)	0.24 - 0.93	2	11 (11, 203)	-
33	*PTPN11*	chr12:112924336 G>A (p.V428M)	0.54	1	3 (3, 3)	-
34	*PTPN11*	chr12:112924336 G>T (p.V428L)	0.52	1	0 (0, 3)	-
35	*PTPN11*	chr12:112926908 C>A (p.Q510K)	0.13 - 0.30	2	3 (3, 21)	-
36	*RET*	chr10:43613906 G>C (p.L790F)	0.15	1	0 (0, 1)	MEN2A
37	*RET*	chr10:43613906 G>T (p.L790F)	0.10 - 0.42	3	0 (0, 1)	MEN2A
38	*RET*	chr10:43615613 G>T (p.D898Y)	0.15	1	0 (0, 1)	MEN2
39	*SOS1*	chr2:39250292 T>G (p.Q426P)	0.37	1	0 (0, 1)	-

Table 3.3: Tier 2 validated variants. Counts from COSMIC v82 refer to identical DNA substitutions, identical amino acid substitutions and total substitutions at the specific amino acid.

Variant number	Gene	Position (hg19) and amino acid substitution	VAF range (%)	Number of positive biopsies	COSMIC v82	Germline disorder
40	*BRAF*	chr7:140453155 C>G (p.D594H)	0.06 - 0.23	2	3 (3, 126)	-
41	*BRAF*	chr7:140453132 T>G (p.K601N)	0.22 - 0.42	2	7 (18, 129)	-
42	*CBL*	chr11:119148991 G>A (p.C404Y)	0.52 - 0.63	2	15 (15, 19)	-
43	*FGFR2*	chr10:123276865 G>C (p.S351C)	0.09 - 0.26	4	0 (0, 1)	Pfeiffer syndrome
44	*FGFR2*	chr10:123276893 A>T (p.C342S)	0.26 - 2.95	7	-	Crouzon syndrome
45	*FGFR2*	chr10:123258034 A>C (p.N549K)	0.14 - 0.34	2	10 (34, 44)	-
46	*FGFR3*	chr4:1808029 C>G (p.R669G)	0.14 - 0.24	2	0 (0, 1)	-
47	*LRP5*	chr11:68115514 C>T (p.A97A)	0.53 - 1.20	4	-	
48	*MAP2K2*	chr19:4110584 A>T (p.C125S)	0.14 - 0.21	2	1 (3, 5)	-
49	*NF1*	chr17:29554264 G>A (p.M760I)	0.87 - 2.07	9	-	-
50	*PTPN11*	chr12:112888166 A>C (p.D61A)	0.26 - 1.02	2	1 (1, 121)	Noonan syndrome
51	*PTPN11*	chr12:112888166 A>G (p.D61G)	0.71 - 0.73	2	6 (6, 121)	Noonan syndrome
52	*PTPN11*	chr12:112891083 G>T (p.E139D)	0.15 - 0.56	2	1 (5, 6)	Noonan syndrome
53	*PTPN11*	chr12:112915455 T>G (p.F285C)	0.09 - 0.24	2	-	Noonan syndrome
54	*PTPN11*	chr12:112915523 A>G (N308D)	0.68 - 0.69	2	4 (4, 7)	Noonan syndrome
55	*PTPN11*	chr12:112926884-112926885TC>AA (p.S502K)	0.25	1	0 (0, 44)	-
56	*PTPN11*	chr12:112926890 A>G (p.M504V)	0.67 - 0.82	2	1 (1, 1)	Noonan syndrome
57	*RAF1*	chr3:12645699 G>A (p.S257L)	0.44 - 0.56	2	14 (14, 15)	Noonan syndrome

3.4.3 Tier 2 variants

The majority (88.7%) of callable sequence was covered by a single amplicon only and therefore remaining variants were investigated using a tiered approach. Tier 2 comprised variants called in a single amplicon in two or more biopsies, at VAF of $\geq 0.2\%$ in at least one of the called samples. Twenty-six tier 2 variants were identified, 18 (69%) of which validated upon resequencing (Table 3.3).

3.4.4 Tier 3 variants

There were 29 variants with a VAF between 0.1 and 0.2%. 22 of these variants were resequenced by PCR or smMIP (Dr Geoffrey Maher), with only 4 (18%) successfully validating. This suggests that below <0.2%, a large proportion of calls are artefacts.

Table 3.4: Tier 3 validated variants. Counts from COSMIC v82 refer to identical DNA substitutions, identical amino acid substitutions and total substitutions at the specific amino acid.

Variant number	Gene	Position (hg19) and amino acid substitution	VAF range (%)	Number of positive biopsies	COSMIC v82	Germline disorder
58	*BRAF*	chr7:140453096 C>G (p.L613F)	0.1	1	-	-
59	*MAP2K1*	chr15:66727451 A>C (p.Q56P)	0.07 - 0.11	3	8 (8, 8)	-
60	*PTPN11*	chr12:112888162 G>C (p.G60R)	0.09 - 0.17	2	6 (6, 55)	-
61	*RAF1*	chr3:12645688 G>C (p.P261A)	0.15	1	0 (0, 8)	Noonan syndrome

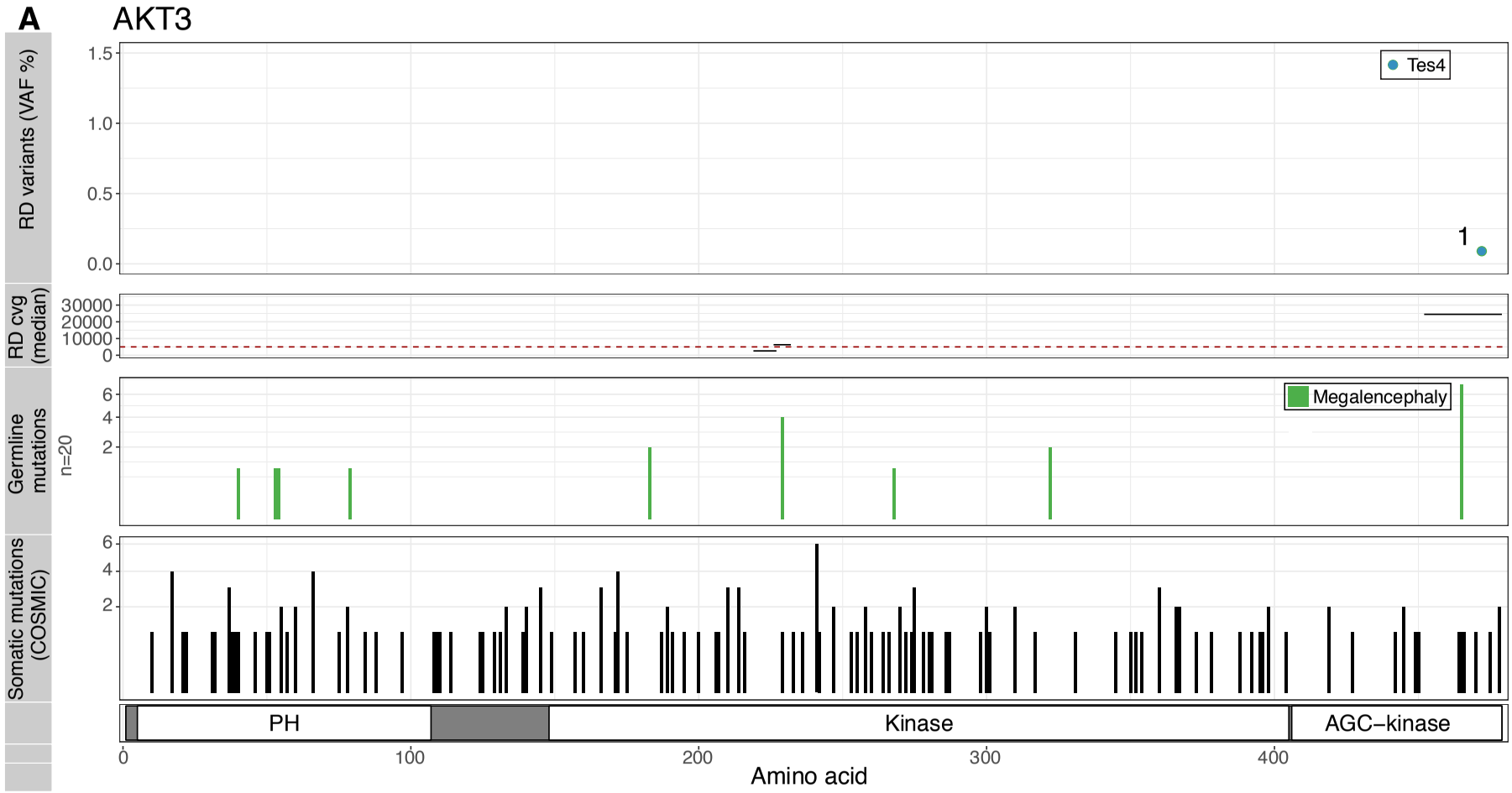
3.4.5 Tier 4 variants

There were 20 calls with a VAF <0.1%. However, as the validation rate of variants in tier 3 was low, no validation was performed for tier 4 calls.

3.4.6 Validated variants

In total, 61 variants in 15 genes were validated in this study, including in five novel genes not previously implicated in SSS (*BRAF, CBL, MAP2K1, MAP2K2, RAF1*). Of these, 51 of the predicted amino acid substitutions have been described in cancer and 25 have been described in a germline disorder (Tables 3.2, 3.3 and 3.4). This overlap with dominant congenital disorders and/or cancer is strongly suggestive that these variants have a functional effect. To visualise the location of the validated variants in the encoded protein and to understand the overlap with the spectrum of previously described germline and somatic mutations, variants were plotted relative to the protein domain structure (Figure 3.7 A-O).

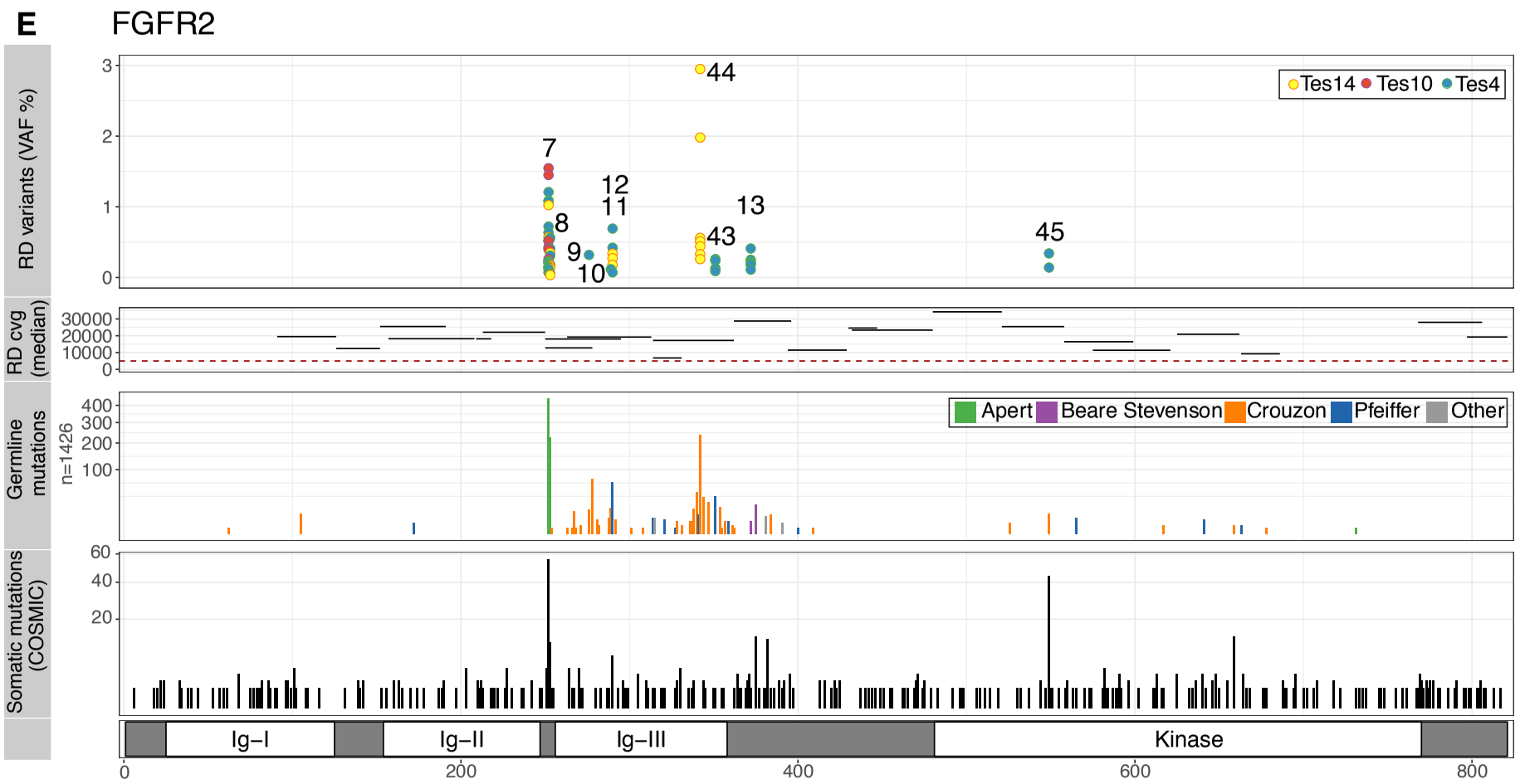
Figure 3.7 *(following page)*: Validated variants in *(A) AKT3 (B) APC (C) BRAF (D) CBL (E) FGFR2 (F) FGFR3 (G) KRAS (H) LRP5 (I) MAP2K1 (J) MAP2K2 (K) NF1 (L) PTPN11 (M) RAF1 (N) RET (O) SOS1.* These plots were produced with the aid of a tool developed by Dr Nils Koelling. (I) Validated variants positioned along the amino acid sequence of the protein, with VAF on the y-axis. Numbers correspond to the mutation number in the tables of validated variants 3.2, 3.3 and 3.4). (II) Relative location and length of amplicons used to sequence hotspots in the gene are plotted on the x-axis. Median coverage per amplicon is plotted on the y-axis. The red dashed line shows the coverage threshold of 5,000x). (III) Number of reported constitutional variants encoding amino acid substitutions associated with developmental disorders (sqrt scale). *AKT3, APC, LRP5* and *NF1* data was obtained from Clinvar (2017-07-10). For all other genes, this data was manually curated from reports in the literature; for *FGFR2*, the track was updated from (Wilkie, 2005). (IV) Number of reported somatic amino acid substitutions reported in cancer (from COSMIC v82). (V) Protein domains from Uniprot.

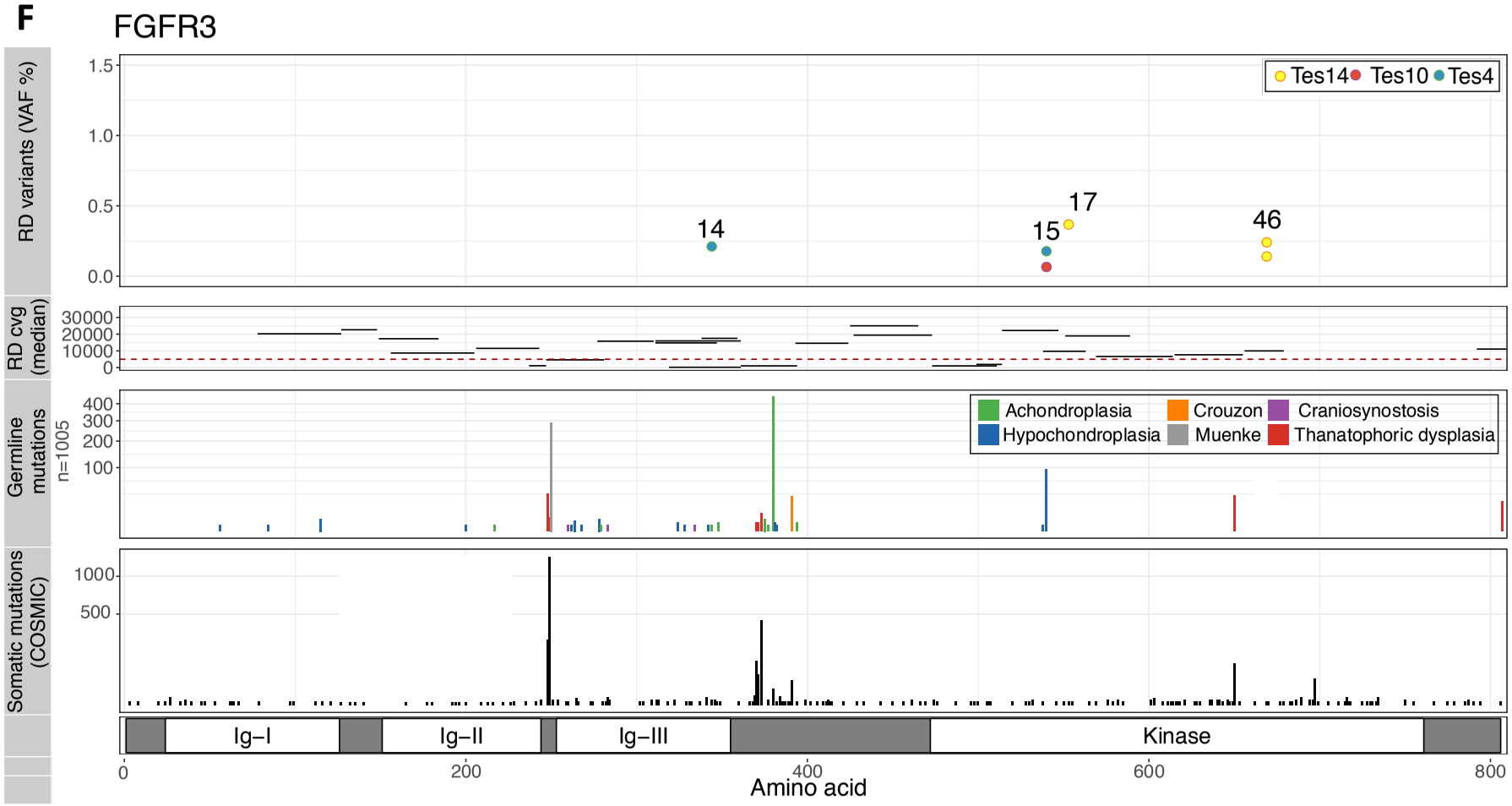

A
AKT3
I
RD variants (VAF %)
1.5
1.0
0.5
0.0
Tes4
1
II
RD cvg (median)
30000
20000
10000
0
III
Germline mutations
n=20
6
4
2
Megalencephaly
IV
Somatic mutations (COSMIC)
6
4
2
V
PH
Kinase
AGC−kinase
0
100
200
300
400
Amino acid

B
APC
I
II
III
IV
V
RD variants (VAF %)
RD cvg (median)
Germline mutations
Somatic mutations (COSMIC)
Tes4
Pathogenic/ Likely Pathogenic
2
1.5
1.0
0.5
0.0
30000
20000
10000
0
4
3
2
1
0
200
100
0
1000
2000
Amino acid

C
BRAF
I
RD variants (VAF %)
Tes14
Tes10
Tes4
1.5
1.0
0.5
0.0
3
4
41
40
58
5
6
II
RD cvg (median)
30000
20000
10000
0
III
Germline mutations
CFCS
Noonan
Other
n=165
40
30
20
10
IV
Somatic mutations (COSMIC)
20000
10000
V
RBD
C1
Kinase
0
200
400
600
Amino acid

D
CBL
RD variants (VAF %)
RD cvg (median)
Germline mutations
Somatic mutations (COSMIC)
I
II
III
IV
V
Tes5
42
n=14
Noonan−like
Cbl−PTB
RING
UBA
Amino acid
0
250
500
750


E
FGFR2
RD variants (VAF %)
RD cvg (median)
Germline mutations
n=1426
Somatic mutations (COSMIC)
Tes14
Tes10
Tes4
Apert
Beare Stevenson
Crouzon
Pfeiffer
Other
Ig−I
Ig−II
Ig−III
Kinase
7
8
9
10
11
12
13
43
44
45
0
200
400
600
800

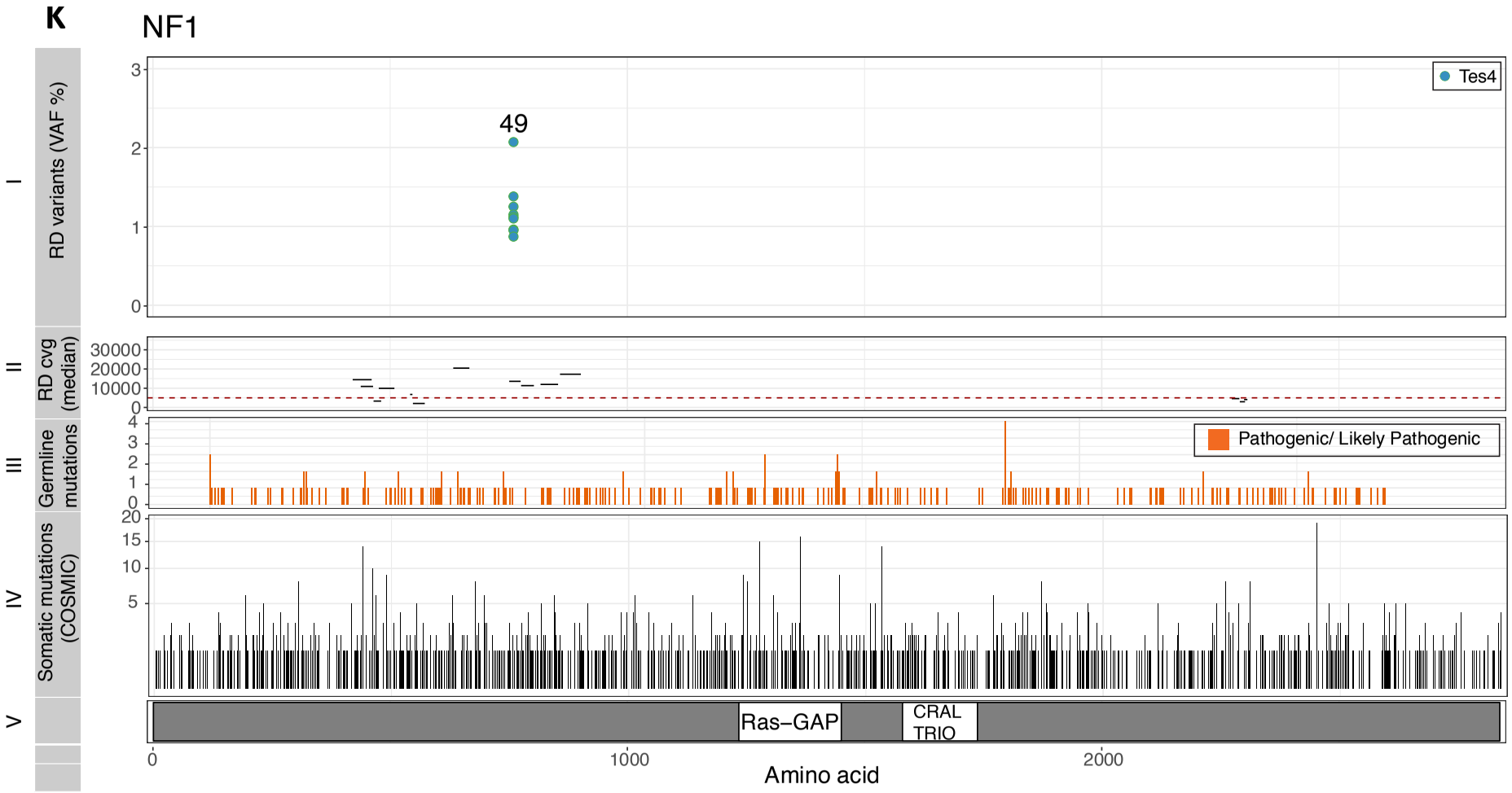

F
FGFR3
I
RD variants (VAF %)
1.5
1.0
0.5
0.0
Tes14
Tes10
Tes4
14
15
17
46
II
RD cvg (median)
30000
20000
10000
0
III
Germline mutations
n=1005
400
300
200
100
Achondroplasia
Crouzon
Craniosynostosis
Hypochondroplasia
Muenke
Thanatophoric dysplasia
IV
Somatic mutations (COSMIC)
1000
500
V
Ig−I
Ig−II
Ig−III
Kinase
0
200
400
600
800
Amino acid

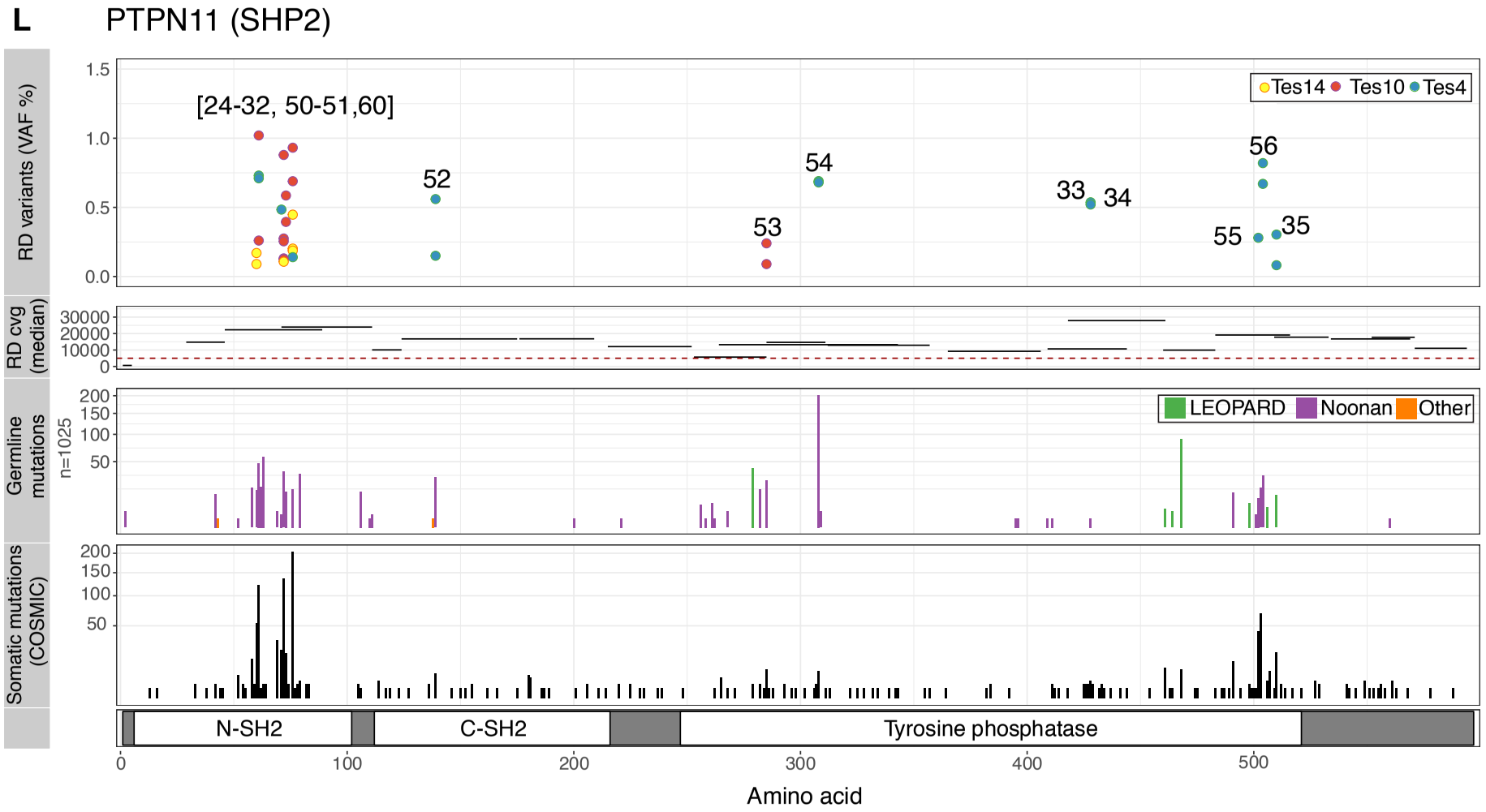
G
KRAS
RD variants (VAF %)
RD cvg (median)
Germline mutations
Somatic mutations (COSMIC)
I
II
III
IV
V
Tes14
Tes10
Tes4
CFCS
Costello
Noonan
19
18
22
21
20
n=50
Amino acid

H
LRP5
RD variants (VAF %)
RD cvg (median)
Germline mutations
Somatic mutations (COSMIC)
I
II
III
IV
V
Tes4
47
Pathogenic/ Likely Pathogenic
Amino acid
0
400
800
1200
1600

MAP2K1
Tes4
59
23
RD variants (VAF %)
RD cvg (median)
Germline mutations
n=49
CFCS
Costello
Melorheostosis
Noonan
Somatic mutations (COSMIC)
Kinase
Amino acid
I
II
III
IV
V

J
MAP2K2
I
RD variants (VAF %)
Tes10
48
II
RD cvg (median)
III
Germline mutations
CFCS
Noonan
n=26
IV
Somatic mutations (COSMIC)
V
Kinase
Amino acid


K
NF1
I
RD variants (VAF %)
II
RD cvg (median)
III
Germline mutations
IV
Somatic mutations (COSMIC)
V
49
Tes4
Pathogenic/ Likely Pathogenic
Ras–GAP
CRAL TRIO
Amino acid
0
1000
2000


L
PTPN11 (SHP2)
RD variants (VAF %)
[24-32, 50-51,60]
52
53
54
33
34
55
56
35
Tes14
Tes10
Tes4
RD cvg (median)
Germline mutations
n=1025
LEOPARD
Noonan
Other
Somatic mutations (COSMIC)
N-SH2
C-SH2
Tyrosine phosphatase
Amino acid

M
RAF1
I
RD variants (VAF %)
Tes4
57
61
1.5
1.0
0.5
0.0
II
RD cvg (median)
30000
20000
10000
0
III
Germline mutations
Costello
Noonan
n=80
40
30
20
10
IV
Somatic mutations (COSMIC)
16
12
8
4
V
RBD
C1
Kinase
0
200
400
600
Amino acid

N
RET
RD variants (VAF %)
RD cvg (median)
Germline mutations
Somatic mutations (COSMIC)
I
II
III
IV
V
Tes10
Tes4
36
38
FMTC
MEN2
MEN2A
MEN2B
n=502
Cadherin
Kinase
Amino acid

O
SOS1
I
RD variants (VAF %)
II
RD cvg (median)
III
Germline mutations
IV
Somatic mutations (COSMIC)
V
Tes4
39
CFCS
Noonan
n=210
DH
PH
Ras–GEF
Ras–GEF
0
500
1000
Amino acid

3.5 Visualisation of VAF in biopsies

In order to visualise the geographic position of the called variants and the relative VAF, heatmaps were produced using the Pheatmap package in R. An example plot is shown in (Figure 3.8). Each testicular slice is constructed as a table and the biopsy numbers are listed in appendix A.1.

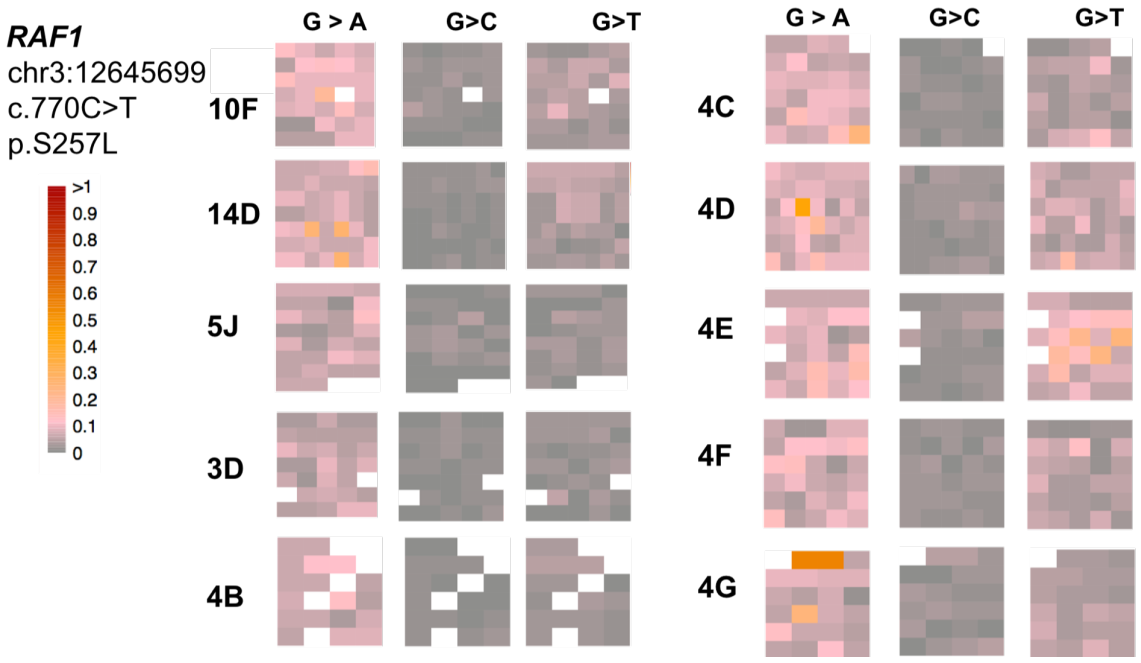


Figure 3.8: Output created by HeatMap Script. The percentage non-reference for each possible nucleotide change at chr3:12645699 is shown for each of the ten testicular slices. This reveals individual biopsies where the mutation frequency is elevated and an understanding of the relative geographic positions between biopsies. It also allows visualisation of the background level of mutation in all samples.

3.6 Synonymous mutations

Of the 61 variants that were validated, only two were synonymous variants. The variant in *AKT3* is a SNP (rs148332912) in 1.1% of the African population. The existence of this variant in the population suggests the mutation is likely neutral. Biopsies were checked for this SNP and no individual was found to be heterozygous, excluding the possibility that the mutation was present as a result of contamination.

The other synonymous variant, LRP5 p.A97A, was observed at a VAF ranging from 0.53% to 1.2%, in testis 4. These biopsies were also positive for a pathogenic variant in KRAS (c.182A>G, p.Q61R) (Figure 3.9). This suggests that the mutation may have arisen in the same seminiferous tubule as the pathogenic *KRAS* mutation and may therefore represent a passenger mutation.

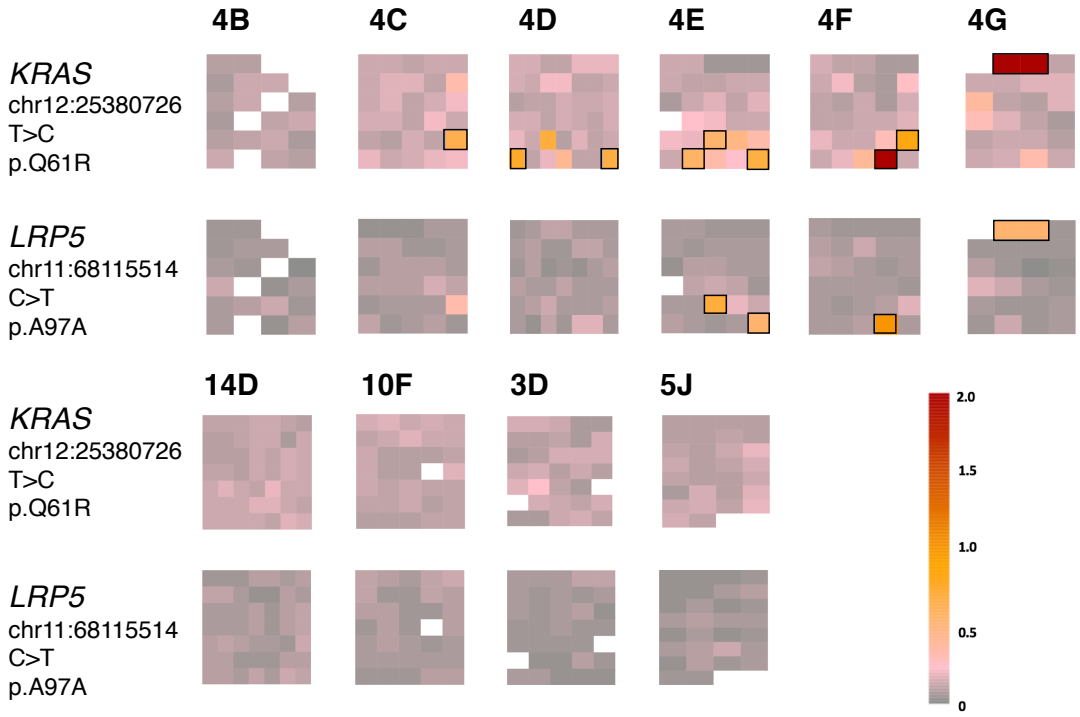


Figure 3.9: Heatmap showing the mutation distribution of *KRAS* c.182A>G (p.Q61R) and LRP5 c.291C>T (p.A97A) in testicular slices. Biopsies in which the variant was called by the statistical prioritisation method are outlined in black.

3.7 Mosaic mutations

A single variant was called in *NF1* c.2280G>A (p.M760I), which displayed a pattern distinct from the other variants that were identified. The variant was originally called in nine biopsies at relatively high VAF (median 1.1%, range 0.9-2.1%), and inspection of the mutation frequency in each sample showed numerous other biopsies in Tes4 with elevated VAFs (Figure 3.10). This wide distribution of biopsies with an elevated VAF is suggestive of the mutation being present in multiple seminiferous tubules and therefore having arisen as an earlier post-zygotic mosaic event. However, no other tissue was available from this individual to test whether the variant was restricted to the single testis in which it was observed.

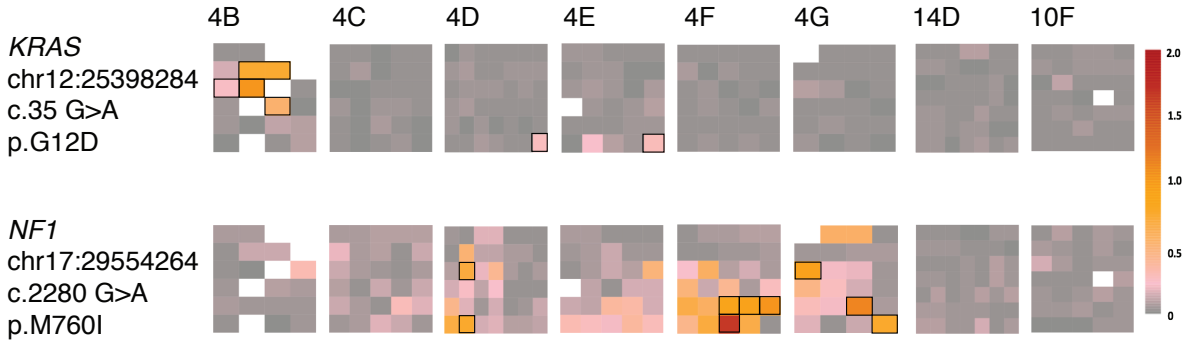


Figure 3.10: Heatmap showing the distribution of G>A variants in *NF1* at chr17:29554264. This variant was initially called in 9 biopsies in Tes4 (outlined in black). Elevated mutation levels can be observed in a number of additional biopsies as well as a strong clustering of the variants in slices E, F and G. Heatmaps of the same variants in Tes14 and Tes10 demonstrate that the higher variant levels are specific to Tes4. For comparison, a variant of the same nucleotide change type is shown (KRAS p.G12D), called in 6 biopsies in Tes4).

3.8 Discussion

3.8.1 Validated variants

In conclusion, 61 distinct variants were identified and validated. The observed variants displayed a striking overlap with mutational hotspots previously associated with cancer

and developmental disorders (Figure 3.7).

Although only 41.9% of the callable sequence of the panel belonged to RTK-RAS-MAPK pathway, all but three of the validated variants represented known or likely pathogenic substitutions within members of this pathway (p value = 4.233e-13, Fisher's two tailed test). This significantly strengthens the association of SSS with the RAS-MAPK pathway.

Of the three variants identified outside the RTK-RAS-MAPK pathway, two were synoynmous (AKT3 and LRP5) and likely neutral (section 3.6). The third variant was identified in *APC*, a tumour suppressor, which controls β-catenin turnover and functions in the Wnt pathway. The variant (c.4471T >G (p.F1491V) was identified in a single biopsy at a VAF of 0.47%. This variant has not been reported in the literature or in COSMIC but may be functional (CADD score = 26.2; Kircher et al. (2014)). However, the significance of this singleton variant remains unclear and requires further investigation.

In this study, the majority of the testes sampled were from older individuals (71 to 90 years). Only one testis (5J) was from a younger individual (34 years old). In this individual, a single variant was identified (*CBL* chr11:119148991 G>A (p.C404Y). This variant validated in two adjacent biopsies and likely represents a single clonal event. This is in line with expectations that the prevalence of mutant clones increases with time. Surprisingly, no variants were identified in the testis of an 87 year old man. It was not possible to determine the state of spermatogenesis in this individual due to the lack of available histological tissue.

3.8.2 Technology

The use of a massively-parallel PCR based discovery screen was successful in that it led to the identification of novel mutations associated with SSS. Although variants as low as 0.06% were validated, this was only possible where there was sequence information from overlapping amplicons. As this comprised only 11.3% of the panel, it was not possible for the majority of variants to be validated using this criteria.

In order to identify variants with high confidence, extensive manual curation and variant validation was required. Other technologies are now available that can aid in variant calling through the incorporation of UMIs at the time of capture. The performance of these will be investigated in the subsequent chapters. This may aid the observation of mutations with more subtle effects that lead to milder clonal expansion.

Chapter 4

Multiplex target discovery of candidate variants with a commercial capture-based UMI technology

4.1 Introduction

Previous studies described in chapter 1 and the work described in the previous chapter have established that, at a molecular level, SSS is strongly associated with the activation of components of the growth factor receptor-RAS/MAPK signalling pathway – a key regulator of stem cell homeostasis in the testis. Additionally, RAS-regulated pathways are important in mediating proliferation and survival in other cellular contexts, including in the brain (Brambilla et al., 1997; Kim et al., 2003). Functionally, the RAS-MAPK and PI3K-AKT pathways overlap and share a number of signalling molecules including mTOR (Reijnders et al., 2017). The mTOR pathway is also a key regulator of normal brain development and hyperactivity of this pathway has been shown to lead to a variety of neurocognitive outcomes, including macrocephaly and seizures (Pun et al., 2013). *De novo* hyperactivating mutations in genes in the mTOR pathway have recently been shown to be linked to brain volume (Reijnders et al., 2017). Brain overgrowth has been significantly associated with neurodevelopmental disorders (de Ligt et al., 2012). An upstream regulator of mTOR (Rheb, Ras Homolog Enriched in Brain) was identified as a candidate ID gene in a cohort of patients (Reijnders et al., 2017). This leads to the intriguing

possibility that dysregulation of the RAS-MAPK pathway may also be relevant to the pathology of neurodevelopmental disorders (Goriely et al., 2013) and that SSS-associated mutations may contribute to neurocognitive outcomes. Further support for this hypothesis also comes from epidemiological observations of the association between APA and complex neurocognitive disorders, such as schizophrenia (Malaspina et al., 2001, 2015) (section 1.4).

As has been demonstrated for a subset of spontaneous human disorders (Table 1.2), the preferential selection of activating mutations in the male germline can explain epidemiological observations of birth prevalence and association with APA. To investigate whether selfish spermatogonial selection (SSS) may be a contributing mechanism explaining the link between APA and neurocognitive disorders, genes proposed to be involved with intellectual disability, autism spectrum disorder and developmental disorders will be screened in the testis.

Detection of multiple variants within a single biopsy of the testis (chapter 3) confirms the success of strategy B (described in Figure 1.6). Plotting the geographic register of the validated variants demonstrates that the variants likely represent single clones expanding along the length of a tubule (Figure 3.8). Therefore, detection of SSS associated mutation may be more readily performed in individual seminiferous tubules where VAF could reach up to 50%. Such an approach may aid the identification of mutations with a more subtle effect that do not promote expansion to the same level as the previously described, highly activating mutations.

As discussed in chapter 1, technologies that incorporate UMI at the time of capture may provide the sensitivity required to detect low-level variants and thus enable the detection of mutations under more subtle selection in the male germline. A commercial target enrichment system, HaloPlex HS (Agilent Technologies), will therefore be assessed (Figure 4.1). UMI information can be used to generate a consensus read, providing additional confidence to distinguish true variation from artefacts arising due to sequencing or PCR error. As well as the incorporation of UMI into the fragments, an additional benefit of the technology is the multiple restriction enzyme digestions performed in step

1 of the method (Figure 4.1). This generates unique overlapping amplicons that provide additional confidence when calling variants, as evidenced by the high validation rate achieved for variants called in overlapping amplicons in chapter 3 (section 3.4.2).



Figure 4.1: Overview of HaloPlex HS target-enrichment method. Step 1: Digestion and denaturation of sample DNA with restriction enzymes. DNA samples are digested with 16 different restriction enzymes to create a library of gDNA fragments. Step 2: Hybridization and circularization of the probe library to DNA targets. During this step, UMI and Illumina sequencing motifs are incorporated into the fragments. Step 3: Ligation and capture of uniquely barcoded targets. Circularized target DNA-probe hybrids contain biotin and are captured on streptavidin beads. Step 4: Amplify captured target fragments by PCR. Figure taken from HaloPlex HS Target Enrichment System Protocol (Version C1, December 2016: Agilent Technologies).

The aims of this chapter are to:

- Multiplex target discovery of candidate variants associated with neurocognitive disorders, in order to investigate whether there is any evidence that mutations linked to neurocognitive disorders are under selection in the male germline.
- Assess the performance of a commercially available target enrichment system that incorporates UMI technology at the time of capture, in order to determine whether

this technology is sensitive enough to detect low-level variants directly in the human testis.

- Address whether SSS associated mutations can be detected within individual seminiferous tubules.

4.2 Gene selection

I performed a review of the literature of exome studies of patients with intellectual disability and autism (O'Roak et al., 2012; Rauch et al., 2012; D'Gama et al., 2015; Sanders et al., 2012; de Ligt et al., 2012; Iossifov et al., 2014; Neale et al., 2012). From this, a long list of candidate genes was generated with priority being given to genes that were recurrently identified across studies. Genes were further prioritised based on reported mutation distributions, function, the size of gene, G/C content, testicular expression and pLi score (a measure of LOF intolerance) (Samocha et al., 2014). As well as genes identified as candidates from the literature, genes associated with syndromic autism (e.g. *FMR1, UBE3A*), overgrowth conditions (e.g. *EZH2, SKI*) and recurrent hotspot mutations in developmental disorders (e.g. *SMAD4, PACS1*) were selected. Genes involved in overgrowth conditions were selected due to their potential to cause increased cellular proliferation when dysregulated. Additionally, genes in the RAS-MAPK pathway were also targeted. In total, 37 genes were covered by the panel (Materials and Methods section 2.6.1).

4.3 Results

To test the performance of the HaloPlex HS capture system and identify novel mutations and pathways under selection in the male germline, systematically biopsied human testes and individual seminiferous tubules were screened. In total, the panel was used to capture 25 testicular biopsies, 19 seminiferous tubules and 5 control samples. These were screened by ultra-deep sequencing on the Illumina HighSeq. Samples were analysed using SureCall

(Agilent Technologies) to trim the sequencing reads, create aligned sequencing files and produce lists of variant calls.

4.3.1 Analysis of panel performance: coverage

The samples were sequenced in three libraries on the Illumina HiSeq. On average, 977182.25 reads passing mapping quality filters were generated per sample (range 502,558 to 1,675,487). Of the total target panel of 61.31 kbp, a median of 83.36% of target regions were covered by at least 1000 consensus reads (Figure 4.2).

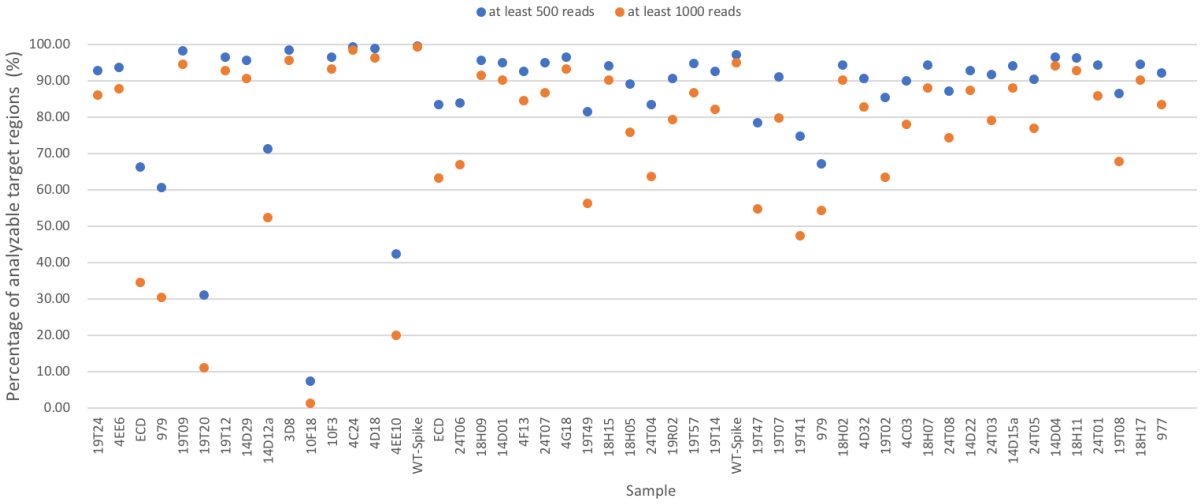


Figure 4.2: Percentage of analyzable target regions (y-axis) covered by at least 500 reads (blue dot) or at least 1000 reads (orange dot) in all samples (x-axis).

For three regions, probes failed to generate any coverage (*UBE3A* exon 14, *MECP2* exon 14 and *PTEN* exon 1). Coverage across all target regions was consistent between samples but was not evenly balanced between regions.

Genes with high G/C content are difficult to capture and sequence, as demonstrated in the previous chapter. In the HaloPlex HS protocol, DNA is digested by 16 restriction enzymes prior to probe capture. This increases accessibility of the probes to the genomic targets. Coverage was high across *FGFR3* in this study (Figure 4.3), suggesting that prior DNA digestion improves target capture.

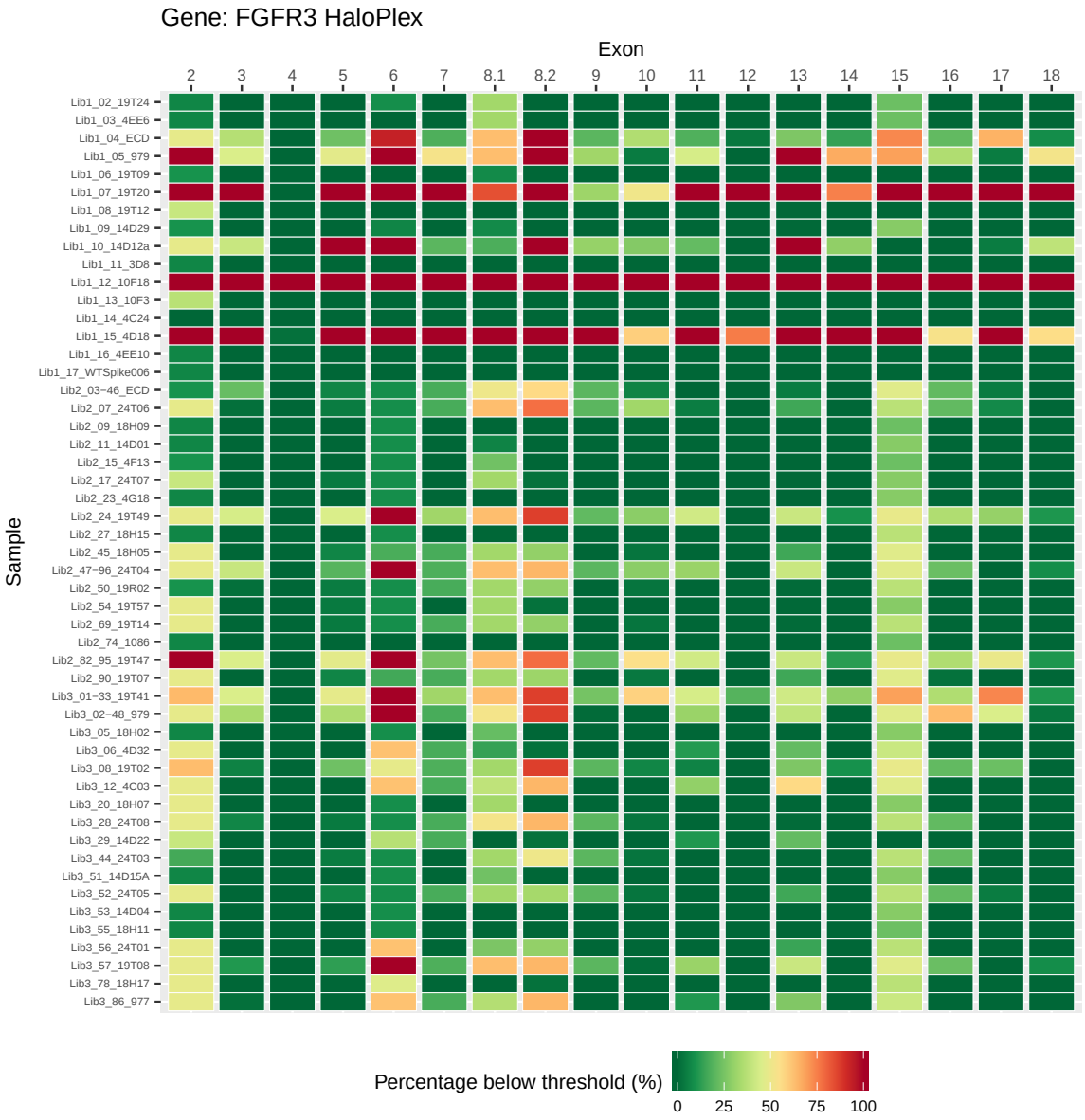


Figure 4.3: *FGFR3* coverage generated by HaloPlex HS panel. For a given target exon of *FGFR3*, the percentage of samples with a consensus coverage below 1000 reads is shown. Although some specific samples have performed poorly (y-axis), there are few regions (x-axis) with poor coverage) The achondroplasia mutation site (c.1138G>A;C) (p.G380R)) is in exon 9.

4.3.2 Variant calling

In total, 3666 (1901 unique) variants were called across all samples (Figure 4.4).

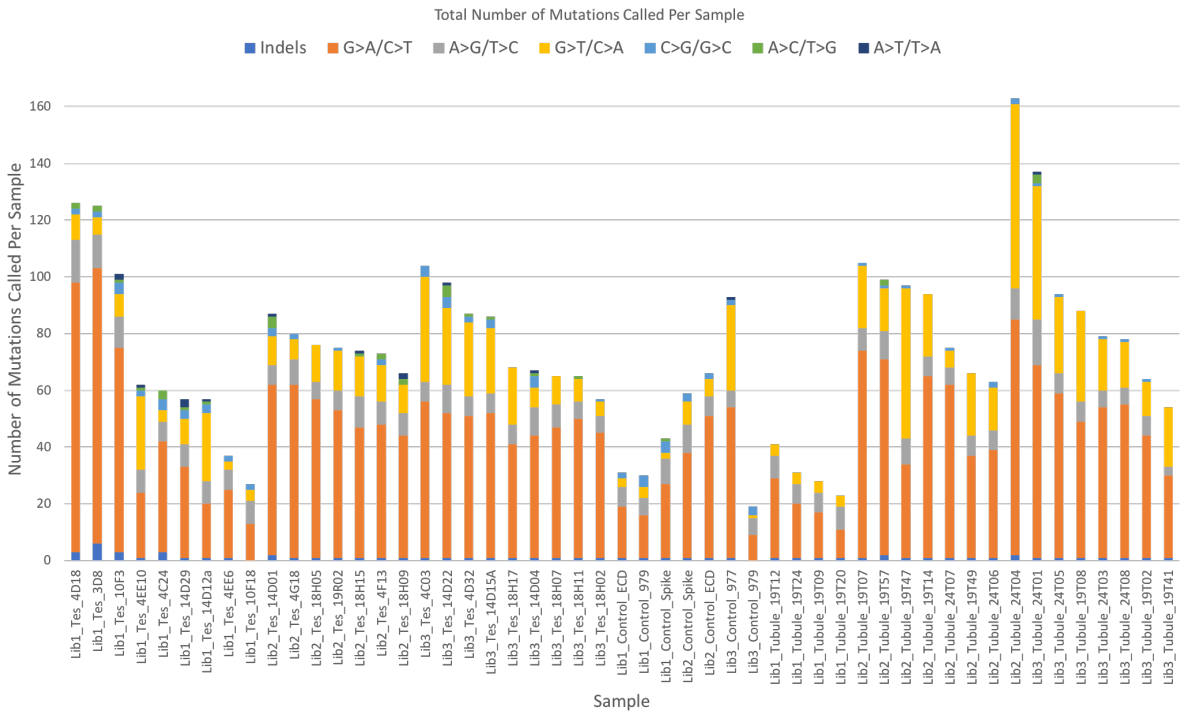


Figure 4.4: Total variants per sample coloured by nucleotide change type.

