Comparative analysis of derived mutations across human populations.



Alison Feder

Magdalen College

University of Oxford

A thesis submitted for the degree of

MSc (Res) Statistics

August 2013

Abstract

Comparative analysis of derived mutations across human populations.

Alison Feder, Magdalen College

Msc(res) Thesis, Trinity Term, 2013

Multiple factors, including the rate of mutation of different types, the action of natural selection, recombination and population history all influence the distribution of genetic variation within and between populations. Here, we seek to explore the relative contributions of different factors to notable aspects of the derived site frequency spectrum in different populations from the 1000 Genomes Project. Specifically, we ask whether individuals from different populations carry different numbers of mutations (indicative of differences in the mutation rate or average generation time of populations) and whether patterns differ between types of mutation in different sequence contexts. Our results indicate that the average mutation rate (per year) in African populations has been slightly, but significantly, higher than in non-African populations which has led to an enrichment in the number of derived alleles harbored in African populations. In a departure from the literature, we find that the non-synonymous mutation load does not vary between populations. We present weak evidence for differential mutational processes at CpG sites in African individuals and estimate a TMRCA of the human population as approximately 3.5 mya. Last, we suggest ways in which our knowledge of derived alleles can allow us to make improvements to existing population genetic methods, such as the sequentially Markov coalescent.

Acknowledgements

Effusive thanks to Gil McVean for his guidance and mentorship in this project, members of the McVean group for their thoughtful feedback and support and my transfer and viva examiners for giving their time to help me grow as a scientist. I am also grateful for friends and family far and near, particularly Frank Thompson for his unrelenting encouragement and editing. Last, thanks to the Thouron Award, without which, study at Oxford would have been impossible.

Contents

1 The Importance of Derived Alleles **1**

1.1 Rare Derived Variation in Genomic Analysis 1

1.1.1 Characterization of Rare Variation 2

1.1.2 Sources of Rare Variation 3

1.1.3 Problems associated with working with rare variation 5

1.1.4 Importance of Rare Variation 6

1.2 Non-synonymous Population Load 7

1.2.1 Number of non-synonymous alleles in a human genome 7

1.2.2 Non-synonymous alleles vary by population 8

1.3 Derived Variation as a Predictor of Time to Most Recent Common Ancestor 9

1.4 Thesis Outline 13

2 Derived Allele Differences between Populations **17**

2.1 Introduction 17

2.2 Theoretical basis for expectations 17

2.3 Data Background 21

2.3.1 Description 22

2.3.2 Masking 22

2.3.3 SNP selection 23

2.3.4 Power Correction 26

2.4 Differences Between Populations in the Derived Allele Frequencies . . . 29

2.4.0.1 Distortion in the Derived Allele Frequency Spectrum . 29

2.4.0.2 High frequency derived alleles 31

2.5 Differing numbers of derived alleles in different populations 35

2.5.0.3 Time to Most Recent Common Ancestor 37

2.5.1 Derived Mutations Differences Driven by CpG Site Imbalance . . 38
2.5.2 Non-synonymous Load Similar in Different Populations 48
2.6 Summary 52

3 Incorporation of Derived Allele Information into Models of Coalescence Prediction 53
3.1 Introduction 53
3.1.1 Standard PSMC 54
3.2 Future Directions 57

References 59

Glossary

BED file - Annotation file used to specify areas of the genome with a chromosome number, start and end position.

Reference SNP Cluster ID (RSID) - Unique SNP identifier used to aggregate reported polymorphisms.

dbSNP - Short polymorphism database which can be accessed here: http://www.ncbi.nlm.nih.gov/projects/SNP/

Variant Call Format (VCF) - Data file that lists polymorphism information across multiple genomes.

FASTA - Data file listing a string one-letter codes in order to represent sequence information.

1

The Importance of Derived Alleles

As mutations enter the human population over time, they leave patterns that can help us understand important features of how humans have evolved. Many of these patterns can be identified in the site frequency spectrum (SFS), which describes the distribution of alleles at different frequencies in the population. I ask how we can use these patterns of derived alleles to better understand relationships between human populations.