4.3.3 Analysis of panel performance: variant detection

One control DNA sample was sequenced twice across two of the libraries. This control sample had been spiked with 5 known mutations in *FGFR2* at an approximate level of 0.06% (for full mutation details, see Table 5.1 on page 133). Despite the presence of a double and triple mutant sequence that was visible with IGV (Figure 4.5), the mutations were called by Surecall as a series of single mutations. This means that the mutation levels predicted by the software are overestimated due to the overlap between the mutant sequences (Table 4.1). The c.755C>G mutation failed to be detected in either replicate. Presence of this mutation in the control sample was confirmed using an independent technology (chapter 5) which detected this mutation at levels between 0.06% to 0.1%. With an average WT UMI depth at this position of 7130 in replicate one and 10898 in replicate two at this position, it is possible that the molecules containing the C>G mutation failed to capture due to the low level.

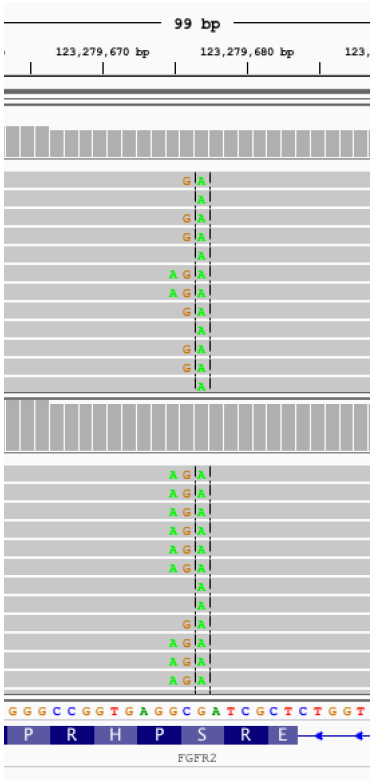


Figure 4.5: Screenshot from IGV showing a single (c.755C>T), double (c.755_756CG>TC) and triple (c.755_757CGC>TCT) mutation in *FGFR2* spiked into a WT *FGFR2* background as a positive control. The upper panel is the sample sequenced in library 1 and the lower panel shows the sample sequenced in library 2.

Table 4.1: Detection of known *FGFR2* variants

Mutation	**Amino Acid Substitution**	**Non Reference UMI (Replicate 1)**	**Percentage non-reference (%) (Replicate 1)**	**Non Reference UMI (Replicate 2)**	**Percentage non-reference (%) (Replicate 2)**
c.758C>G	p.P253R	14	0.13	0	-
c.755C>G	p.S252W	0	-	0	-
c.755C>T	p.S257L	21	0.19	23	0.32
c.755_756 CG>TC	p.S252F	11	0.10	18	0.25
c.755_757 CGC>TCT	p.S252W_ P253S	3	0.32	8	0.11

4.3.4 Analysis of panel performance: known positive mutations

Of the 25 testicular biopsies, 17 had been previously sequenced and analysed as part of the massively-parallel PCR-based discovery screen described in chapter 3. Twenty variants were validated in these biopsies in chapter 3. Of these, 12 were also on the variant list generated by SureCall (Table 4.2). The majority (5/8) of variants that were not detected

were in regions with a low coverage (<120 reads) in the relevant sample.

Table 4.2: Detection of previously validated variants

Gene	**Variant and position (hg19)**	**Amino Acid**	**Sample**	**Raindance Level (%)**	**HaloPlex Level (%)**
FGFR3	chr4:1807488 G>A	p.V553M	14D12a	0.41	na Low coverage
FGFR3	chr4:1808029 C>G	p.R669G	4C3	0.14	0.43
KRAS	chr12:25398284 C>G	p.G12A	4C24	0.18	0.43
NF1	chr17:29554264 G>A	p.M760I	4D32	0.96	0.79
NF1	chr17:29554264 G>A	p.M760I	4D18	0.32	0.30
SOS1	chr2:39250292 T>G	p.Q426P	4E6	0.30	0.36
RET	chr10:43613906 G>T	p.L790F	4E10	0.21	na Low coverage
PTPN11	chr12:112888162 G>C	p.G60R	14D22	0.17	0.08
PTPN11	chr12:112888211 A>C	p.E76A	10F18	0.82	na Low coverage
PTPN11	chr12:112924336 G>A	p.V428M	4E6	0.56	0.34
PTPN11	chr12:112926890 A>G	p.M504V	4E10	0.82	-
FGFR2	chr10:123258034 A>C	p.N549K	4D18	0.14	0.18
FGFR2	chr10:123276893 A>T	p.C342S	14D22	0.51	0.44
FGFR2	chr10:123276893 A>T	p.C342S	14D29	2.95	2.86
FGFR2	chr10:123279562 C>A	p.W290C	4E10	0.34	0.64
FGFR2	chr10:123279677 G>C	p.S252W	14D15a	1.24	-
FGFR2	chr10:123279677 G>C	p.S252W	10F18	0.25	na Low coverage
FGFR2	chr10:123279677 G>C	p.S252W	10F3	1.16	-
BRAF	chr7:140453132 T>G	p.K601N	4D32	0.22	0.23
BRAF	chr7:140453155 C>G	p.D594H	10F18	0.23	na Low coverage

4.3.5 Detection of variants in seminiferous tubules vs biopsies

Seminiferous tubules, ranging in length from 4cm to 15 cm (mean: 8 cm), were isolated from the testes of two individuals (Tes19, aged 58 years, and Tes24, aged 50 years). The rationale for sequencing individual seminiferous tubules was that SSS associated mutations, if present, would be more readily detectable at an elevated level (up to 50%). Therefore, I began by examining all variants called in the tubules only (n = 1439 variants, 944 unique variants) (Table 4.3). I inspected all mutations with an elevated mutation level (>3%). This corresponded to a total of 337 variants at 28 distinct genomic positions. Of these, all but 3 mutations were identified as heterozygous SNPs segregating within all samples of each individual.

Table 4.3: Number of variants, length (cm) and total number of reads passing mapping quality filters per seminiferous tubule.

Tubule Sample	**Number of Variants**	**Length of tubule (cm)**	**Sequencing depth**
19T12	40	4	6656589
19T24	31	15	4799922
19T09	27	8	6200704
19T20	22	15	502558
24T04	161	8	1675487
19T07	105	5	2604316
19T57	98	8	3285187
19T47	96	7	1355065
19T14	94	7	2899559
24T07	75	10	3353408
19T49	66	10	1394243
24T06	62	7	1791580
24T01	136	6	3466644
24T05	94	7	2285278
19T08	88	5	1793927
24T08	78	10	2327235
24T03	78	5	2314222
19T02	64	5	1540330
19T41	53	10	1141148

The remaining three variants ranged in levels from 3.4% to 22.4%. These were visually inspected using IGV and found to result from alignment errors due to repetitive sequences or as a result of improper trimming of the amplicons (Figure 4.6).

In the biopsies, there were eight unique variants called above 3%. Of these, six were due to alignment problems and two were due to low coverage (<350 reads). As there were no variants >3% in the seminiferous tubules worth further investigation, all remaining variants were analysed with the biopsy variant calls.

4.3.6 Variant Prioritisation

After exclusion of known SNPs, a total of 2619 (1856 unique) variants remained. The majority of variants fall within the range 0.1% to 0.2% (Figure 4.7).

Manual inspection of these variants is highly informative and patterns of error can be easily distinguished (Figure 4.8). However, this is not feasible for this long list of variants.

To prioritise the long list of candidate variants, I wrote a pipeline to automate extraction of supporting information about each variant. Each variant position was interrogated in the sequence alignment files by extracting:

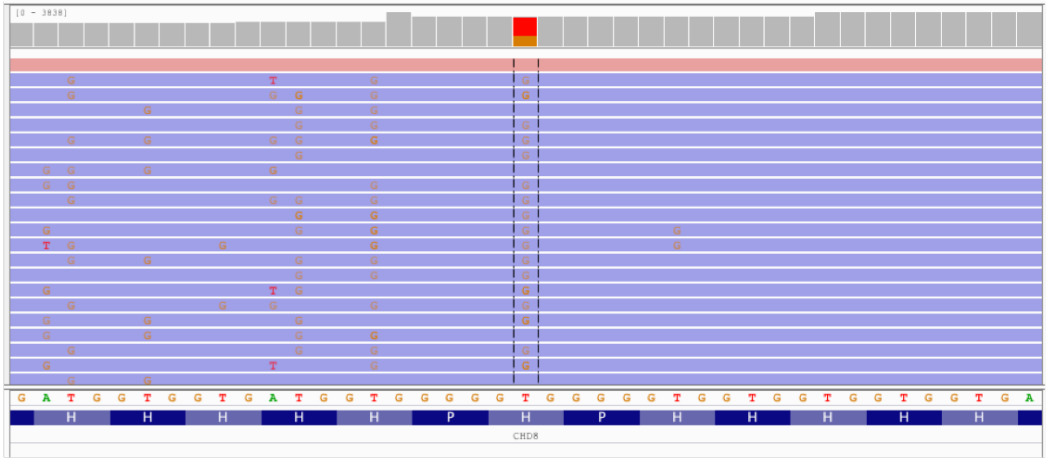
- The reads that overlapped the target position from the sequence alignment files. Only reads that are correctly aligned and contain less than 5 errors across the length of the read were considered. Correct alignment was assessed using the CIGAR value, a string that describes how a read aligns to a reference sequence using a combination of character operators and numbers (e.g. 22M refers to 22 exact matches ('M') to the reference sequence).

- The number of independent overlapping amplicons that supported the variant. This was calculated by creating a vector of unique start positions for the supporting reads.

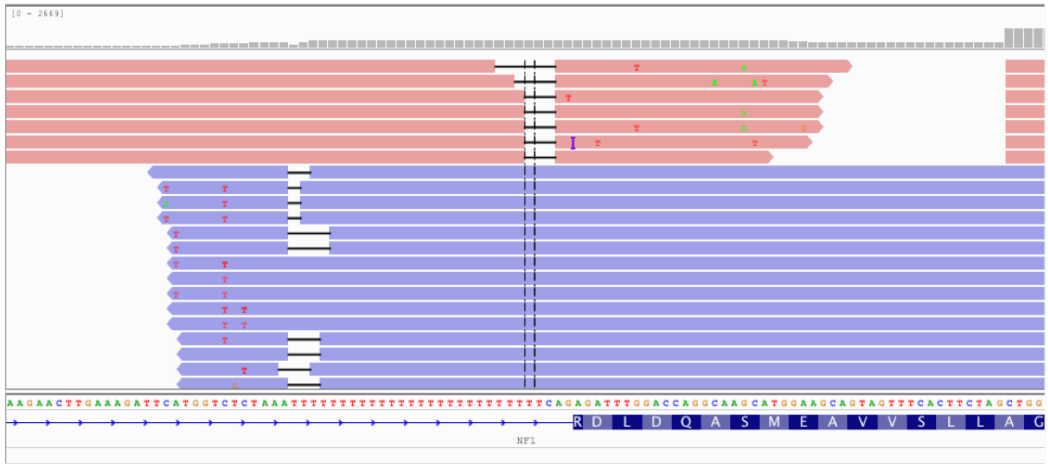
- The number of other non-reference nucleotides present at the same position.

Figure 4.6 *(following page)*: Image from IGV showing three unique mutations in seminiferous tubules present at levels above 3%. (A) CHD8 mutation called at 22.37% in tubule sample 19T57. Note the number of other errors present in these reads. (B) NF1 mutation called at 6.06% in 19T57. (C) RET mutation called at in 24T04 (3.44%, top panel) and 19T47 (3.38%, lower panel).

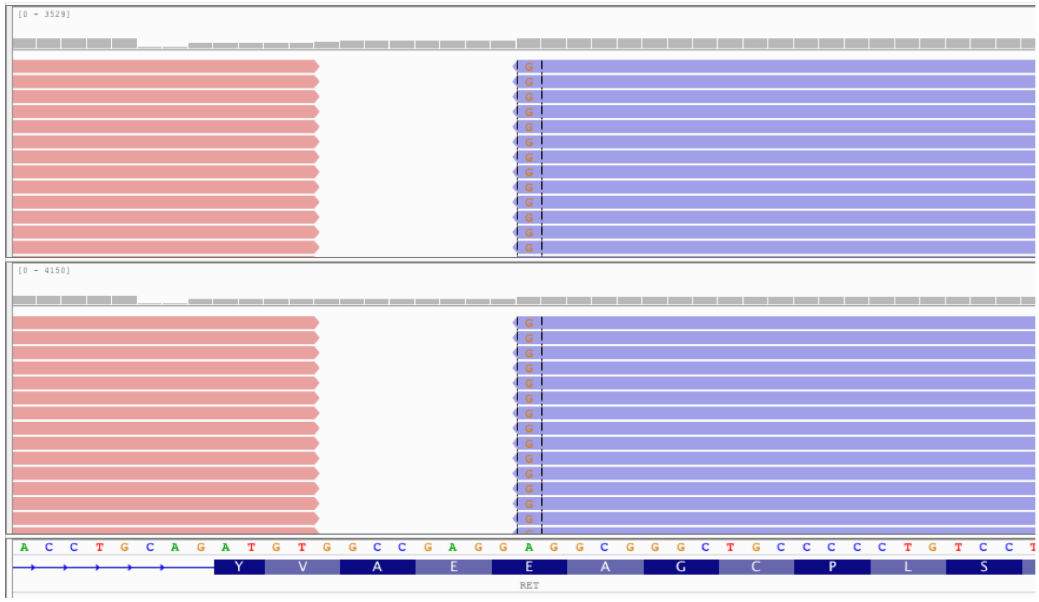
A. *CHD8* chr14:21854019 T>G p.H2221P



B. *NF1* chr17:29559086 TTC>T intronic



C. *RET* chr10:43607559 A>G p.E512G



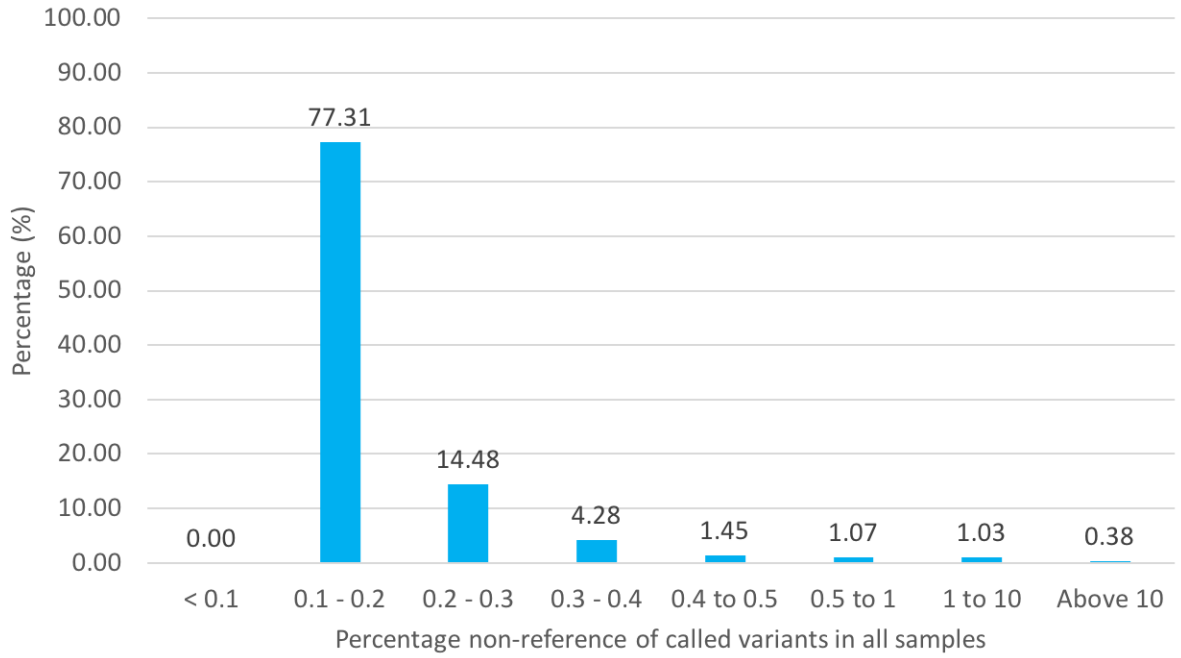


Figure 4.7: Percentage of the called variants (y-axis) that fall within each percentage level of non-reference calls (x-axis). The majority of variants have a non-reference percentage 0.1 and 0.2%.

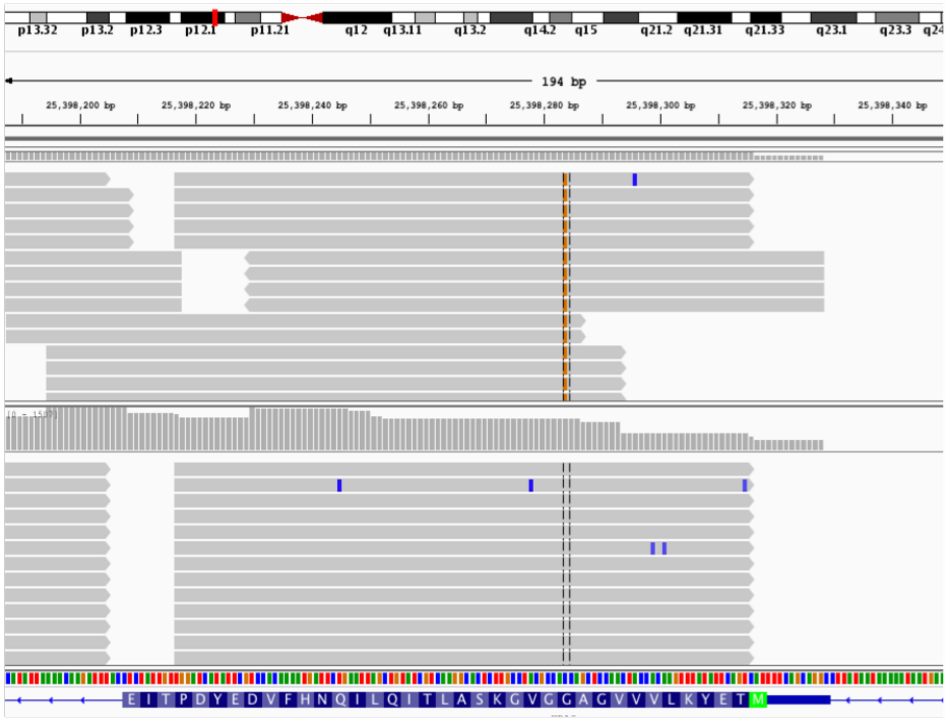
- The number of other samples in which the variant was present. This was calculated as the total number of other samples in which the variant is present and separately as the number of samples in which the variant is present and also supported by more than one amplicon.
- The number of reads that support each consensus UMI.
- The sequence identity of the supporting UMIs.
- The number of forward and reverse reads supporting the variant.
- The base quality (Phred) of the non-reference nucleotide.
- The sequence context of the variant position (5 bp either side of the variant posi-

Figure 4.8 *(following page)*: IGV comparison of two variants identified by SureCall. (A) This variant in *KRAS* is supported by multiple independent amplicons (top panel) and is absent from other samples (bottom panel). This variant has been previously validated in this sample (chapter 3). (B) This variant in *NF1* is at the end of a single amplicon and is present in multiple samples. This is consistent with a pattern of incorrectly trimming the reads to remove the signature of enzyme digestion.

A. KRAS chr12:25398284 C>G p.G12A

Lib1_4C24

Lib2_24T06



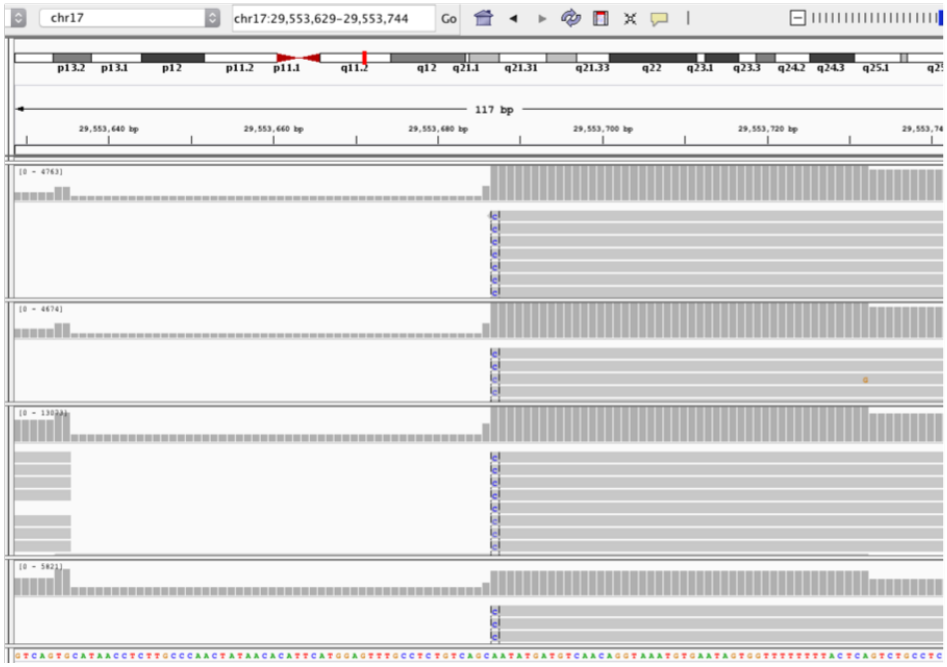
B. NF1 chr17:29553687 A>C p.N746H

Lib1_3D8

Lib1_3D8

Lib1_Spike

Lib1_Control2



tion).

An output table was created using this script and a secondary script was written to produce a summary file (Figure 4.9). Each category tested was annotated as pass/fail. For example, if the variant was supported by multiple amplicons then 'pass' was printed to the summary file while if the variant was supported by a single amplicon then 'fail' was recorded.

Figure 4.9 *(following page)*: Example annotation output for known positive mutations. Note that this information is printed for all alignment files that are provided to the script, but details are only shown here for the sample in which the variant call was made by SureCall. (A) Detailed output for target variants. From left to right, columns show: Name of gene in which variant is located; position in genome (hg19) and nucleotide change; sample in which variant was called; VAF (%); sequence context (5 bp either side of variant position); coverage at target position in sample; total count of reads supporting (in order) nucleotide A, C, G, T; total count of amplicons supporting (in order) nucleotide A, C, G, T; number of reads that form the consensus group for the UMIs supporting the called non-reference nucleotide; number of reads supporting the non-reference nucleotide in the forward and reverse orientation; sequence of all UMIs supporting the non-reference call; base quality of non-reference calls; MD tag. Sequence context, UMI sequences, base quality and MD tag are printed for subsequent scrutiny of the variant but are not used by the second summary script. (B) Summary output for target variants. From left to right, columns show: Name of gene in which variant is located; position in genome (hg19) and nucleotide change; sample in which variant was called; Amplicon Filter - PASS (multiple supporting amplicons) /FAIL (single amplicon); Error filter - PASS (no other non-reference nucleotides at position) / FAIL (other non-reference nucleotides present at position); Soft Sample Filter - PASS (variant is not present in any other sample) / FAIL (variant is present in other samples); Hard Sample Filter - PASS (variant is not present in any other sample after removal of non-reference nucleotides supported by a single amplicon) / FAIL (variant is present in other sample after removal of non-reference nucleotides supported by a single amplicon); Other samples count - The number of other samples the variant was present in, The number of other samples the variant was present in and supported by multiple amplicons; Collapsed reads - The UMI at this position are not supported by singleton reads; Read bias flag - The supporting reads for the non-reference variant are not present entirely in either the forward or reverse orientation).

A

Gene	Variant (hg19)	Sample	VAF (%)	Sequence Context	Coverage	Counts (ACGT)	Amplicons (ACGT)	Collapsed Read Counts	Plus/Minus Count	UMI Sequences	Base Quality	MD Tag
NF1	chr17:29554264 G>A	Lib1_4EE10	0.32	GTGATCGCACT	6587	21,0,6540,11	4,0,6,1	2, 1, 1, 2, 1, 1, 4, 3, 3, 1, 3, 2, 1, 1, 4, 3, 3, 1, 3, 2, 1	8,13	TTAGATGTTA, TTCCTACAGT, GAGATCCTGG, ACACAGTGCC, AGTCGCTTCG, GAGGCGTATC, AGTGATCGTG, CTAACAGTTT, CAAACACCGG, AGGATCTCAA, TGCAGTACGC, ATTCCATCGA, GCAACAACAC, TATACACAGG, AGTGATCGTG, CTAACAGTTT, CAAACACCGG, AGGATCTCAA, TGCAGTACGC, ATTCCATCGA, GCAACAACAC	38, 40, 37, 31, 34, 35, 40, 40, 40, 40, 40, 40, 40, <30, 40, 40, 39, 40, 39, 40, 39	98G0, 98G0, 98G0, 98G0, 98G0, 96G2, 67G31,
FGFR2	chr10:123258034 A>C	Lib1_4EE10	0.18	AGAAGATTTAT	6080	6063,11,0,1	5,4,0,1	4, 2, 1, 1, 1, 1, 1, 1, 1, 4, 2	4,7	TATGTAATAC, GAGACGATAA, AGGACTCTTC, CCGACTCTCA, GACGCTCTTC, AGACGCGTCT, ACCCTTTGTC, AGACGCGTCT, ACCCTTTGTC, TATGTAATAC, GAGACGATAA	40, 39, 40, 40, 40, 40, 40, 40, 40, 40, 40	62A36, 62A36, 26A72, 26A72,
FGFR3	chr4:1808029 C>G	Lib3_4C03	0.32	TTGACCGAGTC	3461	0,3445,11,3	0,3,2,2	7, 5, 6, 7, 2, 7, 5, 6, 7, 2, 4	5,6	AAGACGCGAA, GTAGCCGCGC, AACGAATATA, AGCAGGAGCC, TCTGGAATTA, AAGACGCGAA, GTAGCCGCGC, AACGAATATA, AGCAGGAGCC, TCTGGAATTA, TTGAATAGGA	40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40	13C46, 13C46, 13C46, 13C46,
KRAS	chr12:25398284 C>G	Lib1_4C24	0.43	CCTACGCCACC	9953	0,9900,43,1	0,5,5,1	2, 2, 2, 5, 5, 8, 5, 1, 2, 4, 4, 3, 2, 3, 3, 3, 2, 1, 1, 5, 3, 2, 5, 4, 1, 4, 4, 2, 2, 2, 1, 2, 1, 2, 5, 5, 8, 5, 1, 2, 4, 4, 3	10,33	CGGGCGAGTT, ATCGCAGTAT, CTCCAAGATA, CTACAACAAA, TCTAATCCAC, AATTAAGCTC, GAATACTTAA, ACCCACATAG, AAGTCATATC, CCAACTTACA, GAGCGGCAAG, GGTATTGAAA, CAGGTAGATA, AGCTACATCA, CCAAGTGCGA, TTAAAATTGG, GCTAAAACTA, ATAGAGGTTC, GCCTCGTAGA, GCTTGATACC, CATAGCACGC, TTCCGCCAAG, CCCAGACAAA, CATCGTTAAA, CACATAGCAA, CAGTGCACTA, TTGTAGTCGC, CACCCACGTC, CATGATTAAA, CAGTACGATG, GTTTTTCGAA, CGCATAAAAC, AGCAACCGTA, CTCCAAGATA, CTACAACAAA, TCTAATCCAC	40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 39, 40, 40, 40, 40, 40, 40	96C2, 96C2, 89C9, 89C9, 89C9, 89C9, 89C9, 89C9, 89C9, 89C9, 89C9, 89C9

B

Gene	Variant (hg19)	Sample	Amplicon Number	Other Errors	Other samples (soft)	Other samples (hard)	Other samples	Collapsed Reads	Read Bias
NF1	chr17:29554264 G>A	Lib1_4EE10	PASS	FAIL	FAIL	FAIL	13,9	PASS	PASS
FGFR2	chr10:123258034 A>C	Lib1_4EE10	PASS	FAIL	FAIL	PASS	1,0	PASS	PASS
FGFR3	chr4:1808029 C>G	Lib3_4C03	PASS	FAIL	FAIL	PASS	1,0	PASS	PASS
KRAS	chr12:25398284 C>G	Lib1_4C24	PASS	FAIL	FAIL	PASS	3,0	PASS	PASS

The example variants shown in Figure 4.9 are all known positive variants that were independently validated in chapter 3. The variant described in *NF1* is present in other samples as this is the likely mosaic mutation detailed in section 3.7 of chapter 3). As demonstrated in chapter 3, variants supported by overlapping amplicons have a high validation rate. Based on the pattern observed for known positive mutations, the most informative flags appear to be amplicon count and collapsed read number (Figure 4.9).

The output of the annotation summary script was used first used to filter variants. Variants called in control samples and present in the majority of samples were reasoned to be most likely to be artefacts. Removal of these variants resulted in 2393 total variants. 1798 variants remained after removing those that failed the number of supporting amplicons test. A final 1678 variants remained after removing those that were supported by UMIs with singleton reads (collapsed read flag). An overview of the number of variants at each step is shown in Figure 4.10.

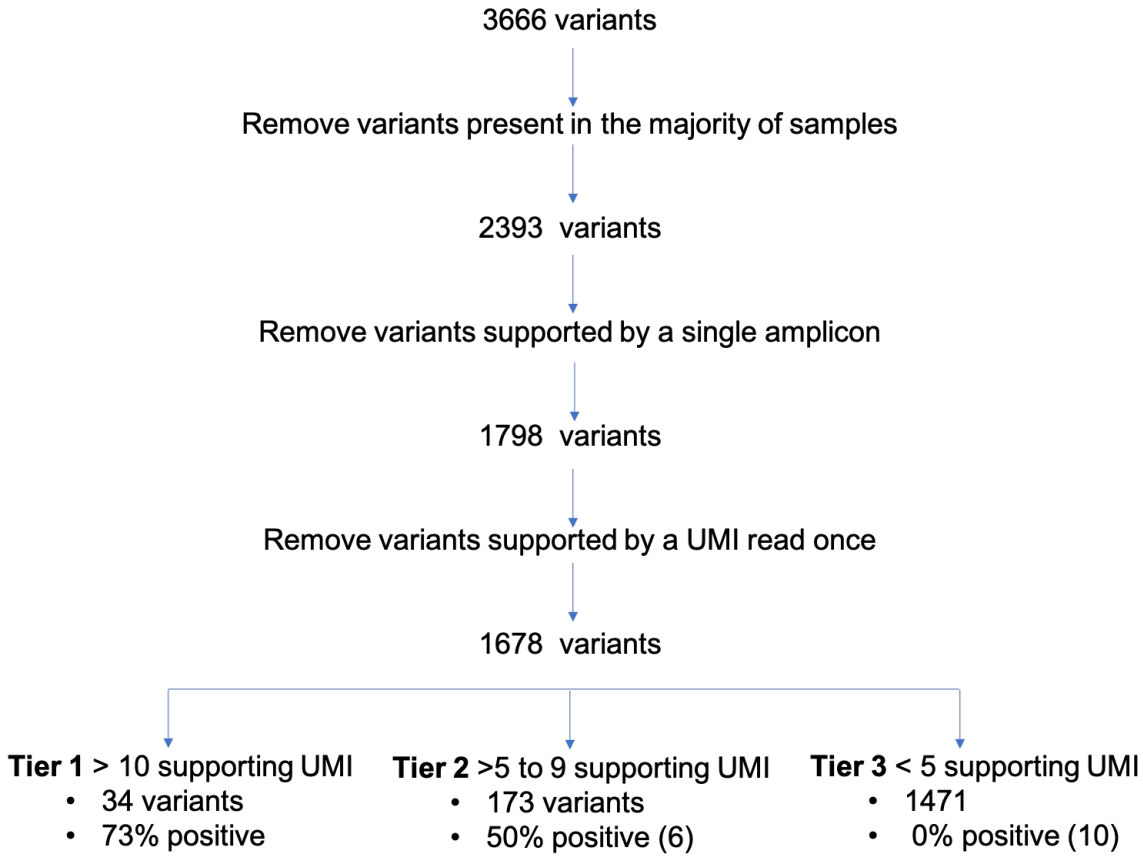


Figure 4.10: An overview of the number of variants present at each filtration step of the analysis. A tiered approach was taken to prioritise the remaining 1678 variants.

To assess the benefit of using UMI to prioritise variants, the remaining variants were analysed in descending order based on the number of UMI supporting the non-reference call (Figure 4.11). The majority of variants are supported by <5 UMI. The output table and summary created by the annotation script, inspection of alignments using IGV, and selective validation was used to prioritise the remaining variants.

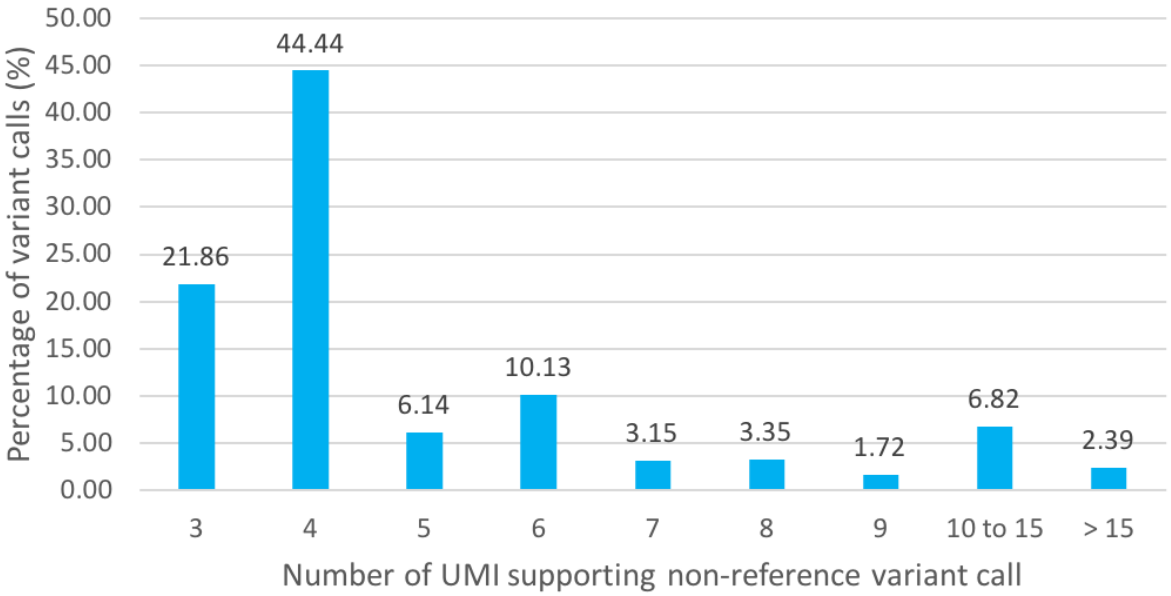


Figure 4.11: Percentage of the called variants (y-axis) that are supported by the number of UMI (x-axis).

4.3.7 Variant validation: Tier 1

34 variants were identified that were supported by at least 10 UMI. These variants include the previously described variants in *FGFR2* in the positive control sample (section 4.3.3) and previously validated variants from chapter 3 (Table 4.3.4). Eight variants were discounted due to low confidence after inspection of alignments. The remaining eight variants were selected for validation by PCR, of which six were positive. The variant validated in *APC* is a technical artefact and is shown in more detail in Figure 4.12. The remaining four changes in genes that previously had not been implicated in SSS (*ADNP, CHD8, NF1, SMAD4*) are shown in Figure 4.13.

Table 4.4: Variants supported by at least 10 UMI

Gene	**Variant and position (hg19)**	**Amino Acid**	**Sample**	**Non-ref UMI**	**Alt %**	**Validation Result**
FGFR2	chr10:123276893 A>T	C342S	14D29	70	2.94	Previously validated
KRAS	chr12:25398284 C>G	G12A	4C24	43	0.43	Previously validated
NRAS	chr1:115258745 C>G	G13R	10F3	34	0.30	Previously validated
FGFR3	chr4:1807370 A>G	N428S	4G18	28	0.31	Previously validated
FGFR2	chr10:123276893 A>T	C342S	14D22	24	0.44	Previously validated
SMAD4	chr18:48604678 A>G	I500M	4F13	21	0.40	Positive PCR
CTNNB1	chr3:41277268 C>T	I579	4G18	19	0.18	Low confidence (alignment)
FGFR2	chr10:123279676 C>G	Spike	WT Spike	18	0.25	Positive control
FGFR3	chr4:1803113 G>A	R155	4C24	18	0.13	Previously validated
NF1	chr17:29554264 G>A	M760I	4D32	17	0.79	Previously validated
FGFR3	chr4:1803578 G>A	R252	4C24	16	0.13	Low confidence (alignment)
NF1	chr17:29554264 G>A	M760I	4D18	15	0.30	Previously validated
APC	chr5:112175772 A>G	E1494G	3D8	15	0.20	Positive PCR
FGFR3	chr4:1808029 C>G	R557G	4C3	14	0.44	Previously validated
FGFR2	chr10:123279674 G>C	P253R	WT Spike	14	0.13	Positive control
ADNP	chr20:49509094 G>T	Y719*	24T05	13	1.03	Positive PCR
PTPN11	chr12:112915523 A>G	N308D	4D18	13	0.16	Previously validated
CHD8	chr14:21866006 T>C	D1397G	4D18	13	0.18	Positive PCR
FMR1	chrX:147026518 G>A	V423M	4D18	12	0.30	Negative PCR
PTPN11	chr12:112856920 C>T	T2I	19T24	12	0.22	Positive PCR
NF1	chr17:29654664 C>T	Q1806*	4G18	12	0.27	Positive PCR
FGFR3	chr4:1808988 G>C	*807S	4C3	12	0.26	Previously validated
RET	chr10:43614996 G>A	V804M	4F13	11	0.28	Previously validated
FGFR2	chr10:123258034 A>C	N432K	4D18	11	0.18	Previously validated
SYNGAP1	chr6:33412384 G>A	R1191Q	18H15	11	0.16	na Low confidence (alignment)
DYRK1A	chr21:38884471 A>G	Q634	24T01	11	0.15	na Low confidence (alignment)
CHD8	chr14:21896124 C>T	R223H	4C24	11	0.11	na Low confidence (alignment)
FGFR2	chr10:123279676 C>G	Spike	Spike	11	0.10	Positive control
RET	chr10:43614996 G>A	V804M	4D18	10	0.23	Previously validated
CHD8	chr14:21876562 G>T	T601K	4F13	10	0.19	na Low confidence (alignment)
FGFR3	chr4:1806099 A>G	Y373C	4G18	10	0.18	Previously validated

KRAS	chr12:25380332 C>T	K42	24T01	10	0.17	na Low confidence (alignment)
MECP2	chrX:153297818 C>T	A73T	10F3	10	0.14	na Low confidence (alignment)
SYNGAP1	chr6:33406025 C>T	A448V	4D18	10	0.11	Negative PCR

Figure 4.12 *(following page)*: (A) IGV screenshot showing an A>G variant in *APC* at chr5:112175772. Top panel shows the mutation in Tes3D8 from HaloPlex HS and the bottom panel shows the mutation in the same sample from the study described in chapter 3. Note that the variant 'G' is present on reads with both alleles of the rs41115 SNP (MAF 0.33). (B) Plot showing positive validation of the variant in Tes3D8. (C) IGV screenshot showing the same region of *APC* in Tes4C21 where the p.F1491V mutation was called and validated. Note that Tes4 is homozygous for the adjacent SNP while Tes3 is heterozygous. (D) Plot showing percentage levels of A>G at chr5:112175772 in all 288 testicular samples from chapter 3. The variant appears elevated in samples from Tes3 only, which is heterozygous for an adjacent SNP.

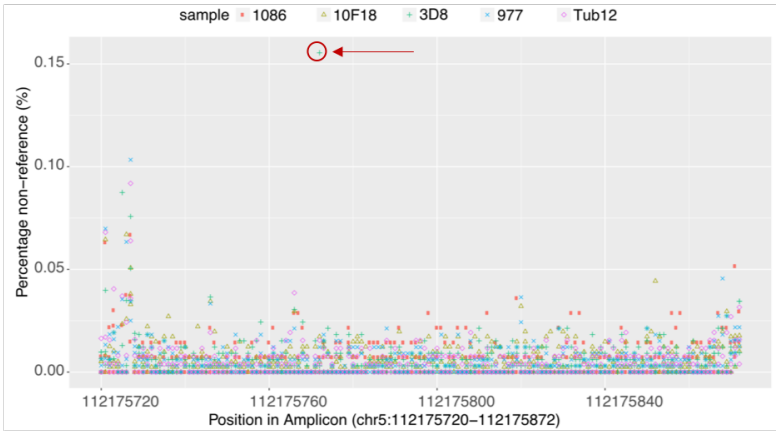
A

3D8 (HaloPlex)

3D8 (RainDance)

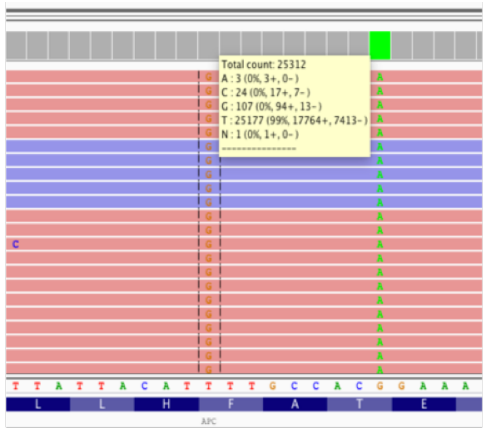


B

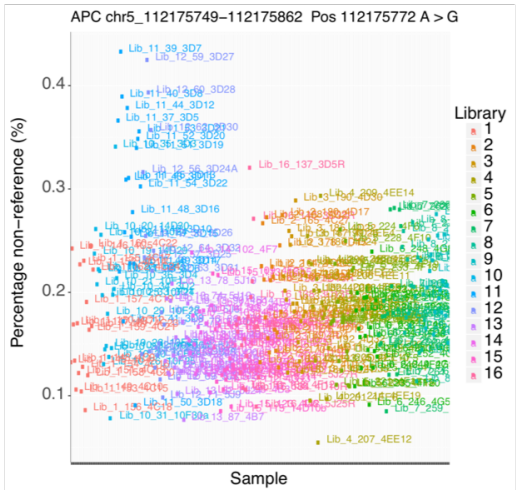


C

4C21 (RainDance)



D



4.3.8 Variant validation: Tier 2

173 variants were identified that were supported by between 5 and 9 UMI. Three previously validated variants were identified in this tier. The remaining variants were inspected using the annotation tables and IGV. Three variants were selected for further validation, all of which were negative (Table 4.5).

Table 4.5: Tier 2 Validation Results

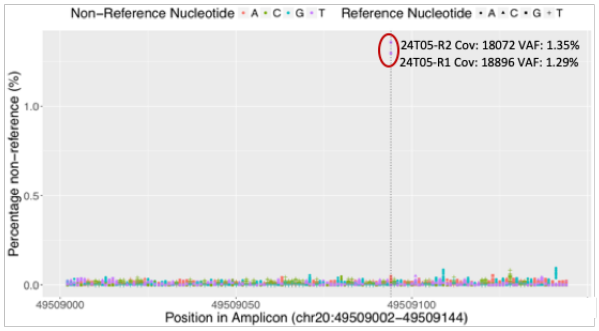
Gene	**Variant and position (hg19)**	**Amino Acid**	**Sample**	**Non-ref UMI/ Ref UMI**	**Validation Result**	**Validation Level (%)**	**Validation Coverage Depth**
PTPN11	chr12:112924336 G>A	V428M	4E6	8, 2336 (0.34%)	Previously validated	2336	12622;31706
PTPN11	12:112910785 G>A	R265Q	4C24	8, 5847 (0.13%)	Previously validated	0.27	10487
SYNGAP1	chr6:33406335 C>T	A509V	4D18	9, 8387 (0.1%)	Negative	0, 0.006	27514, 31388
NF1	chr17:29554264 G>A	M760I	4G18	6, 3096 (0.19%)	Previously validated	0.34	35376
CHD8	chr14:21899652 G>T	Q51K	19T14	6, 5997 (0.1%)	Negative	0.01, 0.01	8408, 6041
CHD8	chr14:21866038 G>T	D1386E	24T05	6, 2521 (0.24%)	Negative	0, 0.02	13281, 11986

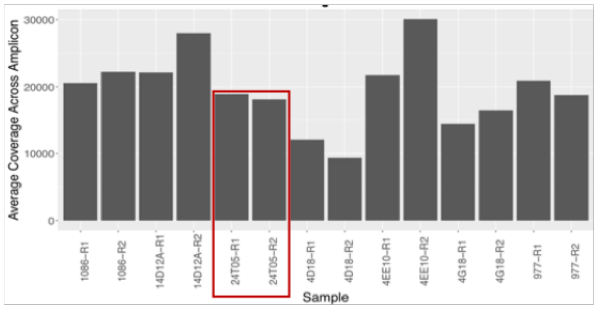
4.3.9 Variant validation: Tier 3

The remaining variants were supported by less than 5 UMI. A subset of variants were visually inspected and selected for validation based on the predicted functional effect of the amino acid substitution and overlap with genes of biological interest. All ten of the selected variants were negative by PCR (Table 4.6). The only previously validated variants in this category were variants that had previously been identified in multiple biopsies: NF1 p.M760I (section 3.7) and FGFR2 p.C342S (Table 3.2 and section 6.4.6).

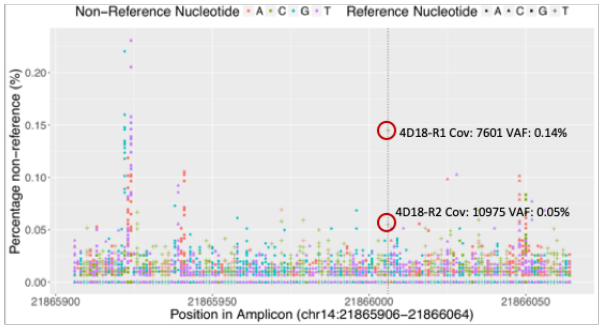
Figure 4.13 *(following page)*: Validated variants in *ADNP, CHD8, NF1* and *SMAD4*. In each case, the left panel shows the non-reference percentage levels across the target amplicon and the right panel shows the coverage in the two replicates of the amplified DNA samples. The putatively positive samples are outlined in red.(A) G>T variant validated in *ADNP* at chr20:49509094 in sample 24T05. (B) T>C variant validated in *CHD8* at chr14:21866006 in 4D18. (C) C>T variant validated in *NF1* at chr17:29654664 in sample 4G18. (D) A>G variant validated in *SMAD4* at chr18:48604678 in sample 4F13.

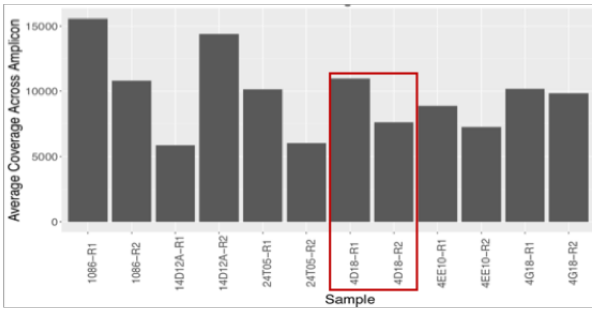
A. *ADNP* chr20:49509094 G>T p.Y719*



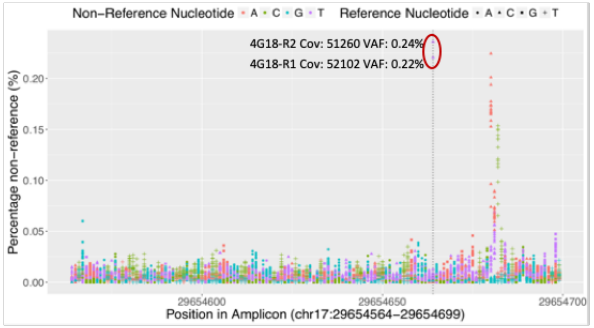


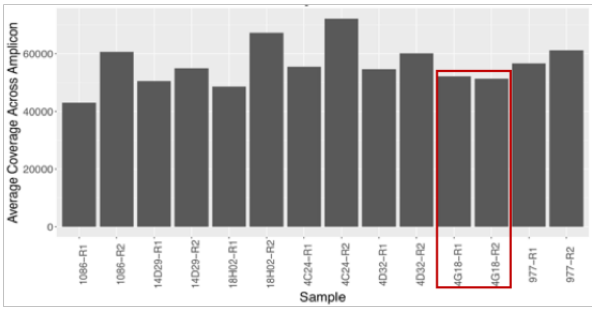
B. *CHD8* chr14:21866006 T>C p.D1397G



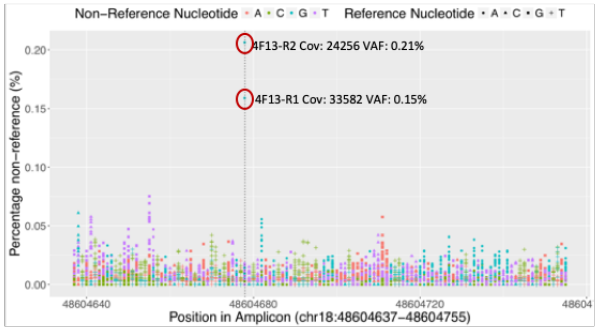


C. *NF1* chr17:29654664 C>T p.Q1785*





D. *SMAD4* chr18:48604678 A>G p.I500M



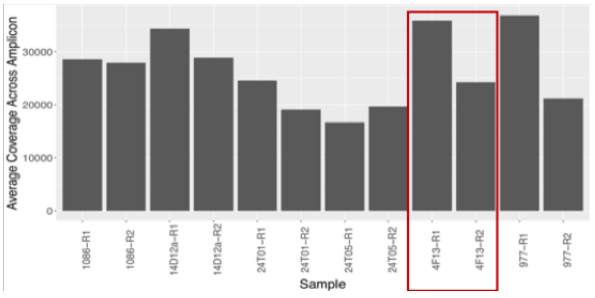


Table 4.6: Tier 3 Validation Results

Gene	**Variant and position (hg19)**	**Amino Acid**	**Sample**	**Non-ref UMI/ Ref UMI**	**Validation Result**	**Validation Level**	**Validation Coverage**
ADNP	chr20:49510129 G>T	N374K	24T07	3, 1810 (0.17%)	Negative	0, 0.004	45997, 49746
CHD8	chr14:21897482 G>A	R286C	24T07	3, 2601 (0.12%)	Negative	0.02, 0.004	14924, 20079
CHD8	chr14:21867863 C>A	E1328*	19T12	4, 3648 (0.11%)	Negative	0, 0.01	10111, 18293
CTNNB1	chr3:41268761 C>A	Y333*	4G18	4, 2258 (0.18%)	Negative	0, 0	49337, 60607
DYRK1A	chr21:38858865 C>T	R196*	14D01	3, 1563 (0.19%)	Negative	0.02, 0.03	96770, 76856
EZH2	chr7:148508803 G>T	L565M	24T05	3, 1823 (0.16%)	Negative	0.005, 0.005	39308, 38813
MECP2	chrX:153296525 C>T	G159S	18H17	4, 3624 (0.11%)	Negative	0.01, 0.01	67696, 35492
RET	chr10:43600523 G>T	R250L	19R02	3, 2720 (0.11%)	Negative	0.005, 0.005	54874, 34052
SYNGAP1	chr6:33409077 G>T	G681C	19T14	4, 3810 (0.1%)	Negative	0, 0	2657, 6076
UBE3A	chr15:25615740 G>T	D507E	4F13	3, 2034 (0.15%)	Negative	0, 0	39021, 24827

Therefore, no further investigation of variants in this tier was performed.

4.3.10 Validated variants

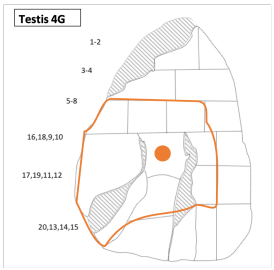
Four variants in genes previously not implicated in SSS were validated. Notably, three of these variants have been associated with germline disorders (Table 4.7). To assess whether the novel variants validated in this study had expanded beyond the biopsy in which the variant was initially identified, surrounding testicular biopsies were screened by PCR. There was no evidence detected that any of the variants had spread to adjacent biopsies as measured non-reference nucleotide levels were not above background (Figure

4.14). This was not possible for the ADNP p.Y719* mutation which was identified in a single seminiferous tubule from Tes24.

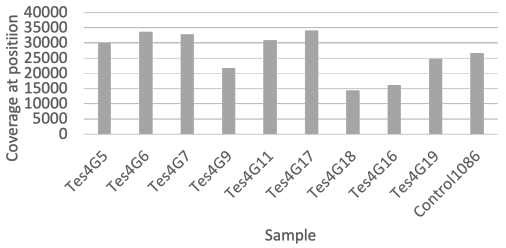
Table 4.7: Novel variants validated. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical DNA substitutions.

Gene	**Variant**	**Amino Acid**	**CADD score**	**COSMIC v86**	**Germline disorder**
ADNP	chr20:49509094 G>T	p.Y719*	22.9	-	Helsmoortel-Van der Aa syndrome
CHD8	chr14:21866006 T>C	p.D1397G	23.9	-	-
NF1	chr17:29654664 C>T	p.Q1806*	41	2	Neurofibromatosis 1
SMAD4	chr18:48604678 A>G	p.I500M	20.1	-	Myhre syndrome

A. *NF1* chr17:29654664 C>T p.Q1806*



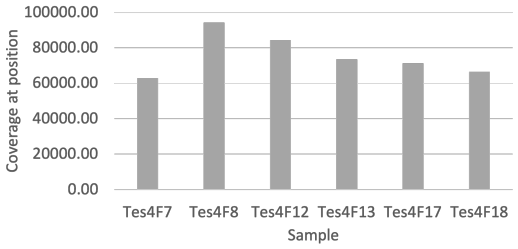
Biopsy	Percentage non-reference level (%)
4G5	0.01
4G6	0.00
4G7	0.01
4G17	0.00
4G19	**0.40**
4G11	0.01
4G20	0.00
4G14	0.00
4G15	0.01



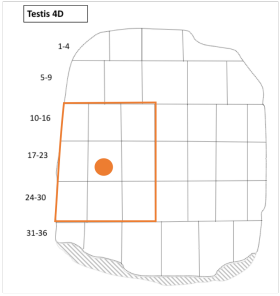
B. *SMAD4* chr18:48604678 A>G p.I500M



Biopsy	Percentage non-reference level (%)
4F7	0.00
4F8	0.01
4F12	0.00
4F13	**0.16**
4F17	0.01
4F18	0.01



C. CHD8 chr14:21866006 T>C p.D1397G



Biopsy	Percentage non-reference level (%)
4D10	0.01
4D11	0.02
4D12	0.02
4D17	0.02
4D18	**0.11**
4D19	0.03
4D24	0.02
4D25	0.01
4D26	0.01

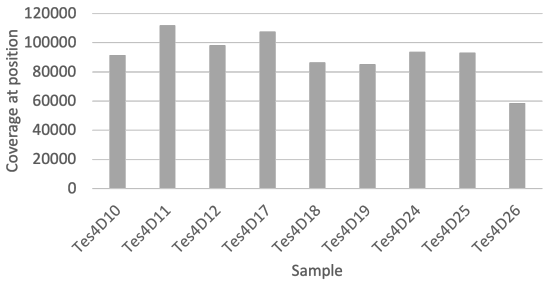


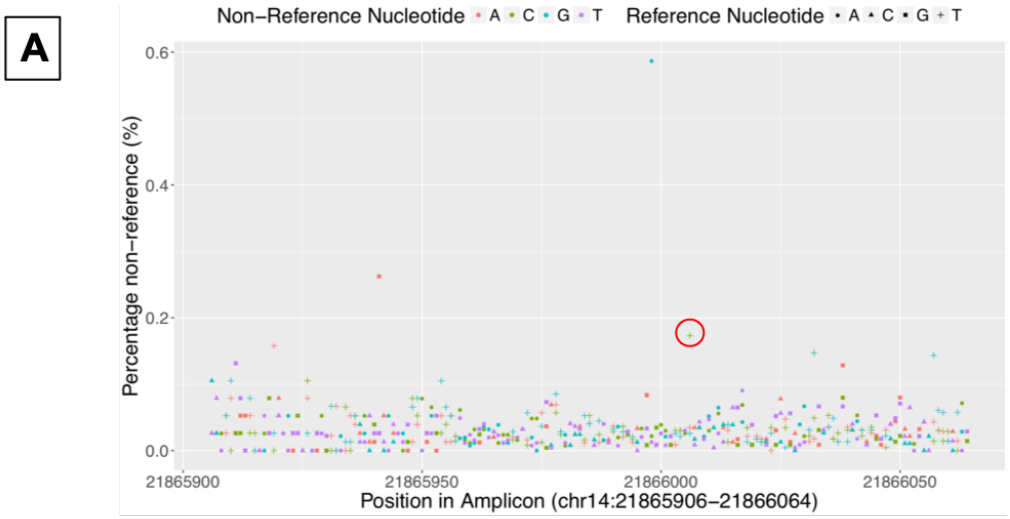
Figure 4.14: Novel variants that were positive in testicular biopsies were screened in surrounding positive pieces. The left panel shows the testicular slices with the biopsies that were rescreened by PCR outlined in orange. The orange dot indicates the original biopsy in which the variant was identified. The middle panel shows a table of the non-reference nucleotide levels measured in each biopsy, with the level in the known positive biopsy shown in bold. The right panel shows the sequencing coverage obtained at the position of interest in each rescreened biopsy. No evidence of the variant spreading to adjacent biopsies was observed.

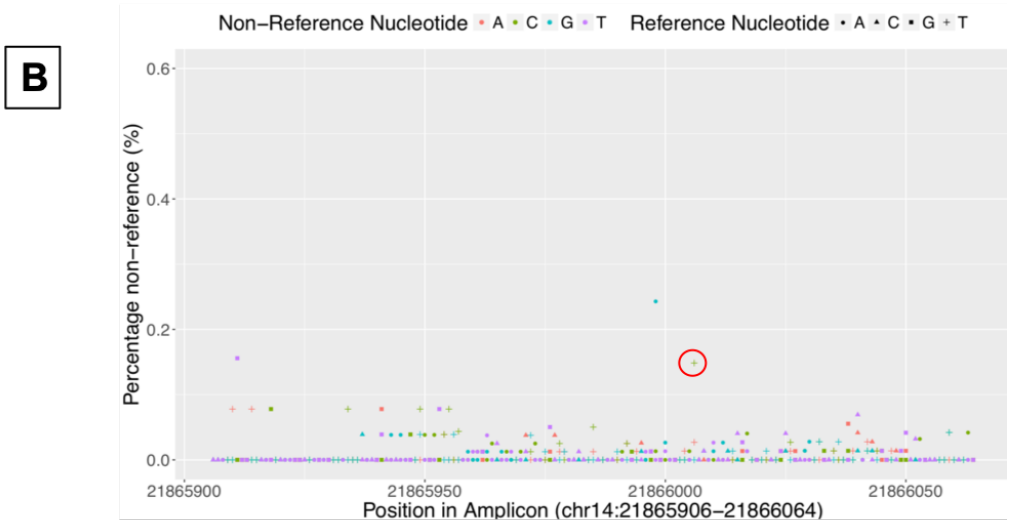
4.4 Benefit of UMI technology

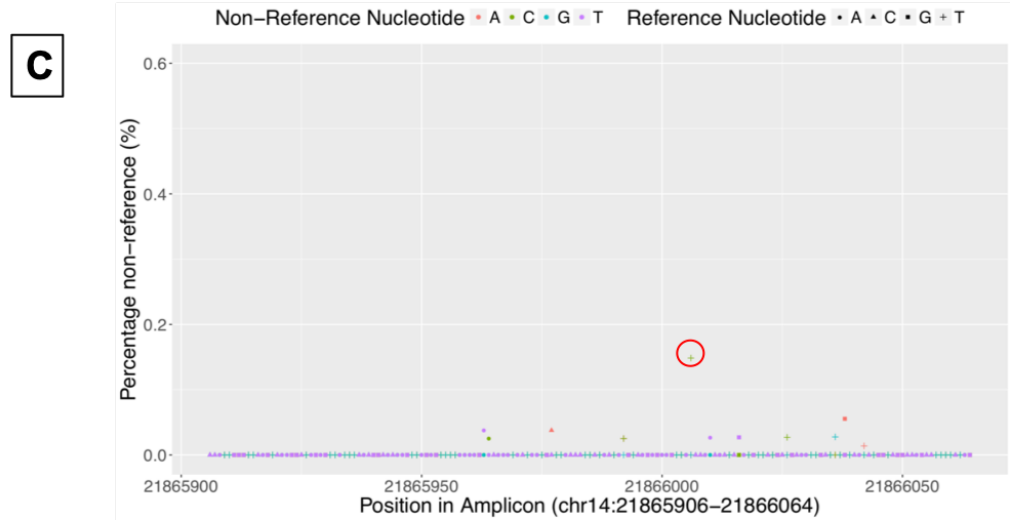
To determine whether the application of UMI technology aided in the calling and validation of low-level variants, the sequencing fastQ files were re-analysed using SureCall to generate non-deduplicated aligned files. Non-reference counts across amplicons in which known positive variants were identified were then extracted and examined to determine if the variant call could have been made without using UMI de-duplication. For exam-

ple, the lowest level variant (0.18%) called by Surecall and independently validated by PCR (average of two replicates = 0.1%) was in *CHD8* (p.D1397G) (Figure 4.15). While this variant was still called by Surecall even without the use of UMIs, there are many additional variants called (62 variants in the UMI-deduplicated alignments versus 383 variants in the non-deduplicated alignments for Tes4D18). The highest variant call that would be identified in Figure 4.15 (A and B) is present at the end of a single amplicon and is therefore a likely technical artefact. Combining UMI information with overlapping amplicons reduces the background in a target amplicon.

Figure 4.15 *(following page)*: Plots showing the comparison of a target region in *CHD8* (chr14:21865906 to 21866064) with and without the use of UMI information. The validated variant in *CHD8* at chr14:21866006 T>C (p.D1397G) is outlined in red. (A) Plot showing the percentage non-reference in target amplicon based on raw read counts (B) Plot showing the percentage non-reference in the amplicon based on UMI counts (C) Plot showing the percentage non-reference in example amplicon based on UMI present in more than one overlapping amplicon.


A
Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
21865900
21865950
21866000
21866050
Position in Amplicon (chr14:21865906−21866064)


B
Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
21865900
21865950
21866000
21866050
Position in Amplicon (chr14:21865906−21866064)


C
Non−Reference Nucleotide A C G T
Reference Nucleotide A C G T
Percentage non−reference (%)
0.6
0.4
0.2
0.0
21865900
21865950
21866000
21866050
Position in Amplicon (chr14:21865906−21866064)

4.5 Discussion

4.5.1 Validated variants

In total, 19 variants in ten genes were validated (Table 4.8). These include variants in seven genes that encode members of the RAS-MAPK pathway, of which six of the genes have previously been implicated. The previously identified variant in *NF1* (section 3.7) is likely to be a mosaic mutation that arose during early development. Thus, the identification of a nonsense mutation in *NF1* represents the first implication of *NF1* in SSS.

Table 4.8: Summary of all validated variants. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical DNA substitutions.

Gene	**Variant and position (hg19)**	**Amino Acid**	**Number of positive samples**	**VAF (%)**	**COSMIC(v86)**	**Germline Disorder**
ADNP	chr20:49509094 G>T	Y719*	1	1.04	-	Helsmoortel-Van der Aa syndrome
CHD8	chr14:21866006 T>C	D1397G	1	0.15	-	-
FGFR2	chr10:123258034 A>C	N549K	1	0.18	34	-
FGFR2	chr10:123276893 A>T	C342S	2	0.45 - 3.02	-	Crouzon syndrome
FGFR2	chr10:123276898 T>C	Y340C	1	0.13	-	Pfeiffer syndrome
FGFR3	chr4:1806099 A>G	Y373C	1	0.18	697	Thanatophoric dysplasia
FGFR3	chr4:1806119 G>A	G380R	3	0.1 - 0.26	33	Achondroplasia
FGFR3	chr4:1807188 G>A	V505M	1	0.12	-	-
FGFR3	chr4:1807370 A>G	N540S	1	0.31	1	Hypochondroplasia
FGFR3	chr4:1808988 G>C	*807S	1	0.22	-	Thanatophoric dysplasia
KRAS	chr12:25398284 C>G	G12A	1	0.44	2256	-
NF1	chr17:29554264 G>A	M760I	2	0.3 - 0.8	-	-
NF1	chr17:29654664 C>T	Q1806*	1	0.25	2	Neurofibromatosis 1
NRAS	chr1:115258745 C>G	G13R	1	0.30	180	-
PTPN11	chr12:112856920 C>T	T2I	1	0.22	-	Noonan syndrome
PTPN11	chr12:112910785 G>A	R265Q	1	0.14	4	Noonan syndrome
PTPN11	chr12:112915523 A>G	N308D	1	0.16	4	Noonan syndrome
RET	chr10:43614996 G>A	V804M	1	0.23	4	FMTC
SMAD4	chr18:48604678 A>G	I500M	1	0.40	-	Myhre syndrome

Five variants in genes not previously implicated in SSS were identified. A variant in *APC* was identified in Tes3D8, nearby a variant (APC p.F1491V) that was validated in chapter 3. Further inspection of this variant found that the variant allele was inconsistently present with the two alleles of an adjacent heterozygous SNP. Examining the levels of non-reference allele at this variant position found that the variant was elevated above background in all samples of the same individual (Figure 4.12). This suggests that

the variant may be a technical artefact.

Mutations in *CHD8*, a chromatin remodelling factor, have been identified in a cohort of patients with ASD. *CHD8* mutation is considered to define a subtype of autism, with patients exhibiting common features such as macrocephaly and gastrointestinal problems (Bernier et al., 2014; Barnard et al., 2015). The majority of mutations reported in this gene in ASD patients are frameshift and truncating mutations (Bernier et al., 2014; Barnard et al., 2015). CHD8 expression is detected in testicular tissue (The Human Protein Atlas, www.proteinatlas.org, Uhlén et al. (2015)). The variant validated in this study is a missense substitution and is predicted to be benign (Polyphen = 0.358; Adzhubei et al. (2015)). There are no reports of this mutation in the literature. Therefore, there is little evidence at present to link *CHD8* mutation with SSS.

Mutations in *NF1* are associated with the overgrowth condition neurofibromatosis (see panel V of figure K in Figure 3.7) for the distribution of mutations throughout the protein). NF1 functions as a regulator in the previously described RAS-MAPK pathway (Figure 1.4). This variant (p.Q1806*) has been reported in a patient with type 1 neurofibromatosis (Fahsold et al., 2000).

Mutations in *ADNP* are among the most commonly identified in ASD (Arnett et al., 2018). In particular, truncating mutations have been shown to cluster in exon 5 (Helsmoortel et al., 2014). Patients with *ADNP* mutation share a number of clinical features including intellectual disability and facial dysmorphism. ADNP p.Y719* mutation was first reported in two unrelated patients with ASD and other shared clinical features by Helsmoortel et al. (2014). Subsequently, the mutation has been identified in additional patients and currently has been reported as the causative mutation in almost a quarter of patients (17/80) with ASD as a result of *ADNP* mutation (Pescosolido et al., 2014; Gozes et al., 2017; Takenouchi et al., 2017). As well as identifying recurrent mutations resulting in p.Y719*, Helsmoortel et al. (2014) identified other nearby mutations in *ADNP* that may result from a common underlying mechanism involving a DNA-repair defect following replication fork pausing at hairpin structures. *ADNP* encodes a zinc finger protein with transcription factor activity. ADNP expression is detected in testicular tissue (The

Human Protein Atlas, www.proteinatlas.org, Uhlén et al. (2015))It has been shown to be essential for brain formation (Mandel and Gozes, 2007) and also has a proposed role in cell survival maintenance (Zamostiano et al., 2001). ADNP has recently been shown to form a complex with the chromatin related proteins CHD4 and HP1 (called the 'ChAHP' complex). Ablation of this complex in mice stem cells resulted in loss of proper cell fate decision and spontaneous differentiation (Ostapcuk et al., 2018). The authors of this studied showed that specific mutations identified in patients with ADNP mutation, including p.Y719*, disrupt the formation of the ChAHP complex.

Finally, a variant was identified in *SMAD4* resulting in the amino acid substitution p.I500M. This mutation has been identified in individuals with Myhre syndrome, a developmental disorder characterized by intellectual disability, short stature and facial dysmorphism. Three distinct mutations affecting the I500 codon (I500M, I500T, I500V) have been reported in individuals with Myhre syndrome (Le Goff et al., 2011). *SMAD4* mutations will be discussed further in section 6.5.

4.5.2 Variant detection in seminiferous tubules

Seminiferous tubules were included in this study in order to test the hypothesis that SSS associated mutations could be more readily distinguished. No mutations above 3% were validated in the seminiferous tubules. Two known pathogenic mutations, ADNP p.Y719* (Helsmoortel et al., 2014) and PTPN11 p.T2I (Sarkozy et al., 2003) were validated in the seminiferous tubules at 1.02% and 0.22% respectively. As well as distinct seminiferous tubules from Tes19, a single biopsy from the same testis and seven biopsies from the second testis of the same individual were also screened. No variants were identified in these biopsies, perhaps due to the comparatively young age of the individual (58 years). The chance of identifying a mutation in a single seminiferous tubule is lower than in a biopsy, which contains cross sections of multiple tubules, and therefore screening more seminiferous tubules is required in order to determine whether they represent a beneficial strategy. The cost for a 48 sample kit (Agilent) to assess a 1-500 kb target region is £7,260.00 (£151.25 per sample). Using a commercial solution such as this limits the

number of samples that can be processed in a cost-effective manner.

4.5.3 Technology Performance

Identification and validation of variants was comparatively straightforward at the higher end of the UMI spectrum. A high rate of validation (24 of 26 variants, 92.3%) was achieved for variants supported by more than 10 UMI when combined with manual inspection to generate high-confidence calls (VAF: 0.18 to 2.94%). The number of supporting UMI for a variant is a useful predictor, especially when combined with other supporting information such as the number of overlapping amplicons and reduces the background of mutation across an amplicon (Figure 4.15). However, a poor rate of validation was achieved for variants supported by <7 UMI, which accounted for 82.57% of variants. The discriminatory power of UMI at the lower end of the spectrum is poor.

A significant strength of the technology is the performance of high G/C content genes, such as *FGFR3*, which was previously poorly assessed due to poor coverage (Figure 3.7). In this study, 36.84% (7 of 19) of the variants validated were in *FGFR3*. The achondroplasia-associated mutations in *FGFR3* (c.1138G>A or c.1138G>C) (p.G380R) are considered to be the most frequently mutated nucleotides in the germline, with a birth prevalence of approximately 1 in 30,000 (Bellus et al., 1995). However, mutations at this site were not detected using the described PCR approach in chapter 3 due to insufficient coverage at these sites (Figure 3.7). The most frequent variant identified using the HaloPlex panel was the FGFR3 p.G380R mutation, identified in 3 independent samples (Table 4.8).

Many aspects of the HaloPlex HS target enrichment system are made simple by the fact it is developed and supported by a commercial company (Agilent). This also has drawbacks, however, as it impacts flexibility. For example, panel design is extremely simple and the only user input required is a list of gene names or coordinates. However, for some regions the program either fails to generate high coverage probes (Table 2.2) or the probes fail to perform (as was the case for several exons, including (*UBE3A* exon 14, *MECP2* exon 14 and *PTEN* exon 1)). As the panel is fixed, there is no opportunity to

replace or redesign probes. Therefore, I will next investigate a technology that retains the benefit of using UMI to call low-level variants but provides more flexibility in panel design - single molecule Molecular Inversion Probes (chapters 5 and 6). smMIPs also have the advantage of reducing the processing cost per sample, a serious limitation of the use of the commercial panel.

Chapter 5

Assessment of single molecule Molecular Inversion Probes for low level variant detection

5.1 Introduction

smMIPs are single stranded DNA oligonucleotides designed to circularise upon hybridisation to a genomic target region (Figure 5.1). Two features of their design warrant investigation of the suitability for low-level variant detection. Firstly, the common backbone sequence facilitates high-throughput multiplexing of target probes, making it possible to assess many genomic sites in a single capture experiment. Secondly, the two hybridisation arms are each flanked by a 5 bp degenerate sequence ("Unique Molecular Identifier, UMI"), which theoretically provides the ability to tag more than 4^{10} (1,048,576) different molecules. Following sequencing, the information provided by the UMI can be used to group reads originating from the same molecule and generate a consensus sequence.

In order to assess the suitability of smMIPs as a strategy to detect low-level spontaneous mutations in the human testis and understand their performance, a proof of principle study was designed.

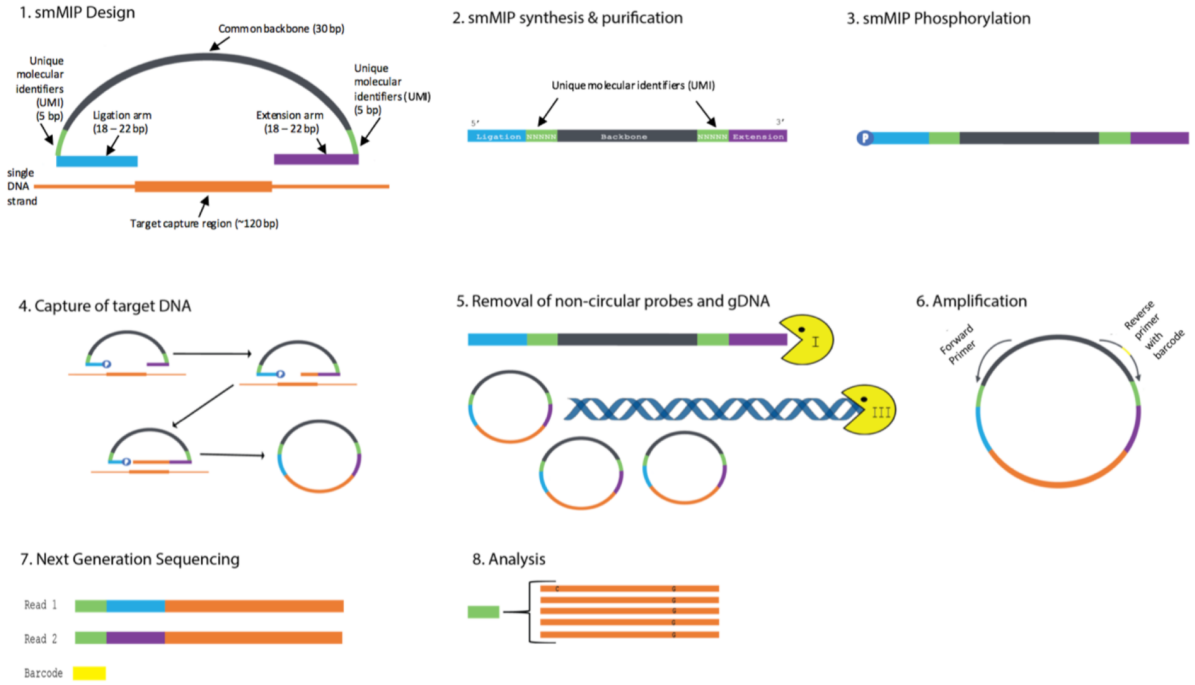


Figure 5.1: Figure 1: Schematic view of smMIP method. (1) Single molecule molecular inversion probes consist of two targeting arms (extension, shown in purple and a ligation arm, shown in blue) joined by a common backbone sequence (black). The arms are both flanked by a 5 bp UMI. Probes are designed such that the arms are complementary to sequences flanking individual target regions. (2) smMIP synthesis and purification. (3) smMIPs are phosphorylated at the 5′ end to facilitate ligation and subsequent circularisation (4) Probes are pooled, hybridised to genomic DNA and polymerase and ligase are added to gap fill and ligate the probe into a single-stranded circle. (5) Non-circular probes and gDNA are removed by exonuclease treatment (6) Captured and circularised smMIPs are amplified from the common backbone sequence using a reverse primer which adds a sample specific barcode and a forward primer which adds sequences specific to the sequencing platform to be used. (7) Next generation sequencing is used to collect three reads, including the barcode sequence that is used to demultiplex samples. In Illumina sequencing, the molecular tags are the first five bases of the read 1 and 2. (8) UMI groups can be used to facilitate consensus calling.

5.2 Effect of gDNA input and probe ratio on smMIP performance

To assess whether detection of low-level mutation, in the range 1% to 0.1%, is achievable and reproducible using smMIPs, a proof of principle study was designed. As a first step in understanding the performance of smMIPs, two probes were manually designed (Professor Anne Goriely) targeting the site of causative mutations of Apert Syndrome in *FGFR2* (chr10:123279674-677).

To assess the method and begin to optimise the protocol, the effect of different genomic DNA input amounts and the number of smMIP copies (per haploid genome) on capture efficiency was investigated. Wild type genomic DNA was used with smMIPs added in amounts corresponding to either 800 or 4000 copies per haploid genome. Reads were demultiplexed and aligned to the reference genome. Using the known sequence of the smMIP arms, a custom script was developed to extract the read pairs belonging to each target smMIP. The UMI, present as the first 5 nucleotides in a sequencing read, was also extracted. The script was used to yield the number of read pairs and UMI for a given probe. Between 111,549 and 2,155,249 (mean: 776,668) total read pairs for both probes were generated.

As a measure of efficiency, the number of UMIs (corresponding to the number of independent molecules captured) per 1000 reads was calculated, in order to account for differences in sequencing read depth between samples (Figure 5.2). Little difference was observed between the two DNA samples, suggesting that performance of probes is reproducible between samples. Additionally, there was little difference between the two probe concentrations suggesting that saturating the system with additional probes offers little advantage to capture more molecules. As expected, an increasing number of UMIs captured as the amount of DNA input increased was observed. This effect was modest and did not show the expected proportional increase. However, the average number of UMIs captured per 1000 reads at 1000 ng (37) is more than 1.5 times higher than at 100

ng (22) suggesting that there may be benefit in increasing the amount of DNA input.

Overall, capture efficiency appears to be low. At 100 ng, 33,000 copies of the haploid genome are theoretically available for capture. At 100 ng, between 712 and 11,422 (mean: 3266) unique molecules are captured by both probes. This pattern of low capture efficiency is observed for all gDNA inputs. For a mutation present at 0.1%, only 33 supporting haploid genome copies would be expected. Based on the mean number of UMIs captured here, only 3 haploid genome copies supporting the mutation would be expected. This means that, at this capture efficiency, if a mutation is not detected, it would not be possible to conclusively prove the true absence of the mutation due to Poisson sampling.

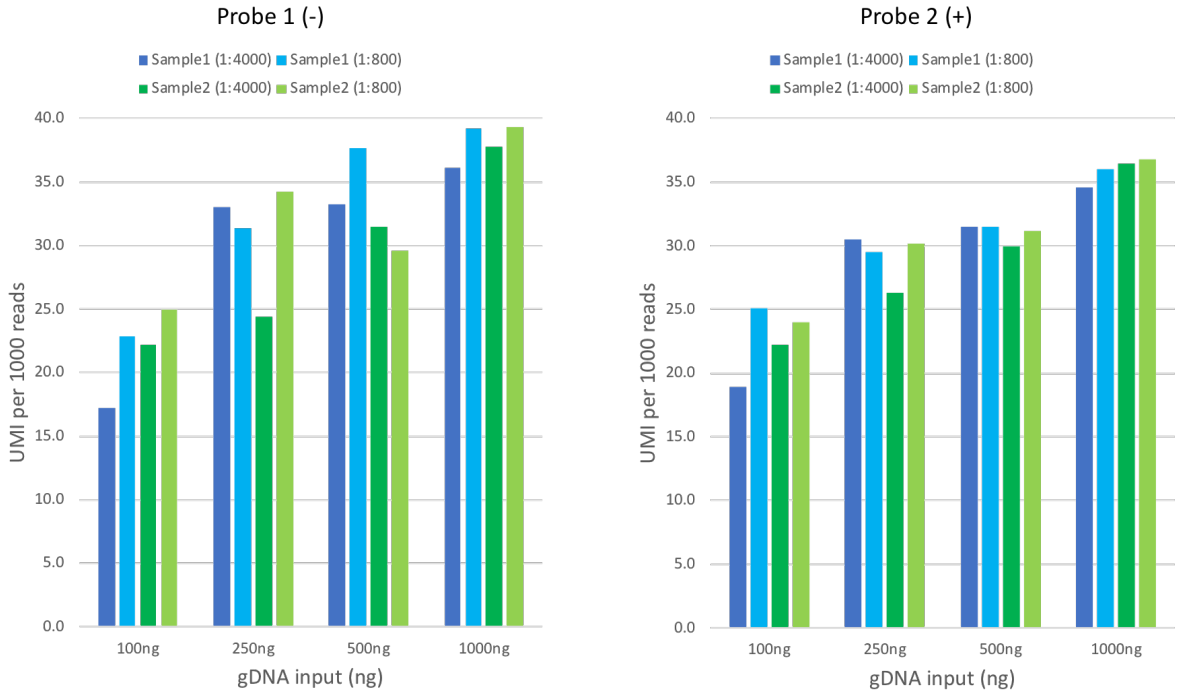


Figure 5.2: Performance of smMIPs targeting *FGFR2* (chr10:123279674-677) with increasing gDNA input and two different smMIP:DNA ratios. The number of unique molecular tags collected per 1000 reads (y-axis) in relation to varying genomic DNA input (x-axis) and two different concentrations (1 in 800 or 1:4000) of smMIP.

5.3 Assessment of low-level mutations in a dilution series

To assess the sensitivity of smMIPs, a dilution series was analysed with DNA from five patients spiked into a wild type background. Each patient had a different mutation affecting c.755-758 of *FGFR2* (Table 5.1). All mutations were analysed at levels from 1% to 0.06% (1%, 0.5%, 0.25% and 0.125% and 0.06%).

Table 5.1: Mutations in *FGFR2* in a dilution series present in a wild type background

Sample	**Mutation**	**Amino Acid Substitution**	**Syndrome**	**Sequence**
WT Background	-	-	-	CGA TCG CCT GCT AGC GGA
Mutation 1	c.758C>G	p.P253R	Apert	CGA TCG CGT GCT AGC GCA
Mutation 2	c.755C>G	p.S252W	Apert	CGA TGG CCT GCT ACC GGA
Mutation 3	c.755C>T	p.S252L	Crouzon/WT	CGA TTG CCT GCT AAC GGA
Mutation 4	c.755_756CG>TC	p.S252F	Apert	CGA TTC CCT GCT AAG GGA
Mutation 5	c.755_757CGC>TCT	p.S252W_P253S	Pfeiffer	CGA TTC TCT GCT AAG AGA

Each sample in the dilution series was captured, barcoded and sequenced in replicate. Between 938,058 and 1,630,773 (mean: 1,287,941.67) total read pairs for both probes were generated. A custom script was developed to automatically extract the sequence of interest (WT or mutant) and the associated UMI.

It was possible to call a consensus for UMI tags supporting wild type and all mutant sequences as low as 0.06% (Figure 5.3). Capture efficiency ranged from 25.87% to 51.99% (mean: 35.46%) (Figure 5.4). However, to achieve this, on average each sample was sequenced to a read depth of more than one million (mean: 1,287,941.67). Additionally, the majority of UMI are sequenced <3 times and are therefore not informative to call a consensus sequence. A UMI sequenced a single time does not not provide any information as to the true sequence of the original capture molecule. Similarly, a UMI sequenced twice and observed once with a particular sequence and once with a different sequence

is not informative. Optimisation of the protocol is required to move from a single, well-understood site with known mutations to a high throughput discovery screen (Figure 5.4).

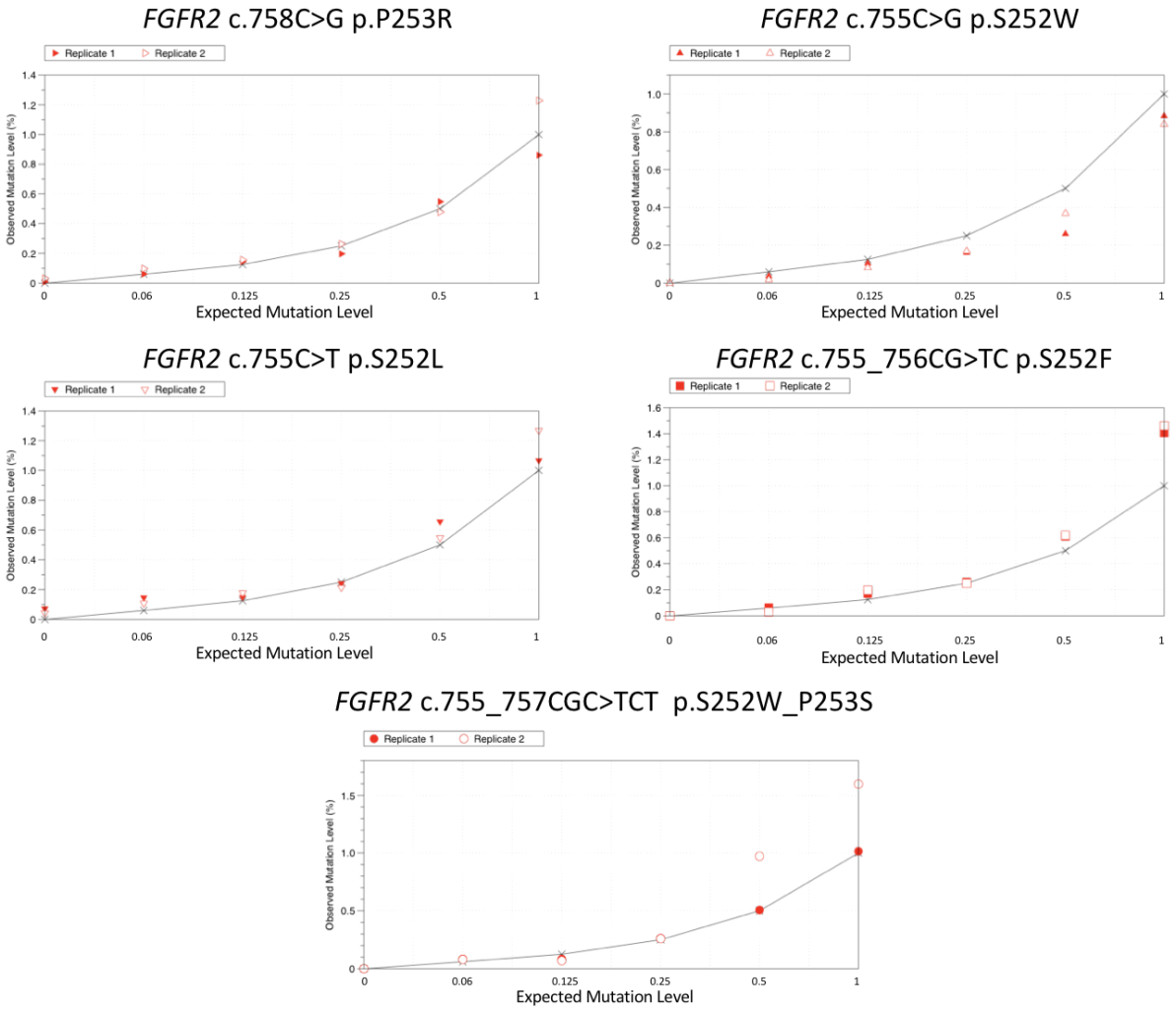


Figure 5.3: Detection of each of the five spiked mutations in two replicates

5.4 Optimisations to smMIP workflow protocol

Two primary steps in the smMIP protocol (Figure 5.1) were identified where optimisations could be made to attempt to improve the efficiency of capture of unique molecules. These are the synthesis of the oligonucleotide itself (section 5.4.1) and during the capture step (section 5.4.2).

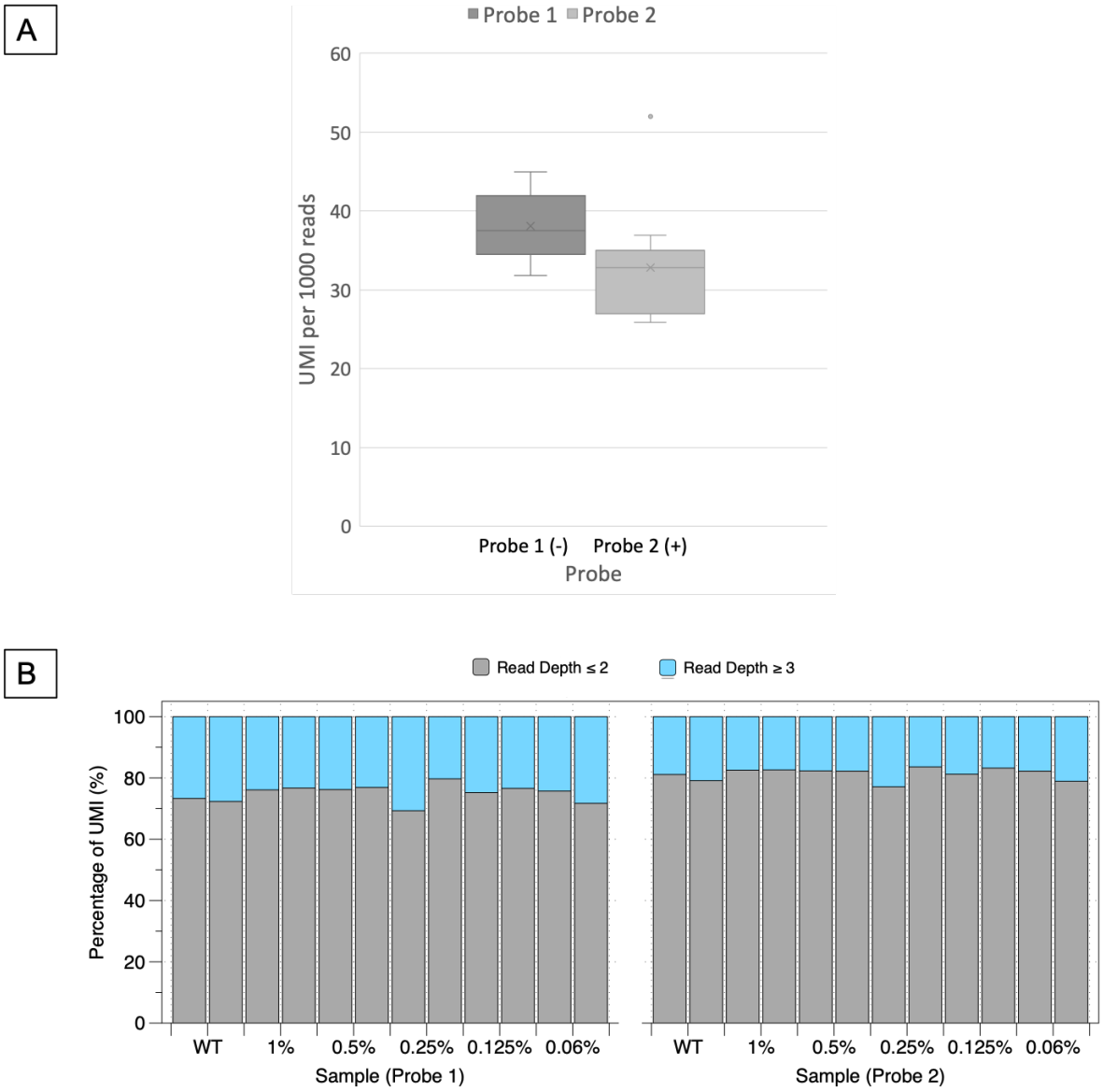


Figure 5.4: Capture efficiency of WT DNA molecules (A) Median capture efficiency for probe 1 and probe 2. (B) The majority of UMI are sequenced <twice and are insufficient to generate a consensus read.

5.4.1 Synthesis quality of the smMIP

smMIPs are long oligos (80 nucleotides) and, as such, their synthesis is challenging (LeProust et al., 2010). It is therefore unclear how many of the smMIPs will be full length. Commercial companies offer different purification solutions which are application-dependent. Depending on the purpose for which the oligonucleotide is to be used, post-production purification may be required as the truncated probes may interfere with the function of the full length probes. This is particularly important in the case of smMIPs,

as 5′ phosphorylation is necessary for ligation and subsequent circularisation (Figure 5.5).



Figure 5.5: Synthesising full length long oligos is challenging. The presence of truncated oligos can interfere with the function of full length probes by occupying target DNA and being unable to undergo ligation and complete circularisation.

Commercial synthesis companies offer two main methods of purification, namely PAGE and HPLC purification. PAGE purification is used to separate the full-length oligos from shorter oligos and has a reported achievable purity of >90% (IDT). HPLC purification results in a higher yield of probes but a lower purity (85%). Although purification may increase the proportion of full length probes, this comes at a significant price. For example, the synthesis cost per base for IDT (Integrated DNA Technologies, Inc) is £0.47 (prices as listed on IDT oligo services webpage in August 2018). This means an 80 bp oligo with standard desalting would cost £37.60. However, the additional step of PAGE purification brings the per probe price up to £93.60.

To investigate whether this investment in smMIP purification yields any benefit, probes were synthesized and purified in collaboration with Dr Afaf El-Sagheer and Professor Tom Brown (Department of Chemistry, University of Oxford). Four probes targeting exon 7 of *FGFR2* were synthesized (Materials and Methods, Table 2.4). Mass spectrometry and capillary electrophoresis was performed by Dr Afaf El-Sagheer to verify molecular weight and oligonucleotide purity were in line with expectations. The smMIP were split into two pools, with one pool being subject to standard desalting and the other undergoing HPLC purification.

The four *FGFR2* probes were used to capture the samples in the previously described dilution series (Table 5.1) alongside the same probes synthesised and received from IDT

with standard desalting. Between 12,370 and 1,330,082 (mean: 293,692) total read pairs for both probes were generated. The performance of HPLC purified probes is similar to the performance of the commercial probes and the standard desalted probes, with no significant difference observed between any probe set (two sided t-tests, $p>0.26$ for all comparisons) (Figure 5.6). The synthesis quality of the probes had little effect on variant detection, with all mutations being detected at all dilution levels and an R^2 value of 0.8194 (Figure 5.7).

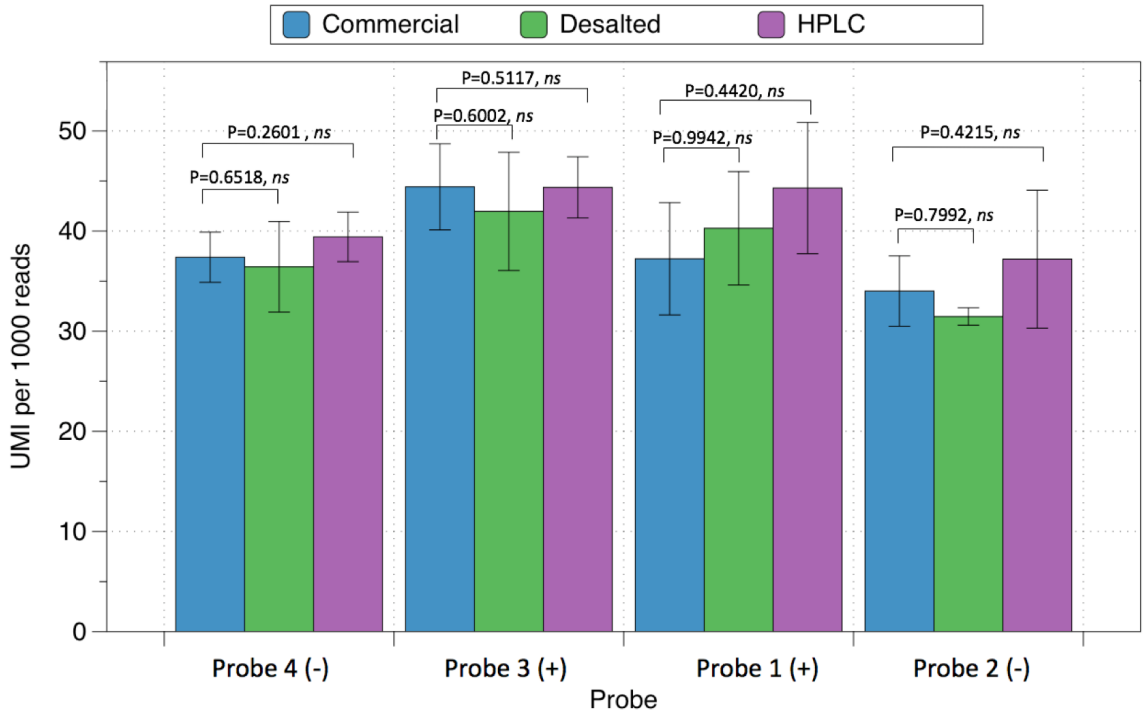


Figure 5.6: The performance of oligonucleotides ordered from IDT (blue), standard desalted (green) and HPLC purified (purple) was similar in all cases. The number of UMIs per 1000 reads was calculated and the average of 6 samples per probe is shown. No significant difference is detected between any probe set (two sided t-tests). Error bars show standard deviation.

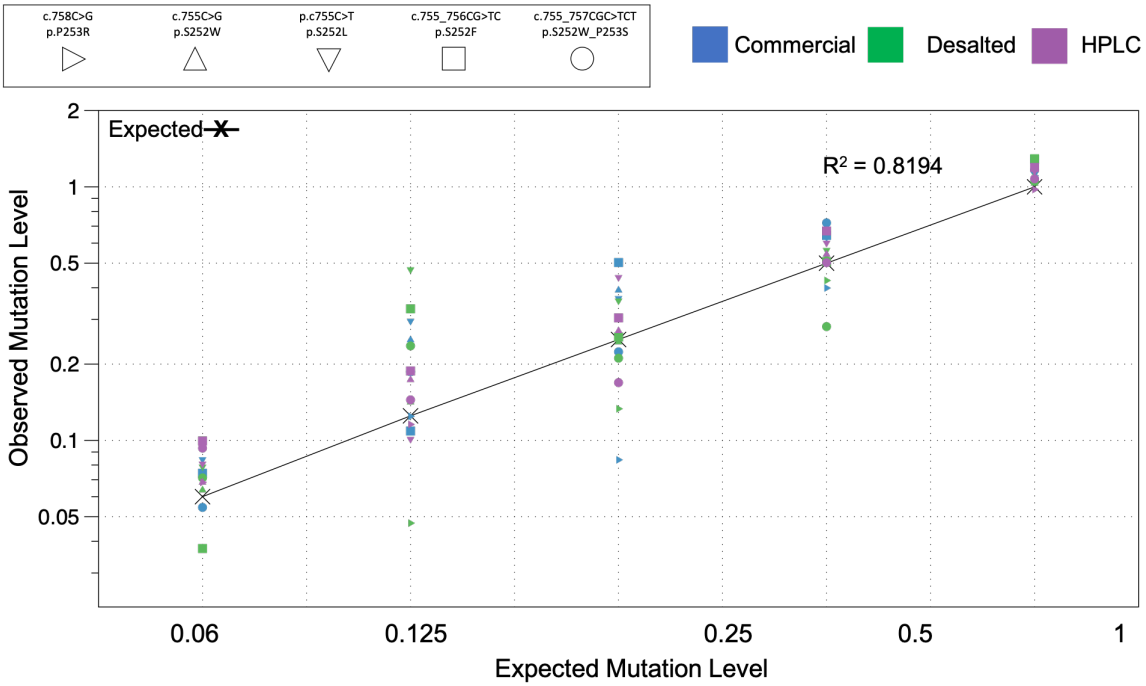


Figure 5.7: The five *FGFR2* mutations in the dilution series were detected by all probe sets. Black line shows expected mutation level in each sample.

5.4.2 Multiple barcoding events for capture

The captured circularised molecules are amplified and barcoded in a single reaction well. Bias in PCR amplification may result in an under representation of the number of unique captured molecules observed at the time of sequencing. To investigate the effect of splitting the capture reaction into multiple wells, each of the four probes was used to capture the previously described dilution series (Table 5.1). The captured product was then split into 4, 8 or 12 wells with a unique barcode for each set.

Each sample was sequenced to an average read depth of 339344.08 (range: 36905 to 774018). The number of unique UMIs was extracted per sample and used to compare the number of capture events at each split (Figure 5.8). Increasing the number of barcoding events reduces the saturation of unique molecules at the time of capture. There is a statistically significant increase in the yield of UMI between each treatment (4, 8 or 12) (two sided t-tests, $p<0.0001$ for all comparisons) (Figure 5.6). Splitting the capture into 12 wells for barcoding gives the most benefit, with an average number of 140 UMI recovered per 1000 reads. Little variation was observed in variant detection, with all five

mutations being detected at all variant levels and an R^2 value of 0.9375 (Figure 5.9).

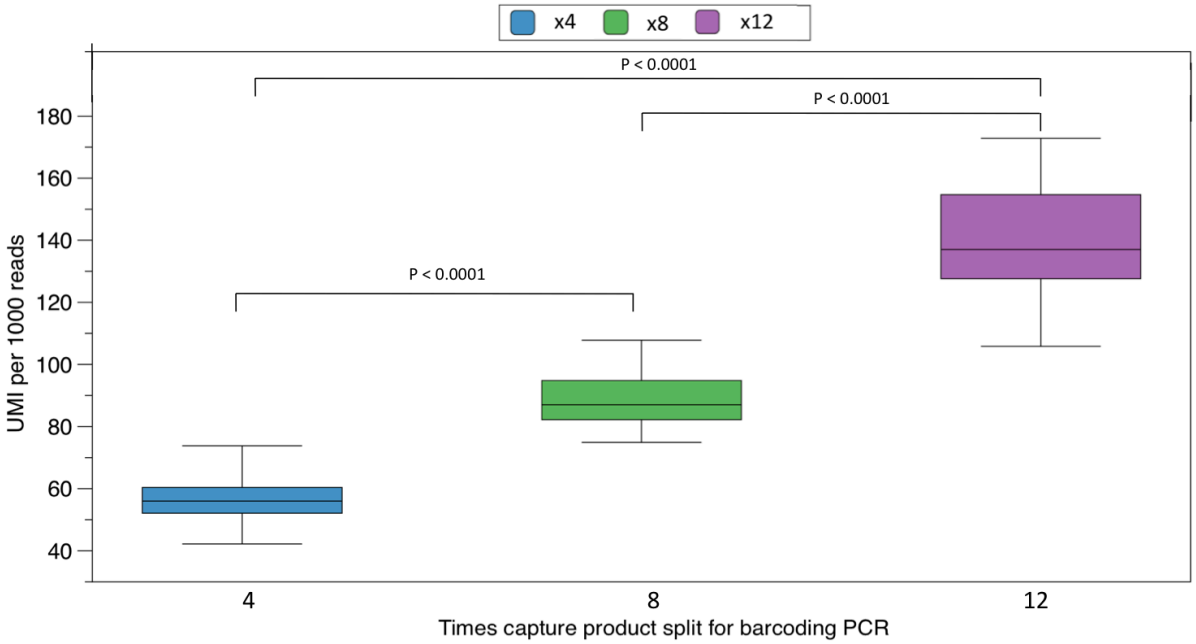


Figure 5.8: Splitting the capture reaction into multiple PCR wells at the time of barcoding improves the yield of UMI returned following sequencing as a result of reduced saturation during capture amplification. To account for the differences in the number of reads per sample, data presented is the number of UMIs per 1000 reads. There is a statistically significant increase in the number of UMIs returned (two sided t test). The box plot shows the median and interquartile ranges. Each box plot is made up of 12 independent samples (6 per probe).

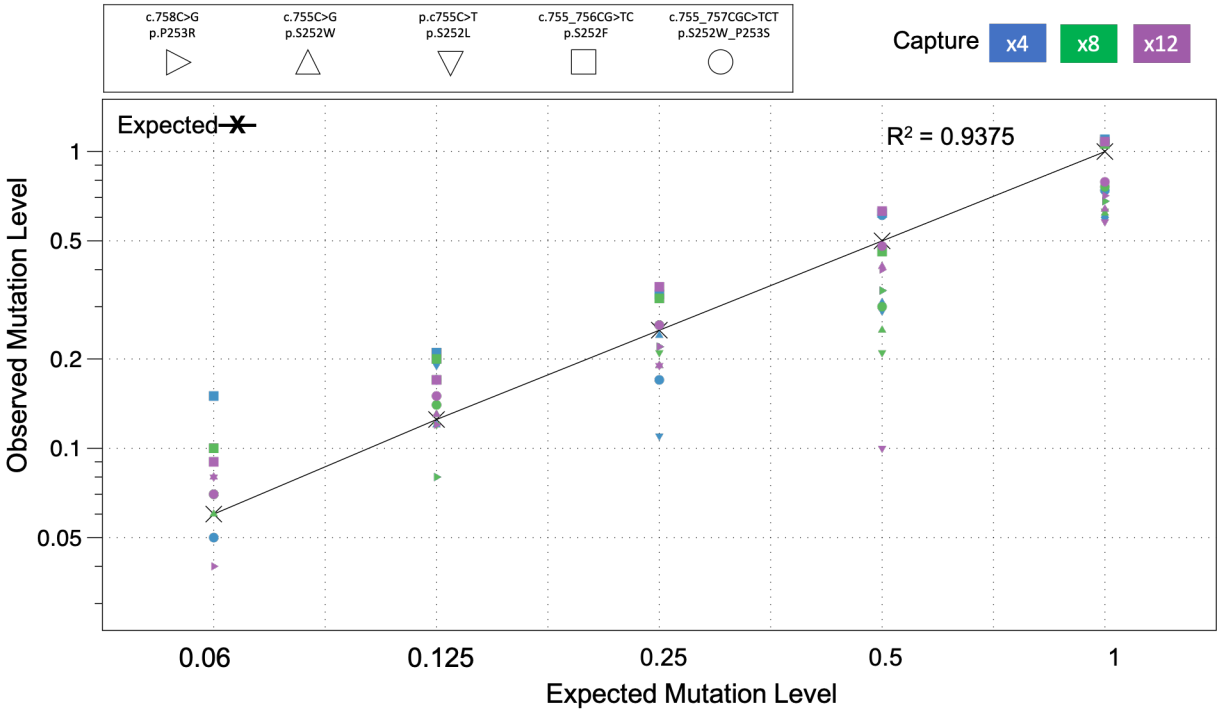


Figure 5.9: Observed mutation level (y-axis) on a log scale. Mutations are detected at all levels and in all treatment conditions. Points are shaped by mutation number (for details see Table in 5.1) and coloured by the number of times the captured product was split for barcoding (blue x4, green x8, purple x12). The five *FGFR2* mutations in the dilution series were detected by all probe sets. Black line shows expected mutation level in each sample.

5.5 Discussion

Initial performance of smMIP in this proof of principle study are encouraging, with mutations being reproducibly detectable down to 0.06%. In order to attempt to improve capture efficiency, two points in the protocol were identified which could be modified. Changing the purification of the smMIP probes offered no detectable benefit (Figure 5.4.1), meaning that subsequent high-throughout panels of smMIP can be investigated using commercially synthesised probes with standard purification. A highly significant increase in capture efficiency was observed by splitting the captured product into multiple wells at the time of the barcoding PCR (Figure 5.8). This could potentially be extended even further by using an emulsion PCR system. However, these tests were conducted in an artificial system with known mutations and little background. The performance of smMIPs in the context of high-throughput discovery screens will therefore be explored in

the following chapter.

Chapter 6

Multiplex target discovery of candidate variants associated with age-related clonal haematopoiesis in the human testes

6.1 Introduction

Selfish spermatognial selection (SSS) is a process of clonal expansion of activating mutations, occurring uniquely in the male germline. Low-level mosaicism as a result of clonal expansion has also been described in somatic tissue of healthy individuals, including the skin (Hafner et al., 2007; Martincorena et al., 2015), gut epithelium (Nicholson et al., 2018) and blood (Genovese et al., 2014; Xie et al., 2014; Jaiswal et al., 2014; Zink et al., 2017; Acuna-Hidalgo et al., 2017). In the blood, haematopoietic stem cells (HSC) may acquire a mutation that confers a competitive advantage over other cells. Over time, a genetically distinct subpopulation of cells is derived. The acquisition of such clonal haematopeotic driver mutations (CHDMs) is believed to represent the early stages in the development of haematologic malignancies. The presence of clonal haematopoiesis confers an increased risk of developing haematologic cancer as well as an increased incidence of coronary heart disease and ischemic stroke (Jaiswal et al., 2014). Haematological cancers, such as acute myeloid leukemia (AML), myelodysplastic syndrome (MDS), and

myeloproliferative neoplasms (MPN), increase in incidence with age (Buscarlet et al., 2017). Correspondingly, expansion of somatic mutations in haematopoetic stem cells is a rare finding in individuals aged under 50 years (<1%) but can be detected in 10% of individuals over the age of 65 years (Genovese et al., 2014), rising to 18% in individuals over the age of 90 years (Jaiswal et al., 2014).

Multiple WES/WGS studies have established that the most commonly mutated genes in clonal haematopoiesis are *DNMT3A, ASXL1* and *TET2*, all of which function as epigenetic regulators. Mutations in *DNMT3A* have been identified in AML with reported frequencies up to 22% (Ley et al., 2010). The most common site of somatic mutation in *DNMT3A* is the p.R882 residue (Ley et al., 2010; Yan et al., 2011). In the germline, mutations affecting this site have been reported as causes of an overgrowth/intellectual disability syndrome, Tatton-Brown syndrome (Kosaki et al., 2017; Shen et al., 2017). These studies have also reported the identification of mutations in genes, such as *PTPN11* and *BRAF* that are known to be subject to SSS. Given this overlap between SSS associated genes and reported mutations in clonal haematopoiesis, a panel of smMIP targeting known clonal haematopoiesis genes was used for target discovery in the testis. This provides the opportunity to address whether age-related clonal haematopoiesis mutations are also relevant in the context of the male germline, as well as to assess the performance of a high-throughput panel of smMIP.

6.2 Identification of a recurrent mutation in *DNMT3A*

Initial interest in the potential overlap between genes identified in clonal haematopoiesis and genes subject to SSS began after the observation of a known recurrent mutation in *DNMT3A* (c.2644C>T p.R882C) in the long list of candidate variants from the multiplex PCR strategy described in chapter 3. Amplification of the variant in the putatively positive sample (Tes14D9), two control samples and 12 testicular samples was performed (Figure 6.1). This identified the variant was present in other samples from the same testes as well as a variant present in the adjacent codon (*DNMT3A* c.2645G>A p.R882H), also

only in Tes14.

In order to determine if these variants were true independent occurrences in the same biopsies, three samples from Tes14 and one sample from an unrelated individual was captured using a smMIP panel designed to target clonal haematopoiesis candidates. This panel was received from a collaborator and covers a total of 104 loci in 74 genes, including 40 regions known to drive clonal haematopoiesis and 64 novel candidate loci (including known hotspots in developmental disease and single hotspots from known SSS genes (*FGFR2, FGFR3, RET*) (Acuna-Hidalgo et al., 2017)).