I start by examining the importance of rare derived variation in genomic analysis and its impact on estimates of demography, population relatedness and the way we understand selection. From here, I look specifically at the distributions of non-synonymous derived alleles between populations, and the potential consequences of those distributions. I then move into a discussion of using derived variation to estimate the time to the most recent common ancestor (TMRCA) of populations. In combining these three topics, I introduce and motivate the use of derived variation in helping us understand human evolution.

1.1 Rare Derived Variation in Genomic Analysis

All derived variants originally enter the populations as single mutations, and thus spend some time as very rare alleles, even if they confer a strong selective advantage. Whether these new mutations rise to higher frequencies or remain rare in the population depends on selection, drift, linkage and demography. We can gain insights into population

processes by tracing their contribution in the site frequency spectrum. Since new variants have not had sufficient time to be affected by ancient selection and demography, they offer additional important information about recent history in ways that other derived variation cannot.

1.1.1 Characterization of Rare Variation

Rare variation has been explored recently in a series of exome deep sequencing efforts conducted by different groups. Tennessen et al (51) sequenced 200 Danish exomes at a 12-fold sequencing depth to create a dataset of over 53,000 coding SNPs. Nelson et al (39) sequenced 202 drug target genes in 14,002 people to at least 20x coverage. Li et al (29) sequenced 15,585 human genes to a depth of 111x in 2,440 individuals of European or African ancestry. Marth et al (33) also report on initial stages of the 1000 Genomes Exon Pilot project, sequencing around 700 samples of 1000 human genes.

These studies are in agreement about a number of facts: rare variation is common, population specific and non-synonymous variants are concentrated in the spectrum of rare alleles.

First, rare variation is common. Nelson et al (39) found that in the drug target genes, 95% of variants occurred at frequency $< 0.5\%$ and as frequently as 1 every 17 and 21 bp in Europeans and Africans respectively. Tennessen et al's estimation was only slightly more conservative, suggesting 86% had a minor allele frequency $< 0.5\%$. Further, Tennessen estimated that 57% of SNPs were singletons (51). Li et al did not attempt to characterize the very lowest of allele frequencies, but performed extensive validation to ensure accurate allele information for the sites they did include (29). An additional targeted re-sequencing effort of 2 genes in 13,715 individuals by Coventry et al (11) found a five fold increase in singletons over neutral expectations.

Second, rare variation is private. Nelson et al (39) found that more than 75% of the alleles in their sample were present in only one or two individuals. Tennessen et al (51) observed that even excluding singletons, 58% of coding SNPs were population specific, and that the frequency of singletons (by definition private) varied between populations, with Africans having only 49.3% singletons compared to Europeans' 50.7% singletons. Marth et al (33) examined allele sharing between populations at low, mid and high frequencies in the coding regions of the 1000 Genomes exon pilot and compared them to the same allele frequency classes in intergenic regions from the HapMap3 ENCODE

resequencing project. They found that while genic and intergenic regions had similar levels of allele sharing in mid and high frequency sites, there was significantly less sharing in low frequency ($< 1\%$) alleles in genic regions than in intergenic regions. This observation suggests that these coding variants are comparatively younger than their low frequency intergenic counterparts, possibly because of purifying selection.

All these studies have agreed that rare variation is of significant functional importance. Non-synonymous variants are enriched at low frequencies in a way that is not seen at higher frequencies, highlighting how functional constraint can prevent functional variants from rising in incidence (39). Li et al found a 1.8 excess of non-synonymous mutations over synonymous mutations in the low-frequency range (2-5%) (29). Tennessen et al estimated that 95.7% of functionally important SNPs in their sample were rare (51) and addressed the importance of purifying selection in creating this pattern. They estimated the functional consequences of non-synonymous SNPs and found that they were negatively correlated with allele frequency, suggesting that rare variation is more functionally important (51). Marth et al (33) found that whereas 63% of missense mutations had allele frequencies $< 1\%$, only 53% of silent variants fell into this category.