6.2.1 Detection of *DNMT3A* mutation using smMIP panel

The smMIP panel was captured on three biopsies from Tes14 (14D4, 14D9, 14D28) and one biopsy from an unrelated individual (18H17). Samples were sequenced on the Illumina MiSeq and analysed using amplimap. The average coverage across all target regions was 1118.19 (range: 14 to 5621). Performance of the panel was largely consistent between the four samples, with a 200-fold difference between the best covered region (*DNMT3A* p.G843 to p.R866) and the poorest (*FGFR3* p.R248 to p.P260). In all samples, the target hotspot (p.R882) in *DNMT3A* was covered by a minimum of 1000 reads (range 1071 to 1829).

Variant levels at the p.R882 codon *DNMT3A* (chr2:25457241 to chr2:25457244) were assessed (Figure 6.2). The p.R882H variant was observed in all three samples from Tes14 and was supported by multiple independent UMI. Similarly, at p.R882C multiple UMI supported the presence of the variant in all samples. The identity of the supporting UMI were examined for all samples and were found to be completely independent.

6.3 Multiplex target discovery in testicular samples

The smMIP panel was next used to multiplex target discovery in 64 testicular samples (49 biopsies and 15 seminiferous tubules) and two control samples. These samples included adjacent biopsies from Tes14D and Tes4F in order to allow tracking of the geographic

distribution of mutations across multiple biopsies. As described in section 5.8, the capture reaction was split into twelve reactions at the time of barcoding with two barcodes being used per sample (2 x 6 reactions).

6.3.1 Coverage and data analysis

The sequencing files for all samples were analysed using amplimap to produce alignment files, coverage data and a pileup file across all targets and samples. I then wrote a script to parse the coverage data and the pileup to:

- Produce plots of the number of supporting UMI for each target region within each target category (Figure 6.3).
- Extract a list of SNP positions and check that the zygosity of the SNP matches the individual group as expected. To facilitate this, a graph per SNP site is produced.
- Remove positions with no non-reference nucleotide from the pileup analysis.
- Produce graphs and tables showing the count and distribution of nucleotide changes at each UMI depth (Figure 6.4).
- Output a table of unique variant calls supported by more than 4 non-reference UMI. The output table contains a summary of the number of samples in which the variant is present, the number of non-reference UMI in each sample and the number of smMIPs overlapping the variant site.

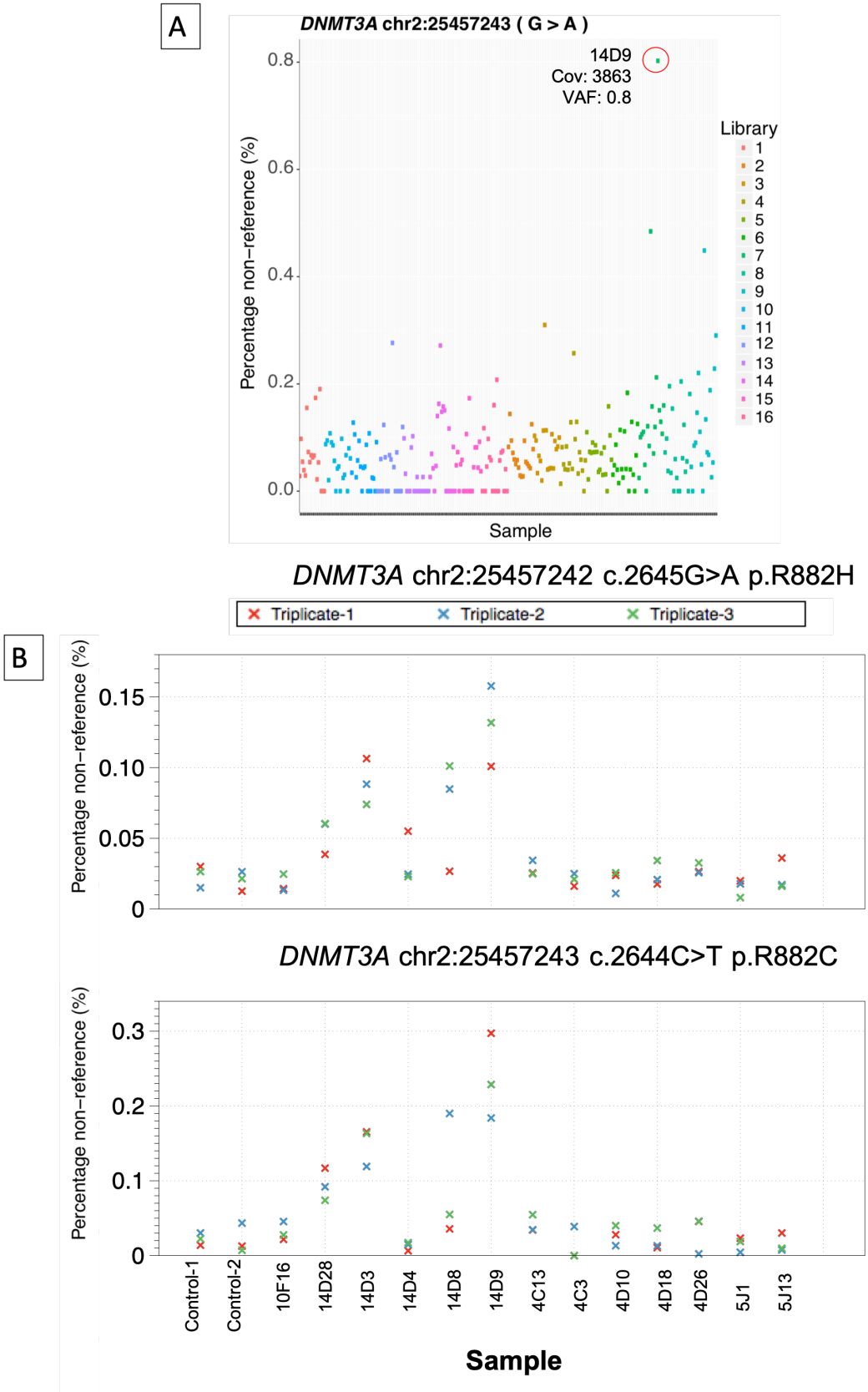


Figure 6.1: *DNMT3A* c.2644C>T p.R882C and c.2645G>A p.R882H variant. (A) A variant in *DNMT3A* (c.2644C>T p.R882C) was present as a singleton hit on a long list of candidate variants from a multiplex PCR discovery screen (chapter 3). (B) Validation in triplicate of the *DNMT3A* variant c.2644C>T p.R882C (bottom panel) identified a variant present in the adjacent codon c.2645G>A p.R882H in the samples from Tes14 (top panel).

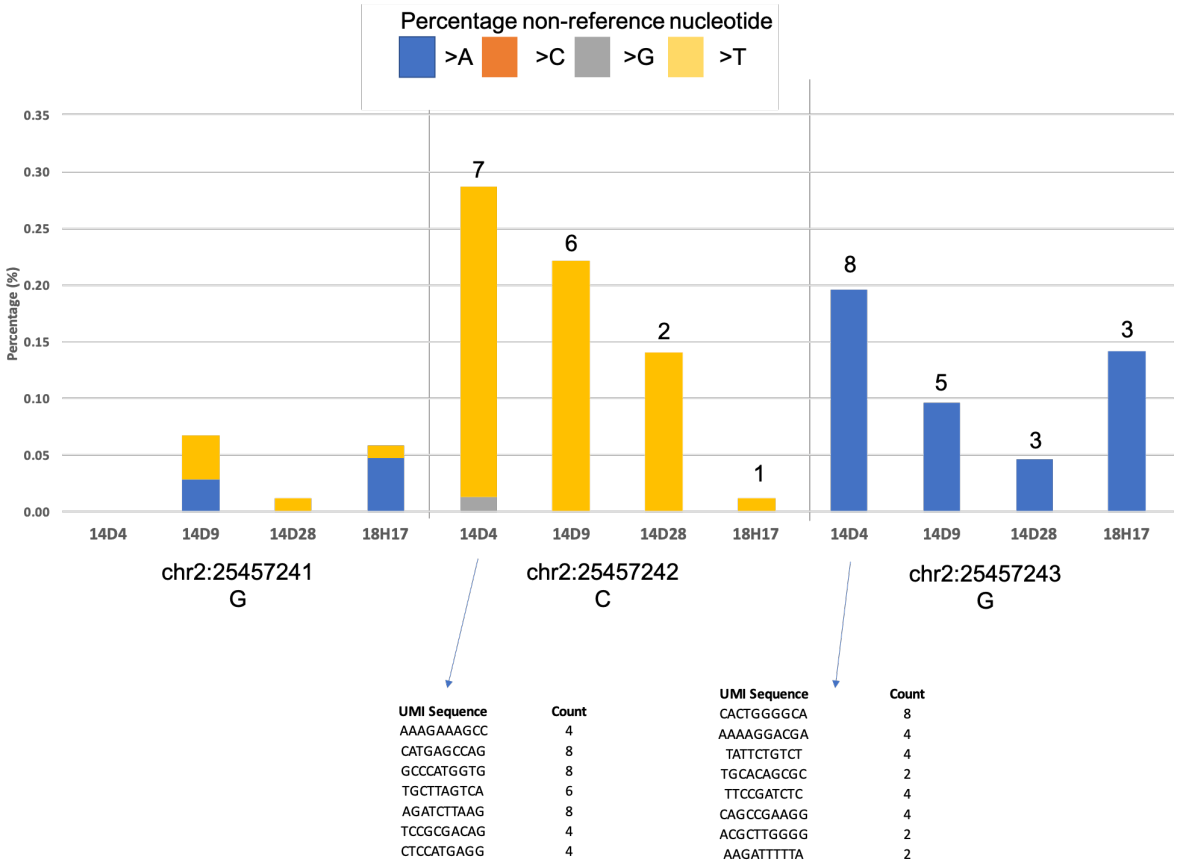


Figure 6.2: Percentage non-reference levels supporting non-reference nucleotide calls at chr2:25457241-25457243 (p.R882). The number of independent UMI supporting chr2:25457242 C>T (p.R882H) in each sample is shown above the bar (middle panel) and chr2:25457243 G>A (p.R882C) (right panel). The sequence identity of the UMI supporting the variant calls at each position in the same samples were verified to be completely independent.

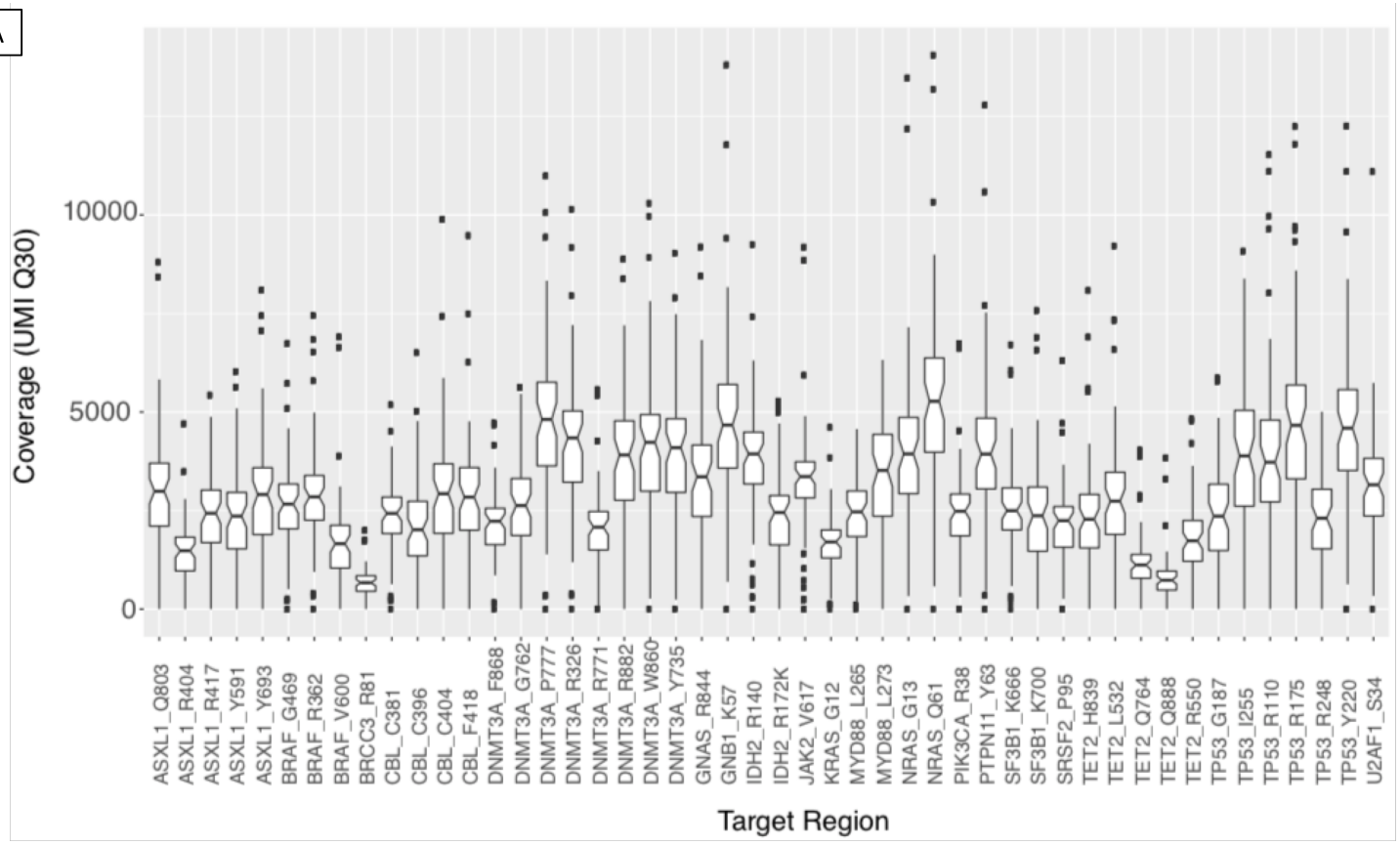
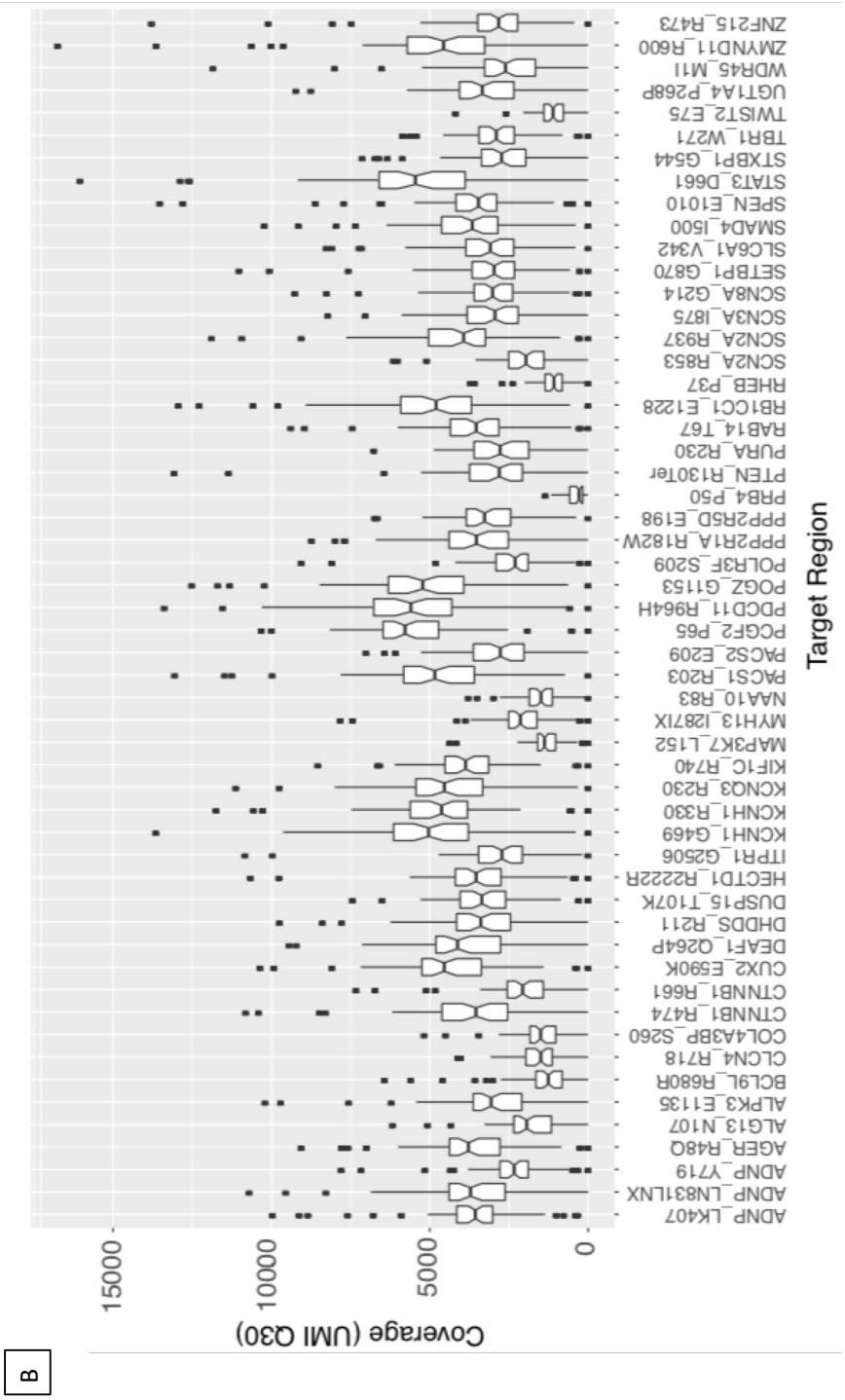


Figure 6.3: Box plots showing the number of UMI supporting a given target region for all samples. (A) Clonal haematopoesis genes (B) Other candidate hotspots (C) Genes in which mutations have been previously associated with SSS. Note that genes that could be considered as SSS-associated or CHD are listed in category A.


B
Coverage (UMI Q30)
15000
10000
5000
0
Target Region
ADNP_LK407
ADNP_LN831LNX
ADNP_Y719
AGER_R48Q
ALG13_N107
ALPK3_E1135
BCL9L_R680R
CLCN4_R718
COL4A3BP_S260
CTNNB1_R474
CTNNB1_R661
CUX2_E590K
DEAF1_Q264P
DHDDS_R211
DUSP15_T107K
HECTD1_R2222R
ITPR1_G2506
KCNH1_G469
KCNH1_R330
KCNQ3_R230
KIF1C_R740
MAP3K7_L152
MYH13_I287IX
NAA10_R83
PACS1_R203
PACS2_E209
PCGF2_P65
PDCD11_R964H
POGZ_G1153
POLR3F_S209
PPP2R1A_R182W
PPP2R5D_E198
PRB4_P50
PTEN_R130Ter
PURA_R230
RAB14_T67
RB1CC1_E1228
RHEB_P37
SCN2A_R853
SCN2A_R937
SCN3A_I875
SCN8A_G214
SETBP1_G870
SLC6A1_V342
SMAD4_I500
SPEN_E1010
STAT3_D661
STXBP1_G544
TBR1_W271
TWIST2_E75
UGT1A4_P268P
WDR45_M11
ZMYND11_R600
ZNF215_R473

C
Coverage (UMI Q30)
15000
10000
5000
0
Target Region
FGFR2_P253
FGFR3_G380
FGFR3_K650
FGFR3_S249
HRAS_G12
RET_C634
RET_M918

6.3.2 Variant prioritisation

The majority of positions in the panel where a non-reference nucleotide was observed are supported by a single UMI only (>86.1%), with a further 11.3% supported by 2 UMI (Figure 6.4). To prioritise variants taking advantage of the discriminatory power provided by UMI, all variants supported by four or more UMI were inspected.

6.3.3 Variant prioritisation: Tier 1

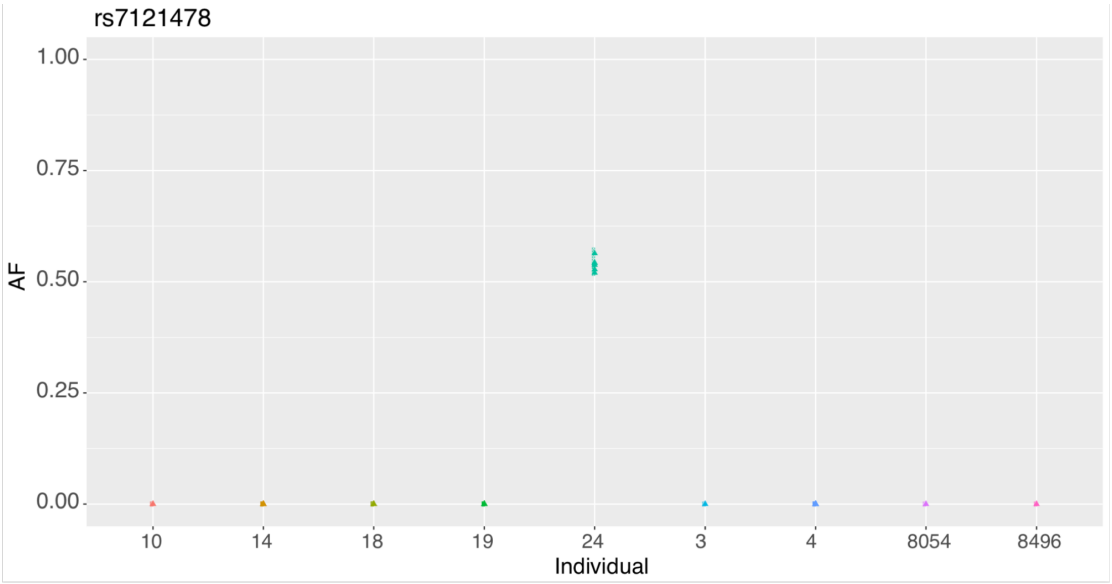
Variant positions that were supported by 10 or more UMI were first inspected. This accounted for 18 unique mutations, including a 3 bp deletion in *PTPN11* which was initially identified in the pileup as a series of single deletions (Figure 6.7). The two variants identified at the highest level (in *PRB4* and *SCN3A*) were excluded based on the presence of the called non-reference nucleotide in isoforms of the respective genes (PRB and SCN). The majority of the variants in this tier (14 of 18) had previously been validated in the previous chapters. The remaining two novel variants (*MYD88* and *TP53*) were selected for further validation by PCR (section 6.3.5).

6.3.4 Variant prioritisation: Tier 2

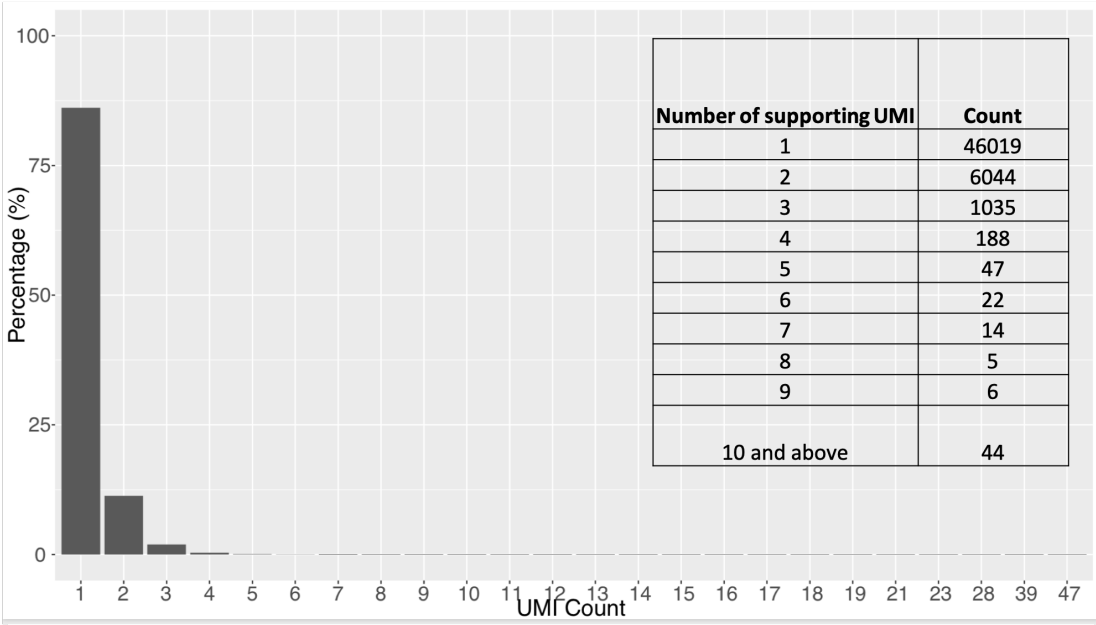
181 unique variants supported by between 4 and 9 UMI were identified and manually inspected. Two previously validated variants and four candidate variants for further validation were identified. Of these, two variants were successfully validated (Table 6.2).

Figure 6.4 *(following page)*: Summary graph output. (A) Graph output for a heterozygous SNP present in a single individual. AF is shown on the y-axis for each individual (x-axis) in which the SNP was present. (B) For all pileup positions in which a non-reference nucleotide was present, the relative UMI count is shown. The majority of non-reference nucleotides are supported by a single UMI only. Overlaid is a table with the absolute counts of positions per supporting UMI (C) The distribution of nucleotide changes at each UMI depth.

A



B



Number of supporting UMI	Count
1	46019
2	6044
3	1035
4	188
5	47
6	22
7	14
8	5
9	6
10 and above	44

C

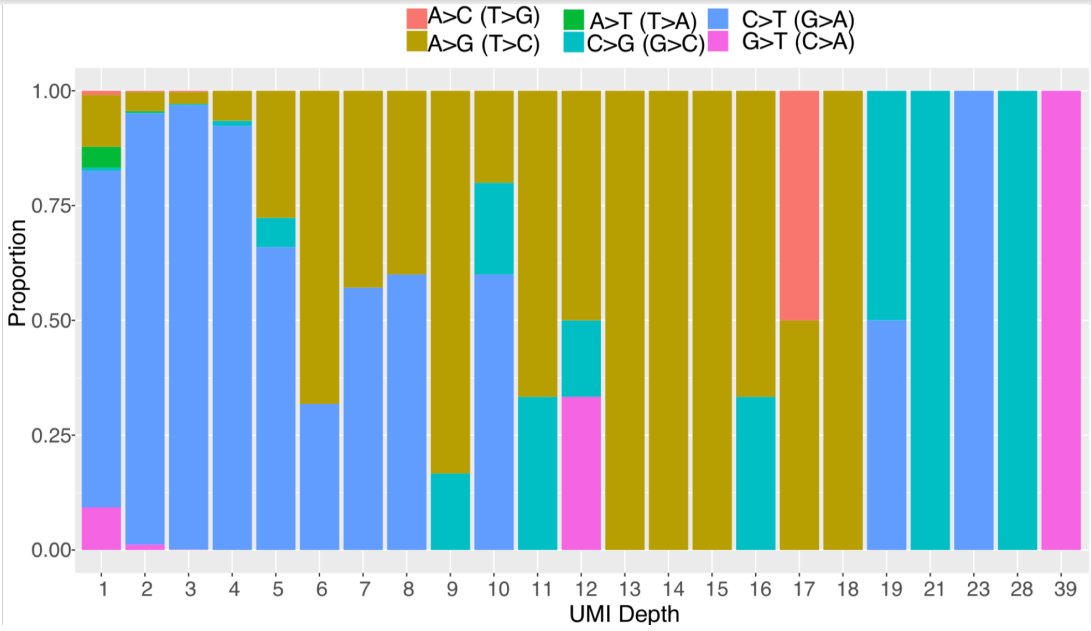


Table 6.1: Tier 1 variants

Gene	Variant (hg19)	Amino Acid	Overlapping smMIP	Percentage non-ref (%)	Sample	Validation
PRB4	chr12:11461745 G>	p.Q58E	2	3.18	Multiple samples - max in 24T04 (21/943)	na - Alignments
SCN3A	chr2:165986769 A>G	p.L819P	1	1.56	Multiple samples - max in 4F17 (17/1074)	na - Alignments
ADNP	chr20:49509094 G>T	p.Y719	2	1.32	24T05 (39/2951)	Previously validated
PTPN11	chr12:112888211 A>C	p.E76A	1	1.08	10F18 (17/1578)	Previously validated
FGFR2	chr10:123279677 G>C	p.S252W	2	0.94	10F3 (28/2989); 14D15a (12/2827)	Previously validated
PTPN11	chr12:112888164_66 TGA>DEL	p.D61 deletion	2	0.93	4F2 (47/5030); 4F6 (39/5418)	Previously validated
KRAS	chr12:25398284 C>G	p.G12A	2	0.73	4C24 (16/2201)	Previously validated
FGFR2	chr10:123279674 G>C	p.P253R	2	0.62	4F7 (19/3073); 4F12 (10/3120)	Previously validated
MYD88	chr3:38182641 T>C	p.L265P	2	0.56	10F3 (15/2669)	Positive PCR
SMAD4	chr18:48604664 C>T	p.R496C	2	0.48	4F5 (19/3962)	Previously validated
NRAS	chr1:115258748 C>A	p.G12C	2	0.37	4F8 (12/3239)	Previously validated
NRAS	chr1:115258745 C>G	p.G13R	2	0.34	10F3 (11/3250)	Previously validated
TP53	chr17:7578396 G>T	p.H178Q	2	0.33	4F4 (12/3676)	Positive PCR
NRAS	chr1:115256529 T>C	p.Q61R	2	0.31	4F3 (18/5805); 4F2 (18/5952); 4F8 (13/4446)	Previously validated
SMAD4	chr18:48604676 A>G	p.I500V	2	0.28	4F6 (13/4574)	Previously validated
CBL	chr11:119148931 G>A	p.C384Y	3	0.27	14D4 (10/4061)	Previously validated
RET	chr10:43609948 T>C	C634R	2	0.26	14D19 (14/5352)	Previously validated
DNMT3A	chr2:25457242 C>T	p.R882H	2	0.21	14D4 (10/4697)	Previously validated

Table 6.2: Tier 2 variants

Gene	Variant (hg19)	Amino Acid	Overlapping smMIP	Percentage non-ref (%)	Sample	Validation
FGFR3	chr4:1803564 C>T	p.R248C	2	0.38	4F11 (8/2127)	Previously validated
CBL	chr11:119148891 T>C	p.Y371H	2	0.26	4F1 (7/2697); 4F5 (6/2821) 4F6 (5/3352); 4F4 (5/2673)	Previously validated
NAA10	chrX:153197817 A>G	p.M98T	1	0.82	19T27 (6/730)	Positive PCR
KIF1C	chr17:4925594 C>T	p.R740C	2	0.23	4F8 (7/3028)	Positive PCR
GNB1	chr1:1747279 A>G	p.V40A	1	0.21	14D23 (5/2406)	Negative PCR
PPP2R1A	chr19:52715946 C>T	p.R171W	2	0.17	24T02 (5/3909)	Negative PCR

6.3.5 Validated Variants

In total, four variants in genes not previously implicated with SSS were validated (*KIF1C, MYD88, NAA10* and *TP53*) (Table 6.3 and Figure 6.5, 6.6).

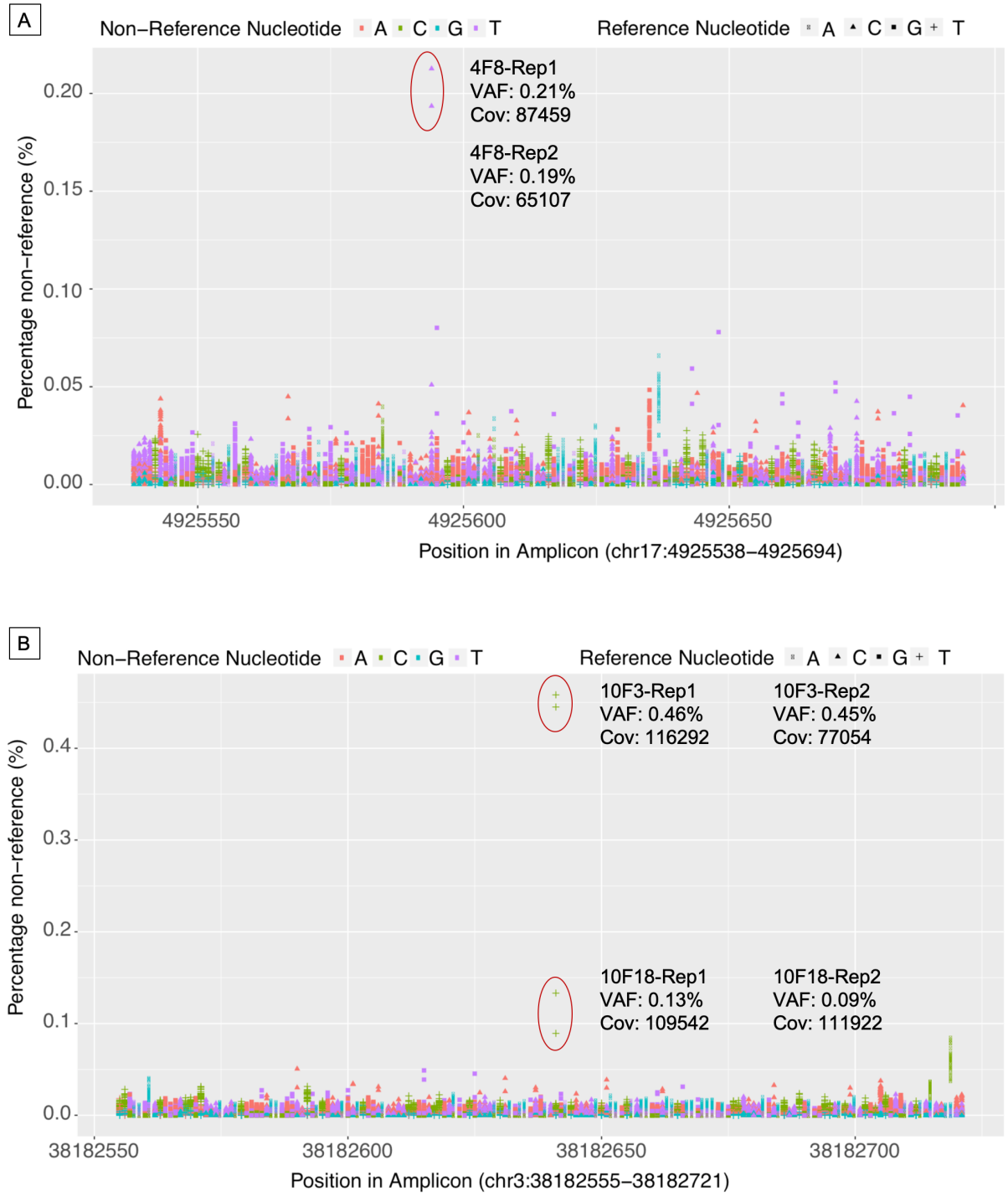


Figure 6.5: Novel variants validated in *KIF1C* and MYD88 (A) Variant validated in biopsy 4F8 in *KIF1C* at chr17:4925594 C>T p.R740C (B) Variant validated in biopsy 10F3 in *MYD88* at chr3:38182641 T>C p.L265P

KIF1C encodes a member of the kinesin family of proteins. Nonsense and missense mutations in *KIF1C* have been reported in autosomal recessive cases of spastic ataxia (Dor et al., 2014). The variant (p.R740C) has not been reported in the literature but

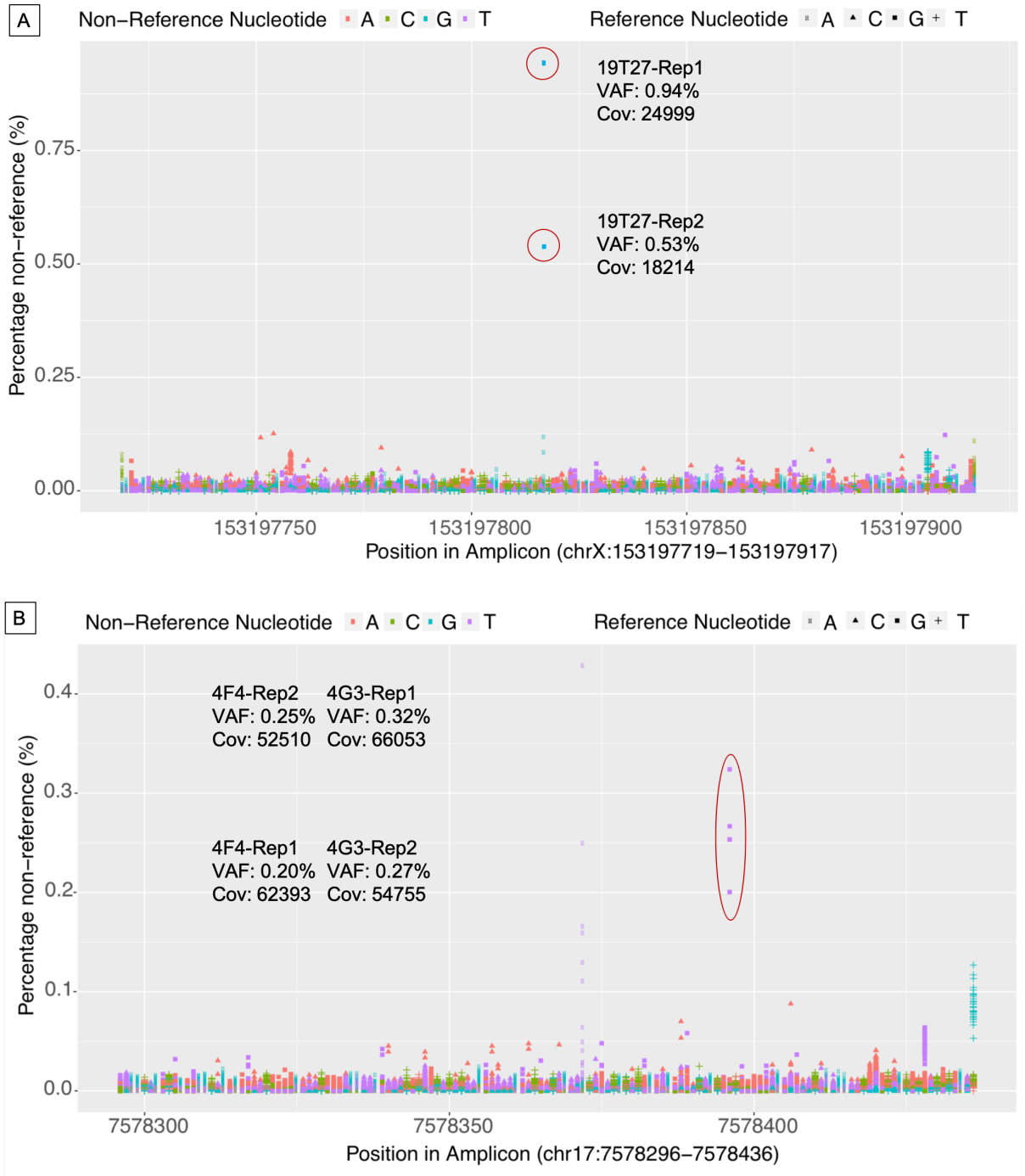


Figure 6.6: Novel variants validated in *NAA10* and TP53 (A) Variant validated in tubule 19T27 in *NAA10* at chrX:153197817 A>G p.M98T (B) Variant validated in biopsy 4F4 and 4G3 at chr17:7578396 G>T p.H178Q

is predicted to be damaging (PolyPhen = 0.988 (sensitivity: 0.73; specificity: 0.96), Adzhubei et al. (2015)). MYD88 p.L265P was validated in two samples from Tes10. *MYD88* encodes a downstream adapter for Toll-like receptors (TLRs) and interleukin-1 receptors (IL1R). This mutation will be discussed further in section 6.4.7

NAA10 encodes the protein ARD1A, which forms part of a protein complex involved in N-alpha-acetylation (a protein modification placed during protein synthesis). A recurrent missense mutation in NAA10 (p.S37P) causes the X-linked lethal condition Ogden syndrome (Alan et al., 2011). Less catalytically active mutations at other sites in the protein have been reported in a novel intellectual disability syndrome with global developmental delay (Popp et al., 2014; Casey et al., 2015). The missense mutation identified here, p.M98T, has not been reported in the literature but is predicted to be damaging (Polyphen score = 0.998,; Adzhubei et al. (2015)). Intriguingly, one of the mutations in a patient with intellectual disability (p.V107F) was noted to be located in a hydrophobic core of the enzyme (Popp et al., 2014). The authors specifically noted the tight interaction between residue V107 and residue M98. The steric clash caused by substitution at residue 107 was predicted to result in a loss of enzyme activity. Given this, the missense mutation identified here could be anticipated to have a similar functional effect.

The variant in *TP53* validated in two adjacent biopsy pieces (Tes 4F4 and 4G3). This call was present on the long list of variants identified in chapter 3. However, biopsy 4F4 was identified as a sample with an excessively high number of mutations. Specifically, G>T nucleotide changes were found to occur 7-fold more commonly than expected in this sample (section 3.2). Thus, while this variant is present in the biopsies, interpretation of the biological relevance is hindered as it may reflect an artefact of sample preparation.

Table 6.3: Details of novel variants validated by PCR. CADD is an aggregate score from multiple functional annotation tools (Kircher et al., 2014). A CADD score of 20 means a variant is in the top 1% of deleterious variants in the human genome. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical amino acid substitutions.

Gene	**Variant**	**Amino Acid**	**CADD score**	**COSMIC Reports**	**Germline disorder**
KIF1C	chr17:4925594 C>T	p.R740C	26.5	-	-
MYD88	chr3:38182641 T>C	p.L265P	33	2112	-
NAA10	chrX:153197817 A>G	p.M98T	26.0	2	-
TP53	chr17:7578396 G>T	p.H178Q	23.4	5	-

6.3.6 *PTPN11* deletion

A 3bp deletion was identified in *PTPN11* in adjacent biopsies from Tes4F (Figure 6.7. This in-frame deletion results in the loss of Aspartate 61, a key residue of the N-terminal loop of the protein (see protein domain structure in Figure 3.7 on page 85). Based on structural analysis, deletion of this residue is predicted to result in low stability of the loop leading to improved access to the active site (Lee et al., 2005). Other activating mutations at this residue of PTPN11 are described in Noonan syndrome (Tartaglia et al., 2001).

The deletion in *PTPN11* was initially identified in two adjacent biopsies of Tes4F. Although these biopsies, and this entire testicular slice, had previously been sequenced and analysed in chapter 3, only SNVs were called as part of this mutation prioritisation pipeline. To identify whether this mutation was present in other biopsies of the same individual, the sequencing files were searched for the presence of the mutant sequence (Table 6.4). In Tes4F7, only 2 UMI supported the deletion (out of 3206 WT, 0.06%).

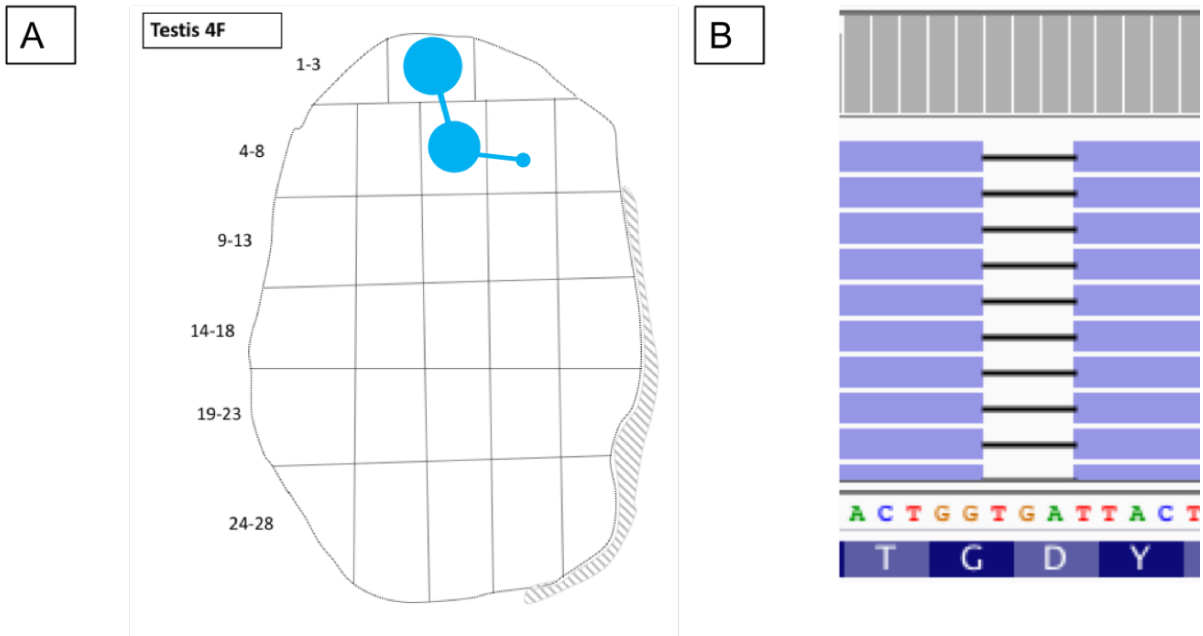


Figure 6.7: (A) Distribution of detected PTPN11 p.D61A deletion (B) IGV screenshot showing the presence of a 3 bp deletion in *PTPN11*

Table 6.4: Presence of reads supporting the deletion of residue p.D61 in PTPN11 in samples from Tes4

Sample	**Number of mutant sequences (CACTGGT-TACTAT)**	**Number of WT sequences (CACTGGT-GATTACTAT)**	**Percentage (%)**
4F2	321	28058	1.14
4F6	51	9349	0.54
4F7	11	23906	0.046
4F16	7	26699	0.02
4F14	1	6897	0.01
4E3	1	29223	0.003

6.3.7 Conclusions

In this screen, no additional variants were validated in other testes samples in *DNMT3A* and no variants in TET2 were identified, despite these genes being the most represented in clonal haematopoesis studies. Although the hotspot position (R882) in *DNMT3A* was covered at median depth of 3910 UMI in all samples (Figure 6.3), no additional variants at this position outwith Tes14 were identified. *TET2*, which has a reported hotspot at p.R550, performed more poorly with a median depth of 1731 UMI in all samples at this position (Figure 6.3). At this level, and given the difficulty in distinguishing true variation below 4 supporting non-reference UMI, this places a limit on the mutation level that can be identified ($\sim$0.4%).

To increase the resolution of mutation detection, individual seminiferous tubules were screened. However, as discussed in chapter 4, the chance of finding a SSS associated mutation in a single seminiferous tubule is much lower than in a biopsy. As a compromise, the size of biopsies screened will be reduced in order to improve resolution but retain the benefits of screening multiple seminiferous tubules at once.

6.4 Improved resolution of multiplex target discovery in testicular samples

6.4.1 Modification of target smMIP panel

To ensure adequate coverage of targets in key clonal haematopoiesis genes, a literature review was undertaken. Five studies of clonal haematopoiesis were reviewed (Table 6.5). From these studies, a total of 1604 mutations in 93 genes were identified.

Table 6.5: Clonal haematopoiesis studies

Study	**Number of Individuals**	**Approach**
Jaiswal et al. (2014)	17,182	WES
Genovese et al. (2014)	2380	WES
Zink et al. (2017)	11,262	WGS
Young et al. (2016)	20	Targeted
Xie et al. (2014)	2728	WES

The most represented genes are *DNMT3A* (n=828), *TET2* (n=221) and *ASXL1* (n=147). Mutations are reported throughout the coding regions of these genes. Across these studies, 15% of mutations in *DNMT3A* occur at a single hotspot (p.R882) (Figure 6.8). As well as mutations in epigenetic modifier proteins (1131 mutations), mutations also cluster within proteins related to the DNA damage response (100 mutations) and splicing factors/ribosome related proteins (94 mutations). Coverage of mutation sites in the most represented genes were checked with additional smMIP designed to improve coverage of key mutation sites (Table 6.6).

Table 6.6: Table of smMIP covering key clonal haematopoiesis related genes. The percentage of reported mutations covered by the previously discussed panel (section 6.2 and 6.3 is shown in brackets.

Gene	**Percentage of reported mutations covered (%)**
ASXL1	39.08 (22.76)
DNMT3A	53.48 (26.27)
PPM1D	69.69 (0)
TET2	35.55 (12)
TP53	58.53 (41.46)

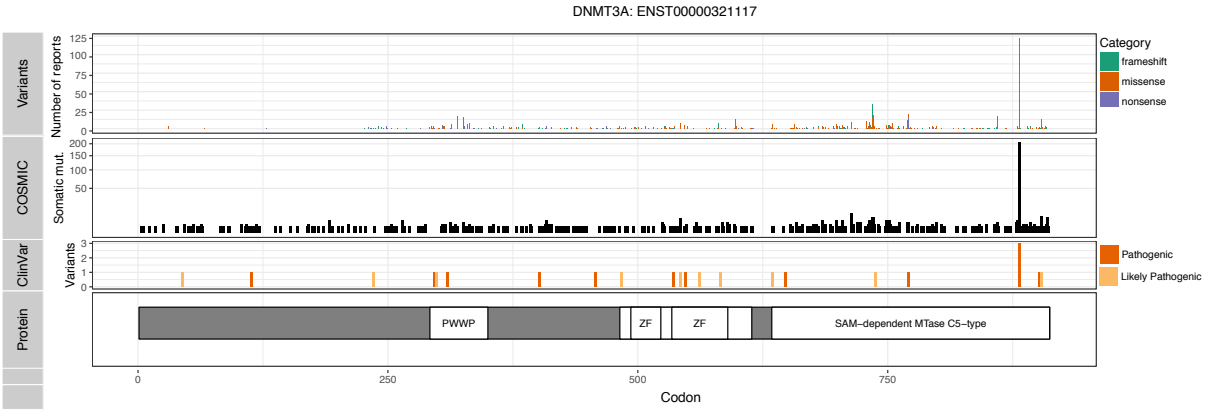


Figure 6.8: Overview of *DNMT3A*. (I) Mutations have been identified throughout *DNMT3A*, with a hotspot of recurrent mutations reported at p.R882 (Jaiswal et al., 2014; Genovese et al., 2014; Zink et al., 2017; Young et al., 2016; Xie et al., 2014). (II) Somatic mutations reported in DNMT3A (Cosmic v82) (III) Germline mutations reported in DNMT3A (Clinvar). (IV) Protein structure of DNMT3A from Uniprot.

Additionally, three recurrent mutations in *SMAD4* associated with the developmental disorder Myhre syndrome have so far been identified in testicular biopsies - at p.I500M (chapter 4, Table 4.4), p.I500V and p.R496C) (chapter 6 Table 6.1). Therefore, additional targets in the TGFβ signalling pathway were also added to the panel. smMIPs were designed to cover 13 hotspots in *ACV2RA, ACVRL1, BMPR1B, BMPR2, FBN1, SMAD2, SMAD3, TGFBR1, TGFBR2* based on either prior disease association or the presence of a hotspot in cancer. In total, 8.33 kb of unique bases are covered by the modified panel (clonal haematopoesis = 4.15 kb, SSS = 1.92 kb, TGF = 1.08 kb and candidate hotspots = 1.18 kb).

6.4.2 Capture and sequence analysis of testicular biopsies

To assess whether decreasing the size of biopsies screened would increase the resolution and allow the detection of more subtle mutations that expand to a lower level, a testicular slice from individual 14 was systematically dissected into 102 biopsies (20 to 60 mm^3, approximately 1/3rd smaller than the size of biopsies that have previously been screened). These biopsies were captured and sequenced alongside complete slices from two additional individuals (Tes10F and Tes4C). The sequencing files for all samples were analysed using amplimap to produce alignment files, coverage data and a pileup file across all target

regions and samples. The UMI coverage across the target regions of all gene categories was extracted and plotted (Figure 6.9).