1.1.2 Sources of Rare Variation

The prevalence of rare variation is often attributed to explosive population growth over the past several thousand years (Figure 1.1), with estimates ranging from a more conservative 0.23 (0.16-0.34)% per generation growth over the past 26,400 (21,700-30,700) years (18) to very rapid 9.4 (4.5-14.5)% per generation growth over the past 1,400 (900 - 2,800) years (11), and many estimates falling in between.

In particular, extreme recent population growth leads to an excess of the very rarest variants, with predictions given that SNPs at frequency of $> 5\%$ have been segregating for more than 10,000 years (25). This rapid growth has resulted in an increase in rare variants, with new variants not yet having time to spread in the population (11, 25). If a population of constant size is given sufficient time, mutation and selection will eventually reach an equilibrium. The dynamics in a growing population are more complicated. The larger the population, the more efficacious selection is when compared to drift. Resultantly, deleterious mutations decrease more quickly in relative frequency when compared to neutral mutations. However, since all individuals in a growing

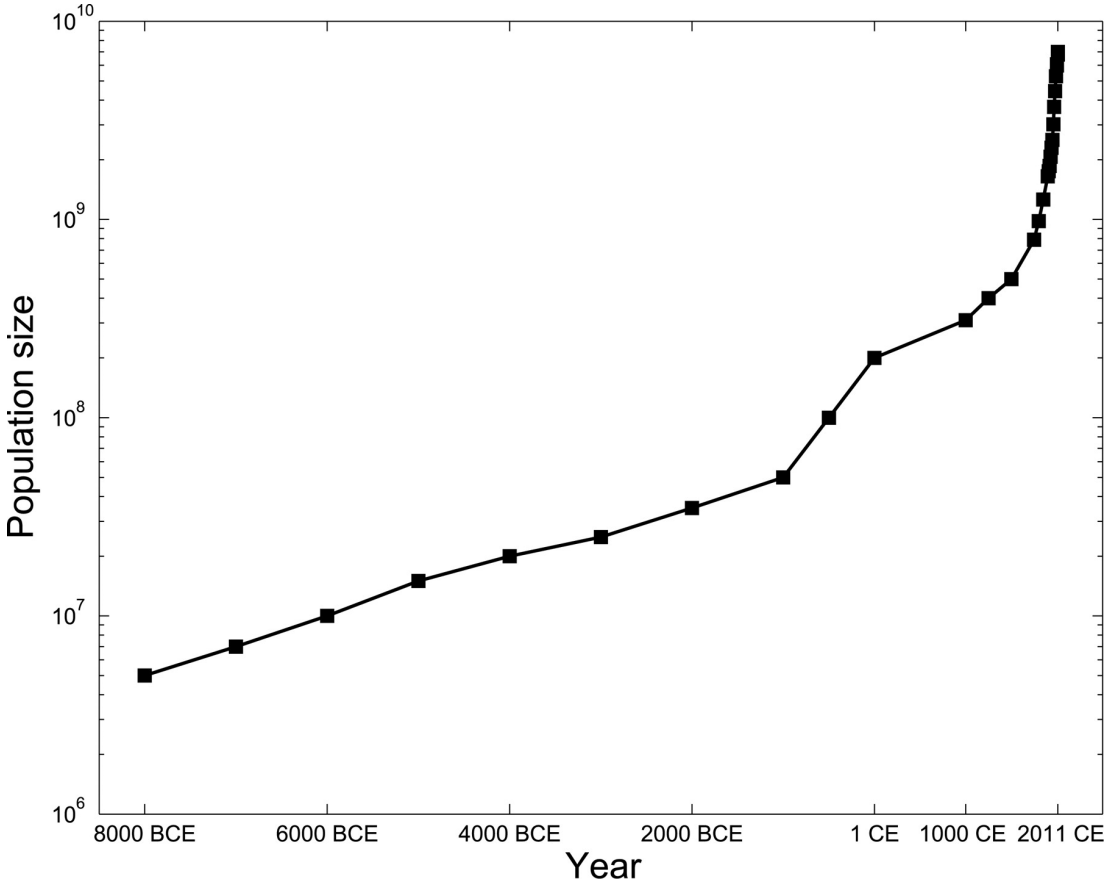


Figure 1.1: Explosive population size estimates as a function of time. - From Keinan A, Clark AG: **Recent explosive human population growth has resulted in an excess of rare genetic variants.** Science 2012, **336(6082):**740-743. Reprinted with permission from AAAS.

population are having more offspring on average, more deleterious alleles are still passed along. In this way, although selection acts more strongly against deleterious alleles in an expanding population, they persist in the population longer given that there are a greater number of alleles in total.

Theoretical models also suggest the conditions under which rare variants can be maintained in the population. Pritchard proposed a model for complex disease susceptibility loci under weak selection and found that there are fewer mid frequency alleles than expected under neutrality (41). Eyre-Walker introduced a model in which mutations have both an effect on selection and also a pleiotropic effect on other alleles. Under this model, pleiotropy can keep mutations at low frequency (13).

1.1.3 Problems associated with working with rare variation

Rare variation has become increasingly important over the past several years. This has represented a shift in perspective from 2006, when Hinds et al's paper characterizing the spectrum of variation suggested that common variants explained more phenotypic variance than rare ones and were therefore more useful diagnostically (22). As genome-wide association studies have failed to find common loci that explain a large proportion of the heritability, rare variation represents a new avenue through which to find disease-associated alleles (46). However, working with rare variation is not easy.

At the sequencing level, current technology requires significant depth on many individuals in different populations in order to ascertain accurate allele frequency estimates. It can be difficult to balance newly discovered genotypes versus false positives generated from sequencing errors. An allele found only once in a population can be difficult to distinguish from a sequencing error. Although this can be resolved by sequencing individuals at greater depths (11), this can be costly. Accurate allele calls must be done across many individuals to permit discovery of very low frequency variants. Further, even if a SNP is present and discovered, there may be some bias against calling a rare allele (2).

Problems finding and calling rare variants in the sequencing stage are also found in the sampling stages of study creation. Because much variation in a population is private (51), populations must be sampled widely and diversely in order to reveal rare variants. The HapMap Consortium attempted to use discovered polymorphism in one population to find new variants in another. Because of the amount of private variation, the study

found low power to use other populations to find new alleles citeAltshuler:2010qd. In some respects, this can be considered a positive result, in that more sampling leads to discovery of more variants. Nelson et al found that the rate of variant discovery remained nearly constant with increasing sample size (39).

On top of sequencing and sampling problems, statistical errors can occur that further exacerbate the difficulty of calling rare variants. To address this, Li et al (29) develop an unbiased estimator of allele frequencies integrating error rates and variation in coverage, which they use to determine allele frequencies in the population. However, their method is only reliable for SNPs at greater than 2% frequency, again highlighting the difficulty of inferring allele frequencies at the extremes of the spectrum. The HapMap Consortium also encountered allele frequency calling difficulties when they found significant differences in called allele frequency distributions between the first and second phases of the HapMap Project (26).

Further, misidentification of ancestral alleles results in having a folded site frequency spectrum and biased analysis. This can occur in situations documented in (21).

1.1.4 Importance of Rare Variation

For the above reasons and because of the previous scarcity of data , rare variants have seldom been used in the analysis of genomes and the study of how populations relate to each other. However, rare variation can provide a variety of insights.

First, although the frequency of common variants is influenced via linkage to nearby SNPs, new variants have not had time to be significantly affected by their background. As such, rare allele frequencies can act well as a proxy for age of mutation (15). However, blindly using rare variation as a molecular clock comes with its own problems, given that a large number of substitutions are unobserved (21), and can lead to biased results when performing analyses such as estimating the time to the most recent common ancestor (4).

Rare variants are useful in the exploration of human demography, as their signal provides information about recent history not readily available in more common variants. Several studies have used shared rare genetic variation in order to disentangle human demographic patterns over the last several thousand years (17, 18, 25, 39). The presence of this rare variation further highlights the need to integrate demography into population genetic methods (39, 40). However, outside of simply estimating

demographic parameters, rare variation has been used more directly though looking at how alleles shared between populations can explain the age of mutations and haplotype sharing (2, 15).

Rare variation is often cited as a potential way to resolve the missing heritability question (32). In Nelson et al's study of drug target genes, rare variants were tested for association with their corresponding drug-targeted disease, but no statistically significant associations were found (39). This provides some evidence that including rare variation in association studies will not resolve all missing heritability concerns.