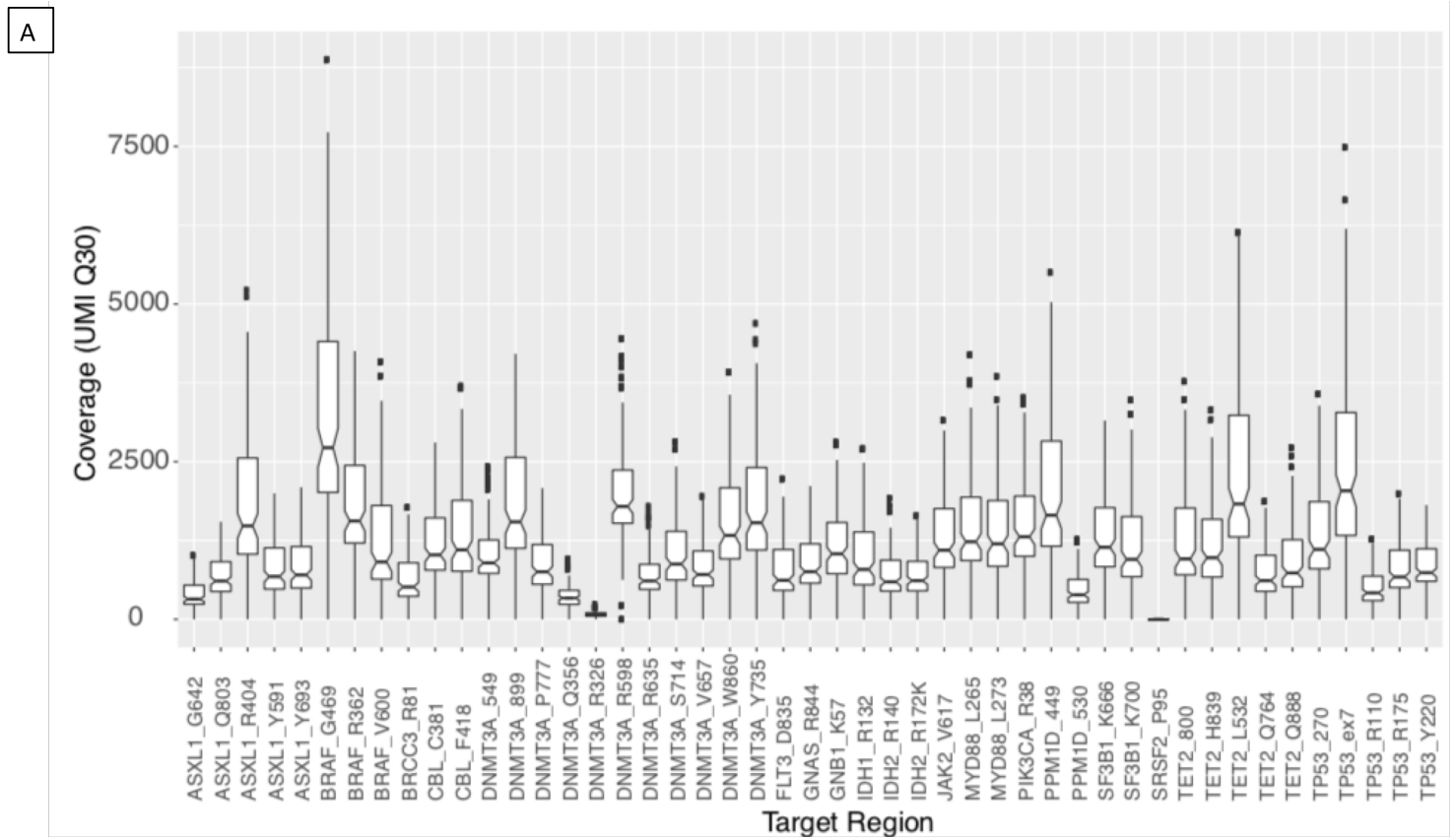
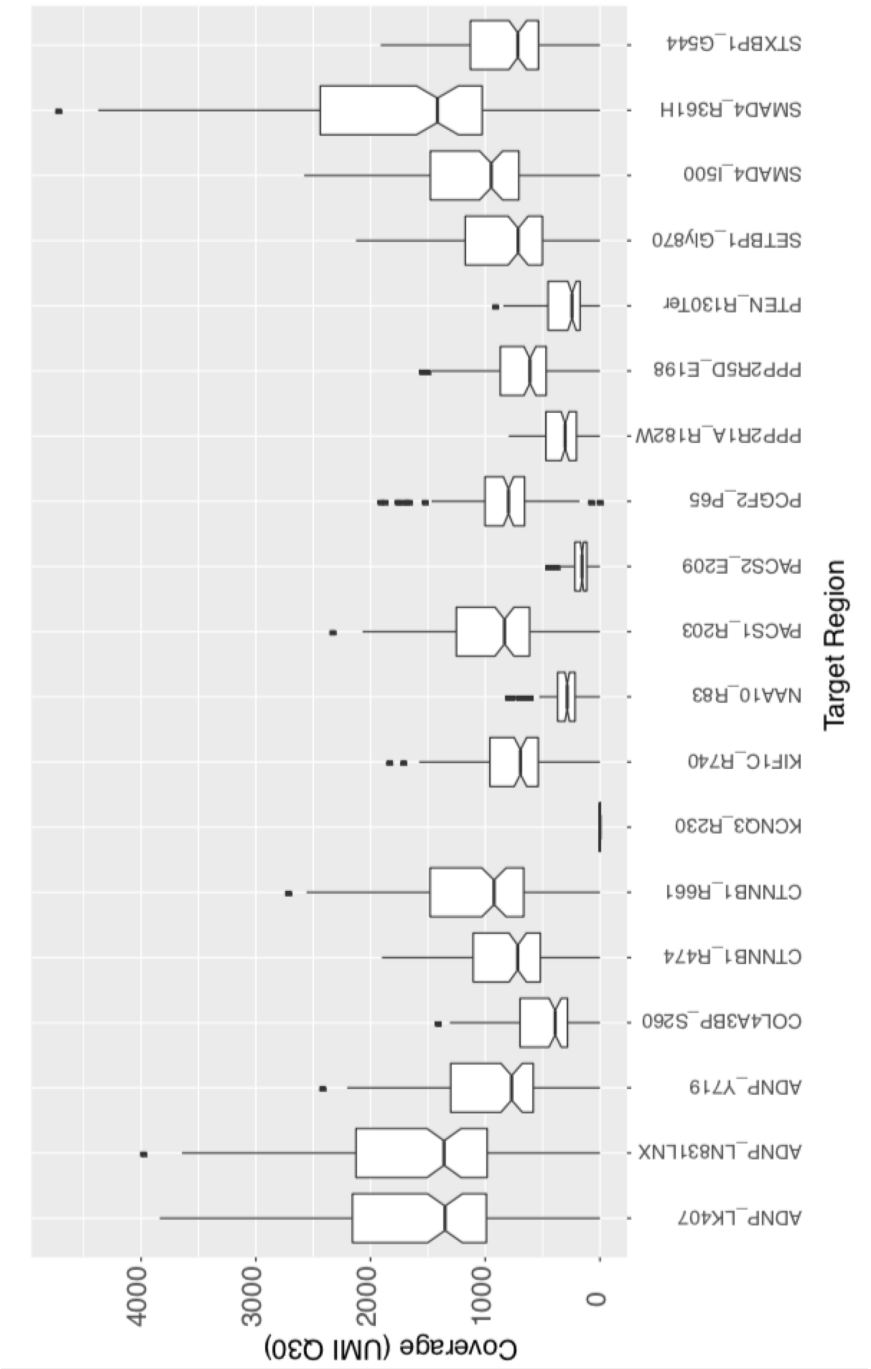
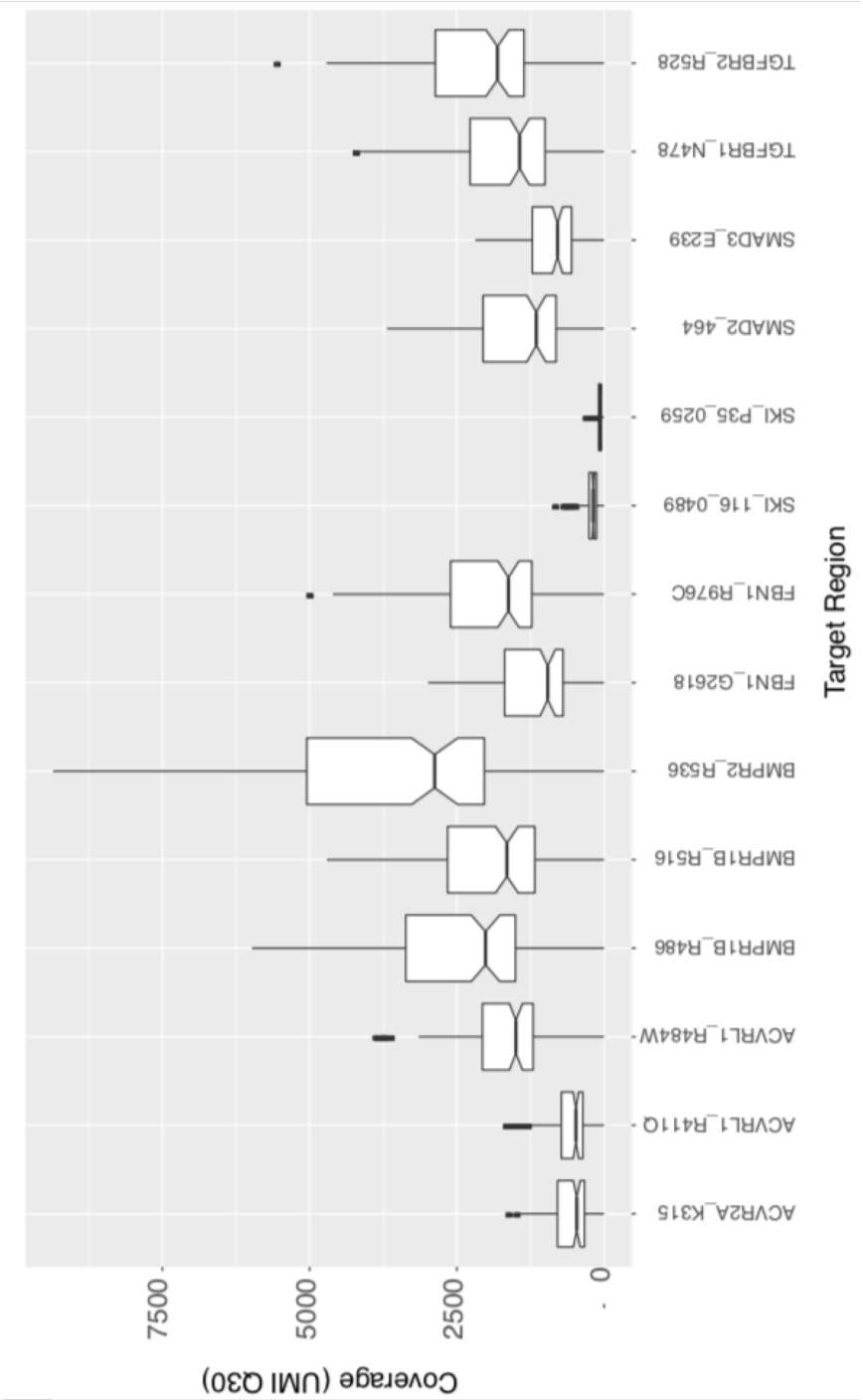


Figure 6.9: Box plots showing the number of UMI supporting a given target region for all samples. (A) Clonal haematopoesis targets (B) Other candidate hotspots (C) TGF-β/BMP signalling pathway (D) Genes in which mutations have been previously associated with SSS. Note that genes that would be considered subject to SSS or CHD are listed in category A


B
Coverage (UMI Q30)
0
1000
2000
3000
4000
Target Region
ADNP_LK407
ADNP_LN831LNX
ADNP_Y719
COL4A3BP_S260
CTNNB1_R474
CTNNB1_R661
KCNQ3_R230
KIF1C_R740
NAA10_R83
PACS1_R203
PACS2_E209
PCGF2_P65
PPP2R1A_R182W
PPP2R5D_E198
PTEN_R130Ter
SETBP1_Gly870
SMAD4_I500
SMAD4_R361H
STXBP1_G544


C
Coverage (UMI Q30)
7500
5000
2500
0
Target Region
ACVR2A_K315
ACVRL1_R411Q
ACVRL1_R484W
BMPR1B_R486
BMPR1B_R516
BMPR2_R536
FBN1_G2618
FBN1_R976C
SKI_116_0489
SKI_P35_0259
SMAD2_464
SMAD3_E239
TGFBR1_N478
TGFBR2_R528

D
Coverage (UMI Q30)
9000
6000
3000
0
Target Region
FGFR2_C342S
FGFR2_N549
FGFR2_P253
FGFR2_Q289
FGFR3_G380
FGFR3_K650
FGFR3_S249
HRASe2
HRASe3
KRASe2
KRASe3
NRASe3
PTPN11e13
PTPN11e3
PTPN11e8
RET_C634
RET_M918

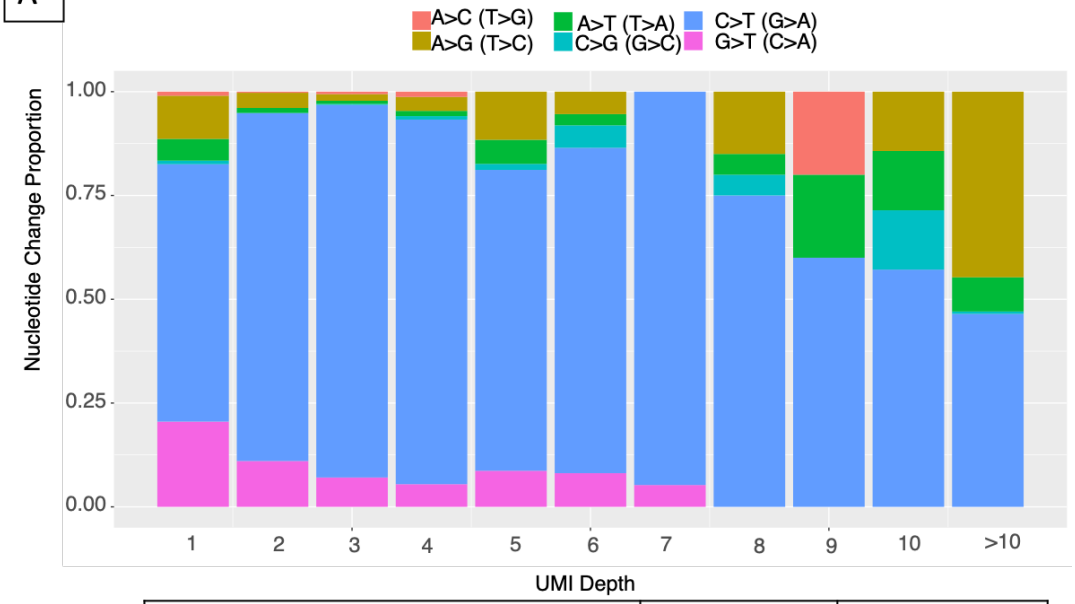
6.4.3 Statistical prioritisation of variant sites

A total of 89661 positions with a non-reference call were identified across all samples and target regions (Figure 6.10). This corresponds to unique mutations at 11850 positions. Positions in which a SNP was identified were removed from further analysis. To aid in variant prioritisation, a statistical test was developed. For a given variant position in a sample, the non-reference level across all other samples is used to estimate the distribution of background error. This is used to try identify outlier variant calls. To do so, a zscore was calculated for each possible nucleotide change and individual (testis sample) combination. The mean and SD of the zscores across all samples were then used to calculate a p-value for the variant call. To adjust for multiple comparisons, p-values were corrected using FDR method described by Benjamini and Hochberg (1995). The adjusted p-values were then used to rank variant positions. 5423 variant calls were identified with a p-value $<10^{-20}$; 17596 variant calls have a p-value $<10^{-5}$.

6.4.4 Top variant sites

To explore the long list of candidate variants, a coverage threshold of 500 and a minimum requirement for 3 UMIs supporting a non-reference call was applied. The capture reaction for all samples was split into 12 wells (as described in section 5.4.2) and two barcodes were used per sample (2 x 6 captures). Although the sequencing files for individual samples were merged for analysis, information about the independent barcodes was retained. Therefore, if all UMIs supporting a non-reference call were present in a single barcode of the sample only, this variant call was discounted. A total of 1411 (564 unique) variants met these criteria with 133 variants with a p-value <E-5. Detailed information for the top 25 variants, by p-value, is presented in Table 6.7. Of these, the majority appear to be biologically functional: 15 have been reported as somatic mutations in cancer (COSMIC) and 11 have been associated with germline disorders. 7 variants were selected for validation by PCR (performed by Dr Geoffrey Maher and Dr Maria Neofytou). A 100% validation rate was achieved (details in section 6.4.5 and figures 6.11 and 6.12).

A



B

Number of supporting UMI	Count	Percentage (%)
1	76217	85.01
2	10560	11.78
3	2056	2.29
4	488	0.54
5	155	0.17
6	75	0.08
7	23	0.03
8	20	0.02
9	9	0.01
10 and above	58	0.06

Figure 6.10: Distribution of non-reference UMIs and nucleotide changes. (A) The distribution of nucleotide changes at each supporting UMI depth. (B) Table of the number of non-reference positions supported by given number of UMI.

Two variants that met the above criteria but had less significant p-values were also selected for validation. These were *TP53* chr17:7577547 C>T (c.734G>A p.G245D) with a p-value of 0.008805919 and *MYD88* chr3:38182647 C>A (c.800C>A p.P267H) with a p-value of 0.007324582. Both were negative by PCR.

A summary of the genes in which the 133 top variants were identified and the mutation classification is provided in Table 6.8. As well as variants in known SSS associated genes

BRAF, FGFR2, KRAS, NRAS, PTPN11, RET, there are a number of novel genes with multiple variants reported. The average non-reference percentage was 0.39 (range: 0.04% to 5.67).

Table 6.7: Top 25 variants. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical DNA substitutions, identical amino acid substitutions and all substitutions at the site respectively.

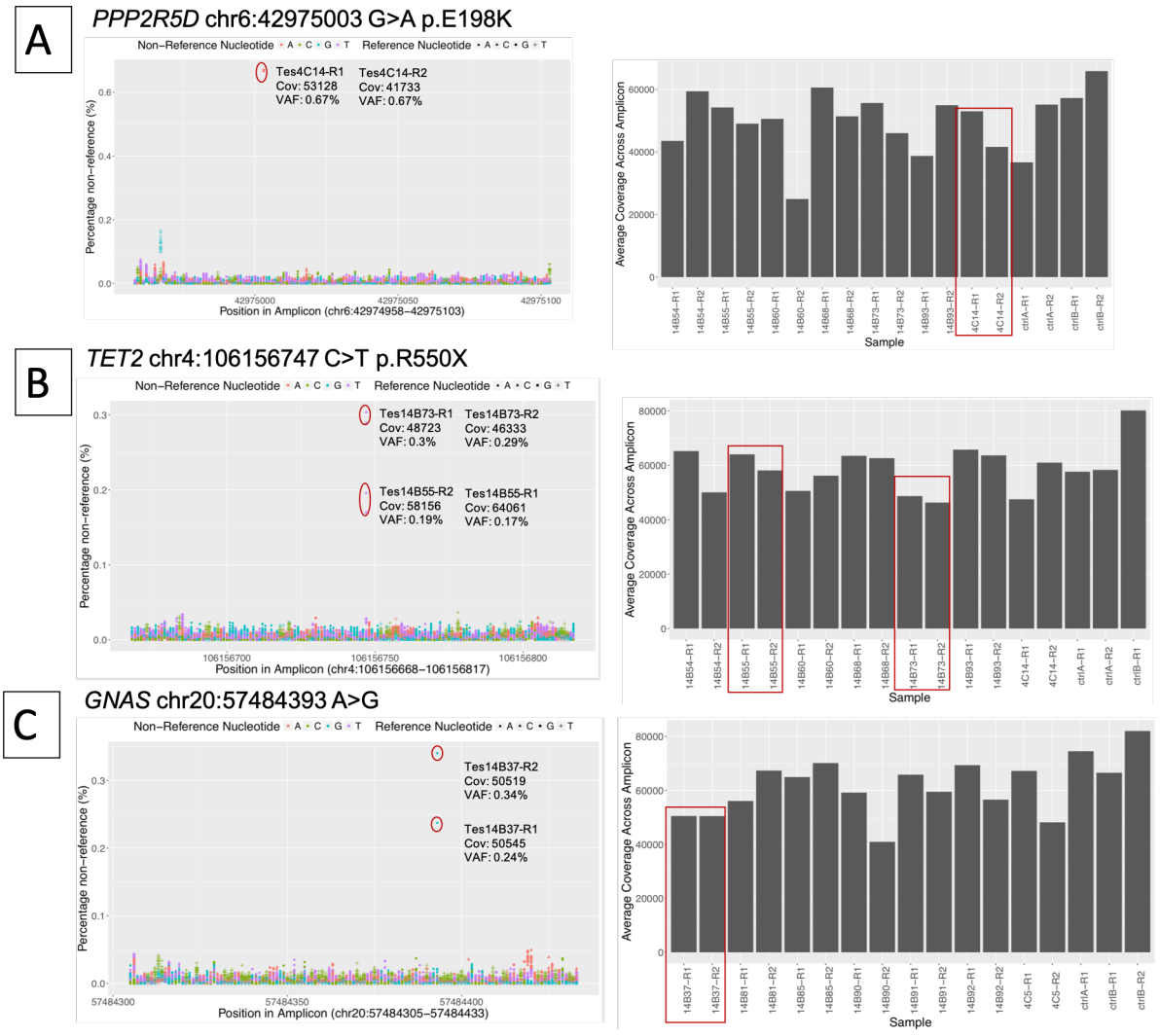
Gene	Position (hg19)	Variant	Sample	Non-ref UMI	COSMIC v86	Germline disorder	P value	Validation
ADNP	chr20:49509062	c.2189G>A (p.R730Q)	Tes14B85	4/622 (0.64%)	0 (0, 1)	-	6.21E-20	Validated
BRAF	chr7:140453132	c.A1803A>C (p.K601N)	Tes4C27	9/1090 (0.83%)	7 (18, 129)	-	1.59E-30	Previously validated
BRAF	chr7:140481402	c.1406G>C (p.G469A)	Tes4C7	13/3149 (0.41%)	52 (52, 123)	-	2.12E-30	Previously validated
BRAF	chr7:140481403	c.1405G>C (p.G469R)	Tes4C12	6/2922 (0.21%)		Cardio-facio-cutaneous syndrome	1.10E-29	Previously validated
BRAF	chr7:140481411	c.1397G>T (p.G466V)	Tes10F28	6/1936 (0.31%)	35 (47, 107)		1.55E-23	-
BRAF	chr7:140481411	c.1397G>C (p.G466A)	Tes14B27	11/6407 (0.17%)	9 (9, 107)		4.37E-22	na
DNMT3A	chr2:25463329	intronic	Tes14B81	18/682 (2.64%)	-	-	3.79E-27	Validated
FGFR2	chr10:123276893	c.1024T>A (p.C342S)	Tes14B22	265/4669 (5.68%)	-	Crouzon syndrome	2.53E-13	Validated (inferred)
FGFR2	chr10:123276898	A1019A>G (p.Y340C)	Tes4C27	10/3119 (0.32%)	-	Pfeiffer syndrome	5.78E-13	Previously validated
FGFR2	chr10:123276898	A1019A>G (p.Y340C)	Tes14B33	20/8617 (0.23%) -	-	Pfeiffer syndrome	8.70E-13	na
FGFR2	chr10:123279677	c.755C>G (p.S252W)	Tes14B11	10/1247 (0.80%)	54 (54, 57)	Apert syndrome	1.14E-11	na
GNAS	chr20:57484393	intronic	Tes14B37	5/863 (0.58%)	-	-	2.98E-29	Validated
GNB1	chr1:1747268	c.130C>A (p.Q44K)	Tes4C5	4/860 (0.47%)	-	-	2.01E-12	Validated
KRAS	chr12:25398284	c.35G>T (p.G12V)	Tes14B3	22/1263 (1.74%)	10253 (10253, 36634)		1.19E-29	na
KRAS	chr12:25398254	c.65A>G (p.Q22R)	Tes14B80	6/2823 (0.21%)	2 (2, 32)	Noonan syndrome	6.45E-22	na
MYD88	chr3:38182645	c.798C>A (p.I266I)	Tes14B72	4/720 (0.56%)	1 (1, 1)		1.28E-10	-
NRAS	chr1:115258745	c.37G>C (p.G13R)	Tes10F3	6/1293 (0.46%)	180 (180, 785)		5.13E-30	Previously validated
PPP2R5D	chr6:42975003	c.592G>A (p.E198K)	Tes4C14	6/555 (1.081%)	2 (2, 2)		8.09E-27	Validated
PTPN11	chr12:112888163	c.179G>T (p.G60V)	Tes14B23	15/1686 (0.89%)		Noonan syndrome	1.66E-28	na
PTPN11	chr12:112915524	c.923A>G (p.N308S)	Tes14B17	13/2908 (0.45%)	2 (2, 8)	Noonan syndrome	2.39E-22	na
PTPN11	chr12:112915523	c.922A>G (p.N308D)	Tes4C21	4/585 (0.68%)	0 (0, 8)	Noonan syndrome	3.48E-11	Previously validated
SMAD4	chr18:48604664	c.1486C>T (p.R496C)	Tes14B92	10/1235 (0.81%)		Myhre syndrome	4.27E-14	Validated
SMAD4	chr18:48604664	c.1486C>T (p.R496C)	Tes14B91	14/1956 (0.72%)		Myhre syndrome	8.72E-11	Validated
TET2	chr4:106156747	c.1648C>T (p.R550X)	Tes14B55	7/1800 (0.39%)	45 (45, 53)	-	1.23E-11	Validated
TP53	chr17:7577588	c.693C>A (p.T231T)	Tes14B21	4/3554 (0.11%)	0 (0, 16)	-	8.32E-12	-

Table 6.8: Summary of top variants by p-value. Note that only 131 variants are listed in this table, the remaining 2 are predicted to be intronic.

Gene	**Nonsynonymous SNV**	**Synonymous SNV**	**Stoploss**	**Stopgain**
ACVRL1	2	1	0	0
ADNP	5	4	0	0
ASXL1	4	0	0	0
BMPR1B	2	1	0	0
BMPR2	1	1	0	0
BRAF	7	3	0	0
DNMT3A	7	4	0	0
FBN1	1	3	0	0
FGFR2	10	2	0	0
GNAS	2	1	0	0
GNB1	4	0	0	1
KIF1C	0	1	0	0
KRAS	2	3	0	0
MYD88	5	0	2	0
NAA10	0	1	0	0
NRAS	5	2	0	0
PIK3CA	2	0	0	0
PPM1D	2	1	0	0
PPP2R5D	3	0	0	0
PTPN11	7	0	0	0
RET	1	0	0	0
SF3B1	2	0	0	0
SMAD2	1	0	0	0
SMAD3	0	1	0	0
SMAD4	4	0	0	0
STXBP1	1	0	0	0
TET2	4	5	0	2
TGFBR2	0	1	0	1
TP53	3	2	0	1

6.4.5 Novel variants validated

Seven novel variants from the top tier of candidate calls (section 6.4.4) were positive upon PCR amplification (figures 6.11 and 6.12). These variants are summarised in Table 6.9.



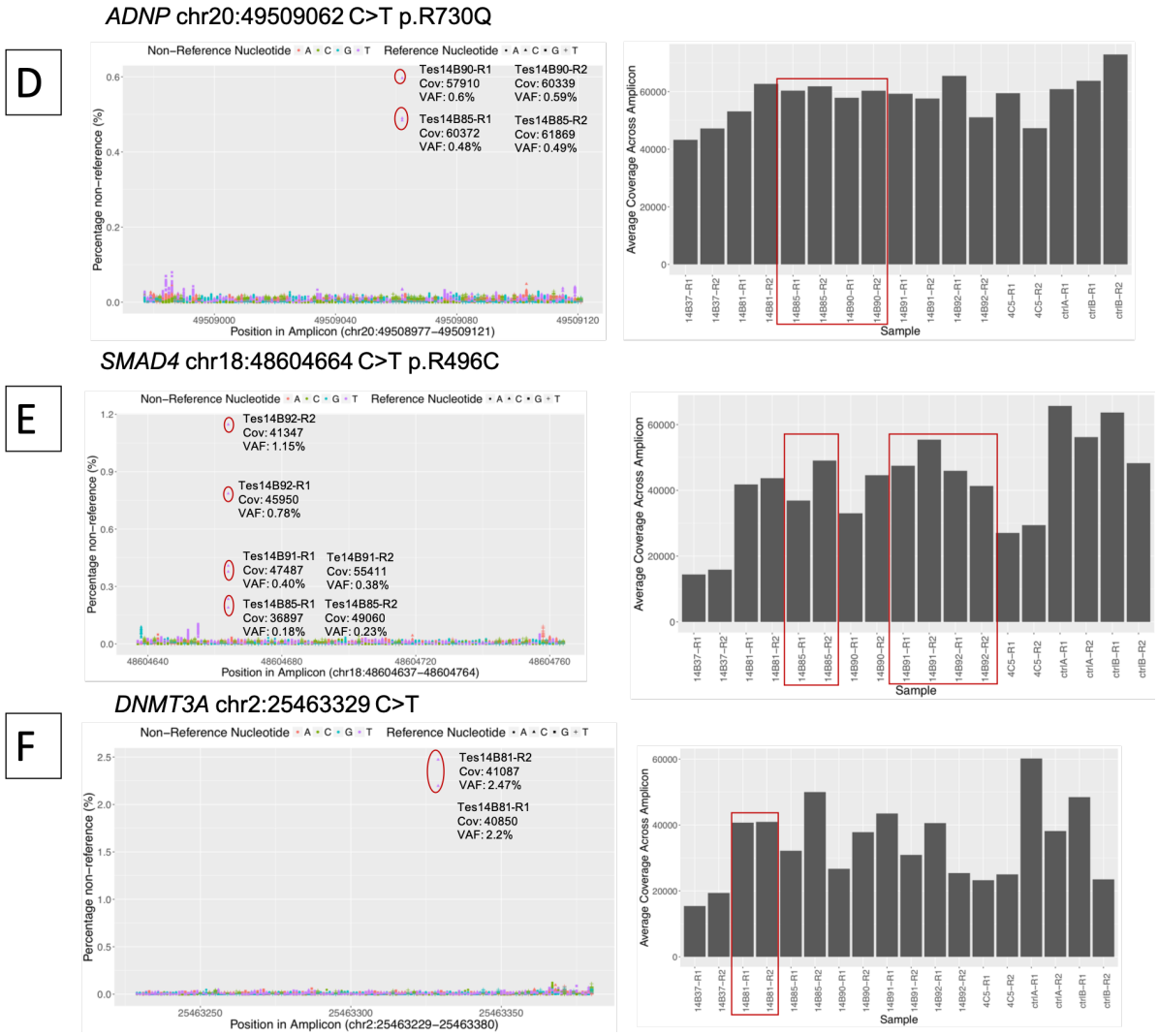


Figure 6.11: Novel variants validated by PCR. Left panel shows the percentage non-reference across the target amplicon and right panel shows average coverage across the target amplicon for the samples of interest (outlined in red) and control DNA samples. (A) Variant validated in *PPP2R5D* at chr6:42975003 G>A p.E198K. (B) Variant validated in *TET2* at chr4:106156747 C>T p.R550X. (C) Variant validated in *GNAS* at chr20:57484393 A>G (intronic) (D) Variant validated in *ADNP* at chr20:49509062 C>T p.R730Q. (E) Variant validated in *SMAD4* at chr18:48604664 C>T p.R496C in three samples from Tes14. (F) Variant validated in *DNMT3A* at chr2:25463329 C>T (intronic).

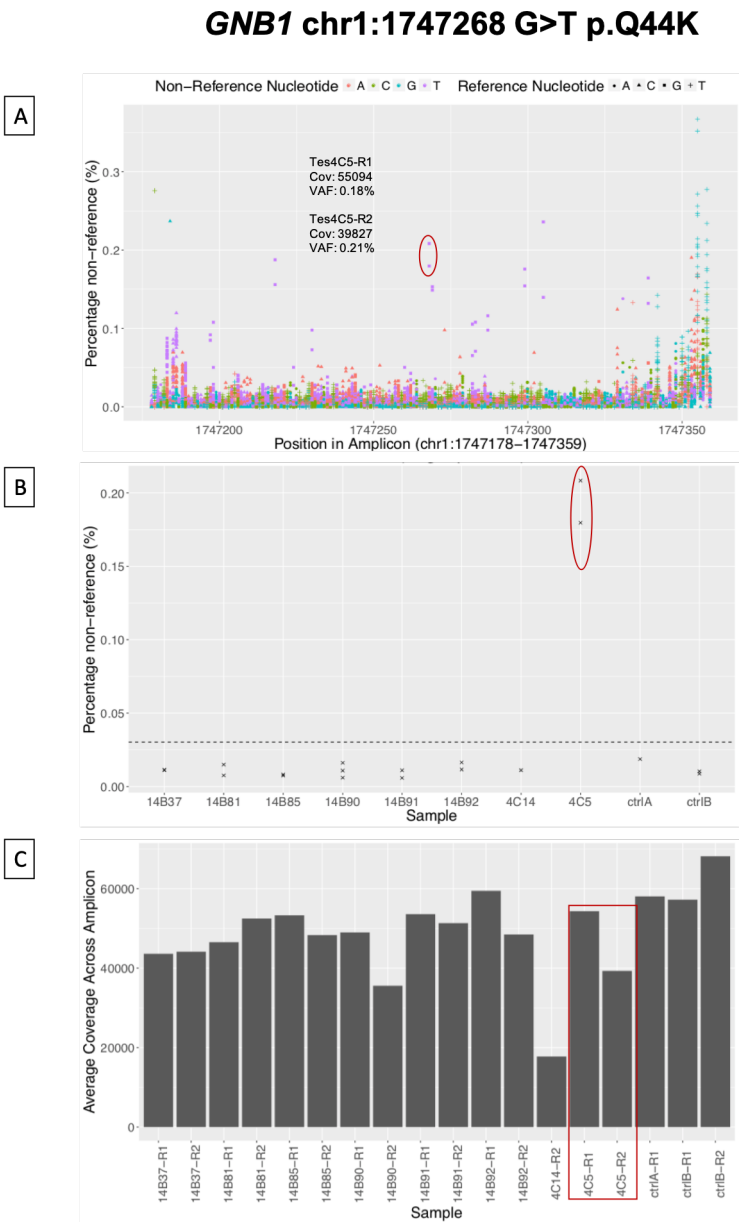


Figure 6.12: Novel variant validated by PCR in *GNB1* at chr1:1747268 G>T. (A) Percentage non-reference (y-axis) across the target amplicon (x-axis) for all nucleotide changes. (B) Percentage G>T at target position for all samples, shown due to high background across target amplicon in (A). Two crosses refer to each PCR replicate. (C) Average coverage across the target amplicon for the samples of interest (outlined in red) and control DNA samples.

Table 6.9: Summary of novel validated variants in *PPP2R5D, TET2, GNAS, ADNP, SMAD4, DNMT3A* and *GNB1*. Details of novel variants validated by PCR. CADD is an aggregate score from multiple functional annotation tools (Kircher et al., 2014). A CADD score of 20 means a variant is in the top 1% of deleterious variants in the human genome. Counts from the COSMIC database (v86) (http://cancer.sanger.ac.uk/cosmic/) refer to the reported number of identical amino acid substitutions.

Gene	**Variant**	**Amino Acid**	**CADD score**	**COSMIC v86**	**Germline disorder**
PPP2R5D	chr6:42975003 G>A	p.E198K	33	2	Intellectual disability
TET2	chr4:106156747 C>T	p.R550X	32	53	-
GNAS	chr20:57484393 A>G	intronic	0.749	-	-
ADNP	chr20:49509062 C>T	p.R730Q	22.5	-	-
SMAD4	chr18:48604664 C>T	p.R496C	34	5	Myhre syndrome
DNMT3A	chr2:25463329 C>T	intronic	17.24	-	-
GNB1	chr1:1747268 G>T	p.Q44K	20.7	-	-

The validated variant in *PPP2R5D* occurs at a reported hotspot for mutation. The gene encodes a regulatory subunit (B56δ) of protein phosphatase2A (PP2A), an intracellular serine/threonine phosphatase. DNMs in *PPP2R5D* (3 at p.E198K and 1 at P201R) were first reported in four unrelated patients with undiagnosed developmental disorders by Deciphering Developmental Disorders Study (2014). Recurrent mutations at residue E198 and at other positions in the highly conserved substrate recognition site of the protein (E198 to W207) were subsequently reported. In 11 patients with intellectual disability, six patients were reported with the p.E198K mutation. All patients with this mutation had severe intellectual disability (Houge et al., 2015). Paternal age information was not reported in this study. The authors transfected mutant B56 subunits into a human embryonic kidney cell line and found impaired complex formation as well as a disruption of B56δ dependent dephosphorylation. They proposed that *PPP2R5D* mutations may act in a dominant-negative manner by decreasing association of the PPP2R5D subunit with the other subunits of the PP2A complex. The mutation was also reported in

patients with intellectual disability and an overgrowth phenotype (Loveday et al., 2015).

A nonsense mutation was validated in *TET2* (p.R550X) in two samples from Tes14. *TET2*, along with *DNMT3A* are the most commonly identified genes with mutations in clonal haematopoiesis (Buscarlet et al., 2017). This particular variant has been reported in this context (Genovese et al., 2014). TET2 and DNMT3A both act as epigenetic regulators. DNMT3A/B establishes *de novo* methylation. TET2 is one of the methylcytosine dioxygenases that catalyzes the conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), leading to the subsequent removal of methylation marks. Thus, LOF in TET2 is associated with DNA hypermethylation. The loss of *TET2* in haematopoietic stem cells has been reported to result in increase stem cell renewal capacity *in vivo* based on transplant competition assays (Moran-Crusio et al., 2011).

The validated variant in *ADNP* results in the substitution p.R730Q. It was identified in two samples from Tes14. Missense mutations at this position have not been reported in the literature, but it may be biologically functional (CADD score = 22.5, (Kircher et al., 2014)). Primarily, nonsense mutations in *ADNP* have been reported in patients with Helsmoortel-Van der Aa syndrome, one of which (p.Y719*) was identified in a seminiferous tubule in chapter 4 (Helsmoortel et al., 2014). Truncating mutations at residue p.R730 have been reported in multiple patients and are reportedly the third most common *ADNP* mutation in Helsmoortel-Van der Aa syndrome (Krajewska-Walasek et al., 2016; Dijck et al., 2018).

The validated variant in *GNB1* has not been reported in the literature. *GNB1* encodes a subunit of heterotrimeric G proteins. Missense, splicing and frameshit mutations in *GNB1* have recently been identified in patients with global developmental delay, intellectual disability and seizures (Petrovski et al., 2016; Lohmann et al., 2017).

Of the seven variants, two were intronic. The variant in an intron of *GNAS* is a known SNP (rs367909148) reported in multiple populations (gnomAD exome freqeuncy = 0.0002381). The variant in the intronic region of DNMT3A has not been reported.

6.4.6 Distribution of p.C342S mutations in *FGFR2*

The highest level mutation identified was *FGFR2* c.1024T>A p.C342S, called at a level of 5.68% in Tes14B22 (p-value = 2.53E-13). This variant has been previously identified in an adjacent testicular slice from the same individual and was also the highest level mutation (2.95%) identified in that study (section 3.4.2). The higher level of mutation identified here is consistent with the increased resolution afforded by the smaller size of biopsies screened from this slice.

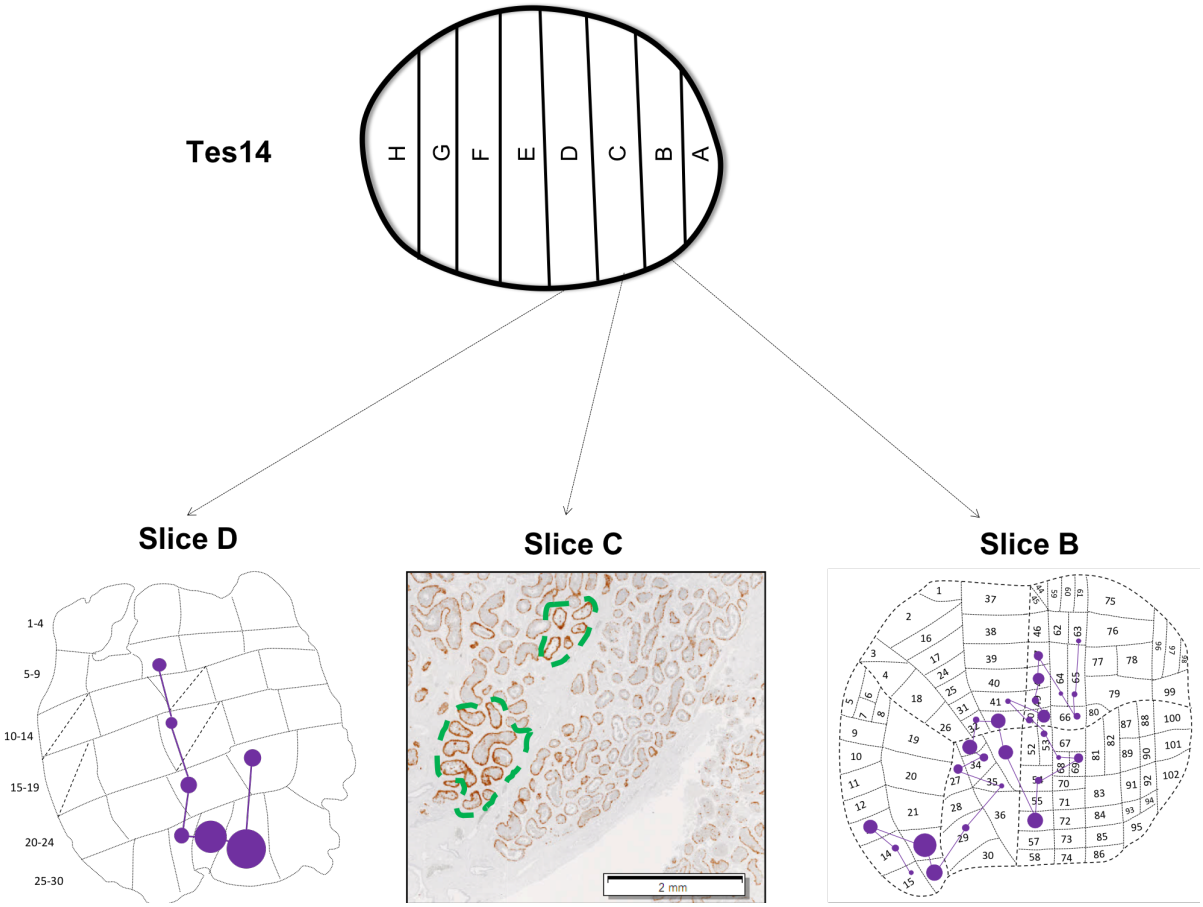


Figure 6.13: Distribution of C342S mutation in *FGFR2* across Tes14. Image for Tes14 Slice C produced by Dr Geoffrey Maher.

Given the prior observation of this mutation in an adjacent testicular slice, it was anticipated that this mutation would be called in multiple biopsies (Figure 6.13). This reflects a limitation of the Z-score ranking and p-value prioritisation strategy, as variant positions which are recurrently mutated across an individual will have a lower adjusted p-value. An assumption is made during the statistical test that only one sample will

have a elevated score. In this data, only one sample (Tes14B22) had a highly significant p-value: the rest of the samples had p-values ranging from 0.038798275 to 0.808856587 (Table 6.10). This is despite the z-score for a number of biopsies being much higher than the mean z-score (z=10.47359643, SD = 31.10878731) for the variant site. Thus, the statistical testing approach is instructive as a first pass to identify variant positions requiring further scrutiny in other samples.

6.4.7 Distribution of p.L265P mutations in *MYD88*

The second highest level variant identified was in MYD88 (p.L265P). This variant was previously validated in two biopsies from Tes10F (10F3 and 10F18) in section 6.3. The remaining biopsies from this testicular slice were captured with the modified smMIP panel and sequenced to assess the distribution of this variant across 10F. The variant is present at a high-level and is distributed across the testicular slice (Figure 6.14). The distribution of mutations across the majority of biopsies in the testicular slice is reminiscent of the pattern observed for the *NF1* p.M760I mutation in section 3.7 and is suggestive that this mutation may have arise as an early post-zygotic event, allowing it to reach elevated mutation levels within this individual.

This mutation (p.L265P) has been reported as a recurrent oncogenic mutation in 29% of activated B cell-like diffuse large B-cell lymphoma (Ngo et al., 2010). The authors demonstrated that cells carrying this mutation exhibited increased survival as a result of signalling changes that included increased IRAK4 kinase activity, JAK kinase activation of STAT3 and secretion of IL6 and IL10. The mutation was also reported in 91% of patients studied with Waldenstrom macroglobulinemia, a type of lymphoplasmacytic lymphoma (Treon et al., 2012).

Table 6.10: Distribution of z-scores and p-values for *FGFR2* c.1024T>A (p.C342S). The mean Z score for position is 10.47359643 and the standard deviation is 31.10878731. Only samples with at least 2 non-reference UMI are listed. The adjusted p-value for all other samples was 0.808856587

Sample	**Reference UMI**	**Non-reference UMI**	**Percentage non-reference (%)**	**Z score**	**p-value**	**FDR adjusted p-value**
Tes14B22	4669	265	5.68	262.095909	3.02E-16	2.53E-13
Tes14B23	7694	204	2.65	122.4223637	0.000159954	0.038798275
Tes14B56	10005	253	2.53	116.7565228	0.000317168	0.072210946
Tes14B42	5061	118	2.33	107.650032	0.000892775	0.182634317
Tes14B33	8572	187	2.18	100.7209458	0.001859725	0.349516835
Tes14B43	9930	210	2.11	97.63969302	0.002539527	0.459258181
Tes14B13	2590	52	2.01	92.69428889	0.004108686	0.63154338
Tes14B51	6931	116	1.67	77.26521944	0.015895289	0.63154338
Tes14B48	3865	52	1.35	62.10628543	0.048482769	0.633964045
Tes14B27	7392	68	0.92	42.45547845	0.151959425	0.808856587
Tes14B69	3326	30	0.90	41.62745192	0.158304958	0.808856587
Tes14B47	6269	56	0.89	41.22565775	0.161445852	0.808856587
Tes14B34	2318	18	0.78	35.83359888	0.207477824	0.808856587
Tes14B66	3251	17	0.52	24.12074838	0.330442644	0.808856587
Tes14B29	6179	32	0.52	23.88831266	0.333154382	0.808856587
Tes14B50	2637	12	0.46	20.98703819	0.367697683	0.808856587
Tes14B49	4178	19	0.45	20.973205	0.367865247	0.808856587
Tes14B53	3762	17	0.45	20.84039277	0.369475301	0.808856587
Tes14B54	4745	21	0.44	20.41013958	0.374706774	0.808856587
Tes14B14	8489	37	0.44	20.10011166	0.378490831	0.808856587
Tes14B32	7497	27	0.36	16.6033688	0.421896899	0.808856587
Tes14B65	3188	10	0.31	14.45732437	0.449051551	0.808856587
Tes14B99	8343	22	0.26	12.14897265	0.478525185	0.808856587
Tes14B35	10417	27	0.26	11.94103563	0.481188386	0.808856587
Tes14B68	7392	17	0.23	10.59186077	0.49848337	0.808856587
Tes14B41	5237	12	0.23	10.5530881	0.498980592	0.808856587
Tes14B15	4859	11	0.23	10.42586224	0.500612148	0.808856587
Tes14B63	5765	13	0.23	10.38497457	0.501136494	0.808856587
Tes14B11	4383	9	0.21	9.453920245	0.513074091	0.808856587
Tes14B75	2572	5	0.19	8.948785193	0.519546509	0.808856587
Tes14B64	9038	17	0.19	8.657527414	0.523276234	0.808856587
Tes14B85	3772	5	0.13	6.092540131	0.555997952	0.808856587
Tes14B74	4127	5	0.12	5.565942296	0.562676177	0.808856587
Tes14B52	6948	8	0.12	5.288271729	0.566190495	0.808856587
Tes14B67	3299	3	0.09	4.170427751	0.580282793	0.808856587
Tes14B80	7316	6	0.08	3.758257587	0.585454017	0.808856587
Tes14B30	4898	4	0.08	3.742276141	0.585654234	0.808856587
Tes14B36	2756	2	0.07	3.322142713	0.59090956	0.808856587
Tes14B77	4539	3	0.07	3.023100508	0.594640321	0.808856587
Tes14B40	4701	3	0.06	2.917910933	0.595950611	0.808856587
Tes14B91	6839	4	0.06	2.671839401	0.599011568	0.808856587
Tes14B60	3624	2	0.06	2.519413288	0.60090461	0.808856587
Tes14B79	9311	5	0.05	2.450705484	0.601757147	0.808856587
Tes14B12	4217	2	0.05	2.161003581	0.605346432	0.808856587
Tes14B57	6341	3	0.05	2.155649472	0.605412684	0.808856587
Tes14B25	7337	3	0.04	1.859035786	0.609078187	0.808856587
Tes14B55	4944	2	0.04	1.838919539	0.609326433	0.808856587
Tes14B78	9025	3	0.03	1.505840554	0.613430282	0.808856587
Tes14B6	7383	2	0.03	1.221731742	0.616920794	0.808856587
Tes14B21	7989	2	0.03	1.126832183	0.618084618	0.808856587
Tes14B10	8449	2	0.02	1.063884904	0.618855999	0.808856587
Tes14B24	9144	2	0.02	0.980792728	0.619873521	0.808856587
Tes14B93	11139	2	0.02	0.799876577	0.62208609	0.808856587

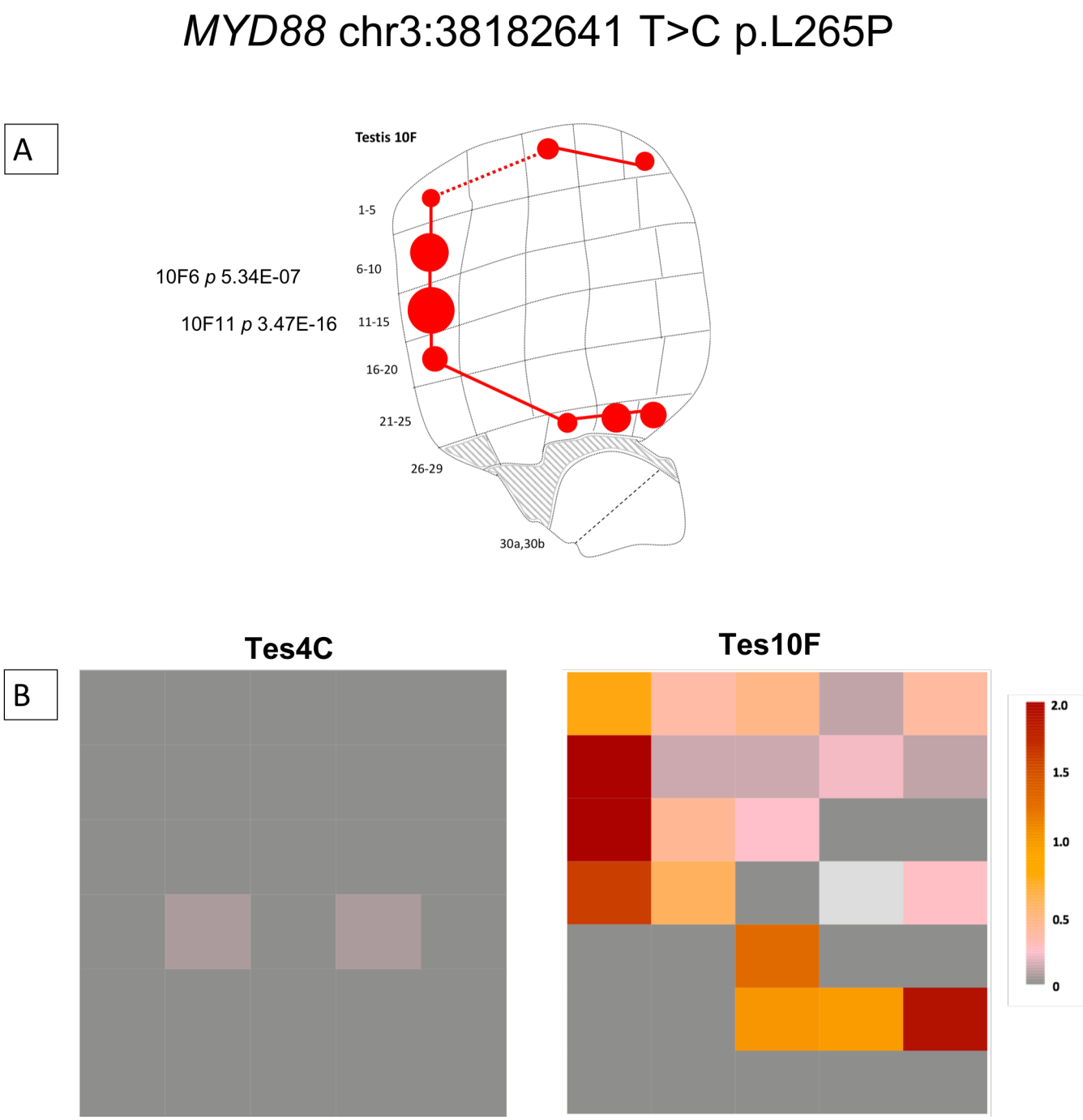


Figure 6.14: Distribution of*MYD88* mutation in Tes10F. (A) The two biopsies with the most significant p-values are indicated. The 9 biopsies in which the mutation was supported by at least 5 UMI are plotted. The size of the circle is proportional to the observed VAF. (B) Heatmaps showing the percentage level of T>C mutation across Tes4C (left panel) and Tes10F (right panel).

6.4.8 Discussion and Future Work

In this study, a panel of smMIP targeting candidate hotspots in clonal haematopoiesis, developmental disorders and SSS associated genes was used to analyse testicular biopsies at a higher resolution than has been done previously.

Decreasing the size of testicular biopsies proved very effective, leading not only to the validation of variants in novel genes *ADNP, TET2, SMAD4* but also to the identification of these variants spreading across adjacent biopsies (Figure 6.15). This provides the first evidence that variants in genes outside the RAS-MAPK pathway could be susceptible to SSS and are clonally expanding within the male germline.

Using a variant prioritisation strategy based on the calculation of z-scores and p-values across the target positions in the panel, 133 variants of interest were identified (Table 6.8). As well as variants in known SSS associated genes *BRAF, FGFR2, KRAS, NRAS, PTPN11, RET*, there are a number of novel genes with multiple variants reported. For example, *GNB1*, in which a single variant (c.130C>A p.Q44K), has thus far been validated, is also represented by three additional missense mutations (c.137G>A p.R46H; c.163C>A p.L55M; c.157G>T p.G53W) and a nonsense mutation (c.121G>T p.G41X). These variants also include the identification of non-synonymous variants in a number of genes that encode members of the TGF-β/BMP signal transduction pathway (*ACVRL1, BMPR1B, BMPR2, SMAD2, SMAD3, SMAD4, TGFBR2*). Further validation work is required as part of the interpretation of these variants, as these exciting results represent potentially novel signalling pathways that are subject to SSS.

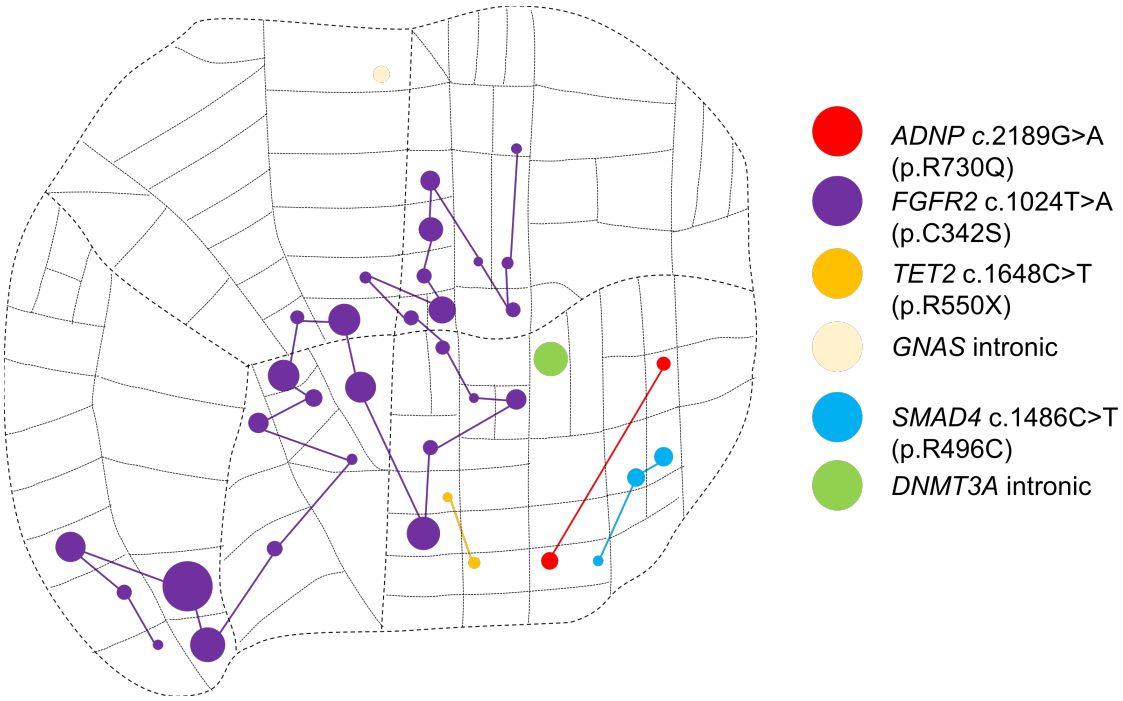


Figure 6.15: Variants validated in Tes14B

6.5 Recurrent mutations in *SMAD4* are identified in testicular biopsies

A total of three recurrent mutations in *SMAD4* associated with the developmental disorder Myhre syndrome have been identified in testicular biopsies - at p.I500M, p.I500V and p.R496C. These mutations overlap the narrow spectrum of dominant mutations in *SMAD4* that have been associated with the developmental disorder Myhre syndrome. *SMAD4* encodes a protein that functions in the TGF-β/Bone Morphogenic Protein (BMP) signal transduction pathway. The described mutations in Myhre syndrome cluster in the Mad Homology (MH2) domain of SMAD4 which is required for SMAD oligomerization and signal transduction (Caputo et al., 2012). Studies using patient fibroblasts have suggested that increased stabilization of the protein occurs as a result of mutations at p.I500 (Le Goff et al., 2011). An increase in phosphorylated SMAD proteins (SMADs 1,2,3 and 5) was observed along with decreased levels of downstream TGFβ targets.

This finding, combined with the validation of SMAD4 p.R496C across three adjacent testicular biopsies of Tes14B (Figure 6.15), provides the most convincing hint thus far that

mutations outwith the classically described RASopathies (Table 1.2) enter the population as a result of their preferential expansion in the male germline.

6.5.1 Distribution of Myhre syndrome associated mutations in *SMAD4*

To try identify additional mutations in *SMAD4* and to assess the distribution of mutations in testicular biopsies, a total of 220 samples from three individuals (Tes4, Tes10F and Tes14D) were screened by high-fidelity PCR. The biopsies were amplified in duplicate and sequenced to a median sequencing depth of 43,827 reads (range: 7674 to 95,511). Variant levels at the sites of four recurrent mutations in *SMAD4* associated with Myhre syndrome were examined and plotted as a heatmap (Figure 6.16). In total, 25 mutations in *SMAD4* were identified. The VAF of these mutations in testicular biopsies and the reported distribution of mutations in *SMAD4* are shown in Figure 6.17.

6.5.2 Candidate variants in TGF-β BMP signaling pathway

In chapter 3, a massively-parallel PCR-based discovery screen was described. Hotspots in *SMAD4* and other members of the TGF-β BMP signalling pathway were targets in this study. Although an amplicon covered the *SMAD4* p.R496 and p.I500 codons in this study, a high level of background was observed for the positions of interest and no mutation could be called. All mutations from the long list of candidates in *ACVR1* (n=30), *SMAD4* (n=42), *TGFBR1* (n=13), *TGFBR2* (n=46) and *FBN1* (n=36) were plotted as described in section 3.3 of chapter 3 and inspected in IGV. *SKI* was also targeted in this study, however, no mutations were called in this gene owing to inadequate coverage.

Only one mutation was considered to appear potentially real (*SMAD4* chr18:48604641 C>A c.1463C>A p.A488D). This was called in a sample from Tes14. However, this was negative by PCR. No other mutation in this dataset was considered to look real enough to be worth pursuing further validation by PCR.

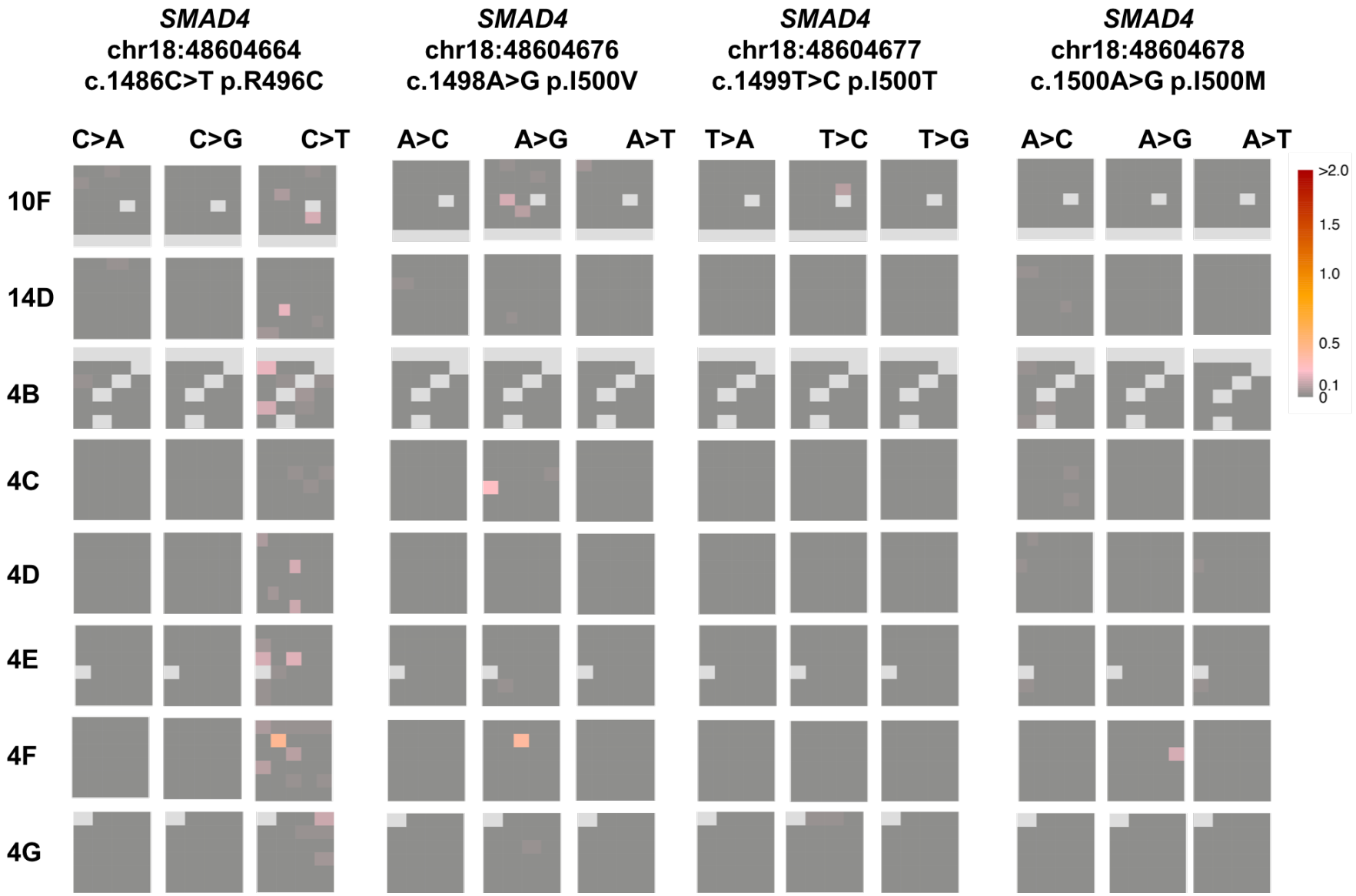


Figure 6.16: Heatmap of all possible nucleotide changes in *SMAD4* at four positions where recurrent mutation has been associated with Myhre syndrome. The percentage non-reference level shown is the average of the two sequencing replicates. Light grey denotes biopsies that were not resequenced due to lack of remaining available DNA material.

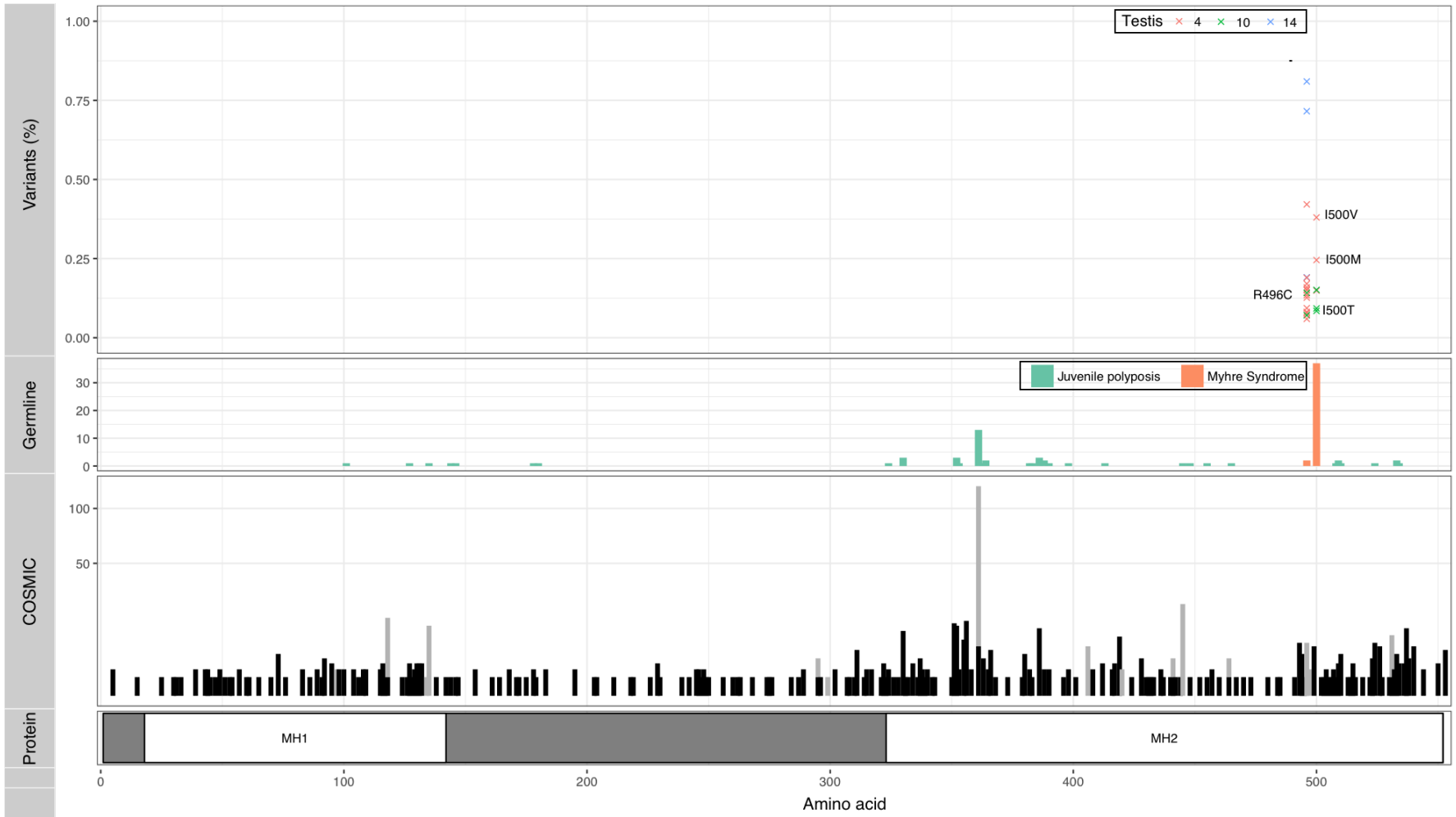


Figure 6.17: Diagram of mutations identified in *SMAD4*. (I) Validated variants positioned along the amino acid sequence of the protein, with VAF on the y-axis. (II) Distribution of germline mutations. Loss of function mutations are distributed throughout *SMAD4* that are associated with Juvenile polyposis, a hereditary condition characterised by the presence of benign polyps in the intestinal tract. A narrow spectrum of mutation in the MH2 domain of the protein are associated with Myhre syndrome. Germline mutations were curated from the literature. (III) Number of reported somatic amino acid substitutions in cancer (from COSMIC v86). (IV) Protein domains from Uniprot.

Table 6.11: Variants in *SMAD4* at p.I500 and p.R496 codons

Gene	**Codon**	**Amino Acid**	**Sample**	**Percentage non-reference (%)**
SMAD4	c.1500A>G	I500M	4F13	0.15
SMAD4	c.1499T>C	I500T	10F14	0.09
SMAD4	c.1498A>G	I500V	4F6	0.38
SMAD4	c.1498A>G	I500V	10F17	0.15
SMAD4	c.1498A>G	I500V	10F23	0.08
SMAD4	c.1498A>G	I500V	4C14	0.25
SMAD4	c.1486C>T	R496C	4F5	0.42
SMAD4	c.1486C>T	R496C	4G2	0.13
SMAD4	c.1486C>T	R496C	14D16	0.19
SMAD4	c.1486C>T	R496C	4B3	0.19
SMAD4	c.1486C>T	R496C	4E9	0.17
SMAD4	c.1486C>T	R496C	4E7	0.16
SMAD4	c.1486C>T	R496C	4B14	0.15
SMAD4	c.1486C>T	R496C	4D13	0.15
SMAD4	c.1486C>T	R496C	10F24	0.14
SMAD4	c.1486C>T	R496C	4D34	0.13
SMAD4	c.1486C>T	R496C	4F14	0.09
SMAD4	c.1486C>T	R496C	4F11	0.08
SMAD4	c.1486C>T	R496C	4D1	0.08
SMAD4	c.1486C>T	R496C	4F1	0.07
SMAD4	c.1486C>T	R496C	10F12	0.07
SMAD4	c.1486C>T	R496C	4G10	0.05
SMAD4	c.1486C>T	R496C	14B91	0.72
SMAD4	c.1486C>T	R496C	14B92	0.80
SMAD4	c.1486C>T	R496C	14B85	0.21

6.6 Discussion

6.6.1 Technology performance

The use of smMIP panels enabled the flexible, rapid and high-throughput screening of candidate sites across a large number of testicular slices. As was the case in chapter 4, the incorporation of UMIs provides confidence when calling variants. The majority of non-reference positions are supported by a single UMI only and distinguishing any signal at this level is not possible. Even with the support of up to 5 UMI, identifying high-confidence variant calls was challenging and low rates of validation were achieved (section 6.3.4). Combining UMI information with a statistical prioritisation method significantly improved validation rates. For example, a variant validated in *ADNP* (p.R730Q) (section 6.4.5) was supported by only 4 UMI but had a highly significant p-value (6.21E-20) and thus was easily identified. Variant interpretation would be improved by increasing capture efficiency. The biggest improvement to capture efficiency was made by splitting the capture reaction during barcoding amplification to reduce saturation (section 5.4.2).

The median UMI depth for some genes (e.g. *FGFR3* and *HRAS*) was consistently too low to call variants (figures 6.3 and 6.9). There is a little that can be done to improve the performance of a poor probe: adding additional aliquots of poorly performing probes was attempted during initial balancing tests of the panels, with little success. One solution is to redesign the probes. In the second multiplex study described (section 6.4), additional smMIPs covering HRAS were added. However, the poor performance is more likely to be related to DNA accessibility rather than probe design. For such problematic regions, there may be a benefit in sonicating or fragmenting the DNA with restriction enzymes prior to capture in order to improve access of the probes to the target DNA.

6.6.2 Identified variants

Initial interest in the overlap between clonal haematopoiesis and SSS resulted from the identification of a *DNMT3A* mutation in Tes14. The two variants encoding changes at

p.R882 were only ever identified in biopsies from this single individual. In the absence of identifying this mutation in other individuals, the significance of this mutation remains unclear. The presence of this mutation in Tes14 may be a result of the presence of blood vessels or somatic contamination in these biopsies. This is a general point of caution in the interpretation of variants that have been typically associated with clonal haematopoiesis. Improving the resolution of variant detection by using smaller testicular biopsies and combining this with statistical prioritisation yielded a high rate of validation. Within the top tier of candidate variants are non-synonymous variants in *DNMT3A*, *TET2*, *ASXL1*, *TP53* and *PPM1D1*, all of which are genes recurrently identified in clonal haematopoiesis studies (Jaiswal et al., 2014; Genovese et al., 2014; Zink et al., 2017; Acuna-Hidalgo et al., 2017). Validation of these variants is required in order to fully understand whether there is a functional overlap between clonal haematopoiesis and SSS.

6.6.3 Recurrent mutations in *SMAD4*

In total, four variants in *SMAD4*, a key mediator of the TGF-β BMP signalling pathway, were identified in 24 testicular biopsies from 3 men. The distribution of mutations observed in the testis does not fully recapitulate the spectrum of germline mutations observed in patients (Figure 6.17). In patients, most mutations have been reported affecting the p.I500 codon (p.I500V=21, p.I500T=14, I500M=2). While the p.I500T mutation is the second most commonly reported mutation in patients (Le Goff et al., 2011), it was only identified once in the testicular biopsies screened. Given that the SMAD4 p.R496C mutation arises at a CpG dinucleotide, it would be expected that this mutation would arise more frequently. Indeed, this is the most common variant identified in the testicular biopsies (figures 6.16 and 6.17; Table 6.11). However, this is not the most common variant associated with Myhre syndrome in patients (Figure 6.17 and Le Goff et al. (2011); Caputo et al. (2012)). This mutation may be subject to reduced penetrance, as evidenced by the observation of this mutation in the general population (gnomAD exome frequency = 8.123e-6).

To broaden these exciting findings, more testicular biopsies should be screened. Screen-

ing smaller biopsies of a testicular slice, as described in section 6.4.2, allowed the identification of a SMAD4 variant (p.R496C) within 2 adjacent biopsies, strongly suggesting that the mutation is present within a single seminiferous tubule. Continuing with such an approach may help strengthen the evidence that *SMAD4* is subject to SSS. Additional evidence needs to be gathered to fully address whether *SMAD4* mutation is associated with SSS. In line with the evidence gathered for the best characterised SSS mutations (e.g. *FGFR2* p.S252W, Table 1.2), this includes: demonstration that the mutations are inherited on the paternal allele and that the father of affected children are older than the population average. This would be best gathered in collaboration with groups who study Myhre syndrome (Dr Tartaglia and Prof Cormier-Daire). Validation of the candidate variants identified in the TGF-β BMP signalling pathway in section 6.4.4 may also add additional weight to the hypothesis that this signalling pathway is vulnerable to SSS.

Chapter 7

Conclusion and Future

7.1 Technologies to identify variants associated with Selfish Spermatogonial Selection

Distinguishing the signal of low-level variation from noise is a significant challenge when working with PCR-based methods and NGS technology (Salk et al., 2018). Thus, while 61 novel variants were identified and validated using a massively-parallel PCR-based discovery screen (RainDance Technologies) (chapter 3), this required significant manual curation, filtering and validation of a long list of candidate variants.

In an attempt to improve sensitivity, two different capture-based technologies incorporating unique molecular identifiers (UMIs) were investigated (chapter 4, HaloPlex HS; chapters 5 and 6, smMIPs). Using UMI based technology provided confidence when calling variants and led to the identification of novel mutations in *ADNP, GNB1, NF1, SMAD4, PPP2R5D* and *TET2*. The flexibility and low-cost of smMIP make them an attractive option for mutation detection. In a comparison of the technologies assessed in this thesis, 12 variants were identified in regions and samples that were analysed using all three technologies. Of these, only one variant (*KRAS* c.35G>A (p.G12D) in Tes4C24) was successfully identified by all approaches. 11 of the 12 were detected by using smMIPs; 6 by PCR (chapter 3) and smMIP; 5 by HaloPlex HS (chapter 4) and smMIP.

A summary of the benefits and limitations of the technologies focused on in this thesis is provided in Table 7.1.

Table 7.1: Comparison of the benefits and limitations of each of the three technologies used in this thesis

Technology	**Benefits**	**Limitations**	**Suggested Usage**
PCR (Chapter 3)	• Single molecule level assessment is possible with emulsion PCR • Easy to test PCR primers to ensure adequate target coverage	• Distinguishing errors introduced by PCR amplification from true variation is a significant and time-consuming challenge	Performs well for a small, targeted screen
HaloPlex HS (Chapter 4)	• UMI incorporation enables error correction • Overlapping amplicons produced by RED step provides additional support to variant call and increases performance of difficult regions • Easy and rapid probe design • Built in analysis software does not require any bioinformatics expertise	• Lack of flexibility in design of probe panel • Inability to improve poor coverage or failed regions • High cost per sample limits throughput • Proprietary software impairs own analysis • Protocol not optimised for use of high input amounts of gDNA, limiting capture efficiency	Performs extremely well for a screen of a small number of samples, particularly if targets include high G/C regions.
smMIP (Chapter 5 and 6)	• UMI incorporation enables error correction • Highly customizable panel • Target both DNA strands independently • Simple experimental workflow • Low per sample cost	• Capture efficiency limits detection of truly low-level variants • Poor performance on hard to capture regions (e.g. G/C rich) • Design of probes requires manual scrutiny	Performs well for a high-throughput discovery screen of mutations with a VAF >0.5%

The most reliable detection of low-level variants (<1%) was achieved when using sequence information from overlapping amplicons or probes. This is possible using all three technologies and is most easily achieved with the HaloPlex HS panel, which produces overlapping gDNA fragments through RED as a step in the experimental protocol (Figure 4.1). Without this, the detection of such low-level variants was extremely challenging and so other technologies may provide a more reliable alternative. The performance of other recent technological developments should therefore be investigated for their ease of use and sensitivity. For example, lockdown probes incorporate UMIs and 5 -biotinylation for capture target enrichment and may offer a benefit to detect low-level variation.

Duplex-Sequencing uses adaptors (with primer sequences and UMI) that are ligated onto the template and amplified using PCR. As with the previously discussed use of UMI, duplex sequencing results in an extremely low error rate. High levels of accuracy are reported with this technology, with estimates that it is more than 10,000-fold more accurate than conventional NGS (Ahn et al., 2015). However, the increased accuracy and error correction capability afforded by all of the discussed technologies is redundant if the capture efficiency of the technology is not high enough to pick up the mutant molecule in a WT background. A study has reported that duplex-sequencing can be used to detect a single point mutation among $>10^7$ nucleotides in a method that used sequential rounds of hybridization with biotinylated oligonucleotides to enrich the target genomic regions (Schmitt et al., 2015). This additionally offers the advantage of circumventing the need for PCR amplification prior to sequencing. The technology warrants investigation to assess whether this reported sensitivity can be achieved in testicular biopsies and if the benefits are sufficient to outweigh the laborious process of library preparation.

7.2 Variants identified in this thesis

Target panels were developed to assess hotspots in candidate genes, including genes involved in cancer, intellectual disability and clonal haematopoiesis. The majority of variants identified cluster within the RAS-MAPK pathway, demonstrating the importance

of this pathway in SSS. The geographic distribution of all validated variants in the testis slices is presented in figure 7.1. A summary of the validated variants and associated variant number is listed in Table 7.2.

I have described the identification and validation of the first variants outwith the RAS-MAPK pathway that are potentially subject to SSS. The most promising is the identification of mutations in *SMAD4*, a key mediator of the TGF-βBMP signalling pathway, that are associated with the developmental disoder Myhre syndrome. In summary, my work has strengthened the association of the RAS-MAPK pathway with SSS. Using refined methods to increase sensitivity and scalability, I also report the first hints that SSS may impact additional signalling processes in the testis.

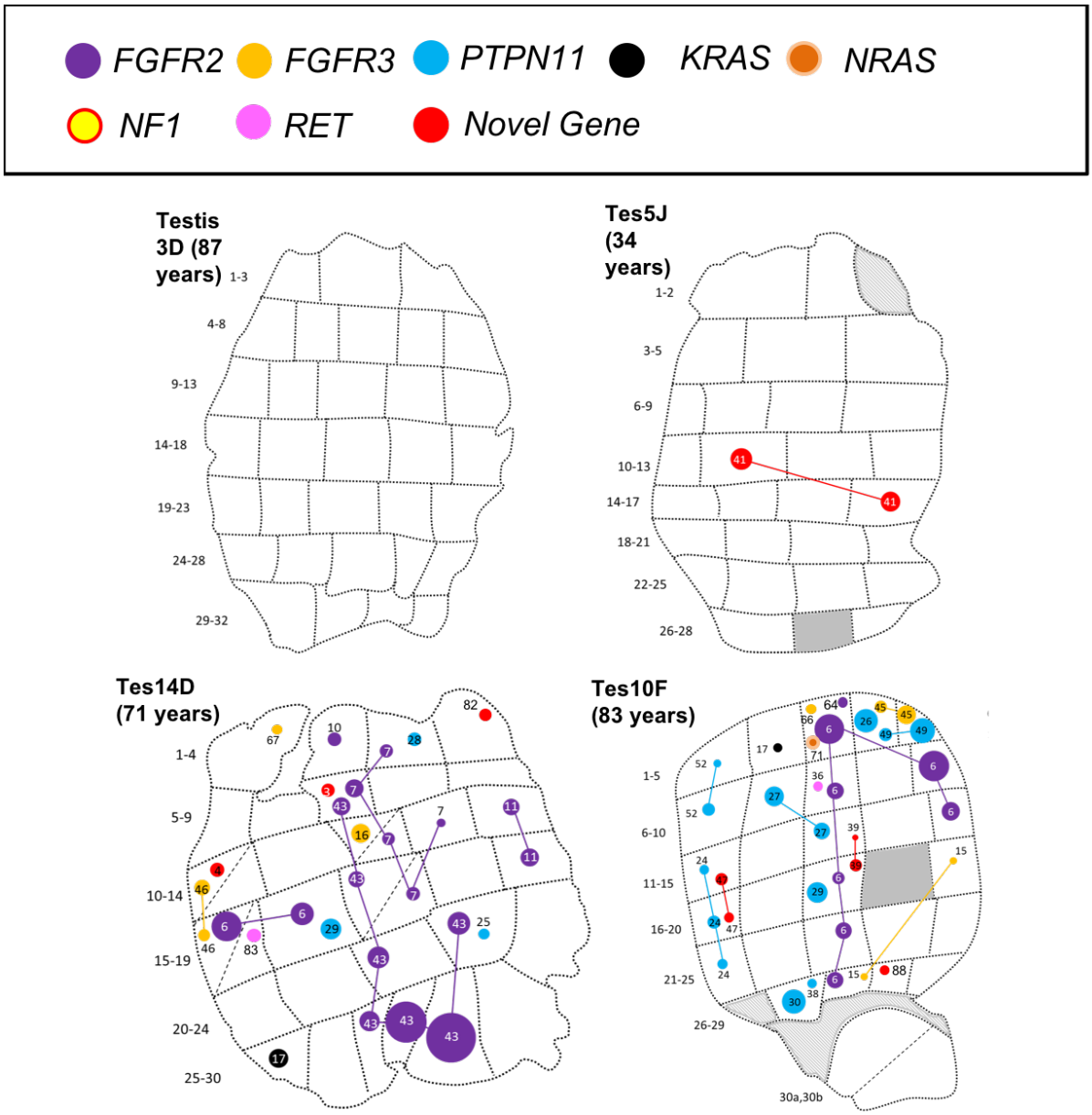
Table 7.2: Summary of validated variants

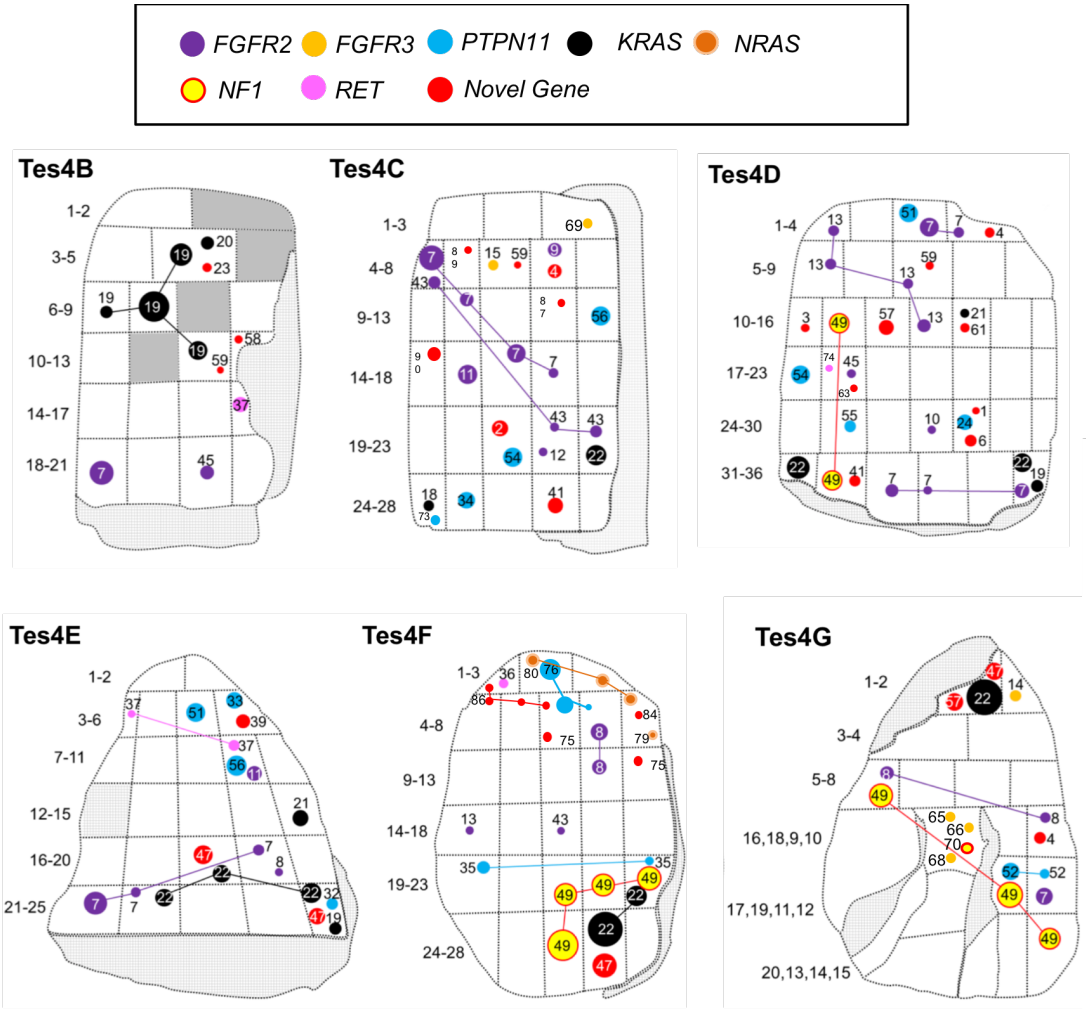
Variant number	**Gene**	**Variant position (hg19) and predicted amino acid substitution**	**COSMIC v82**	**Germline disorder**
1	*AKT3*	chr1:243668575 G>C (p.S472S)	-	-
2	*APC*	chr5:112175762 T>G (p.F1491V)	-	-
3	*BRAF*	chr7:140482928 G>C (p.P403A)	-	-
4	*BRAF*	chr7:140481402 C>G (p.G469A)	52	-
5	*BRAF*	chr7:140449196 T>G (p.Q628P)	-	-
6	*BRAF*	chr7:140439614 G>T (p.Q709K)	-	-
7	*FGFR2*	chr10:123279677 G>C (p.S252W)	54	Apert syndrome
8	*FGFR2*	chr10:123279674 G>C (p.P253R)	9	Apert syndrome
9	*FGFR2*	chr10:123279605 A>C (p.F276C)	1	-
10	*FGFR2*	chr10:123279566 T>G (p.Q289P)	-	Crouzon syndrome
11	*FGFR2*	chr10:123279562 C>A (p.W290C)	-	Pfeiffer syndrome
12	*FGFR2*	chr10:123279562 C>G (p.W290C)	7	Pfeiffer syndrome
13	*FGFR2*	chr10:123274803 G>C (p.S372C)	1	Beare Stevenson syndrome
14	*FGFR3*	chr4:1805519 C>T (p.S344F)	1	-
15	*FGFR3*	chr4:1807371 C>A (p.N540K)	-	Hypochondroplasia
16	*FGFR3*	chr4:1807371 C>G (p.N540K)	-	Hypochondroplasia
17	*FGFR3*	chr4:1807488 G>A (p.V553M)	-	-
18	*KRAS*	chr12:25398284 C>G (p.G12A)	2255	-
19	*KRAS*	chr12:25398284 C>T (p.G12D)	14126	-
20	*KRAS*	chr12:25380282 G>C (p.A59G)	8	-
21	*KRAS*	chr12:25380282 G>T (p.A59E)	6	-
22	*KRAS*	chr12:25380276 T>C (p.Q61R)	115	-
23	*MAP2K1*	chr15:66727455 G>C (p.K57N)	1	-
24	*PTPN11*	chr12:112888197 T>G (p.F71L)	1	Noonan syndrome
25	*PTPN11*	chr12:112888198 G>C (p.A72P)	-	Noonan syndrome
26	*PTPN11*	chr12:112888199 C>A (p.A72D)	11	-
27	*PTPN11*	chr12:112888199 C>G (p.A72G)	2	Noonan syndrome
28	*PTPN11*	chr12:112888199 C>T (p.A72V)	72	-
29	*PTPN11*	chr12:112888202 C>T (p.T73I)	19	Noonan syndrome
30	*PTPN11*	chr12:112888210-112888211 GA>CT (p.E76L)	-	-
31	*PTPN11*	chr12:112888211 A>C (p.E76A)	20	-
32	*PTPN11*	chr12:112888211 A>T (p.E76V)	11	-
33	*PTPN11*	chr12:112924336 G>A (p.V428M)	3	-

34	*PTPN11*	chr12:112924336 G>T (p.V428L)	-	-
35	*PTPN11*	chr12:112926908 C>A (p.Q510K)	3	-
36	*RET*	chr10:43613906 G>C (p.L790F)	-	MEN2A
37	*RET*	chr10:43613906 G>T (p.L790F)	-	MEN2A
38	*RET*	chr10:43615613 G>T (p.D898Y)	-	MEN2
39	*SOS1*	chr2:39250292 T>G (p.Q426P)	-	-
40	*BRAF*	chr7:140453155 C>G (p.D594H)	3	-
41	*BRAF*	chr7:140453132 T>G (p.K601N)	7	-
42	*CBL*	chr11:119148991 G>A (p.C404Y)	15	-
43	*FGFR2*	chr10:123276865 G>C (p.S351C)	-	Pfeiffer syndrome
44	*FGFR2*	chr10:123276893 A>T (p.C342S)	-	Crouzon syndrome
45	*FGFR2*	chr10:123258034 A>C (p.N549K)	10	-
46	*FGFR3*	chr4:1808029 C>G (p.R669G)	-	-
47	*LRP5*	chr11:68115514 C>T (p.A97A)	-	
48	*MAP2K2*	chr19:4110584 A>T (p.C125S)	1	-
49	*NF1*	chr17:29554264 G>A (p.M760I)	-	-
50	*PTPN11*	chr12:112888166 A>C (p.D61A)	1	Noonan syndrome
51	*PTPN11*	chr12:112888166 A>G (p.D61G)	6	Noonan syndrome
52	*PTPN11*	chr12:112891083 G>T (p.E139D)	1	Noonan syndrome
53	*PTPN11*	chr12:112915455 T>G (p.F285C)	-	Noonan syndrome
54	*PTPN11*	chr12:112915523 A>G (N308D)	7	Noonan syndrome
55	*PTPN11*	chr12:112926884-112926885TC>AA (p.S502K)	-	-
56	*PTPN11*	chr12:112926890 A>G (p.M504V)	1	Noonan syndrome
57	*RAF1*	chr3:12645699 G>A (p.S257L)	14	Noonan syndrome
58	*BRAF*	chr7:140453096 C>G (p.L613F)	-	-
59	*MAP2K1*	chr15:66727451 A>C (p.Q56P)	8	-
60	*PTPN11*	chr12:112888162 G>C (p.G60R)	6	-
61	*RAF1*	chr3:12645688 G>C (p.P261A)	-	Noonan syndrome
62	*ADNP*	chr20:49509094 G>T (p.Y719*)	-	Helsmoortel-Van der Aa syndrome
63	*CHD8*	chr14:21866006 T>C (p.D1397G)	-	-
64	*FGFR2*	chr10:123276898 T>C (p.Y340C)	-	Pfeiffer syndrome
65	*FGFR3*	chr4:1806099 A>G (p.Y373C)	697	Thanatophoric dysplasia
66	*FGFR3*	chr4:1806119 G>A (p.G380R)	33	Achondroplasia
67	*FGFR3*	chr4:1807188 G>A (p.V505M)	-	-
68	*FGFR3*	chr4:1807370 A>G (p.N540S)	1	Hypochondroplasia
69	*FGFR3*	chr4:1808988 G>C (p.*807S)	-	Thanatophoric dysplasia
70	*NF1*	chr17:29654664 C>T (p.Q1806*)	2	Neurofibromatosis 1
71	*NRAS*	chr1:115258745 C>G (p.G13R)	180	-
72	*PTPN11*	chr12:112856920 C>T (p.T2I)	-	Noonan syndrome
73	*PTPN11*	chr12:112910785 G>A (p.R265Q)	4	Noonan
74	*RET*	chr10:43614996 G>A (p.V804M)	4	FMTC
75	*SMAD4*	chr18:48604678 A>G (p.I500M)	-	Myhre syndrome
76	*PTPN11*	chr12:112888164_66 TGA>DEL	-	-
77	*MYD88*	chr3:38182641 T>C (p.L265P)	2112	-
78	*NRAS*	chr1:115258748 C>A (p.G12C)	184	-
79	*SMAD4*	chr18:48604678 A>G (p.R496C)	5	Myhre syndrome
80	*NRAS*	chr1:115256529 T>C (p.Q61R)	1616	-
81	*SMAD4*	chr18:48604676 A>G (p.I500V)		Myhre syndrome
82	*CBL*	chr11:119148931 G>A (p.C484Y)	7	-
83	*RET*	chr10:43609948 T>C (p.C634R)	19	MEN2A
84	*KIF1C*	chr17:4925594 C>T (p.R740C)	-	-
85	*NAA10*	chrX:153197817 A>G (p.M98T)	2	-
86	*CBL*	chr11:119148891 T>C (p.Y371H)	27	Leukaemia, juvenile myelomonocytic

87	*BRAF*	chr7:140481403 G>C (p.G469R)	27	Cardio-facio-cutaneous syndrome
88	*BRAF*	chr7:140481411 G>T (p.G466V)	35	-
89	*GNB1*	chr1:1747268 G>A (p.Q44K)	-	-
90	*PPP2R5D*	chr6:42975003 G>A (p.E198K)	2	Intellectual disability

Figure 7.1 *(following page)*: Distribution of validated variants in testis slices (3D; 5J; 10F; 14D; 4B-G). Variants are coloured by the gene (*FGFR2* = purple; *FGFR3* = yellow; *PTPN11* = blue; *KRAS* = black; *NRAS* = orange; *NF1* = yellow with red outline; *RET* = pink; *Novel gene* = red). Modified from diagrams produced by Geoffrey Maher for the work in chapter 3; (Maher et al., 2018).


FGFR2
FGFR3
PTPN11
KRAS
NRAS
NF1
RET
Novel Gene
Testis 3D (87 years)
1-3
4-8
9-13
14-18
19-23
24-28
29-32
Tes5J (34 years)
1-2
3-5
6-9
10-13
14-17
18-21
22-25
26-28
Tes14D (71 years)
1-4
5-9
10-14
15-19
20-24
25-30
Tes10F (83 years)
1-5
6-10
11-15
16-20
21-25
26-29
30a,30b


FGFR2
FGFR3
PTPN11
KRAS
NRAS
NF1
RET
Novel Gene
Tes4B
Tes4C
Tes4D
Tes4E
Tes4F
Tes4G

7.3 Future Work

Seminiferous tubules represent the clonal units in which selfish mutations arise and expand over time. Seminiferous tubules were screened in this thesis with the hypothesis that SSS associated mutations may be more readily detectable at an elevated level. A good compromise in resolution and screening power was achieved using smaller testicular biopsies, with the identification of the first clones in novel genes spreading across multiple biopsies. Biopsying testis samples in this way, combined with UMI-based capture technologies and statistical prioritisation, should be continued to identify additional of novel mutations, genes and pathways that are subject to SSS. Finally, with the knowledge of specific validated variant sites, mutation specific assays can be developed to estimate mutation levels directly in sperm samples. This should be combined with assessment of the origin of the mutation on the paternal allele in affected children. This will help build evidence that novel SSS associated mutations, outwith the RAS-MAPK pathway, have an important role in human disease, potentially resulting in intellectual disability and developmental disorders.

The previously documented SSS associated mutations (Table 1.2) have been proposed to represent extreme examples of a process that may be much more common (Goriely and Wilkie, 2012). For the first time, mutations in pathways other than RAS-MAPK, as well as mutations with a more subtle activating effect, have been identified and validated directly in the male germline. An ongoing challenge will be to establish how many sites are subject to SSS to truly gain an understanding of how positive selection in the male germline shapes our genome and health on a population level.

Appendix A

Appendix

A.1 Testis Table Layout

Tes3D

3D1	3D2	3D2	3D3	3D3
3D4	3D5	3D6	3D7	3D8
3D9	3D10	3D11	3D12	3D13
3D14	3D15	3D16	3D17	3D18
3D19	3D20	3D21	3D22	3D23
3D24	3D25	3D26	3D27	3D28
3D29	3D29	3D30	3D31	3D32

Tes5J

5J1	5J1	5J1	5J2
5J3	5J3	5J4	5J5
5J6	5J7	5J8	5J9
5J10	5J11	5J12	5J13
5J14	5J15	5J16	5J17
5J18	5J19	5J20	5J21
5J22	5J23	5J24	5J25
5J26	5J26	5J27	5J28

Tes10F

10F1	10F2	10F3	10F4	10F5
10F6	10F7	10F8	10F9	10F10
10F11	10F12	10F13	10F14	10F15
10F16	10F17	10F18	10F19	10F20
10F21	10F22	10F23	10F24	10F25
10F26	10F26	10F27	10F28	10F29
10F30a	10F30a	10F30a	10F30b	10F30b

Tes14D

14D1	14D1	14D1	14D2	14D2	14D3	14D4
14D5	14D5	14D6	14D7	14D7	14D8	14D9
14D10 a	14D10 a	14D11	14D12 a	14D12 b	14D13	14D14
14D10 b	14D10 b	14D11	14D12 a	14D12 b	14D13	14D14
14D15 b	14D15 a	14D16	14D17 a	14D17 b	14D18	14D19
14D20	14D20	14D21	14D22	14D22	14D23	14D24
14D25	14D25	14D26	14D27	14D28	14D29	14D30

Tes4B

4B1	4B1	4B2	4B2
4B3	4B4	4B4	4B5
4B6	4B7	4B8	4B9
4B10	4B11	4B12	4B13
4B14	4B15	4B16	4B17
4B18	4B19	4B20	4B21

Tes4C

4C1	4C1	4C2	4C3	4C3
4C4	4C5	4C6	4C7	4C8
4C9	4C10	4C11	4C12	4C13
4C14	4C15	4C16	4C17	4C18
4C19	4C20	4C21	4C22	4C23
4C24	4C25	4C26	4C27	4C28

Tes4D

4D1	4D2	4D3	4D3	4D4	4D4	4D4
4D5	4D6	4D7	4D7	4D8	4D8	4D9
4D10	4D11	4D12	4D13	4D14	4D15	4D16
4D17	4D18	4D19	4D20	4D21	4D22	4D23
4D24	4D25	4D26	4D27	4D28	4D29	4D30
4D31	4D32	4D33	4D34	4D35	4D35	4D36

Tes4E

4E1	4E1	4E2	4E2	4E2
4E3	4E4	4E5	4E6	4E6
4E7	4E8	4E9	4E10	4E11
	4E12	4E13	4E14	4E15
4E16	4E17	4E18	4E19	4E20
4E21	4E22	4E23	4E24	4E25

Tes4F

4F1	4F2	4F2	4F3	4F3
4F4	4F5	4F6	4F7	4F8
4F9	4F10	4F11	4F12	4F13
4F14	4F15	4F16	4F17	4F18
4F19	4F20	4F21	4F22	4F23
4F24	4F25	4F26	4F27	4F28

Tes4G

	4G1	4G1	4G2
4G3	4G3	4G4	4G4
4G5	4G6	4G7	4G8
4G16	4G18	4G9	4G10
4G17	4G19	4G11	4G12
4G20	4G13	4G14	4G15

Figure A.1: Table layout of testicular slices used to create heatmaps of nucleotide frequencies. Blank cells represent a square where no colour is plotted in order to reconstitute the approximate shape of the testis slice

A.2 smMIP Sequences

∗ indicates a smMIP received from Dr Alexander Hoischen

- indicates a smMIP designed by Dr Marie Bernkopf

Gene	Target (hg19)	Strand	Sequence	
ACVR2A	chr2:148676091-148676150	+	ATATCTCACAGGTAGACTNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNTCATA TTGCAGAAACCATGGCT	
ACVR2A	chr2:148676093-148676152	-	CTAGCCATGGTTTCTGCAATATGANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNAGTCTACCTGTGAGAT	
ACVRL1	chr12:52309958-52310017	-	TCCACTTGTAGGACTCAAAGCAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGTAGGGTGGGCCCTCAC	
ACVRL1	chr12:52309964-52310023	+	GCCCACCCTACACAGGGTANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNAGT CCTACAAGTGGACTGACA	
ACVRL1	chr12:52314580-52314639	+	TTAGCAACAGTCCAGAGAAGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTC AGATGATGCGGGAGTGCT	
ACVRL1	chr12:52314582-52314641	-	CCAGCACTCCCGCATCATCTGANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC TTCTCTGGACTGTTGCT	
ADNP	chr20:49510000-49510053	+	GGCATTAGCAGACTGCAGGGAGTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTGACCTAACACTCTGGA	*
ADNP	chr20:49510012-49510065	-	CAGTCACAGGCATCCAGAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCCA GCAAGATACTCCCTGCAG	*
ADNP	chr20:49508723-49508776	+	CCCAGCAACACGCCAGGCTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTCA TGATTTTCAAATAGCCAC	*
ADNP	chr20:49508733-49508786	-	TTGATGCTGAGTGGCTATTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGTG AAAAGTACAAGCCTGGCG	*
ADNP	chr20:49509059-49509112	+	GCCAGACTTGGAGACTGATNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGAA TCACTATCATCATCTAAC	*
ADNP	chr20:49509079-49509132	-	CCCTTACTGAAAAAACGAAAGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA AATGCACCCTCTCGGCTT	*
AGER	chr6:32151637-32151690	+	GGGGGCCCCCTTACACTTNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGTCT CCCTGGAAGTTGGGAGGC	*
AGER	chr6:32151655-32151708	-	AGCGGGGCTCCTGTTGCAGCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTT GGCGAGCCACTGGTGCTG	*
ALG13	chrX:110928231-110928284	+	GCAAAGCAGCTACACAAAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGAG ACTCTGGAAAAAGGAAAG	*

ALG13	chrX:110928241-110928294	-	CTACGAGTGGCTTTCCTTTTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGAGATGACCCTCTTTGTGTA	*
ALPK3	chr15:85400746-85400799	+	GGCGGACCCCAGGTCCCCGGANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNACTCAGAGTCACCCACCAG	*
ALPK3	chr15:85400733-85400786	-	CTGAGTGGGGGTCCCCAGCCCCANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCGCCCCTCCACATCTGG	*
ASXL1	chr20:31022891-31022944	+	CAATGGTCCCATTCTGTCTCTAGTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTGAATGTGAGTCTGG	*
ASXL1	chr20:31022903-31022956	-	CAGGAAGTGGTGCCAGACTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTCATCTCCCACTAGAGACAGA	*
ASXL1	chr20:31021200-31021253	+	CCAGAGCCAGAAGGAATCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGTGCGTATACAGCGTGGTCCAG	*
ASXL1	chr20:31021185-31021238	-	TACGCACTGATTCTCCTGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTCTGGTTCGGAGATCTGGCC	*
ASXL1	chr20:31021228-31021281	+	ACAGGAGTCAGAACAAGCAGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCAGCGAGATGGGCATTTTAA	*
ASXL1	chr20:31021238-31021291	-	AGAGCGTTTCTTAAAATGCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTTAGCAACCCCTGCTTGTTC	*
ASXL1	chr20:31022261-31022314	+	CACGGAGTCCTCCTGCCGGGGTTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCACGTATCAAACCACC	*
ASXL1	chr20:31022246-31022299	-	GAAAGTTGAATCTGCAAAAAAAGANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGTGGGGATGATCCGG	*
ASXL1	chr20:31022573-31022626	+	GGAGCATACCCAGGCCGGAACTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCCCGAGCACCCCTGGAAA	*
ASXL1	chr20:31022584-31022637	-	CTGACGTACACTTTCCAGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTGGACATGGCAGTTCCGGCCT	*
ASXL1	chr20:31022386-31022445	+	GGGGTGGCCCGGGTGGAGGTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTAAAGCCCGTGCTCTGCAGG	
ASXL1	chr20:31022417-31022476	-	CTATGGCAGTGGTGACCTCTCGCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCACCTCCCTCATCGGTG	
BCL9L	chr11:118772398-118772451	+	TCCGCCTCACCCTGCCCANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTGCATGCCCATCGACCGCTTC	*
BCL9L	chr11:118772407-118772460	-	GCTGCTGGAGAAGCGGTCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTGCCCTACCCAGGTGGGCAGGG	*
BMPR1B	chr4:96073851-96073910	-	AAGAGGGGTCACTGGGCACTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCCTTACCTCATCACTGCTCC	
BMPR1B	chr4:96073860-96073919	+	ATGAGGTAAGGCTTGAGGTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGTGACCCCTCTTATGAGGACA	

BMPR1B	chr4:96075722-96075781	+	AACACTTGCCAAAATGTCAGAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN AAGGCAGATGGGAAAACT	
BMPR1B	chr4:96075722-96075781	-	AGTTTTCCCATCTGCCTTAGACNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG ACATTTTGGCAAGTGTT	
BMPR2	chr2:203419986-203420045	-	GTGACAGGTTGCTTAAAGANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCAG TATGATGGATAGAGTCTT	
BMPR2	chr2:203419985-203420044	+	GAAGACTCTATCCATCATACTGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT CTTTAAGCAACCTGTCA	
BRAF	chr7:140481373-140481426	+	GTCCCACTGTAATCTGCCCATCAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCACCACATTACATACT	*
BRAF	chr7:140481386-140481439	-	AAGTGGCATGGTAAGTATGTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGG GAGATTCCTGATGGGCAG	*
BRAF	chr7:140494133-140494186	+	CGATGATCTTCATCTGCTGGTCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT TGACAGGTTCTATTGTG	*
BRAF	chr7:140494148-140494201	-	CAATGTGCATATAAACACAATNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC ACAGCCCTTCCGACCAGC	*
BRAF	chr7:140453103-140453156	+	CCTATTTTTACTGTGAGGTCTTCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TGGATCCAGACAACTGT	*
BRAF	chr7:140453115-140453168	-	CCATCAGTTTGAACAGTTNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNTATTT CTTCATGAAGACCTCAC	*
BRCC3	chrX:154305462- 154305515	+	AGTAGAAATTTCTCCAGAGCAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GGTTGATGCCGTCAGAAT	*
BRCC3	chrX:154305473- 154305526	-	GAATGTGAACAATTCTGACNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTG CAGACAGCTGCTCTGGAG	*
CBL	chr11:119148888- 119148941	+	GATAAGGATGTAAAGATTGAGCCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCAAAGGAACAATATGAA	*
CBL	chr11:119148879- 119148932	-	TTCCTTTGATTAAAAAAAATTAAANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCCTTATCATTTTCAGC	*
CBL	chr11:119148987- 119149040	+	GTAATAACCTTGGAAAATTCGGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GGACACCTCATGTGCACA	*
CBL	chr11:119148977- 119149030	-	AGGTGTCCACAGGGCTCAATCTNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GTTATTACATAGCTGAAA	*
CBL	chr11:119148937- 119148990	+	GTCTTACATCCTGGCAGGTACGGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NATGTAAAATATGTGCTG	*
CBL	chr11:119148928- 119148981	-	TTTTACATAGTTGGAATGTGGAGCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGTAAGACAGGATGTGC	*
CBL	chr11:119149232- 119149285	+	GATCCGTTTGATCCTAGAGGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTT CTGCAGGAATCAGAAGGT	*

CBL	chr11:119149232-119149285	+	GATCCATTTGATCCTAGAGGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTT CTGCAGGAATCAGAAGGT	*
CBL	chr11:119149194-119149247	-	AGATGCATCTGAAAATACTTAANNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CTTTAATTTCACATCGGC	*
CBL	chr11:119149242-119149295	-	GACAGCCCTGACCTTCTGANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGGC TGCCACTCCCTCTAGGAT	*
CLCN4	chrX:10188842-10188895	+	AGTGCCTGGTGACGCGGAGCGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA GCCCGTTTACAGTGACAG	*
CLCN4	chrX:10188860-10188913	-	TTTCCATCGGAGTGTGGTCTGTCANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCCGGCTACTCACCCGC	*
COL4A3BP	chr5:74722228-74722281	-	GCTACTCTGCAACATCCACCTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGA ATCTGGATATGGATCTG	*
COL4A3BP	chr5:74722256-74722309	+	AAAAAAGTAAACTATGTGAAAAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGACACCAGGGACACCAT	*
CTNNB1	chr3:41275230-41275283	+	GCAGTTCGCCTTCACTATGGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAC ATCACTGAGCCTGCCATC	*
CTNNB1	chr3:41275219-41275272	-	CAGTGATGTCTTCCCTGTCACCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG AACTGCATTCTGGGCCA	*
CTNNB1	chr3:41278072-41278125	+	CAAGATTACAAGAAACGGCTTTCANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGCCTTTTTCTCCTTGT	*
CTNNB1	chr3:41278084-41278137	-	TGTCGCTAAGAGACAAGGANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCAG CTCAACTGAAAGCCGTTT	*
CUX2	chr12:111748319-111748372	+	AGCCCTGGCGCAAGCTCACGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG GGCAGCGGGTGTTTGGGC	*
CUX2	chr12:111748350-111748403	-	ACCGAGCCCTGCGACAGCCCCAGCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCTTGATGAAGGGCTCC	*
DEAF1	chr11:686862-686915	+	CTGCTGGCTCTTCCTGCCATGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG GTCACGGACGATACCTGG	*
DEAF1	chr11:686837-686890	-	GTGCTCAGCACACAGTTCAGACGTNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGAAAAGAAGCATTCGC	*
DHDDS	chr1:26784338-26784391	+	GGCAGGTAGGTCATTTCCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTATA CCAACCGCTCTCCTCATC	*
DHDDS	chr1:26784348-26784401	-	AAGATGTCAGGATGAGGAGAGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT AATAGTACTTGGAAATGA	*
DNMT3A	chr2:25457257-25457310	+	GGGCAGGGTGAGTGCTGGCCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCC AAGCGGCTCATGTTGGAG	*
DNMT3A	chr2:25457267-25457320	-	ATACTGACGTCTCCAACATNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCAGC CTGGTCTGGCCAGCACT	*

DNMT3A	chr2:25463189-25463242	+	GGGGCGATCATCTCCCTCCTTGGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCGAGAAATCGCGAGAT	*
DNMT3A	chr2:25463177-25463230	-	CTCGAGGTATAGCCAGCAANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGAT CGCCCCTTCTTCTGGCTC	*
DNMT3A	chr2:25462038-25462091	+	ACAGGAGACGGTTTGAAGATGAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCCCAGAAGTAGCGGGCC	*
DNMT3A	chr2:25462051-25462104	-	CAGCTGCACACAGGGCCCGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTC CTCCTTGGCTCATCTTCA	*
DNMT3A	chr2:25470463-25470516	+	GCTCCGGCCCGTCATCCACCAAGANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCCCCACAACTTACCAC	*
DNMT3A	chr2:25470478-25470531	-	GACGGCAAATTCTCAGTGGTANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC CGCATTGTGTCTTGGTGG	*
DNMT3A	chr2:25463154-25463207	+	CCCATGGCCACCACATTCNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNCCATT AGTGAGCTGGCCAAACC	*
DNMT3A	chr2:25463141-25463194	-	CACTAATGGCTTCTACCTTGGACNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TGGGCGTTAGTGACAAG	*
DNMT3A	chr2:25457199-25457252	+	GGAGACGTCAGTATAGTGGACTGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNAGAGGTGGCGGATGAC	*
DNMT3A	chr2:25457216-25457269	-	GGTCATGGAGCGTGCCAGTCANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG TATTTGGTTTCCCAGTCC	*
DNMT3A	chr2:25457216-25457269	-	GGTCATGGAGCGTGCCGGTCANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG TATTTGGTTTCCCAGTCC	*
DNMT3A	chr2:25458569-25458622	+	GACAGGAAAATGCTGGTCTTTGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CCAGATGCCAGCACAACC	*
DNMT3A	chr2:25458580-25458633	-	AAAGGTACCCGGGTTGTGCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTAA AGCAGGGCAAAGACCAGC	*
DNMT3A	chr2:25463255-25463308	+	GCCAGTGCCCTCTGAGAGGTCGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GAAGAAGGGGCGATCATC	*
DNMT3A	chr2:25463280-25463333	-	GCATGATGCGCGGCCCAAGGAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CCAGCTGATGGCTTTCTC	*
DNMT3A	chr2:25467408-25467467	-	TGAGGCTGTTGTGGCCTCCAGTGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCGGCTACCAGTCCTAC	
DNMT3A	chr2:25467408-25467467	+	GTAGGACTGGTAGCCGTCGTCGTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAGGCCACAACAGCCTCA	
DNMT3A	chr2:25457133-25457192	+	GATGACTGGCACGCTCCATGACCGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGTGTCGCTACCTCAGT	
DNMT3A	chr2:25457140-25457199	-	GGGGCAAACTGAGGTAGCGACACANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGGTCATGGAGCGTGCC	

DNMT3A	chr2:25469921-25469980	-	GGTGAGTGTCCCTGCTGGGANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAG CTCGTTTTGCAGTGCGTT	
DNMT3A	chr2:25469929-25469988	+	GCAAAACGAGCTCAGCGGCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNAGG GACACTCACCTGCAGGAC	
DNMT3A	chr2:25467025-25467084	-	TGTGAGTGCTGGGCCTGGGGCGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTACGGGCTGCTGCGGCG	
DNMT3A	chr2:25467028-25467087	+	CGCAGCAGCCCGTAGGTACCCTTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGCCCAGCACTCACAAAT	
DNMT3A	chr2:25466744-25466803	-	TGCTGGCCGCCTTGTCCCAGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA GCTGAGAAGAGGAAGCCC	
DNMT3A	chr2:25466750-25466809	+	CCTCTTCTCAGCTGGGACAGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA AGGCGGCCAGCACCTCTT	
DNMT3A	chr2:25463495-25463554	+	GGGACTGCCCCCAATCACCAGANNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TGAGAAGGTGGAGGGGAC	
DNMT3A	chr2:25463531-25463590	-	CCTGCTCGCAAGGGCCTCTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTTA TCCTCCCAGATCCAGGAG	
DNMT3A	chr2:25464509-25464568	-	TCCATCACGGTGGGCATGGTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTG GCTGGTGCAGGGCTCCTG	
DNMT3A	chr2:25464515-25464574	+	CCCTGCACCAGCCAGCAGACAGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NACCGTGATGGAGTCCTC	
DNMT3A	chr2:25470526-25470585	-	GGCCGGAGCCGAGCAGCTGAANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT GGCATTGGGGAGCTGGTG	
DUSP15	chr20:30450452-30450505	+	CGCTGTCACAATCGTGGTGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGTT GGCGATGGGCCTGGTGGC	*
DUSP15	chr20:30450464-30450517	-	GAAGCCATCAAGGCCACCAGGCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN ATCTCTCGCAGCACCACG	*
FBN1	chr15:48707883-48707942	+	GCGCTGAGGCATTCGTTTTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTGT TCATACTGGAAGCCGGCG	
FBN1	chr15:48707862-48707921	-	GTTCAGTGGAGGATGCCAAGACANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NATCTGCGGAGGAGCCTC	
FBN1	chr15:48782168-48782227	+	TGCACTCCTCGTCCTCGTACCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTC GCATTCCTCAGTACCCC	
FBN1	chr15:48782176-48782235	+	TCGTCCTCGTACCTCAGGNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNTTCCT CAGTACCCCAGGCTGCC	
FGFR2	chr10:123279627- 123279680	+	GCTCTGGTGGAGAGAGGGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNAACT CTACGTCTCCTCCGACCA	*
FGFR2	chr10:123279673- 123279726	-	CACCGGCCCATCCTCCAAGCCGGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TTATTTATTGGTCTCTC	*