1.2 Non-synonymous Population Load

As noted in the introduction, the proportion of non-synonymous variants is higher among very rare alleles. Non-synonymous mutations sometimes result in changes that affect the amino acid composition of a protein and therefore they can have damaging effects on protein function. While not all non-synonymous substitutions are deleterious, these functional changes represent an opportunity to understand the role that purifying selection plays in human evolution.

An interest in deleterious non-synonymous mutations has existed in the literature for many years, since Moreton predicted in 1956 that a human genome carries several lethal mutations in heterozygote form (37). Since then, much work has been dedicated to establishing the extent of non-synonymous and deleterious mutations.

1.2.1 Number of non-synonymous alleles in a human genome

Multiple approaches have yielded similar estimates for the non-synonymous mutation load in a human genome. Sunyaev et al (47) estimate around 2,000 deleterious mutations in a human diploid genome, reasoning that non-synonymous mutations happen about once every 1,000 amino acids. Given an estimate 45,000 proteins with on average 500 amino acids per protein, 20,000 amino acids would be altered. With an estimated 10% of SNPs resulting in a deleterious change, the average number of deleterious mutations in a human diploid genome is around 2000 (47). Empirical estimates have found between 796 and 837 deleterious alleles in a sample of three individuals (9).

1.2.2 Non-synonymous alleles vary by population

It has been noted that there are between population differences in the numbers of non-synonymous and presumably deleterious alleles in a human genome. Lohmueller et al (31) examined a cohort of 15 African American and 20 European American individuals and categorized synonymous and non-synonymous SNPs as shared or private. They found that although there were more derived alleles among both synonymous and non-synonymous mutations in the African American individuals than in the European Americans, European Americans had a higher proportion of private alleles categorized as non-synonymous (55.4% compared to 47%). Further, this was paired with decreased heterozygosity within the European American sample, suggesting that more of these variants are present in multiple copies. They then performed a forward simulation and found that in an out-of-Africa model with uniform selective effects, simulated European populations had more deleterious alleles "surf" to high frequency as the population quickly grew, resulting in overall accrual of deleterious mutations similar to those seen in the data (31).

The results of other studies support Lohmueller et al (31). Boyko et al concentrated mainly on estimating the distribution of fitness effects in different populations, but they reached similar conclusions about the number of non-synonymous mutations in European populations. Interestingly, they also found no evidence to suggest that the distribution of fitness effects differed between African Americans and European Americans. As in Lohmueller et al, this suggests the important role differing population sizes play in evolutionary dynamics (7), as opposed to just selective effects.

Chun et al (9) examined the number of derived alleles in two individuals of European ancestry and one of Asian ancestry. They found that the estimated number of deleterious alleles among the three individuals ranged from 796 to 837, with the Asian individual presenting the highest number of derived deleterious alleles in their sample. While not analogous to the result posited by Lohmueller et al, this estimate suggests that different-populations may carry different numbers of deleterious alleles. However, with this small a sample, it may not be possible to draw a rpbust conclusion. Fu et al found that in their sample of 6515 exomes, 54.7% of European American genes and only 37.5% of African American genes harbor a new SNP deleterious to the function of the gene (15).

Chun and Fay (10) explored the effect of hitchhiking on the number of derived alleles in a population and found evidence that both supports and contradicts Lohmueller et al's results. Their analysis focused on the ratio of non-synonymous alleles that were deleterious to those that were non-deleteirous in regions linked to positively selected sites. They hypothesized that there would be an enrichment in deleterious alleles in these regions, since purifying selection would compete against positive selection. In using differently sized iHS (44) haplotype regions as filters for surrounding SNPs, they observed that there was an enrichment in the ratio of deleterious to non-deleterious non-synonymous SNPs around genes under positive selection in European populations, but no such enrichment existed in Asian and African populations. This suggests that hitchhiking has played a greater role in raising the number of derived alleles in European populations, compared with African populations.

However, when using a different parametrization for the area linked to positive selection, Chun and Fay found that all three populations were enriched for deleterious alleles. Further, they go on to note that while hitchhiking and positive selection do influence deleterious alleles, the effect is relatively minor (10).

This and other studies were equivocal about differences in the numbers of non-synonymous deleterious alleles in populations. In an early study of SNP patterns, Halushka et al (19) noted that estimates of nucleotide diversity were greater in African American samples than European American samples at synonymous, non-synonymous and non-coding SNPs. However, because this analysis was done only in a sample of 75 human genes related to blood pressure homeostasis and nearby noncoding regions, it may not be legitimate to generalize this result to the rest of the genome.

1.3 Derived Variation as a Predictor of Time to Most Recent Common Ancestor

The time to the most recent common ancestor or TMRCA (Figure 1.2) is an important population parameter that is of interest to geneticists, anthropologists, archaeologists and other researchers of human populations. Variation arising since the most recent common ancestor offers important clues into the timescale of human polymorphism.

Archaeological records date the oldest fossils of anatomically modern humans to about 196 kya ($\pm$ 2 ky) (30). It is not possible to tell, however, if this fossil is a most

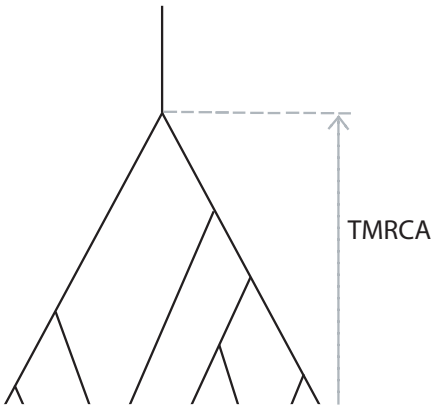


Figure 1.2: Time to Most Recent Common Ancestor - The grey line represents the time at which all sequences in a sample will have coalesced.

recent common ancestor for humanity or even an ancestor at all. In estimating the time to the most recent common ancestor, genetic data can be of particular use.

Most methods of TMRCA estimation rely on integrating sequence divergence, an estimate of recombination rate and the molecular clock hypothesis, which notes the linear accrual of mutations over time, and was used by Kimura in the seminal paper on neutral theory (27). Although the molecular clock is used in most of these estimates, a recent survey of mitochondrial DNA has noted that divergence from the molecular clock can sometimes be strong (4).

Caution must be used when generalizing predictions made from chromosome-specific segregating sites to the rest of the genome. For instance, very different TMRCAs have been estimated using mitochondrial DNA and DNA from the Y chromosome. However, these sections of the genome make a good starting point for developing the technique used to estimate TMRCAs between chromosomes, because they are passed on from parent to offspring without recombination. This simplifies estimates made using coalescent techniques, as recombination often necessitates computational intensity. Additionally, at least in the mitochondrial DNA, (henceforth mtDNA), there is a higher rate of mutation than the autosomal chromosomes (4). This allowed early researchers without access to modern sequencing capabilities to capture more polymorphism in smaller samples.

Many TMRCA estimates center around building a tree from the sample, estimat-

ing a rate of divergence between sequences and then using this estimate to predict a timescale. This is particularly feasible on the Y chromosome and in the mtDNA, because there is no recombination to break up haplotypes and complicate tree building. One early study to do this was that by Cann et al (8), which sampled 133 human mtDNA sites. They found the most parsimonious tree across the 133 variable sites, and estimated a rate of divergence by looking at archaeologically based patterns of human demography. Given divergence between Australians and New Guineans from other populations, as these settlement times are approximately known, they estimated a mutation rate. They then estimated a TMRCA for all living human mtDNA between 140-290 kya.

Thomson et al (52) used the program GENETREE (http://www.stats.ox.ac.uk/∼griff/software.htm to infer coalescent trees along with their TMRCA and probability of occurring based on the input data. Thomson et al computed the expected mean TMCRA by multiplying the probability of a tree based on the data by its estimated TMRCA and summing over the most likely trees. In doing this, they found that the TMRCA for the Y chromosome was 84 kya, with a 95% confidence interval from between 55 and 149 kya. They performed this analysis separately in each of the three genes in their sample and found similar estimates.