FGFR2	chr10:123276890-123276949	+	CTTTGTCCGTGGTGTTAACACCGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNATAGAATTACCCGCCA	
FGFR2	chr10:123276889-123276948	-	GGCGGGTAATTCTATTGGGATATNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAACACCACGGACAAAGA	
FGFR2	chr10:123258015-123258074	+	CCATCTCTGACACCAGATCAGAANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTCCTACTCACCATCCT	
FGFR2	chr10:123258025-123258084	-	GCCTGCACACAGGATGGTGAGTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGACCTTTCTGATCTGGTG	
FGFR2	chr10:123279555-123279614	+	CTCCTCCGACCACTGTGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTTACTGCCGTTCTTTTCCACGT	
FGFR2	chr10:123279543-123279602	-	ACGGCAGTAAATACGGGCCCGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGAGGAGACGTAGAGTTTG	
FGFR3	chr4:1806082-1806135	+	GTTCATCCTGGTGGTGGCGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGAGGAGCTGGTGGAGGCTGA	*
FGFR3	chr4:1806092-1806145	-	GCCCGCCTCGTCAGCCTCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAGAGCGTCACAGCCGCCACCAC	*
FGFR3	chr4:1807841-1807894	+	ACCAACGTGAGCCCGGCCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCGAGGACAACGTGATGAAGATC	*
FGFR3	chr4:1807870-1807923	-	TGCACGTCCCGGGCCAGCCCGAAGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGCATGACCCCCACCC	*
FGFR3	chr4:1803523-1803576	+	GGCCCATCCTGCAGGCGGGGCTGCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAGTGGCGGTGGTGGTG	*
FGFR3	chr4:1803535-1803588	-	CACCCCCTCCCTCACCACCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGTTGGCCGGCAGCCCCGCCT	*
FLT3	chr13:28592602-28592661	-	GAGGCTGCTATTTCCTACTTANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGAAAGTGGTGAAGATATG	
FLT3	chr13:28592605-28592664	+	ATCTTCACCACTTTCCCGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAGTAGGAAATAGCAGCCTCACA	
GNAS	chr20:57484390-57484443	+	GAGACCAAGTTCCAGGTGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTTTGGTGAGATCCATTGACC	*
GNAS	chr20:57484414-57484467	-	AAGCAGGTCCTGAAACAAAATTGANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCTTACTGGAAGTTGAC	*
GNB1	chr1:1747228-1747281	+	GGGTCGATGTTGTTTGTGATCTTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGTGCATGGCGTAGATC	*
GNB1	chr1:1747219-1747272	-	CATGCACTGGGGCACAGACTCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNACATCGACCCAGTGGGAAG	*
GRIA1	chr5:153144040-153144093	+	AGGATGGTGTCTCCCATTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGCGTCTGGTGGTTCTTCACC	*

GRIA1	chr5:153144058-153144111	-	TGAGGAGATGATGATTAAGGTGAANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGGTCCTCTGCACTCTC	*
HECTD1	chr14:31576370-31576423	+	ATATCCAGGAGGTTTCAATCCACCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTGATCCTTTCAAGCTC	*
HECTD1	chr14:31576383-31576436	-	ACAGGATAGTGATGAGCTTGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGA TCTTGGAGGTGGATTGAA	*
HRAS	chr11:534241-534294	+	CCACCACCACCAGCTTATNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNGTGGG GTCGTATTCGTCCACAA	*
HRAS	chr11:534282-534335	-	GGGCAAGAGTGCGCTGACNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTAGG AGGACCCCGGGCCGCAGG	*
HRAS	chr11:534235-534314	-	ACGAATACGACCCCACTATAGAGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGAGGAGCGATGACGGA	-
HRAS	chr11:534263-534349	+	TCCTCCTACAGGGTCTCCTGCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGG TTCTGGATCAGCTGGAT	-
HRAS	chr11:533874-533931	-	GGAGGAGTACAGCGCCATGCGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC TGCAGGATTCCTACCGGA	-
HRAS	chr11:533874-533928	+	GCTTCCGGTAGGAATCCTGCAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN ATGGCGCTGTACTCCTCC	-
IDH1	chr2:209113072-209113131	-	GTGATTTTGCATGTTTTTTNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCGGC TTGTGAGTGGATGGGTA	
IDH1	chr2:209113080-209113139	+	CACTCACAAGCCGGGGGATATTNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GCAAAATCACATTATTGC	
IDH2	chr15:90631915-90631968	+	CAGCTTGAACTCTGTGAGGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGAT GATGGGCTCCCGGAAGAC	*
IDH2	chr15:90631928-90631981	-	CCTGGGGGGGACTGTCTTCCGGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TATTATCTCTGTCCTCAC	*
IDH2	chr15:90631804-90631863	+	GCCAGGGACTAGGCGTGGGATGTTNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGTCAGTGGATCCCCTC	
IDH2	chr15:90631797-90631856	-	TCCACTGACCTGGGCACCCCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCCT AGTCCCTGGCTGGACCA	
ITPR1	chr3:4856190-4856243	+	GAGGTAAATTAATCCCGAGGGGAANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNTGTCACTGTGCTGAGT	*
ITPR1	chr3:4856176-4856229	-	CAATGCACATCAGCAGCGTCTCACNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTCTTTGGACGGCTTCC	*
ITPR1	chr3:4856176-4856229	-	CAATGCACATCAGCAGCGTCTCACNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTCTTTGGACGGTTTCC	*
JAK2	chr9:5073728-5073781	+	GAGAGTAAGTAAAACTACAGGCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GCAGCAAGTATGATGAGC	*

JAK2	chr9:5073748-5073801	-	AAATGCTTGTGAGAAAGCTTGCTCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNAAAGGCATTAGAAAGC	*
KCNH1	chr1:210977447-210977500	+	GAGAAGTGCTAGAGGTGAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCAT CTCATGGTATCTGTTGGT	*
KCNH1	chr1:210977458-210977511	-	TGTATGCCAACACCAACAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTTT GCTCCTCCTCCTCACCTC	*
KCNH1	chr1:210977458-210977511	-	TGTATGCCAACACCAACAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTTT GCTCCTCGTCCTCACCTC	*
KCNH1	chr1:211093343-211093396	+	CAGGCTGCTGATGCCCTGGGANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA TTCAATGTAGTGGTCCAG	*
KCNH1	chr1:211093355-211093408	-	GTGGCCCGTAAGCTGGACCANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTG TTCCTCTTCTCCCAGGGC	*
KCNH1	chr1:211093355-211093408	-	GTGGCCCGGAAGCTGGACCANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTG TTCCTCTTCTCCCAGGGC	*
KCNQ3	chr8:133192489-133192542	+	ACCACTGGCACAGAGGCAATCAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTGCGCAGGATCTGCAGG	*
KCNQ3	chr8:133192479-133192532	-	CTGCGCATGCTGCGGATGGANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTG CCAGTGGTTGCTGTGGGA	*
KIF1C	chr17:4925589-4925642	+	GCTACGAGGTGGCCCTGGCTGACTNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNAGGCTGCAGGGCAAAG	*
KIF1C	chr17:4925569-4925622	-	CGCTGGGGGATCTGATAAACCCTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTCTCCTTCACCGCCTGC	*
KRAS	chr12:25398233-25398286	+	GCTCCAACTACCACAAGTNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNTATTG TTGGATCATATTCGTCC	*
KRAS	chr12:25398260-25398313	-	GATACAGCTAATTCAGAATNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNATA AGGCCTGCTGAAAATGAC	*
KRAS	chr12:25398196-25398275	-	TTAATATGCATATTACTGGTGCAGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGGAGCTGGTGGCGTAG	-
KRAS	chr12:25398191-25398278	+	CGCCACCAGCTCCAACTACCANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCT GCACCAGTAATATGCAT	-
KRAS	chr12:25398274-25398353	-	AAGAGTGCCTTGACGATACAGCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CATGTTCTAATATAGTCA	-
KRAS	chr12:25380227-25380281	+	GCTGTGTCGAGAATATCCAAGAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGCAAATACACAAAGAAA	-
KRAS	chr12:25380252-25380330	+	GCTTCCTGTAGGAATCCTGAGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC CCAGTCCTCATGTACTGG	-
KRAS	chr12:25380294-25380353	-	TCTCGACACAGCAGGTCAAGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAT CCAGACTGTGTTTCTCCC	-

KRAS	chr12:25380293-25380353	+	GGGAGAAACACAGTCTGGATTNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC TTGACCTGCTGTGTCGAG	-
MAP3K7	chr6:91269785-91269838	+	GCTGTGAAGATAAGCCACTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGGC TGAGGGGTGTCACATTAT	*
MAP3K7	chr6:91269805-91269858	-	AACCACCAAAGTAAGTTATAATAANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTTACAGTGTTCCCAAG	*
MYD88	chr3:38182285-38182338	+	GGTAAGCTCAACCCTGCTNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGGTG GTTGTCTCTGATGATTAC	*
MYD88	chr3:38182294-38182347	-	GCTCTGCAGGTAATCATCANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTTC TCTTGCCAGAGCAGGGTT	*
MYD88	chr3:38182614-38182667	+	AGAAAGAGTTCCCCAGCATCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAT GCCAGCATGGCACCCCTT	*
MYD88	chr3:38182624-38182677	-	CCTGCAAGCCAAGGGGTGCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNATG AACCTCAGGATGCTGGGG	*
MYH13	chr17:10258231-10258284	+	CGTCACTCTGGATTTTTCTAACAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTTAGTTCTGGCTTCTT	*
MYH13	chr17:10258207-10258260	-	CTCTTGGAATACCTTTCTTATNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAT TATCCAGTGAGAGAAGC	*
NAA10	chrX:153197834-153197887	+	GTGGGAGAAGAGCAGAGATGGGGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NATGGCTCGAGAGGCCTG	*
NAA10	chrX:153197813-153197866	-	GAGAACTTCAATGCCAAATATGTCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNTGTGAAGCGTTCCCAC	*
NRAS	chr1:115258695-115258748	+	TGCTCCAACCACCACCAGNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNCTATG GTGGGATCATATTCATC	*
NRAS	chr1:115258719-115258772	-	CAGCTAATCCAGAACCACTTTGTANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGTGTGAAATGACTGAG	*
NRAS	chr1:115256481-115256534	+	GCTGTATCCAGTATGTCCAACAANNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CAAATACACAGAGGAAG	*
NRAS	chr1:115256525-115256578	-	GAGTACAGTGCCATGAGAGACCAANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNTTACAGAAAACAAGTG	*
NRAS	chr1:115258652-115258731	-	CGCTGACCTGATCCTGTCTCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGCA GGTGGTGTTGGGAAAAG	-
NRAS	chr1:115258719-115258781	-	CAGCTAATCCAGAACCACTTTGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG GTTCTTGCTGGTGTGAA	-
NRAS	chr1:115258732-115258785	+	CACCAGCAAGAACCTGTTGGAAANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGCTGGATTGTCAGTGCG	-
NRAS	chr1:115256452-115256531	+	CCAGCTGTATCCAGTATGTCCANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT ATCCGCAAATGACTTGC	-

NRAS	chr1:115256519-115256605	+	GTGTGGAGGGTAAGGGGGCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGTA TTGGTCTCTCATGGCACT	-
PACS1	chr11:65978641-65978694	+	AAGACCTTGGCCGTGGGACNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNAAG CGAGATGCCAACAAGCTG	*
PACS1	chr11:65978651-65978704	-	GCATGATCTGCAGCTTGTTGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCA TGTTGATGAGTCCCACGG	*
PACS2	chr14:105834413- 105834466	+	GACGCCGTGCAGGGGCAGGTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTG TCTGTTCATTCCGCAGAT	*
PACS2	chr14:105834431- 105834484	-	CTCCTCCTCGGAGTAGTTANNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGGAG CCCCGGCCCCAGGTCAC	*
PCGF2	chr17:36895844-36895897	+	GTTGGTCTCCAGGTAGCGCACGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GTGCCAAGCCCACCTGAT	*
PCGF2	chr17:36895834-36895887	-	CTTGGCACGGCACCTCCTGTTCCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GACCAACAAATACTGCC	*
PCGF2	chr17:36895834-36895887	-	CTTGGCACGGCACCTCCTGTCCCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GACCAACAAATACTGCC	*
PDCD11	chr10:105184841- 105184894	+	GACAGGGTGTCTCCCTAACCCTCANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCTGGCAGCTTTCTCCC	*
PDCD11	chr10:105184853- 105184906	-	GGTGAGAGGTCAGGGAGAANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTG TGGTCTTGAGGGTTAGGG	*
PIK3CA	chr3:178916710-178916763	+	AAAGAAGCAAGAAAATACCCCCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CCAAATGGAATGATAGTG	*
PIK3CA	chr3:178916722-178916775	-	GCATTCTAAAGTCACTATCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNAAGT TGATGGAGGGGGTATTT	*
POGZ	chr1:151378011-151378064	+	GTCTGCAGGGCATTCTCCTTTCGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGGGTGGGAAGGACAGTG	*
POGZ	chr1:151378048-151378101	-	CTTGGTGTGATGTAGTCCTAGCCANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCTGGATACAGAGGTGC	*
POLR3F	chr20:18461061-18461114	+	GAAGTGTGGAAATATATCTGCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA TAGGCAGAAACAGCACGA	*
POLR3F	chr20:18461076-18461129	-	GTTCTGTTTGCTTTCTCGTGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTG ATTCCCAATTCGCAGAT	*
POLR3F	chr20:18461076-18461129	-	GTTTTGTTTGCTTTCTCGTGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTG ATTCCCAATTCGCAGAT	*
PPM1D	chr17:58740422-58740481	+	GGTGTAGTCATACCCTCANNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTATA CCTGCCCTGGTTCGTAGC	
PPM1D	chr17:58740422-58740481	-	GCTACGAACCAGGGCAGGTATATGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNAGGGTATGACTACACC	

PPM1D	chr17:58740561-58740620	+	ACCAAAAAAATTTGAAGATGTCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CTTTAAGGATACATGATT	
PPM1D	chr17:58740518-58740577	+	CCAATTGGCCTTGTGCCTACTAANNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN AAGATCCAGAACCACTT	
PPP2R1A	chr19:52715958-52716011	+	GCCAAGGTGCTGGAGCTGGANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGG TACTTCCGGAACCTGTGC	*
PPP2R1A	chr19:52715945-52715998	-	AAGTACCTGGGAGTCGGGGACGGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGGCAAACTCCCCCAGC	*
PPP2R5D	chr6:42974969-42975022	+	AGCTGCTTGGCCACATCTCCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTTC CGGACGCTGCCACCTTC	*
PPP2R5D	chr6:42974979-42975032	-	GGGATTCGATGAAGGTGGCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCCC TGGTACCTGGAGATGTGG	*
PRB4	chr12:11461739-11461792	+	GTGGGCGTCGTCCTTCTGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTGGT TTCCTCCTTGTGGGGGTG	*
PRB4	chr12:11461739-11461792	+	GTGGGCGTCATCCTTCTGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTGGT TTCCTCCTTGTGGGGGTG	*
PRB4	chr12:11461721-11461774	-	ACCAGTCCCAAGGTCCCCCANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNACA AGGAGGAAACCAGCCCC	*
PTEN	chr10:89692869-89692922	+	GTGCATATTTATTACATCGGGGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAAGTGAAGATGACAATC	*
PTEN	chr10:89692902-89692955	-	CCTTTCCAGCTTTACAGTGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNAATC TAGGGCCTCTTGTGCCT	*
PTPN11	chr12:112888146-112888199	+	CACTTTGGCTGAGTTGGTCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNAAA TGGAGCTGTCACCCACAT	*
PTPN11	chr12:112888158-112888211	-	TTCTGAATCTTGATGTGGGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTCC ATGTAATACTGGACCAAC	*
PTPN11	chr12:112926892-112926972	+	AGAGCAGGTACCAGCCTGAGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCG GTCTCAGAGGTCAGGGAT	-
PTPN11	chr12:112926927-112926991	-	TATAGATAAATCGGTACTGTGCTNNNNNCTTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAATCCGCATGCCAGCCC	-
PTPN11	chr12:112888134-112888203	-	GCTCCATTTCTTCTAAAATAGTCCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGGACCAACTCAGCCAA	-
PTPN11	chr12:112888111-112888164	+	GATTACTATGACCTGTATGGAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TTGCCTCCCTTTCCAATG	-
PTPN11	chr12:112888277-112888333	-	GCTCAATGACATCTCCATTCTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAG CAGCAGACTTTGTGGTC	-
PTPN11	chr12:112915476-112915555	+	GTTTTCTGACCATACATTTCTAGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCATACCAGGGTTGTCC	-

PTPN11	chr12:112915483-112915542	-	TCGTGTAGGACAACCCTGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCAGAAAACACTGTGAAAAGCA	-
PURA	chr5:139494432-139494485	+	CGGCGTGTTTATGCGAGTGAGCGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGAGCTGCCCGAGGGCAC	*
PURA	chr5:139494444-139494497	-	ACAGTCAAGGAGGTGCCCTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGGCTTCACCTCGCTCACTCGC	*
RAB14	chr9:123952915-123952968	+	ACCAAATTCAACACCAATTGTGTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNATCGCTCCTGTCCTGC	*
RAB14	chr9:123952891-123952944	-	GTTACACGGAGCTACTACAGAGGANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNAATCGAAGTTAGTGGC	*
RB1CC1	chr8:53568661-53568714	+	TGCTCCTGGCTGCTGACCAATTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAATTCTTTTAGGGCAGTC	*
RB1CC1	chr8:53568672-53568725	-	GAAGCTATTCAGACTGCCCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGACAGACAAAAATTGGTCAGC	*
RET	chr10:43609935-43609988	+	CGTCTCGGTGCTGCTGTCTGCCTTNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCACCCACAGATCCACT	*
RET	chr10:43609926-43609979	-	GTGGGTGGGGGTGGTGTGAGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCACCGAGACGATGAAGGAG	*
RET	chr10:43617395-43617448	+	CCACGCAAAGTGATGTGTAAGTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTCCATCTTCTCTTTAGG	*
RET	chr10:43617407-43617460	-	CTGGAATCCGACCCTAAAGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGCAACACCCACACTTACACAT	*
RHEB	chr7:151187994-151188047	+	GTAGGAGTCCACAAATTGGCCTTCNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTTACACAAAAGAAGGC	*
RHEB	chr7:151188026-151188079	-	AGTATTGTTTTCAAGTACTTAANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNAATCCTCATTGACGATTC	*
SCN2A	chr2:166198934-166198987	+	ACTGGGAGTGTTCATAAAATGTACNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNGTTTAATGGAACTTGG	*
SCN2A	chr2:166198954-166199007	-	ATCCTTCCACATTTGCCAAACCAANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNATTAATTACAAAGTAC	*
SCN2A	chr2:166201269-166201322	+	GGAGAGTGGATAGAGACCATGTGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNATTGTGAACTCCCACGC	*
SCN2A	chr2:166201283-166201336	-	TCATGCATGTGCCAGCGTGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNCCATACAGTCCCACATGGTC	*
SCN3A	chr2:165986705-165986758	+	TAGCATATTTAGTGTGGGCCAGGNNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNATGAAGACGATGATGGC	*
SCN3A	chr2:165986717-165986770	-	ACCTTGGTGTTGGCCATCANNNNNCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNTTGGCAAAATCCTGGCCCACA	*

SCN8A	chr12:52082531-52082584	+	CGCACTTTCAGGGTACTGAGGGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NACAAGTTAACTTTGGTT	*
SCN8A	chr12:52082543-52082596	-	CCTGCAGAATCAAACCAAANNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNAGTT TTCAAAGCCCTCAGTAC	*
SETBP1	chr18:42531879-42531932	+	CAGCACTTCTGACCAAGCGGAGAANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNTGTCCCCTGTGAGCGA	*
SETBP1	chr18:42531891-42531944	-	TCACTGTGGGACTCGCTCANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTCT GAGCTCTTCTCCGCTTGG	*
SF3B1	chr2:198267310-198267363	+	TACCAGTGTGTCTCGCTTGCCAGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNTCAACTAAACTTCTAA	*
SF3B1	chr2:198267322-198267375	-	TCTTGCCACATCTTAGAAGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGCAA GAAGTCCTGGCAAGCGA	*
SF3B1	chr2:198266793-198266846	+	CACAAGACCTACAAAACCAAACACNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAAGGAGTTGCTGCTTC	*
SF3B1	chr2:198266826-198266879	-	ACCATCAGTGCTTTGGCCATTGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GGTAATGTTGGGGCATAG	*
SKI	chr1:2160533-2160592	-	GGTCTCGCCTTCCAGTACGGTCTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NGAGAAGTCGCGCAGCAC	
SKI	chr1:2160526-2160585	+	CAACTCGGTGCTGCGCGANNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGCTG CGAGACCGTACTGGAAGG	
SKI	chr1:2160225-2160284	+	GCTCCATGAGCTCGCTGGGCGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA TGGAGGCGGCGGCAGGCG	
SKI	chr1:2160240-2160299	+	TGGGCGGCCCGGCCGCTTTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCAG GCGGCCGCGGCTGTTTCC	
SKI	chr1:2160275-2160334	-	CTGCTCCAGCGTCTTCTGCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCTCC TTCTTGTAGGCCTCCTG	
SLC16A2	chrX:73740903-73740956	+	GCTCCTTCACCAGGTAAGGCTNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT TCACTGACCGTTTGGGCT	*
SLC16A2	chrX:73740889-73740942	-	GTGAATATACTCACAATGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGAAG GAGCTGGTATGGAGGCCA	*
SLC6A1	chr3:11067962-11068015	+	AGAGGTCCATTGCTGATGTGGCGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCAATTCGTGCACCAGCA	*
SLC6A1	chr3:11067976-11068029	-	GAATCCTGCGAACATGCTGGTNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG ACCTGAGGCCGCCACATC	*
SMAD2	chr18:45368203-45368262	+	GTAGAGGTCCATTCAGATGAAGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GATTGGTGAAGCTTTATG	
SMAD2	chr18:45368209-45368268	-	GCATGTCATAAAGCTTCACCAATCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCTTCATCTGAATGGAC	

SMAD3	chr15:67473585-67473644	+	AGCGCGTCGGGGAGACATNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCACC CTGTGTCCACAGACCTGC	
SMAD3	chr15:67473590-67473649	-	TGGCTGCAGGTCTGTGGACACAGGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGGAATGTCTCCCCGAC	
SMAD4	chr18:48604640-48604693	+	GTGAAAGGCTGGGGACCGGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTG TTAGGTCTGTCAGCTGCT	*
SMAD4	chr18:48604650-48604703	-	CAATTCCAGCAGCAGCTGANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGTC TTGGGTAATCCGGTCCCC	*
SMAD4	chr18:48591866-48591925	+	TTTGGGTCAACTCTCCAANNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNGAGAG ACATTTAAGGTTCCTTC	
SMAD4	chr18:48591869-48591928	-	CTTGAAGGAACCTTAAATGTCTCTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NATTGGAGAGTTGACCC	
SPEN	chr1:16256027-16256080	+	CTGCAGGTATTAGATGATCAAGGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCAGGTGAATCTGTGGAA	*
SPEN	chr1:16256039-16256092	-	GACTTCTTGATTTTCCACAGNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNCTC TCTGGTCCTTGATCATC	*
SRSF2	chr17:74732951-74733004	+	GCACGGCCCCGTCCATGGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGGTC CCCGGCGGCTGTGGTGTG	*
SRSF2	chr17:74732933-74732986	-	GACCGCCACCCCGCAGGTACGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC TGGACGGCCGCGAGCTGC	*
STAT3	chr17:40474385-40474438	+	GATGATTTCAGCAAATGACATGNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN GGGAATGTCAGGATAGAG	*
STAT3	chr17:40474398-40474451	-	TCCACTGGTCTATCTCTATCCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAG CTGAACAACATGTCATT	*
STXBP1	chr9:130444720-130444773	+	GCCTGAATGAGATGCGCTGCGCCTNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCATAAGAACAAGGCCC	*
STXBP1	chr9:130444748-130444801	-	AGGCGGGGGCCACTGCGGTANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTT CCGTTGGCCTGGGTCACC	*
TBR1	chr2:162274269-162274322	+	GCGGACACCAATGTGCAAGGCANNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN TTTGTGGATGTGATTTTG	*
TBR1	chr2:162274280-162274333	-	TTGGGATCCGCCAAAATCANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTGG AAGGACTTGCCTTGCACA	*
TET2	chr4:106157586-106157639	+	GACTACACATCCTGAACTTTTTGCNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAATCAAGTGCATGCAA	*
TET2	chr4:106157602-106157655	-	ACAAGAAACCTGTATTTTGCATGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTTGTTTCCTGCAAAAAG	*
TET2	chr4:106156688-106156741	+	AAACACGAGATCTTGTGCCNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNGAG AGCTACAGGACAACTGCC	*

TET2	chr4:106156676-106156729	-	GTAGCTCTCCACTGCTACCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCTCG TGTTTGCTCCTTGTCTC	*
TET2	chr4:106157369-106157422	+	TCTTTGGCCAGACTAAAGTGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA AGAATAAAGAGGAAATAC	*
TET2	chr4:106157359-106157412	-	TTTATTCTTTATTTGTAATTTTTGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCAAAGAATGATCCTTC	*
TET2	chr4:106157723-106157776	+	CTTCAGTTCTACAGGGATATANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC CAAATAATGTGATCCCAA	*
TET2	chr4:106157712-106157765	-	TTATTTGGAAAATATTGCATGTGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCTGAAGCTTGTTGTGAC	*
TET2	chr4:106156709-106156762	+	CAACACAGCACTATCTGAAACCAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTTGATGAGAAACAAAG	*
TET2	chr4:106156698-106156751	-	ATCAACTGCTGGCAGTTGTNNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNTGC TGTGTTGGGGGCACAAGA	*
TET2	chr4:106157446-106157505	+	CAAATGGGACTGGAGGAAGTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTT GGCCAGACTAAAGTGGAA	
TET2	chr4:106157453-106157512	-	AACATTCTTCCACTTTAGTCTGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT CTGTACTTCCTCCAGTC	
TET2	chr4:106157824-106157883	-	ACCAGACATATCTTGGTTTCTNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCT CCCTGGTCAGGCACAGG	
TET2	chr4:106157768-106157827	+	AAGCTGCGCAACTTGCTCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNTTTC AAGAACAGGAGCAGAAGT	
TET2	chr4:106157760-106157819	-	TCCTGTTCTTGAAAGCACCTGTGANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCGCAGCTTGTTGACCA	
TGFBR1	chr9:101911493-101911552	+	CGCAACTCAGTCAACAGGNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNGTAA TGGCTAAAATTATGAGAG	
TGFBR1	chr9:101911479-101911538	-	GCCATTACTCTCAAGGCCTACAAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCGATAATGTTTTCTTA	
TGFBR2	chr3:30732925-30732984	+	GTGTGGCAGAACGCTTCAGTGAGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NACAGGGCATCCAGATGG	
TP53	chr17:7577514-7577567	+	CACATGTAGTTGTAGTGGATGGTGNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCCTGGAGTCTTCCAGT	*
TP53	chr17:7577519-7577572	-	ATCACACTGGAAGACTCCAGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC ACCATCCACTACAACTAC	*
TP53	chr17:7577460-7577513	+	GTGATGATGGTGAGGATGGGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAG GCAAGCAGAGGCTGGGGC	*
TP53	chr17:7577473-7577526	-	ACTGGCCTGCTGTGCCCCAGCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNA TGAACCGGAGGCCCATCC	*

TP53	chr17:7579327-7579380	+	GGTTTTCTGGGAAGGGACAGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNCT GACCGTGCAAGTCACAGA	*
TP53	chr17:7579340-7579393	-	TGGGACAGCCAAGTCTGTGACNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNC CCTGTCATCTTCTGTCCC	*
TP53	chr17:7578374-7578427	+	GCTGTGACTGCTTGTAGATGGCCANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCCCAGCTGCTCACCAT	*
TP53	chr17:7578408-7578461	-	CGCTGCCCCCACCATGAGCGCTGCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNCCCGCCCGGCACCCGC	*
TP53	chr17:7578379-7578432	-	TAGCGATGGTGAGCAGCTGGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT GGCCATCTACAAGCAGTC	*
TP53	chr17:7577558-7577611	+	GGAGATAACACAGGCCCAAGANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNG GTTCATGCCGCCCATGCA	*
TP53	chr17:7577559-7577612	-	CTGCATGGGCGGCATGAACCNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNTC TTGGGCCTGTGTTATCTC	*
TP53	chr17:7578157-7578210	+	CGAAAAGTGTTTCTGTCATCCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT TAACCCCTCCTCCCAGAG	*
TP53	chr17:7578157-7578210	+	CGAAAAGTGTTTCTGTCATCCNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT TAACCCCCCCTCCCAGAG	*
TP53	chr17:7578173-7578226	-	GGTTTGCAACTGGGGTCTCTGGGANNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGTGGAGTATTTGGATG	*
TP53	chr17:7577078-7577137	-	GAATCTCCGCAAGAAAGGGGAGCCNNNNNCT TCAGCTTCCCGATATCCGACGGTAGTGTNNN NNGTAATCTACTGGGACG	
TP53	chr17:7577079-7577138	+	GTCCCAGTAGATTACCACNNNNNCTTCAGCTT CCCGATATCCGACGGTAGTGTNNNNNTCCCC TTTCTTGCGGAGATTCT	
TWIST2	chr2:239757080-239757133	+	TCCCCACGCTGCCCTCTGACANNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAT CCTGGCCAACGTGCGCG	*
TWIST2	chr2:239757056-239757109	-	GGCTCTGCAGCTCCTCGAAGGNNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT TGCGCAGCGCCGCGAAGG	*
U2AF1	chr21:44524447-44524500	+	GAACAGTTGACTCTGTAAGGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNGG CTAAACGTCGGTTTATTG	*
U2AF1	chr21:44524436-44524489	-	CGTTTAGCCAGGTTTGTTTGCCTNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCTGTTCATTTTATTTCA	*
UGT1A4	chr2:234628240-234628293	+	TGGGGGCATCAACTGTGCCAACGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NTCCGTGTGGCTGTTCCG	*
UGT1A4	chr2:234628228-234628281	-	ACGGATGCATAGCTGACAANNNNNCTTCAGC TTCCCGATATCCGACGGTAGTGTNNNNNATG CCCCCAATGAAGACCATG	*
WDR45	chrX:48935701-48935754	+	GGTGCAGGATTGTTCCTCTGNNNNNCTTCAG CTTCCCGATATCCGACGGTAGTGTNNNNNAG GCCCAATCCTCTCTCACT	*

WDR45	chrX:48935730-48935783	-	TGACCAGCCTGCGTTTCAACCANNNNNCTTCA GCTTCCCGATATCCGACGGTAGTGTNNNNNT CTTCTCATCTTTATCCC	*
ZMYND11	chr10:298359-298412	+	CTGGCCCTTCCCGGAGTCNNNNNCTTCAGCT TCCCGATATCCGACGGTAGTGTNNNNNCCTA CTGCTCCATCAAGTGCCA	*
ZMYND11	chr10:298397-298450	-	GGCAGGTGCGCTTGTGCTCCGCNNNNNCTTC AGCTTCCCGATATCCGACGGTAGTGTNNNNN CAAAAACCGCTGTGTCTG	*
ZNF215	chr11:6977578-6977631	+	CCTCTCTTATTCGACACCAAATGANNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NAATCCAACACTCCATT	*
ZNF215	chr11:6977619-6977672	-	GGATTTTCCACAGTTAACACATTGNNNNNCTT CAGCTTCCCGATATCCGACGGTAGTGTNNNN NCCTTACATTTGAATGG	*

A.3 Primer Sequences

Gene	Target Coordinates (hg19)	Forward Primer	Reverse Primer	Product Size (bp)	Annealing Temp °C
ABL1	chr9:133730269-133730488	CAAGCAACTACATCACGCCAGTC	CCTACCTTGCCATCAGAAGCAGT	220	62
ADNP	chr20:49508975-49509121	TGACCCTTGGGGTCTAAAGC	AAGTCTGGCACCTGTGAAGC	147	61
ADNP	chr20:49510383-49510503	CAGGTCACTGCCATGATTGG	GGGAACCGATCCTTGGTGG	121	64
ADNP	chr20:49510452-49510642	ATCACTCAATGGGGCAGTCC	GTTTGGATCGGGGAACCACT	191	64
ADNP	chr20:49510331-49510471	AGTGGTTCCCCGATCCAAAC	AGTCGATTCACCATCTGCTGT	141	64
ADNP	chr20:49510034-49510193	GCAACGCACCAGTTTCCATT	TGGCCCGATGAGAGAGAAGA	160	64
ADNP	chr20:49510219-49510676	TGGCAGCACCTTACATAGCAA	ACCAACACTGTAACCCTGGC	458	64
ADNP	chr20:49507872-49508041	CAGTGAGGATGGGGAACAGT	AGCTTTGCAGTCACACTGGAT	170	64
ADNP	chr20:49510174-49510367	TCCGGTCTTTACCATCACAGC	AATGGAAACTGGTGCGTTGC	194	64
ADNP	chr20:49509000-49509144	CCCTCTCGGCTTAATCAGTCTC	CAACAGGCTCTTCAGGCTTC	145	64
AKT	chr14:105246386-105246601	CCCCAAATCTGAATCCCGAGA	GTCTGACGGGTAGAGTGTGC	216	64
AKT3	chr1:243668433-243668630	CTGGTCTGGGATGTCGGAAG	AGGATGGTATGGACTGCATGG	198	64
ALK	chr2:29432643-29432815	TTTACTCACCTGTAGATGTCTCGG	TCCTAGTGATGGCCGTTGT	173	60
APC	chr5:112174971-112175256	GGCAGAATCAGCTCCATCCA	TCGCTCACAGGATCTTCAGC	286	64
BRAF	chr7:140453061-140453214	CCTTTACTTACTACACCTCAGATATATTTC	GCCTCAATTCTTACCATCCACAAAATG	154	60
BRAF	chr7:140439650-140439736	TGGTGGGACGAGGATACCTG	GGCACTCTGCCATTAATCTCTTC	87	64
BRAF	chr7:140453958-140454095	ACTTCGCAGACAAATTTCAGGA	GTAGATTTCGAGGCCAGAGTC	138	64
CBL	chr11:119148828-119149025	TGCAGTTATTTATTCAACTAATAGTCTTT	CGCTGTTTAGATCCGTACCTGCC	198	60
CBL	chr11:119148899-119149097	GATGGGCTCCACATTCCAAC	AGGCCACCCCTTGTATCAGTA	199	64
CBL	chr11:119148790-119149026	GGAAACAAGTCTTCACTTTTTCTGT	TCGCTGTTTAGATCCGTACCTG	273	64
CHD8	chr14:21876455-21876632	GGCATCCATGGTCCCTTCTT	AAGACATGCAACTCACCCGT	178	64
CHD8	chr14:21865906-21866086	GGCAGACCCAGCCTTATGTT	CCATTGGGTTCAGTCAAGGC	181	64
CHD8	chr14:21865904-21866067	TTCCTAGAAAAGGCTGGCCG	TCCCATTGGGTTCAGTCAAGG	164	64
CHD8	chr14:21863038-21863218	GGCTTCGGCGTCTAGTAACA	ACAGATCAGCACAAGTTCCCA	181	64
CHD8	chr14:21853761-21853948	GCCTCTGGTACTACCTTGCG	ATGCTGGGGCTTCAGTCATC	188	64

CHD8	chr14:21867693-21867905	TCCTGCATTTCTGTGCTGACT	TGCTAAATGTTCCATCCGTGA	213	64
CHD8	chr14:21873352-21873555	CTTGGCACCCAAACAGGAAAC	CCAAGCATTCCCTCACCATT	204	64
CHD8	chr14:21862092-21862238	TTGGAAGATCGGCTGGCATT	GTCTGGGTCCTGCATGATGT	147	64
CHD8	chr14:21899454-21899693	TTGGACTGCCAAGCTCTCTG	ATGTTGGCGTCGATGTCTGT	240	64
CHD8	chr14:21897438-21897616	GTTGTACTGTGTCTTCTGACCTG	TTGGGGACCTCCAGACTGTG	179	64
CHD8	chr14:21863901-21864126	AGGTCTTCATCCTAGGTTCCC	AGATTCAGAGAATGGTTGAAGACT	226	64
CHD8	chr14:21867693-21867805	ACAGTGAGGCTGACAAGTCG	TGCTAAATGTTCCATCCGTGA	113	64
COL4A3BP	chr5:74722170-74722413	CACCCAGACTGCCAATTGTT	ACTTAAGGCATCCATTAACTACACA	244	64
CSF1R	chr5:149434706-149434954	TTGGACAGGGAAAGATCAGG	GGGAATGGGGAGAAGACAAT	249	60
CTNNB1	chr3:41266403-41266583	TGCTGAACTGTGGATAGTGAGTG	TCAGCCAAACGCTGGACATT	181	64
CTNNB1	chr3:41277105-41277352	AAGTTACGGGGAACTTCGGG	ACTTCACTTAAAACATACCTGCACA	248	64
CTNNB1	chr3:41278103-41278270	TCCGAATGTCTGAGGACAAGC	AGAGGAAAGAACAAGCTGCCA	168	64
CTNNB1	chr3:41280601-41280828	TTGTTGACACCCTGACTCTTCT	GGTCAGTATCAAACCAGGCCA	228	64
CTNNB1	chr3:41268706-41268889	TACTGGCTAGTGGTGGACCC	ACTGCAGATGCTATACACAAGAC	184	64
DNMT3A	chr2:25463209-25463380	CCATGGCCACCACATTCTCA	AGCCACACCACTGTCCTATG	172	61
DNMT3A	chr2:25457218-25457313	CCCAGCAGTCTCTGCCTCGCCA	CCCTCTCTGCCTTTTCTCCCCCAG	96	64
DNMT3A	chr2:25457169-25457366	GCTGTGTGGTTAGACGGCTT	CTTCAGCGGAGCGAAGAGG	198	64
DNMT3A	chr2:25456873-25457449	GAACTAAGCAGGCGTCAGAGGA	AAAAAGGGAAGGGGGAGGAAGG	577	64
DNMT3A	chr2:25461930-25462126	GCTCCTCCTTGGCTCATCTT	TCCCCACTATGGGTCATCCC	197	64
DYRK1A	chr21:38858787-38858941	CGTGTGGAGCAAGAATGGGT	CTGCGCTGTTTGCCTGTTT	155	64
DYRK1A	chr21:38884236-38884508	CTCTTGGTTGGTCAGGCACT	TCATGGAGTGGTGACTGTGG	273	64
DYRK1A	chr21:38868474-38868670	CTTGACCAAGCACCAAAAGCA	ATCGTTAACTGAAACACACTGATT	197	64
DYRK1A	chr21:38878384-38878568	TGTGTTGTGATATTTCAGGTGGC	AATGGCAGGTGACACAAGCA	185	64
DYRK1A	chr21:38844965-38845107	ACCTCACTTATCTTCTTGTAGGAG	GCGACGGTCACTGTACTGAT	143	64
EZH2	chr7:148523485-148523681	CCTGAATGTACCCCCAACAT	AGCACTCTCCAAGCTGCTTTA	197	64
EZH2	chr7:148508705-148508871	TGTCCCCAGTCCATTTTCACC	TCAGTGCCTTACCTCTCCAC	167	64
EZH2	chr7:148506368-148506590	AGGCAAACCCTGAAGAACTGT	AGGGAGTTCCAATTCTCACGTC	223	64
EZH2	chr7:148510958-148511155	CGAGAGTGTGACCCTGACCT	TTTGCCCCAGCTAAATCATC	198	64

EZH2	chr7:148512577-148512764	AGGGCATGACTCTTGGCTTT	GAGGAATGGAAAGATGCTAACCC	188	64
FBN1	chr15:48755276-48755400	CACCTGTACTTGGGATGGGAC	TGACAACCAGACCTGTGATGG	125	64
FGFR1	chr8:38285425-38285714	GCACACACACGTACCTTGTAGC	CAGATGGGCAAGACACCTCCAG	290	64
FGFR2	chr10:123276832-123277024	CATTCCTGTGTCGTCTAGCCT	CCTGGCAGAACTGTCAACCA	193	60
FGFR2	chr10:123257928-123258119	ATGATGCCACAGAGAAAGACCT	ATCTGGAAGCCCAGCCATTT	192	64
FGFR2	chr10:123274653-123274849	GTGATCTGCAATCTAGCGCC	ATACGTTTGGTCAGCTTGTGC	197	64
FGFR2	chr10:123324942-123325191	AACCAGAAGTGTACGTGGC	GCCAACTCACCTGTGACATTC	250	64
FGFR2	chr10:123310838-123311031	TACTCATGGAGGGGAAGCTGT	TCCTTCCCGTTTTTCAGCCA	194	64
FGFR2	chr10:123323996-123324081	CCGGAGATGATGAGGATGACAC	GGAGCCGGGCAGTTACTTAC	86	64
FGFR2	chr10:123323999-123324198	TCCTGGGTTGTTGACTTTGC	GCCGGGCAGTTACTTACTCTT	200	64
FGFR2	chr10:123298051-123298196	AGTGTGGTCCCATCTGACAAG	AGTCAAGCAAGAATGGGCTG	146	64
FGFR2	chr10:123260288-123260468	CATTTAGGCTGACACTGGGC	TGATAAGACTCTCCACCCAGC	181	64
FGFR2	chr10:123243174-123243371	TAACCACGTACCCAGTGCAT	CAACCAAGGACAAGGGGCTT	198	64
FGFR2	chr10:123263259-123263431	ATGAACTCCAACACCCCGC	ATAGCTGAGTCTCCATCCTGGG	173	64
FGFR2	chr10:123239419-123239612	AGTCAGCCTCTCGAACCG	TTCGTAAGGCATGGGGTCTG	194	64
FGFR2	chr10:123256007-123256195	CTCCGAGAATACCTCCGAGC	TTGGATTCCACCCAGCCAAG	189	64
FGFR2	chr10:123353156-123353359	AAGAGGACCGGGGATTGGTA	AGGCCTTAACAATCTGCCCC	204	64
FGFR2	chr10:123247477-123247621	CATCGAGATTTAGCAGCCAGA	GCCCACTGTGTTACTGCCAT	145	64
FGFR2	chr10:123279576-123279766	GACAGCCCTCTGGACAACAC	GCTGGGCATCACTGTAAACC	191	64
FGFR2	chr10:123274584-123274842	GCAATCTAGCGCCTGGAAGA	GGGTCATGGGGGAGGAGTAA	259	64
FGFR2	chr10:123279452-123279736	TTGGTCTCTCATTCTCCCATCCCCACTC	GACCCCAGTTGTGGGTACCTTTAGATTCAG	285	64
FGFR2	chr10:123276902-123277101	GTCGGAATCTCCCAGTGGC	CCCCAGCGTCCTCAAAAGTT	200	64
FGFR2	chr10:123279427-123279624	GTCGGAGGAGACGTAGAGTTT	AACCTGTGGCCAAACCCAT	198	64
FGFR3	chr4:1807887-1808064	AGAAGACGACCAACGTGAGC	GGACACGTACACGTCACTCT	178	64
FGFR3	chr4:1807391-1807533	AGGGCGGTAGGTGCGGTA	GCAGAAACTCCCGCAGGTTA	143	64
FGFR3	chr4:1805436-1805666	CACCGACAAGGAGCTAGAGG	GGCCGTAAGTCACAGGATTC	231	64
FGFR3	chr4:1807225-1807464	GTGCAGAGCAGGGCTGGGGGC	GCTCCTCAGACGGGCTGCCAG	240	64
FGFR3	chr4:1806083-1806282	GAGGCGGGCAGTGTGTAT	AGGCAGCTCAGAACCTGGTAT	200	64

FGFR3	chr4:1807283-1807412	CAGACGATGCCACTGACAAG	CCGCTACCGCACCTACC	130	64
FGFR3	chr4:1801086-1801284	CTGTCTGGGTCAAGGATGGCA	TCCTTTCTGTAGCTGGCGTGG	199	64
FGFR3	chr4:1806560-1806765	GAGTCCAACGCGTCCATGA	TGACTGGTGGCTGTTTCACC	206	64
FMR1	chrX:147013875-147014073	TGGCTGGCCTAAATACAGTTG	CAATAGCAGTGACCCCAGGT	199	64
FMR1	chrX:147026429-147026696	TGACCACAGAGACGAACTCAG	TGGTATCAGCACAAAACGGGA	268	64
FMR1	chrX:147010194-147010360	TGCAGCATGTGATGCAACT	TCCAAGTTCTCAGCATTTCTTG	167	64
FMR1	chrX:147011659-147011777	GTCACCTCAAAGCGAGCACAT	GGAAAGTGACATACCTCCAGC	119	64
FMR1	chrX:147025559-147026682	GGTCATTAGCTCCAACAGAGGA	AACGGGAGAGATGGGATCTTG	1124	64
GAPDH	chr12:6646553-6646949	CATGGTATGAGAGCTGGGGAATGG	ACTGACACGTTGGCAGTGG	397	64
GAPDH	chr12:6646599-6646889	ATCCAAGACTGGCTCCTCCC	GGGATGACCTTGCCCACAG	291	64
GNAS	chr20:57484305-57484438	AGCCCTCTTTCCAAACTACTCC	TTCCAGAAGTCAGGACACGG	134	61
GNB1	chr1:1747178-1747369	CTCATGCCCACGCCTACCTG	AAGGATGGGTGATGAGGAGAATCG	192	61
GNB1	chr1:1747144-1747372	GACAAGGATGGGTGATGAGGAG	AGAGCAGCCCCTGAATGTAAC	229	64
GRIA1	chr5:153144021-153144209	CGTCTGGTGGTTCTTCACCTT	ACAAAGGATGTCCCTACCCTG	189	64
HRAS	chr11:534143-534333	CCTGAGGAGCGATGACGGAATA	CCTATCCTGGCTGTGTCCTGG	191	64
HRAS	chr11:533841-534079	AGGGGAGTCCCTCGTCTCAG	ATGTACTGGTCCCGCATGGC	239	64
IDH1	chr2:209113027-209113156	AGTTGGAAATTTCTGGGCCATG	CCCCCGGCTTGTGAGTGGATG	130	64
KCNQ3	chr8:133192370-133192568	GTGCTGATTGCCTCTGTGCC	CCATGTGGTCCTGCAGTTTCT	199	64
KIF1C	chr17:4925538-4925739	CCCCTCGCAGGGTTTATCAG	TTGCCATAGGTGCGACACAG	202	64
KIT	chr4:55594115-55594299	CATGCGCTTGACATCAGTTT	ACACGGCTTTACCTCCAATG	185	64
KRAS	chr12:25368448-25368569	TATTGTCGGATCTCCCTCACCA	CTTGCACATGGCTTTCCCAGTA	122	64
KRAS	chr12:25379940-25380263	CCTAGGTTTCAATCCCAGCA	TGCAATGAGGGACCAGTACA	324	64
LRP5	chr11:68115442-68115618	CAAGGGAGCCGTGTACTGG	ATGCGGTTGGTCTCTGAGTC	177	64
MAP2K1	chr15:66727328-66727588	GTGACAGTATTGACTTGTGCTCCCCACT	AAGGCAAACTCACCTTTCTGGCCATG	261	60
MAP2K2	chr19:4110454-4110659	CATGCTCTGTTCCGTGGAG	TTGCAGCTGATCCACCTTGA	206	60
MAP2K2	chr19:4117177-4117465	CCTCCAGGGCAGAAAGTAGTC	GGTCACCAAAGTCCAGCACA	289	64
MDH1	chr2:63816172-63816373	TCAATCATGGTGAGTGTGGGC	CCGCAAGGTGGAAATCAATCC	202	64
MECP2	chrX:153297953-153298130	AGATCTGAGTGTATGATGGCCT	ACTTGAGGGGTTTGTCCTTGA	178	64

MECP2	chrX:153296148-153296334	GGTCAAGGAAGTGGTGAAGC	CTTTGGGGACTCTGAGTGGT	187	64
MECP2	chrX:153296476-153296632	GTCCTGGGAAGCTCCTTGTC	CGTTTCTTGGGAATGGCCTG	157	64
MED12	chrX:70348797-70349078	CTGCTCTACCTCGCTTTCAATA	CAGGGTCACTTACTTCCATCAG	282	60
MED12	chrX:70339194-70339366	AACAACTAAACGCCGCTTTCC	CTTCCAGCCTTTTCCTTCAGCCT	173	64
MED12	chrX:70347671-70347869	CCTTTCCTGACCATCCCTCGACC	GTAGGTGAAGGTGTGAGCTGCA	199	64
MTOR	chr1:11190713-11190871	AGCTCCAGTGTCTGTCCTTG	GTGGCGTTGGTGATGTTGG	159	64
MYD88	chr3:38182555-38182721	TTGAAGACTGGGCTTGTCCC	GGTTGGTGTAGTCGCAGACA	167	61
NAA10	chrX:153197719-153197930	CTCAAGGTCCCAGTCCCTCA	AAGGCAGTATCTCTCCCCCT	212	64
NF1	chr17:29533128-29533389	TGGTGTGTGTTTGCATGGTC	CGGTGCCATTCGTATTGCTG	262	60
NF1	chr17:29654503-29654699	CCACCACTTTCCAGGTTGGTTC	ACACTCCTGGTGCATGAAGGT	197	64
NOTCH1	chr9:139399287-139399546	TAGGGGAAGATCATCTGCTGGC	GACCAGTACTGCAAGGACCACT	260	62
NOTCH2	chr1:120458221-120458550	GGATGATAGGCTGGGAGAATGGTC	TGAGATGTTTGGTATGGTCCTGGCT	330	64
NOTCH3	chr19:15303119-15303390	CCAGGACAGGGTGAGTTTAGG	TTCTTTGTGTCTGGGGCCATC	272	60
PACS1	chr11:65978619-65978767	TAAGCGAGATGCCAACAAGC	TCAAAGGCCAACAGTCTGGAG	149	64
PCGF2	chr17:36895698-36895937	AGTCTGCAAAACCTGCATCGTG	AAGGTCGGGGAGTAGCCCA	240	64
PDCD11	chr10:105184703-105184952	GTAAGGAAAGCAGGGATGGTG	AAAAGAAGGCCAGTCACTCC	250	64
PIK3CA	chr3:178951957-178952227	GATGCTTGGCTCTGGAATGC	GGTCTTTGCCTGCTGAGAGT	271	64
PPP2R1A	chr19:52715952-52716097	CTGTGCTCAGATGACACCCC	GCAAAACTCACCTGCTCGTC	146	64
PPP2R1A	chr19:52715886-52716057	CTCCTCTCTGCCATACTGC	TGGGGATGATCTCACTCTTGA	172	64
PPP2R5D	chr6:42974958-42975122	CTGCCACCTTCATCGAATCC	GATGTTCCCAAGGACTTGCAG	165	61
PTEN	chr10:89720686-89721017	CATACCAGGACCAGAGGAAACC	CTGCTACGTAAACACTGCTTCG	332	60
PTPN11	chr12:112891011-112891204	ACATCTCTCTGGGAAAGAAGCAG	ACTGGAGATTTACCTGACAGCG	194	60
PTPN11	chr12:112888130-112888374	GAGCTGTCACCCACATCAAG	TCTGACACTCAGGGCACAAG	245	60
PTPN11	chr12:112926818-112927055	TCCTTCGTAGGTGTTGACTGC	CCTGCTCAAAAGGAGAGCGT	238	64
PTPN11	chr12:112888172-112888359	ATGACCTGTATGGAGGGGAGA	ACAAGGGAGCAGCAGACTTT	188	64
PTPN11	chr12:112888082-112888358	TATTTGTCCCCTTGCCTCCCTTTCCAATGGA	CAAGGGAGCAGCAGACTTTGTGGTCACT	277	64
PTPN11	chr12:112883999-112884295	TGCGTCTTAGCTGTGTTGTTG	CTCGAAATGCAGGCAGCAAG	297	64
PTPN11	chr12:112888090-112888357	CCCTTGCCTCCCTTTCCAATG	AAGGGAGCAGCAGACTTTGTG	268	64

PTPN11	chr12:112926859-112927015	CATCCAGATGGTGCGGTCTC	GAATGAGAATCCGCATGCCAG	157	64
PTPN11	chr12:112888081-112888274	TTATTTGTCCCCTTGCCTCC	TCAATGACATCTCCATTCTTCTC	194	64
PTPN11	chr12:112888086-112888245	TGTCCCCTTGCCTCCCTTTC	TTGCCCGTGATGTTCCATGT	160	64
PTPN11	chr12:112856865-112857149	GGAGCCTGAGCAAGGAGC	CTCGGGGAGGCAGGAAATG	285	64
RAF1	chr3:12645604-12645763	TGACATTACCACCCCCAAGGTGC	CACGCCTTCACCTTTAACACCTCCA	160	60
RB1CC1	chr8:53568701-53568841	TGGATTCAGAATTGAGTGCTCTTG	ACTGTTCTCTGTCTTGCTCCTG	141	64
RET	chr10:43600469-43600705	AGCAGCGGGAGAAGTACGA	ACCGAGAAACGAACTGTGGC	237	64
RET	chr10:43613793-43613942	TTGCAACCTGCTCTGTGCTG	GCAGCTGGCCTTACCATCCT	150	64
RET	chr10:43619139-43619325	TGTGGGAGATCGTGACCCTA	GGGAGGGAATGCACACAGAT	187	64
RHO	chr3:129251132-129251355	ACTACTACACGCTCAAGCCG	GTCCAGACCATGGCTCCTC	224	64
SCN2A	chr2:166198960-166199129	TTCTCCGATCATTCCGGCTG	ACTACCATTTGGTGGTATGGCA	170	64
SF3B1	chr2:198267078-198267386	TCCCCAAATCAGTAGCCCAAA	CTGGCAAGCGAGACACACT	309	61
SMAD4	chr18:48604637-48604774	GCTGCTGGAATTGGTGTTGAT	GGCCCGGTGTAAGTGAATTT	138	64
SOS1	chr2:39285657-39285903	AGCATCCCTTCTCACCACAT	GCAATAGCTGATGCCCAATC	247	60
SOS1	chr2:39250195-39250422	TGAGCTCTAGGTTTTCTGTCATCT	CTCATGTTTGGCTCCTACACG	228	60
SOS1	chr2:39249682-39249869	AAAGTGACAGGCAACTGCATA	TGAGGCTGCCTAGTGCTGAT	188	64
SYNGAP1	chr6:33406256-33406418	ATATTCCGCGAGAACACGCT	CTGTGGAAGGGTGGACTTGG	163	64
SYNGAP1	chr6:33411538-33411786	GGGGCAAGTCTCAGCAGTT	CTTCCCTCCCTGTTGTGACC	249	64
SYNGAP1	chr6:33409337-33409447	GTAACAGCCTCACCCTTCCAG	GCTTGGCTGCCTTTGGATGT	111	64
SYNGAP1	chr6:33409038-33409192	TCCAATCTGGACACGCTAACC	TGGAGCCATCTGGGGATAGG	155	64
SYNGAP1	chr6:33411087-33411273	CAACTGCGAATCCCCCTCTC	AGCCATGGAATGGGGCAAAG	187	64
SYNGAP1	chr6:33391266-33391404	CCTCTATGCACCGAACCCAAT	CGTTTTCCCAGCCTTCTCAC	139	64
SYNGAP1	chr6:33411318-33411787	GCTGCCTCCATCCTTCATAGC	TCTTCCCTCCCTGTTGTGACC	470	64
SYNGAP1	chr6:33405400-33405708	TCTGTCCCTTCCCCCTTTAAC	ATAGCCTGCCTTGTCCTTCTTG	309	64
TET2	chr4:106156571-106156833	AGACTGCAGGGACAATGACTG	ATGGGATTCCGCTTGGTGAA	263	64
TP53	chr17:7576708-7576940	ACAACCAGGAGCCATTGTCT	TGCCTCTTTCCTAGCACTGC	233	60
TP53	chr17:7577315-7577533	GGTCCCAAAGCCAGAGAAA	CCCATCCTCACCATCATCA	219	60
TP53	chr17:7576839-7577032	GAGCACTAAGCGAGGTAAGC	CCAAGACTTAGTACCTGAAGGG	194	64

TP53	chr17:7577470-7577615	AGTGTGCAGGGTGGCAAGT	CTCCTAGGTTGGCTCTGACTG	146	61
TP53	chr17:7578282-7578437	CAGTCACAGCACATGACGGA	GGCCAGACCTAAGAGCAATCA	156	64
TP63	chr3:189582099-189582324	CCTCAGGGAGCTGTTATCCG	CTGATCTGCCATCGGAGTGC	226	64
UBE3A	chr15:25615696-25615833	TCGCATGTACAGTGAACGAAGA	CTAGCAGCCCAACTTACCCG	138	64
UBE3A	chr15:25616668-25617031	AGAACTTTTTGCAACAGAGTAAACA	TCTCCCTGTGAGCTATCACCT	364	64
UBE3A	chr15:25601823-25601962	TTGGCATAGTACTGGGTCTGG	TGACAAAGAACTTACTGGGTGA	140	64
UBE3A	chr15:25616374-25616610	AAGGGTCTACACCAGATTGCTC	GCAAGGGGTAGCTTGCTCATC	237	64
ZBTB16	chr11:113934666-113934927	AGGGAACTCTTCAGCCACCT	CTCTCCTCTCGCCCATAGTG	262	60
ZBTB16	chr11:113935010-113935278	CATCTGGGCATCTACTCCGTG	ACAGCCTCGTTATCAGGAAGC	269	60
ZBTB16	chr11:114027055-114027296	ACAGGAAGCTGCACAGTGG	GAGGGTCCAGGGAACAAGA	242	60
ZBTB16	chr11:113934667-113934851	GGGAACTCTTCAGCCACCTGC	TGCCTCTTTCCTCAACCTT	185	64
ZBTB16	chr11:114117935-114118098	AGTGTGAGTTCTGTGGCAGC	GCCTTCGGTACCTGTGTGC	164	64
ZBTB16	chr11:114112895-114113221	ACATGGCCGTCTTCTGTCTGC	GTCCACATTTACCTGCACGACG	327	64

Bibliography

R. Acuna-Hidalgo, J. A. Veltman, and A. Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17(1):241, 2016.

R. Acuna-Hidalgo, H. Sengul, M. Steehouwer, M. van de Vorst, S. H. Vermeulen, L. A. Kiemeney, J. A. Veltman, C. Gilissen, and A. Hoischen. Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *American Journal of Human Genetics*, 101(1):50–64, 2017.

I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, 7:76: 7.20.1–7.20.41, 2015.

E. H. Ahn, K. Hirohata, B. F. Kohrn, and E. J. Fox. Detection of Ultra-Rare Mitochondrial Mutations in Breast Stem Cells by Duplex Sequencing. *PLoS One*, (8):1–17, 2015.

F. R. Alan, K. Wang, R. Evjenth, Jinchuan Xing, J. J. Johnston, J. J. Swensen, W. E. Johnson, B. Moore, C. D. Huff, L. M. Bird, J. C. Carey, J. M. Opitz, C. A. Stevens, T. Jiang, C. Schank, H. D. Fain, R. Robison, B. Dalley, S. Chin, S. T. South, J. P. Theodore, L. B. Jorde, H. Hakonarson, J. R. Lillehaug, L. G. Biesecker, M. Yandell, T. Arnesen, and G. J. Lyon. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *American Journal of Human Genetics*, 89(1):28–43, 2011.

L. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, A. Boot, K. R. Covington, D. A. Gordenin, E. Bergstrom, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, G. Getz,

S. G. Rozen, and M. R. Stratton. The repertoire of mutational signatures in human cancer. *bioRxiv*, 2018.

L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47:1402–1407, 2015.

J. Anderson, H. Burns, P. Enriquez-Harris, A. O. M. Wilkie, and J. Heath. Apert syndrome mutations in Fibroblast Growth Factor receptor 2 exhibit increased affinity for FGF ligand. *Human Molecular Genetics*, 7:1475–1483, 1998.

Y. Aoki, T. Niihori, H. Kawame, K. Kurosawa, H. Ohashi, Y. Tanaka, M. Filocamo, K. Kato, Y. Suzuki, S. Kure, and Y. Matsubara. Germline mutations in HRAS proto-oncogene cause Costello syndrome. *Nature Genetics*, 37:1038–1040, 2005.

A. B. Arnett, C. L. Rhoads, K. Hoekzema, T. N. Turner, J. Gerdts, A. S. Wallace, S. Bedrosian-Sermone, E. E. Eichler, and R. A. Bernier. The autism spectrum phenotype in ADNP syndrome. *Autism Research*, 11(9):1300–1310, 2018.

R. A. Barnard, M. B. Pomaville, and B. J. O'Roak. Mutations and modeling of the chromatin remodeler CHD8 define an emerging autism etiology. *Frontiers in Neuroscience*, 9:477, 2015.

G. A. Bellus, T. W. Hefferon, R. I. Ortiz de Luna, J. T. Hecht, W. A. Horton, M. Machado, I. Kaitila, I. McIntosh, and C. A. Francomano. Achondroplasia is defined by recurrent G380R mutations of FGFR3. *American Journal of Human Genetics*, 56(2):368–373, 1995.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

R. Bernier, C. Golzio, B. Xiong, H. A. Stessman, B. P. Coe, O. Penn, K. Witherspoon, J. Gerdts, C. Baker, A. T. V. van Silfhout, J. H. Schuurs-Hoeijmakers, M. Fichera,

P. Bosco, S. Buono, A. Alberti, P. Failla, H. Peeters, J. Steyaert, L. E. Vissers, L. Francescatto, H. C. Mefford, J. A. Rosenfeld, T. Bakken, B. J. ORoak, M. Pawlus, R. Moon, J. Shendure, D. G. Amaral, E. Lein, J. Rankin, C. Romano, B. B. de Vries, N. Katsanis, and E. E. Eichler. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell*, 158(2):263 – 276, 2014.

L. G. Biesecker and N. B. Spinner. A genomic view of mosaicism and human disease. *Nature Reviews Genetics*, 14:307–320, 2013.

E. A. Boyle, B. J. O'Roak, B. K. Martin, A. Kumar, and J. Shendure. MIPgen: Optimized Modeling and Design of Molecular Inversion Probes for Targeted Resequencing. *Bioinformatics*, 30(18):1–3, 2014.

R. Brambilla, N. Gnesutta, L. Minichiello, G. White, A. J. Roylance, C. E. Herron, M. Ramsey, D. P. Wolfer, V. Cestari, C. Rossi-Arnaud, S. G. Grant, P. F. Chapman, H. P. Lipp, E. Sturani, and R. Klein. A role for the Ras signalling pathway in synaptic transmission and long-term memory. *Nature*, 390:281–286, 1997.

M. Buscarlet, S. Provost, Y. F. Zada, A. Barhdadi, V. Bourgoin, G. Lépine, L. Mollica, N. Szuber, M. P. Dubé, and L. Busque. DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. *Blood*, 130(6):753–762, 2017.

V. Caputo, L. Cianetti, M. Niceta, C. Carta, A. Ciolfi, G. Bocchinfuso, E. Carrani, M. L. Dentici, E. Biamino, E. Belligni, L. Garavelli, L. Boccone, D. Melis, G. Andria, B. D. Gelb, L. Stella, M. Silengo, B. Dallapiccola, and M. Tartaglia. A restricted spectrum of mutations in the SMAD4 tumor-suppressor gene underlies Myhre syndrome. *American Journal of Human Genetics*, 90(1):161–169, 2012.

P. Carrière, P. Baade, and L. Fritschi. Population Based Incidence and Age Distribution of Spermatocytic Seminoma. *Journal of Urology*, 178(1):125–128, 2007.

J. P. Casey, S. I. Støve, C. McGorrian, J. Galvin, M. Blenski, A. Dunne, S. Ennis, F. Brett, M. D. King, T. Arnesen, and S. A. Lynch. Naa10 mutation causing a novel intellectual

disability syndrome with Long QT due to n-terminal acetyltransferase impairment. *Scientific Reports*, 5:16022 EP –, 2015.

J. Y. Choi, K. M. Lee, S. K. Park, D. Y. Noh, S. H. Ahn, K. Y. Yoo, and D. Kang. Association of paternal age at birth and the risk of breast cancer in offspring: a case control study. *BMC Cancer*, 5(1):143, 2005.

S. K. Choi, S. R. Yoon, P. Calabrese, and N. Arnheim. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):10143–10148, 2008.

S. K. Choi, S. R. Yoon, P. Calabrese, and N. Arnheim. Positive Selection for New Disease Mutations in the Human Germline: Evidence from the Heritable Cancer Syndrome Multiple Endocrine Neoplasia Type 2B. *PLoS Genetics*, 8(2):e1002420, 2012.

A. Cohen, S. Wilson, J. Trinh, and X. Ye. Detecting somatic mosaicism: considerations and clinical implications. *Clinical Genetics*, 87(6):554–562, 2015.

M. M. Cohen, S. Kreiborg, E. J. Lammer, J. F. Cordero, P. Mastroiacovo, J. D. Erickson, P. Roeper, and M. L. Martínez-Frías. Birth prevalence study of the Apert syndrome. *American Journal of Medical Genetics*, 42(5):655–659, 1992.

J. F. Crow. There's something curious about paternal-age effects. *Science*, 301(5633): 606–607, 2003.

M. Culty. Gonocytes, the forgotten cells of the germ cell lineage. *Birth Defects Research Part C: Embryo Today: Reviews*, 87(1):1–26, 2009.

M. Dakouane Giudicelli, V. Serazin, C. R. Le Sciellour, M. Albert, J. Selva, and Y. Giudicelli. Increased achondroplasia mutation frequency with advanced age and evidence for G1138A mosaicism in human testis biopsies. *Fertility and Sterility*, 89(6):1651–1656, 2008.

J. David Sweatt. The neuronal MAP kinase cascade: A biochemical signal integration system subserving synaptic plasticity and memory. *Journal of Neurochemistry*, 76: 1–10, 2001.

J. de Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. de Vries, C. Gilissen, M. del Rosario, A. Hoischen, H. Scheffer, B. B. de Vries, H. G. Brunner, J. A. Veltman, and L. E. Vissers. Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, 367(20):1921–1929, 2012.

Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228, 2014.

Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542(7642):433–438, 2017.

A. M. D'Gama, S. Pochareddy, M. Li, S. S. Jamuar, R. E. Reiff, A. T. Lam, N. Sestan, and C. A. Walsh. Targeted DNA sequencing from Autism Spectrum Disorder brains implicates multiple genetic mechanisms. *Neuron*, 88(5):910 – 917, 2015.

A. V. Dijck, A. T. V. van Silfhout, E. Cappuyns, I. M. van der Werf, G. M. Mancini, A. Tzschach, R. Bernier, I. Gozes, E. E. Eichler, C. Romano, A. Lindstrand, A. Nordgren, M. Bakshi, M. Wilson, Y. Berman, R. Dickson, E. Fransen, and C. Helsmoortel. Clinical presentation of a complex neurodevelopmental disorder caused by mutations in ADNP. *Biological Psychiatry*, 2018.

J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), 2008.

T. Dor, Y. Cinnamon, L. Raymond, A. Shaag, N. Bouslam, A. Bouhouche, M. Gaussen, V. Meyer, A. Durr, A. Brice, A. Benomar, G. Stevanin, M. Schuelke, and S. Edvardson. KIF1C mutations in two families with hereditary spastic paraparesis and cerebellar dysfunction. *Journal of Medical Genetics*, 51(2):137–142, 2014.

J. Eboreime, S. K. Choi, S. R. Yoon, N. Arnheim, and P. Calabrese. Estimating exceptionally rare germline and somatic mutation frequencies via next generation sequencing. *PLoS ONE*, 11(6):1–24, 2016.

R. Fahsold, S. Hoffmeyer, C. Mischung, C. Gille, C. Ehlers, N. Kücükceylan, M. Abdel-Nour, A. Gewies, H. Peters, D. Kaufmann, A. Buske, S. Tinschert, and P. Nürnberg. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. *American Journal of Human Genetics*, 66(3):790 – 818, 2000.

A. P. Fayomi and K. E. Orwig. Spermatogonial stem cells and spermatogenesis in mice, monkeys and men. *Stem Cell Research*, 29:207 – 214, 2018.

E. Frans, S. Sandin, A. Reichenberg, P. Lichtenstein, N. Lingström, and C. Hultman. Advancing paternal age and bipolar disorder. *Archives of General Psychiatry*, 65(9): 1034–1040, 2008.

M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, and M. C. O'Donovan. De novo mutations in schizophrenia implicate synaptic networks. *Nature*, 506(7487):179–84, 2014.

G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landén, M. Höglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H. Grönberg, C. M. Hultman, and S. A. McCarroll. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, 371(26):2477–2487, 2014.

A. A. Ghazarian, B. Trabert, B. I. Graubard, S. M. Schwartz, S. F. Altekruse, and K. A. McGlynn. Incidence of testicular germ cell tumors among US men by census region. *Cancer*, 121(23):4181–4189, 2015.

E. Giannoulatou, G. McVean, I. B. Taylor, S. J. McGowan, G. J. Maher, Z. Iqbal, S. P. Pfeifer, I. Turner, E. M. M. Burkitt Wright, J. Shorto, A. Itani, K. Turner, L. Gregory, D. Buck, E. Rajpert-De Meyts, L. H. J. Looijenga, B. Kerr, A. O. M. Wilkie, and A. Goriely. Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. *Proceedings of the National Academy of Sciences of the United States of America*, 110:20152–7, 2013.

E. Giannoulatou, G. J. Maher, Z. Ding, A. J. M. Gillis, L. C. J. Dorssers, A. Hoischen, E. Rajpert-De Meyts, G. McVean, A. O. M. Wilkie, L. H. J. Looijenga, and A. Goriely. Whole-genome sequencing of spermatocytic tumors provides insights into the mutational processes operating in the male germline. *PLoS One*, 12(5):e0178169, 2017.

R. L. Glaser, W. Jiang, S. A. Boyadjiev, A. K. Tran, A. A. Zachary, L. Van Maldergem, D. Johnson, S. Walsh, M. Oldridge, S. A. Wall, A. O. Wilkie, and E. W. Jabs. Paternal origin of FGFR2 mutations in sporadic cases of Crouzon syndrome and Pfeiffer syndrome. *American Journal of Human Genetics*, 66(3):768–777, 2000.

R. L. Glaser, K. W. Broman, R. L. Schulman, B. Eskenazi, A. J. Wyrobek, and E. W. Jabs. The paternal-age effect in Apert syndrome is due, in part, to the increased frequency of mutations in sperm. *American Journal of Human Genetics*, 73(4):939–947, 2003.

J. Glass. *Testes and epididymes. In Gray's Anatomy: The anatomical basis of clinical practice*. Churchill Livingston, Edinburgh, UK, 39th edition, 2005.

J. M. Goldmann, W. S. W. Wong, M. Pinelli, T. Farrah, D. Bodian, A. B. Stittrich, G. Glusman, L. E. L. M. Vissers, A. Hoischen, J. C. Roach, J. G. Vockley, J. A.

Veltman, B. D. Solomon, C. Gilissen, and J. E. Niederhuber. Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, 48:935–939, 2016.

J. M. Goldmann, V. B. Seplyarskiy, W. S. W. Wong, T. Vilboux, L. Bodian, B. D. Solomon, J. A. Veltman, J. F. Deeken, and C. Gilissen. Germline de novo mutation clusters arise during oocyte aging in genomic regions with increased double-strand break incidence. *Nature Genetics*, 50:487–492, 2018.

A. Goriely and A. O. M. Wilkie. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *The American Society of Human Genetics*, 90(2):175–200, 2012.

A. Goriely, G. A. T. McVean, M. Röjmyr, B. Ingemarsson, and A. O. M. Wilkie. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science*, 301(5633):643–646, 2003.

A. Goriely, G. A. T. McVean, A. M. M. van Pelt, A. W. O'Rourke, S. A. Wall, D. G. de Rooij, and A. O. M. Wilkie. Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17):6051–6056, 2005.

A. Goriely, R. M. S. Hansen, I. B. Taylor, I. A. Olesen, G. K. Jacobsen, S. J. McGowan, S. P. Pfeifer, G. A. T. McVean, E. Rajpert-De Meyts, and A. O. M. Wilkie. Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nature genetics*, 41(11):1247–1252, 2009.

A. Goriely, J. J. McGrath, C. M. Hultman, A. O. Wilkie, and D. Malaspina. Selfish spermatogonial selection: A novel mechanism for the association between advanced paternal age and neurodevelopmental disorders. *American Journal of Psychiatry*, 170 (6):599–608, 2013.

I. Gozes, M. C. Patterson, A. Van Dijck, R. F. Kooy, J. N. Peeden, J. A. Eichenberger, A. Zawacki-Downing, and S. Bedrosian-Sermone. The eight and a half year journey of

undiagnosed AD: Gene sequencing and funding of advanced genetic testing has led to hope and new beginnings. *Frontiers in Endocrinology*, 8:107, 2017.

R. F. Green, O. Devine, K. S. Crider, R. S. Olney, N. Archer, A. F. Olshan, and S. K. Shapira. Association of Paternal Age and Risk for Major Congenital Anomalies From the National Birth Defects Prevention Study, 1997 to 2004. *Annals of Epidemiology*, 20(3):241–249, 2010.

J. K. Grether, M. C. Anderson, L. A. Croen, D. Smith, and G. C. Windham. Risk of autism and increasing maternal and paternal age in a large North American population. *American Journal of Epidemiology*, 170(9):1118–1126, 2009.

S. Gulsuner, T. Walsh, A. Watts, M. Lee, A. Thornton, S. Casadei, C. Rippey, H. Shahin, D. Braff, K. Cadenhead, M. Calkins, D. Dobie, R. Freedman, M. Green, T. Greenwood, R. Gur, R. Gur, L. Lazzeroni, G. Light, K. Nuechterlein, A. Olincy, A. Radant, A. Ray, N. Schork, L. Seidman, L. Siever, J. Silverman, W. Stone, C. Sugar, N. Swerdlow, D. Tsuang, M. Tsuang, B. Turetsky, T. Aduroja, T. Allen, L. D. Bradford, M. E. Calkins, B. Devlin, N. B. Edwards, R. Ganguli, R. C. Go, R. E. Gur, R. C. Gur, J. Kwentus, A. C. Lahti, P. Lyons, K. Mathos, R. May, S. McLeod-Bryant, J. P. McEvoy, L. Montgomery-Barefield, V. L. Nimgaonkar, D. L. Braff, M. C. King, and J. M. McClellan. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, 154(3):518 – 529, 2013.

C. Hafner, E. López-Knowles, N. M. Luis, A. Toll, E. Baselga, A. Fernández-Casado, S. Hernández, A. Ribé, T. Mentzel, R. Stoehr, F. Hofstaedter, M. Landthaler, T. Vogt, R. M. Pujol, A. Hartmann, and F. X. Real. Oncogenic PIK3CA mutations occur in epidermal nevi and seborrheic keratoses with a characteristic mutation pattern. *Proceedings of the National Academy of Sciences*, 104(33):13450–4, 2007.

C. Hafner, A. Toll, S. Gantner, A. Mauerer, I. Lurkin, F. Acquadro, A. Fernández-Casado, E. C. Zwarthoff, W. Dietmaier, E. Baselga, E. Parera, A. Vicente, A. Casanova, J. Cigudosa, T. Mentzel, R. M. Pujol, M. Landthaler, and F. X. Real. Keratinocytic

epidermal nevi are associated with mosaic RAS mutations. *Journal of Medical Genetics*, 49(4):249–253, 2012.

J. B. S. Haldane. The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317, 1935.

T. Hassold and P. Hunt. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics*, 2:280–291, 2001.

T. Hassold, M. Merrill, K. Adkins, S. Freeman, and S. Sherman. Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *American Journal of Human Genetics*, 57(4):867–874, 1995.

C. G. Heller and Y. Clermont. Spermatogenesis in man: an estimate of its duration. *Science*, 140:184–186, 1963.

C. Helsmoortel, A. T. Vulto-van Silfhout, B. P. Coe, G. Vandeweyer, L. Rooms, J. van den Ende, J. H. M. Schuurs-Hoeijmakers, C. L. Marcelis, M. H. Willemsen, L. E. L. M. Vissers, H. G. Yntema, M. Bakshi, M. Wilson, K. T. Witherspoon, H. Malmgren, A. Nordgren, G. Annerén, M. Fichera, P. Bosco, C. Romano, B. B. A. de Vries, T. Kleefstra, R. F. Kooy, E. E. Eichler, and N. Van der Aa. A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. *Nature Genetics*, 46(4):380–384, 2014.

I. Herskowitz. Functional inactivation of genes by dominant negative mutations. *Nature*, 329(6136):219–222, 1987.

M. S. Hestand, J. V. Houdt, F. Cristofoli, and J. R. Vermeesch. Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutation Research*, 784-785: 39 – 45, 2016.

J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, and J. Shendure. Single molecule molecular inversion probes for targeted, high accuracy detection of low frequency variation. *Genome research*, 23:843–54, 2013.

G. Houge, D. Haesen, L. E. Vissers, S. Mehta, M. J. Parker, M. Wright, J. Vogt, S. McKee, J. L. Tolmie, N. Cordeiro, T. Kleefstra, M. H. Willemsen, M. R. Reijnders, S. Berland, E. Hayman, E. Lahat, E. H. Brilstra, K. L. van Gassen, E. Zonneveld-Huijssoon, C. I. de Bie, A. Hoischen, E. E. Eichler, R. Holdhus, V. M. Steen, S. O. Døžskeland, M. E. Hoppurles, D. R. FitzPatrick, the Deciphering Developmental Disorders (DDD) study, and V. Janssens. B56-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *The Journal of Clinical Investigation*, 125(8):3051–3062, 2015.

O. A. Ibrahimi, F. Zhang, A. V. Eliseenkova, N. Itoh, R. J. Linhardt, and M. Mohammadi. Biochemical analysis of pathogenic ligand-dependent FGFR2 mutations suggests distinct pathophysiological mechanisms for craniofacial and limb abnormalities. *Human Molecular Genetics*, 13(19):2313–2324, 2004.

I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, J. D. Smith, B. Paeper, D. A. Nickerson, J. Dea, S. Dong, L. E. Gonzalez, J. D. Mandell, S. M. Mane, M. T. Murtha, C. A. Sullivan, M. F. Walker, Z. Waqar, L. Wei, A. J. Willsey, B. Yamrom, Y. h. Lee, E. Grabowska, E. Dalkic, Z. Wang, S. Marks, P. Andrews, A. Leotta, J. Kendall, I. Hakker, J. Rosenbaum, B. Ma, L. Rodgers, J. Troge, G. Narzisi, S. Yoon, M. C. Schatz, K. Ye, W. R. McCombie, J. Shendure, E. E. Eichler, M. W. State, and M. Wigler. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221, 2014.

K. Ishii, M. Kanatsu-Shinohara, S. Toyokuni, and T. Shinohara. FGF2 mediates mouse spermatogonial stem cell self-renewal via upregulation of Etv5 and Bcl6b through MAP2K1 activation. *Development*, 139(10):1734–43, 2012.

E. W. Jabs, Li X, S. AF, G. Meyers, W. Chen, M. Eccles, J. Mao, L. R. Charnas, J. C. E, and J. M. Jackson-Weiss and Crouzon syndromes are allelic with mutations in fibroblast growth factor receptor 2. *Nature*, 8(4):275–279, 1994.

S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C. H. Mermel, N. Burtt, A. Chavez, J. M. Higgins, V. Moltchanov, F. C. Kuo, M. J. Kluk, B. Henderson, L. Kinnunen, H. A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B. F. Banahan, S. Gabriel, S. Kathiresan, H. M. Stringham, M. I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman, L. Groop, G. Atzmon, J. G. Wilson, D. Neuberg, D. Altshuler, and B. L. Ebert. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine*, 371(26):2488–2498, 2014.

K. J. Johnson, S. E. Carozza, E. J. Chow, E. E. Fox, S. Horel, C. C. Mclaughlin, B. a. Mueller, S. E. Puumala, P. Reynolds, J. Von, and L. G. Spector. Parental age and risk of childhood cancer: A pooled analysis. *Epidemiology*, 20(4):475 – 483, 2009.

A. Jørgensen and E. Rajpert-De Meyts. Regulation of meiotic entry and gonadal sex differentiation in the human: Normal and disrupted signaling. *Biomolecular Concepts*, 5(4):331–341, 2014.

Y. S. Khandwala, C. A. Zhang, Y. Lu, and M. L. Eisenberg. The age of fathers in the USA is rising: An analysis of 168,867,480 births from 1972 to 2015. *Human Reproduction*, 32(10):2110–2116, 2017.

J. H. Kim, H. k. Lee, K. Takamiya, and R. L. Huganir. The role of synaptic GTPase-activating protein in neuronal development and synaptic plasticity. *The Journal of Neuroscience*, 23(4):1119–1124, 2003.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46:310–315, 2014.

W. P. Kloosterman, L. C. Francioli, F. Hormozdiari, T. Marschall, J. Y. Hehir-Kwa, A. Abdellaoui, E.-W. Lameijer, M. H. Moed, V. Koval, I. Renkens, M. J. van Roosmalen, P. Arp, L. C. Karssen, B. P. Coe, R. E. Handsaker, E. D. Suchiman, E. Cuppen,

D. T. Thung, M. McVey, M. C. Wendl, G. of the Netherlands Consortium, A. Uitterlinden, C. M. van Duijn, M. A. Swertz, C. Wijmenga, G. B. van Ommen, P. E. Slagboom, D. I. Boomsma, A. Schönhuth, E. E. Eichler, P. I. de Bakker, K. Ye, and V. Guryev. Characteristics of de novo structural changes in the human genome. *Genome Research*, 25(6):792–801, 2015.

A. S. Kondrashov. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation*, 21(1):12–27, 2003.

A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412): 471–475, 2012.

D. A. Korona, K. G. LeCompte, and Z. F. Pursell. The high fidelity and unique error signature of human DNA polymerase. *Nucleic Acids Research*, 39(5):1763–1773, 2011.

R. Kosaki, H. Terashima, M. Kubota, and K. Kosaki. Acute myeloid leukemia-associated DNMT3A p.Arg882His mutation in a patient with Tatton-Brown–Rahman overgrowth syndrome as a constitutional mutation. *American Journal of Medical Genetics, Part A*, 173(1):250–253, 2017.

M. Krajewska-Walasek, D. Jurkiewicz, D. Piekutowska-Abramczuk, M. Kucharczyk, K. H. Chrzanowska, A. Jezela-Stanek, and E. Ciara. Additional data on the clinical phenotype of Helsmoortel Van der Aa syndrome associated with a novel truncating mutation in adnp gene. *American Journal of Medical Genetics Part A*, 170(6):1647–1650, 2016.

H. Kubota, M. R. Avarbock, and R. L. Brinster. Culture conditions and single growth factors affect fate determination of mouse spermatogonial stem cells. *Biology of Reproduction*, 71(3):722–731, 2004.

O. Kwon, E. Hwang, B. Jae, H. Park, J. Lee, J. Youn, and J. Chung. Seborrheic keratosis in the Korean males: causative role of sunlight. *Photodermatology, Photoimmunology & Photomedicine*, 19(2):73–80, 2003.

N. E. Lamb, S. B. Freeman, A. Savage-Austin, D. Pettay, L. Taft, J. Hersey, Y. Gu, J. Shen, D. Saker, K. M. May, D. Avramopoulos, M. B. Petersen, A. Hallberg, M. Mikkelsen, T. J. Hassold, and S. L. Sherman. Susceptible chiasmate configurations of chromosome 21 predispose to non–disjunction in both maternal meiosis I and meiosis II. *Nature Genetics*, 14:400–405, 1996.

M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499:214–218, 2013.

C. Le Goff, C. Mahaut, A. Abhyankar, W. Le Goff, V. Serre, A. Afenjar, A. Destrée, M. di Rocco, D. Héron, S. Jacquemont, S. Marlin, M. Simon, J. Tolmie, A. Verloes, J.-L. Casanova, A. Munnich, and V. Cormier-Daire. Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. *Nature Genetics*, 44: 85–88, 2011.

D. F. Lee, J. Lu, S. Chang, J. J. Loparo, and X. S. Xie. Mapping DNA polymerase errors by single-molecule sequencing. *Nucleic Acids Research*, 44(13):e118, 2016.

J. Lee, M. Kanatsu-Shinohara, K. Inoue, N. Ogonuki, H. Miki, S. Toyokuni, T. Kimura, T. Nakano, A. Ogura, and T. Shinohara. Akt mediates self-renewal division of mouse spermatogonial stem cells. *Development*, 134(10):1853–1859, 2007.

W. H. Lee, A. Raas-Rotschild, M. A. Miteva, G. Bolasco, A. Rein, D. Gillis, D. Vidaud, M. Vidaud, B. O. Villoutreix, and B. Parfait. Noonan syndrome type I with PTPN11 3 bp deletion: Structure and function implications. *Proteins*, 58(1):7–13, 2005.

E. M. LeProust, B. J. Peck, K. Spirin, H. B. McCuen, B. Moore, E. Namsaraev, and M. H. Caruthers. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research*, 38(8):2522–2540, 2010.

T. J. Ley, L. Ding, M. J. Walter, M. D. McLellan, T. Lamprecht, D. E. Larson, C. Kandoth, J. E. Payton, J. Baty, J. Welch, C. C. Harris, C. F. Lichti, R. R. Townsend, R. S. Fulton, D. J. Dooling, D. C. Koboldt, H. Schmidt, Q. Zhang, J. R. Osborne, L. Lin, M. O'Laughlin, J. F. McMichael, K. D. Delehaunty, S. D. McGrath, L. A. Fulton, V. J. Magrini, T. L. Vickery, J. Hundal, L. L. Cook, J. J. Conyers, G. W. Swift, J. P. Reed, P. A. Alldredge, T. Wylie, J. Walker, J. Kalicki, M. A. Watson, S. Heath, W. D. Shannon, N. Varghese, R. Nagarajan, P. Westervelt, M. H. Tomasson, D. C. Link, T. A. Graubert, J. F. DiPersio, E. R. Mardis, and R. K. Wilson. DNMT3A Mutations in Acute Myeloid Leukemia. *New England Journal of Medicine*, 363(25):2424–2433, 2010.

H. Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, 2014.

J. Lim, G. J. Maher, G. D. H. Turner, W. Dudka-Ruszkowska, S. Taylor, E. R. D. Meyts, A. Goriely, and A. O. M. Wilkie. Selfish spermatogonial selection: Evidence from an immunohistochemical screen in testes of elderly men. *PLoS ONE*, 7(8), 2012.

A. Logié, C. Dunois-Lardé, C. Rosty, O. Levrel, M. Blanche, A. Ribeiro, J. M. Gasc, J. Jorcano, S. Werner, X. Sastre-Garau, J. P. Thiery, and F. Radvanyi. Activating

mutations of the tyrosine kinase receptor FGFR3 are associated with benign skin tumors in mice and humans. *Human Molecular Genetics*, 14(9):1153–1160, 2005.

K. Lohmann, I. Masuho, D. N. Patil, H. Baumann, E. Hebert, S. Steinrücke, D. Trujillano, N. K. Skamangas, V. Dobricic, I. Hüning, G. Gillessen-Kaesbach, A. Westenberger, D. Savic-Pavicevic, A. Münchau, G. Oprea, C. Klein, A. Rolfs, and K. A. Martemyanov. Novel GNB1 mutations disrupt assembly and function of G protein heterotrimers and cause global developmental delay in humans. *Human Molecular Genetics*, 26(6):1078–1086, 2017.

C. Loveday, K. Tatton-Brown, M. Clarke, I. Westwood, A. Renwick, E. Ramsay, A. Nemeth, J. Campbell, S. Joss, M. Gardner, A. Zachariou, A. Elliott, E. Ruark, R. van Montfort, T. C. O. Collaboration, and N. Rahman. Mutations in the PP2A regulatory subunit B family genes PPP2R5B, PPP2R5C and PPP2R5D cause human overgrowth. *Human Molecular Genetics*, 24(17):4775–4779, 2015.

M. Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, 107(3):961–968, 2010.

G. J. Maher, A. Goriely, and A. O. M. Wilkie. Cellular evidence for selfish spermatogonial selection in aged human testes. *Andrology*, 2(3):304–314, 2014.

G. J. Maher, S. J. McGowan, E. Giannoulatou, C. Verrill, A. Goriely, and A. O. M. Wilkie. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proceedings of the National Academy of Sciences*, 113(9): 2454–2459, 2016a.

G. J. Maher, E. Rajpert-De Meyts, A. Goriely, and A. O. M. Wilkie. Cellular correlates of selfish spermatogonial selection. *Andrology*, 4(3):550–553, 2016b.

G. J. Maher, H. K. Ralph, Z. Ding, N. Koelling, H. Mlcochova, E. Giannoulatou, P. Dhami, D. S. Paul, S. H. Stricker, S. Beck, G. Mcvean, A. O. M. Wilkie, and A. Goriely. Selfish mutations dysregulating RAS-MAPK signaling are pervasive in aged human testes. *Genome Research*, (28):1779–1790, 2018.

D. Malaspina, S. Harlap, S. Fennig, D. Heiman, D. Nahon, D. Feldman, and E. S. Susser. Advancing paternal age and the risk of schizophrenia. *Archives of General Psychiatry*, 58(4):361–367, 2001.

D. Malaspina, C. Gilman, and T. M. Kranz. Paternal age and mental health of offspring. *Fertility and Sterility*, 103(6):1392–1396, 2015.

S. Mandel and I. Gozes. Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. *Journal of Biological Chemistry*, 282(47):34448–34456, 2007.

L. J. Manley, D. Ma, and S. S. Levine. Monitoring error rates in Illumina sequencing. *Journal of Biomolecular Techniques*, 27(4):125–128, 2016.

L. A. Martin, N. Assif, M. Gilbert, D. Wijewarnasuriya, and M. Seandel. Enhanced fitness of adult spermatogonial stem cells bearing a paternal age-associated FGFR2 mutation. *Stem Cell Reports*, 3(2):219–226, 2014.

I. Martincorena, A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M. R. Stratton, P. H. Jones, and P. J. Campbell. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, 2015.

A. Memon, J. Tomenson, J. Bothwell, and P. Friedmann. Prevalence of solar damage and actinic keratosis in a Merseyside population. *British Journal of Dermatology*, 142 (6):1154–1159, 2001.

X. Meng, M. Lindahl, M. E. Hyvönen, M. Parvinen, D. G. de Rooij, M. W. Hess, A. Raatikainen-Ahokas, K. Sainio, H. Rauvala, M. Lakso, J. G. Pichel, H. Westphal, M. Saarma, and H. Sariola. Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. *Science*, 287(5457):1489–1493, 2000.

M. Mills, R. R. Rindfuss, P. McDonald, and E. te Velde. Why do people postpone

parenthood? Reasons and social policy incentives. *Human Reproduction Update*, 17 (6):848–860, 2011.

A. E. Minoche, J. C. Dohm, and H. Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, 12(11):R112, 2011.

H. Moch, A. L. Cubilla, P. A. Humphrey, V. E. Reuter, and T. M. Ulbright. The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. *European Urology*, 70(1):93 – 105, 2016.

D. M. Moloney, S. R. Slaney, M. Oldridge, S. A. Wall, P. Sahlin, G. Stenman, and A. O. M. Wilkie. Exclusive paternal origin of new mutations in Apert syndrome. *Nature Genetics*, 13:48–53, 1996.

K. Moran-Crusio, L. Reavie, A. Shih, O. Abdel-Wahab, D. Ndiaye-Lobry, C. Lobry, M. Figueroa, A. Vasanthakumar, J. Patel, X. Zhao, F. Perna, S. Pandey, J. Madzo, C. Song, Q. Dai, C. He, S. Ibrahim, M. Beran, J. Zavadil, S. Nimer, A. Melnick, L. Godley, I. Aifantis, and R. Levine. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*, 20(1):11 – 24, 2011.

L. Murray, P. McCarron, K. Bailie, R. Middleton, G. Davey Smith, S. Dempsey, A. McCarthy, and A. Gavin. Association of early life factors and acute lymphoblastic leukaemia in childhood: historical cohort study. *British Journal Of Cancer*, 86:356–361, 2002.

K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13): e90, 2011.

C. K. Naughton, S. Jain, A. M. Strickland, A. Gupta, and J. Milbrandt. Glial cell-line derived neurotrophic factor-mediated RET signaling regulates spermatogonial stem cell fate1. *Biology of Reproduction*, 74(2):314–321, 2006.

B. M. Neale, Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C.-f. Lin, C. Stevens, L.-s. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, L. Lewis, Y. Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, E. H. Cook Jr, B. Devlin, R. A. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe, and M. J. Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, 2012.

V. N. Ngo, R. M. Young, R. Schmitz, S. Jhavar, W. Xiao, K.-H. Lim, H. Kohlhammer, W. Xu, Y. Yang, H. Zhao, A. L. Shaffer, P. Romesser, G. Wright, J. Powell, A. Rosenwald, H. K. Muller-Hermelink, G. Ott, R. D. Gascoyne, J. M. Connors, L. M. Rimsza, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, R. I. Fisher, R. M. Braziel, R. R. Tubbs, J. R. Cook, D. D. Weisenburger, W. C. Chan, and L. M. Staudt. Oncogenically active MYD88 mutations in human lymphoma. *Nature*, 470:115 EP –, 2010.

A. M. Nicholson, C. Olpe, A. Hoyle, A. S. Thorsen, T. Rus, M. Colombé, R. Brunton-Sim, R. Kemp, K. Marks, P. Quirke, S. Malhotra, R. ten Hoopen, A. Ibrahim, C. Lindskog, M. B. Myers, B. Parsons, S. Tavaré, M. Wilkinson, E. Morrissey, and D. J. Winton. Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell*, 22(6):909–918.e8, 2018.

A. Nikolic, V. Volarevic, L. Armstrong, M. Lako, and M. Stojkovic. Primordial germ cells: Current knowledge and perspectives. *Stem Cells International*, 2016:1741072, 2016.

J. M. Oatley, M. R. Avarbock, A. I. Telaranta, D. T. Fearon, and R. L. Brinster. Identifying genes important for spermatogonial stem cell self-renewal and survival. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9524–9, 2006.

M. Oldridge, P. W. Lunt, E. H. Zackai, D. M. McDonald-McGinn, M. Muenke, D. M. Moloney, S. R. Twigg, J. K. Heath, T. D. Howard, G. Hoganson, D. M. Gagnon, E. W. Jabs, and A. O. M. Wilkie. Genotype-phenotype correlation for nucleotide substitutions in the IgII- IgIII linker of FGFR2. *Human Molecular Genetics*, 6(1):137–143, 1997.

B. J. O'Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, 2012.

B. J. O'Roak, H. A. Stessman, E. A. Boyle, K. T. Witherspoon, B. Martin, C. Lee, L. Vives, C. Baker, J. B. Hiatt, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature Communications*, 5(5595):1–6, 2014.

V. Ostapcuk, F. Mohn, S. H. Carl, A. Basters, D. Hess, V. Iesmantavicius, L. Lampersberger, M. Flemr, A. Pandey, N. H. Thomä, J. Betschinger, and M. Bühler. Activity-dependent neuroprotective protein recruits HP1 and CHD4 to control lineage-specifying genes. *Nature*, 557(7707):739–743, 2018.

B. L. Parsons, K. E. Marchant-Miros, R. R. Delongchamp, T. L. Verkler, T. A. Patterson, P. B. McKinzie, and L. T. Kim. ACB-PCR Quantification of K-RAS Codon 12 GAT and GTT Mutant Fraction in Colon Tumor and Non-Tumor Tissue. *Cancer Investigation*, 28(4):364–375, 2010.

L. S. Penrose. Parental age in achondroplasia and mongolism. *American Journal of Human Genetics*, 9(1912):167–9, 1957.

M. F. Pescosolido, M. Schwede, A. Johnson Harrison, M. Schmidt, E. D. Gamsiz, W. S. Chen, J. P. Donahue, N. Shur, B. A. Jerskey, C. Phornphutkul, and E. M. Morrow. Expansion of the clinical phenotype associated with mutations in activity-dependent neuroprotective protein. *Journal of Medical Genetics*, 51(9):587–589, 2014.

S. Petrovski, S. Kury, C. T. Myers, K. Anyane-Yeboa, B. Cogné, M. Bialer, F. Xia, P. Hemati, J. Riviello, M. Mehaffey, T. Besnard, E. Becraft, A. Wadley, A. R. Politi, S. Colombo, X. Zhu, Z. Ren, I. Andrews, T. Dudding-Byth, A. L. Schneider, G. Wallace, A. B. Rosen, S. Schelley, G. M. Enns, P. Corre, J. Dalton, S. Mercier, X. Latypova, S. Schmitt, E. Guzman, C. Moore, L. Bier, E. L. Heinzen, P. Karachunski, N. Shur, T. Grebe, A. Basinger, J. M. Nguyen, S. Bézieau, K. Wierenga, J. A. Bernstein, I. E. Scheffer, J. A. Rosenfeld, H. C. Mefford, B. Isidor, and D. B. Goldstein. Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability, Hypotonia, and Seizures. *American Journal of Human Genetics*, 98(5):1001 – 1010, 2016.

B. Popp, S. I. Støve, S. Endele, L. M. Myklebust, J. Hoyer, H. Sticht, S. Azzarello-Burri, A. Rauch, T. Arnesen, and A. Reis. De novo missense mutations in the NAA10 gene cause severe non-syndromic developmental delay in males and females. *European Journal Of Human Genetics*, 23:602 EP –, 2014.

V. Potapov and J. L. Ong. Examining Sources of Error in PCR by Single- Molecule Sequencing. *PLoS One*, 12(7):e0181128, 2017.

R. Pun, J. D. Uhl, S. N. Schmeltzer, C. Faulkner, S. L. Bronson, B. L. Murphy, D. A. Richards, K. D. Holland, and S. C. Danzer. Excessive activation of mTOR in postnatally-generated granule cells is sufficient to cause epilepsy. *Neuron*, 75(6): 1022–1034, 2013.

J. Qin, P. Calabrese, I. Tiemann-Boege, D. N. Shinde, S. R. Yoon, D. Gelfand, K. Bauer, and N. Arnheim. The molecular anatomy of spontaneous germline mutations in human testes. *PLoS Biology*, 5(9):1912–1922, 2007.

R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton, and M. E. Hoppurles. Timing, rates and spectra of human germline mutation. *Nature Genetics*, 48(2):126–133, 2016.

S. V. Rannan-Eliya, I. B. Taylor, I. M. de Heer, A. M. van den Ouweland, S. A. Wall, and A. O. Wilkie. Paternal origin of FGFR3 mutations in Muenke-type craniosynostosis. *Human Genetics*, 115(3):200–207, 2004.

A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Endele, T. Schwarzmayr, B. Albrecht, D. Bartholdi, J. Beygo, N. Di Donato, A. Dufke, K. Cremer, M. Hempel, D. Horn, J. Hoyer, P. Joset, A. Röpke, U. Moog, A. Riess, C. T. Thiel, A. Tzschach, A. Wiesener, E. Wohlleber, C. Zweier, A. B. Ekici, A. M. Zink, A. Rump, C. Meisinger, H. Grallert, H. Sticht, A. Schenck, H. Engels, G. Rappold, E. Schröck, P. Wieacker, O. Riess, T. Meitinger, A. Reis, and T. M. Strom. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: An exome sequencing study. *The Lancet*, 380(9854):1674–1682, 2012.

W. Reardon, R. M. Winter, P. Rutland, L. J. Pulleyn, B. M. Jones, and S. Malcolm. Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. *Nature Genetics*, 8(1):98–103, 1994.

A. Reichenberg, R. Gross, M. Weiser, M. Bresnahan, J. Silverman, S. Harlap, J. Rabinowitz, C. Shulman, D. Malaspina, G. Lubin, H. Knobler, M. Davidson, and E. Susser. Advancing paternal age and autism. *Archives of General Psychiatry*, 63(9):1026–1032, 2006.

M. R. F. Reijnders, M. Kousi, G. M. van Woerden, M. Klein, J. Bralten, G. M. S. Mancini, T. van Essen, M. Proietti-Onori, E. E. J. Smeets, M. van Gastel, A. P. A. Stegmann, S. J. C. Stevens, S. H. Lelieveld, C. Gilissen, R. Pfundt, P. L. Tan, T. Kleefstra, B. Franke, Y. Elgersma, N. Katsanis, and H. G. Brunner. Variation in a range of mTOR-related genes associates with intracranial volume and intellectual disability. *Nature Communications*, 8(1):1052, 2017.

N. Risch, E. W. Reich, M. M. Wishnick, and J. G. McCarthy. Spontaneous mutation and parental age in humans. *American Journal of Human Genetics*, 41(2):218–48, 1987.

J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

B. J. O. Roak, L. Vives, W. Fu, J. D. Egertson, I. B. Stanaway, I. G. Phelps, G. Carvill, A. Kumar, C. Lee, K. Ankenman, J. Munson, J. B. Hiatt, E. H. Turner, R. Levy, D. R. O. Day, N. Krumm, B. P. Coe, B. K. Martin, E. Borenstein, D. A. Nickerson, H. C. Mefford, D. Doherty, J. M. Akey, R. Bernier, E. E. Eichler, and J. Shendure. Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science*, 338:1619–1623, 2012.

F. Rousseau, J. Bonaventure, L. Legeai-Mallet, A. Pelet, J. M. Rozet, P. Maroteaux, M. Le Merrer, and A. Munnich. Mutations in the gene encoding fibroblast growth factor receptor-3 in achondroplasia. *Nature*, 371(6494):252–254, 1994.

J. J. Salk, M. W. Schmitt, and L. A. Loeb. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics*, 19: 269 EP –, 2018.

K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, and M. J. Daly. A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9):944–950, 2014.

S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Šestan, R. P. Lifton, M. Günel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State. De novo

mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012.

A. Sarkozy, R. Conti, D. Seripa, M. Digilio, N. Grifone, C. Tandoi, V. Fazio, V. Di Ciommo, B. Marino, A. Pizzuti, and B. Dallapiccola. Correlation between PTPN11 gene mutations and congenital heart defects in Noonan and LEOPARD syndromes. *Journal of Medical Genetics*, 40(9):704–708, 2003.

K. Scheffzek, M. R. Ahmadian, W. Kabsch, L. Wiesmüller, A. Lautwein, F. Schmitz, and A. Wittinghofer. The Ras-RasGAP Complex: Structural Basis for GTPase Activation and Its Loss in Oncogenic Ras Mutants. *Science*, 277(5324):333–339, 1997.

M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):1–15, 2016.

M. W. Schmitt, E. J. Fox, M. J. Prindle, K. S. Reid-Bayliss, L. D. True, J. P. Radich, and L. A. Loeb. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods*, 12:423–425, 2015.

L. Ségurel, M. J. Wyman, M. Przeworski, S. Laure, M. J. Wyman, and M. Przeworski. Determinants of Mutation Rate Variation in the Human Germline. *Annual review of genomics and human genetics*, (15):47–70, 2014.

M. Sharland, M. Burch, W. M. McKenna, and M. A. Paton. A clinical study of Noonan syndrome. *Archives of Disease in Childhood*, 67:178–183, 1992.

W. Shen, J. M. Heeley, C. M. Carlston, R. Acuna-Hidalgo, W. M. Nillesen, K. M. Dent, G. V. Douglas, K. L. Levine, P. Bayrak-Toydemir, C. L. Marcelis, M. Shinawi, and J. C. Carey. The spectrum of DNMT3A variants in Tatton–Brown–Rahman syndrome overlaps with that in hematologic malignancies. *American Journal of Medical Genetics, Part A*, 173(11):3022–3028, 2017.

R. Shiang, L. M. Thompson, Y. Z. Zhu, D. M. Church, T. J. Fielder, M. Bocian, S. T. Winokur, and J. J. Wasmuth. Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell*, 78(2):335–342, 1994.

D. N. Shinde, D. P. Elmer, P. Calabrese, J. Boulanger, N. Arnheim, and I. Tiemann-boege. New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. *Human Molecular Genetics*, 22(20):4117–4126, 2013.

R. A. J. Tagelenbosch and D. G. de Rooij. A quantitative study of spermatogonial multiplication and stem cell renewal in the C3H/101 F1 hybrid mouse. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 290(2):193–200, 1993.

T. Takenouchi, T. Miwa, Y. Sakamoto, Y. Sakaguchi, T. Uehara, T. Takahashi, and K. Kosaki. Further evidence that a blepharophimosis syndrome phenotype is associated with a specific class of mutation in the ADNP gene. *American Journal of Medical Genetics Part A*, 173(6):1631–1634, 2017.

M. Tartaglia, E. L. Mehler, R. Goldberg, G. Zampino, H. G. Brunner, H. Kremer, I. van der Burgt, A. H. Crosby, A. Ion, S. Jeffery, K. Kalidas, M. A. Patton, R. S. Kucherlapati, and B. D. Gelb. Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. *Nature Genetics*, 29:465 EP –, 2001.

P. L. Tavormina, R. Shiang, L. M. Thompson, Y.-Z. Zhu, D. J. Wilkin, R. S. Lachman, W. R. Wilcox, D. L. Rimoin, D. H. Cohn, and J. J. Wasmuth. Thanatophoric dysplasia (types I and II) caused by distinct mutations in fibroblast growth factor receptor 3. *Nature Genetics*, 9(3):321–328, 1995.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

I. Tiemann-Boege, W. Navidi, R. Grewal, D. Cohn, B. Eskenazi, A. J. Wyrobek, and

N. Arnheim. The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. *Proceedings of the National Academy of Sciences*, 99(23): 14952–14957, 2002.

S. P. Treon, L. Xu, G. Yang, Y. Zhou, X. Liu, Y. Cao, P. Sheehy, R. J. Manning, C. J. Patterson, C. Tripsas, L. Arcaini, G. S. Pinkus, S. J. Rodig, A. R. Sohani, N. L. Harris, J. M. Laramie, D. A. Skifter, S. E. Lincoln, and Z. R. Hunter. MYD88 L265P Somatic Mutation in Waldenström's Macroglobulinemia. *New England Journal of Medicine*, 367(9):826–833, 2012.

T. N. Turner, B. P. Coe, D. E. Dickel, K. Hoekzema, B. J. Nelson, M. C. Zody, Z. N. Kronenberg, F. Hormozdiari, A. Raja, L. A. Pennacchio, R. B. Darnell, and E. E. Eichler. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*, 171(3): 710–722.e12, 2017.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

J. A. Veltman and H. G. Brunner. De novo mutations in human genetic disease. *Nature Reviews Genetics*, 13(8):565–575, 2012.

F. Vogel and R. Rathenberg. Spontaneous mutation in man. *Advances in Human Genetics*, 5:223–318, 1975.

J. C. Walser and A. V. Furano. The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Research*, 20(7):875–882, 2010.

W. Weinberg. Zur Vererbung des Zwergwuchses. *Arch. Rassen-u Gesell. Biol.*, 9:710–718, 1912.

J. Weischenfeldt, O. Symmons, F. Spitz, and J. O. Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14:125–138, 2013.

A. O. M. Wilkie. Bad bones, absent smell, selfish testes: The pleiotropic consequences of human FGF receptor mutations. *Cytokine and Growth Factor Reviews*, 16(2):187–203, 2005.

A. O. M. Wilkie, S. F. Slaney, M. Oldridge, M. D. Poole, G. J. Ashworth, A. D. Hockley, R. D. Hayward, D. J. David, L. J. Pulleyn, P. Rutland, S. Malcolm, R. M. Winter, and W. Reardon. Apert syndrome results from localized mutations of FGFR2 and is allelic with Crouzon syndrome. *Nature Genetics*, 9:165–172, 1995.

D. J. Wilkin, J. K. Szabo, R. Cameron, S. Henderson, G. A. Bellus, M. L. Mack, I. Kaitila, J. Loughlin, A. Munnich, B. Sykes, J. Bonaventure, and C. A. Francomano. Mutations in Fibroblast Growth-Factor Receptor 3 in Sporadic Cases of Achondroplasia Occur Exclusively on the Paternally Derived Chromosome. *American Journal of Human Genetics*, 63(3):711–716, 1998.

K. P. Willführ and S. Klüsener. The Evolution of Mean Paternal Age from 1900 to the Present Day – Are today's father really older than back in the days? *European Population Conference, Mainz, Germany. Presented at session 111: Before, during and after the fertility transition.*, 2016.

M. Xie, C. Lu, J. Wang, M. D. McLellan, K. J. Johnson, M. C. Wendl, J. F. McMichael, H. K. Schmidt, V. Yellapantula, C. A. Miller, B. A. Ozenberger, J. S. Welch, D. C. Link, M. J. Walter, E. R. Mardis, J. F. Dipersio, F. Chen, R. K. Wilson, T. J. Ley, and L. Ding. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine*, 20(12):1472–1478, 2014.

X.-J. Yan, J. Xu, Z.-H. Gu, C.-M. Pan, G. Lu, Y. Shen, J.-Y. Shi, Y.-M. Zhu, L. Tang, X.-W. Zhang, W.-X. Liang, J.-Q. Mi, H.-D. Song, K.-Q. Li, Z. Chen, and S.-J. Chen. Ex-

ome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nature Genetics*, 43(4):309–315, 2011.

J. Yeatman, M. Kilkenny, and R. Marks. The prevalence of seborrhoeic keratoses in an Australian population: does exposure to sunlight play a part in their frequency? *British Journal of Dermatology*, 137(3):411–414, 2008.

S.-R. Yoon, J. Qin, R. L. Glaser, E. Wang Jabs, N. S. Wexler, R. Sokol, N. Arnheim, and P. Calabrese. The ups and downs of mutation frequencies during aging can account for the Apert syndrome paternal age effect. *PLoS Genetics*, 5(7):1–9, 2009.

S. R. Yoon, S. K. Choi, J. Eboreime, B. D. Gelb, P. Calabrese, and N. Arnheim. Age-dependent germline mosaicism of the most common Noonan syndrome mutation shows the signature of germline selection. *American Journal of Human Genetics*, 92(6):917–926, 2013.

A. L. Young, G. A. Challen, B. M. Birmann, and T. E. Druley. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nature Communications*, 7:1–7, 2016.

R. Zamostiano, A. Pinhasov, E. Gelber, R. A. Steingart, E. Seroussi, E. Giladi, M. Bassan, Y. Wollman, H. J. Eyre, J. C. Mulley, D. E. Brenneman, and I. Gozes. Cloning and characterization of the human activity-dependent neuroprotective protein. *Journal of Biological Chemistry*, 276(1):708–714, 2001.

S. Zhang, B. Wang, L. Wan, and L. M. Li. Estimating Phred scores of Illumina base calls by logistic regression and sparse modeling. *BMC Bioinformatics*, 18(1):1–14, 2017.

F. Zink, S. N. Stacey, G. L. Norddahl, M. L. Frigge, O. T. Magnusson, I. Jonsdottir, T. E. Thorgeirsson, A. Sigurdsson, S. A. Gudjonsson, J. Gudmundsson, J. G. Jonasson, L. Tryggvadottir, T. Jonsson, A. Helgason, A. Gylfason, P. Sulem, T. Rafnar, U. Thorsteinsdottir, D. F. Gudbjartsson, G. Masson, A. Kong, and K. Stefansson. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, 130(6):742–752, 2017.