Thomson et al (52) also proposed a separate unbiased estimator for the divergence times. They proposed that for each of n sequences, to determine x_i, the sequence divergence of the ith sequence from the MRCA. Since $\mu * T_{MRCA}$ should be an estimate for any sequence x_i, $n\mu T_{MRCA}$ should equal the sum of all x_is. Therefore, they propose the following estimator:

$$T_{MRCA} = \sum_{i=1}^{n} \frac{x_i}{n\mu}$$

Applying this estimator, they predicted that the TMRCA is 70 kya (SE 30ky) (52).

Tang et al (48) used a frequentist approach to estimate the TMRCA based on the molecular clock hypothesis. In particular, they attempted to create an estimator divorced from inputed knowledge about population parameters. Their study highlight the importance of finding the TMRCA not just between any pair of individuals, but among the entire population. Their proposed estimator builds on previous work done by Templeton (50). The method works by partitioning the sequences based on the sample coalescence that occurs the furthest back in time. The l sequences on one side

of the tree are grouped together into a set L and the remaining r sequences are grouped into R. Note, $l + r = n$. They do not explicitly outline a suggestion on how to perform the group partition, but in their analysis, they determine the best fit tree and then use the final coalescence subdivision to create their groups.

They then compute D, the average divergence between sequences in L and R. In doing this, they ensure the maximal divergence since the MRCA between pairs. If they were not to partition between the two groups, they would be averaging pairwise divergence between sequences that coalesce much more recently.

$$D = \frac{1}{rl} \sum_{i \in L} \sum_{j \in R} d_{ij}$$

Since the average divergence is the number of mutations that would occur on independently mutating sequence of length l,

$$D = 2l\mu T_{MRCA}$$

So,

$$T_{MRCA} = \frac{D}{2l\mu}$$

While this method has the disadvantage of requiring the initial partition to be created, it creates a test that circumvents problems associated with underlying demography. Applying this estimator to the Y chromosome with mutation rate $3 * 10^{-8}$ per site per generation and generation lengths 25 and 30 years, they estimated TMRCAs of 91kya (60-130kya) and 109 kya (72-156 kya), respectively. When using this estimator on mtDNA with mutation rate of $2.43 * 10^{-7}$ per site per generation and generation lengths 25 and 30, they estimated TMRCAs of 238kya (200-281kya) and 190kya (160-225 kya) respectively (48).

Other groups have used more straightforward approaches. Ingman et al (23) fit a tree using a neighbor-joining approach and made their estimate of the TMRCA of mtDNA by looking at two the most diverged sequences. They estimated approximately 171.5 kya $\pm$ 50ky.

Other more computational approaches have become available as data access and computer resources have increased. Tavare et al (49) present methods that rely on informed priors on demographic parameters and mutations rate, with a simulation

and rejection scheme for generating the densities of population size, mutation rate and TMRCA. Pritchard et al (42) apply these methods to show the importance of accounting for demography. They used 8 human Y chromosome genes in 445 males, specifically looking at previously identified highly variable loci and found estimates for TMRCA on the order of 46kya (16-126 ky).

As seen from the variation in the estimates for mtDNA and Y-chromosomal DNA, there are considerable differences between the TMRCAs of different parts of the genome (Table 1.1). That the TMRCA for mtDNA is twice as long as the Y chromosomal TMRCA was noted in a paper by Favrea & Sornettea (14). The authors used agent based modeling to explore scenarios that could generate this enrichment. Conditions that required males to compete for mates and women to breed without competition generated a strong sex-biased TMRCA, although not as strong as the 2:1 ratio observed in real data.

As more genomes have been sequenced, the TMRCA for Y-chromosomal DNA has been pushed continually backward. A study sequencing 2,204 200 kb samples in the male specific Y region (MSY) estimated a TMRCA 141.5 kya ($\pm$ 15.6 ky) (12). Recently, a very low frequency African American lineage originating in sub-Saharan Africa has been sequenced with a drastically diverged Y-chromosome. This brings the estimate for the TMRCA to 338 kya (237-581 ky) (35). This estimate predates predictions for the mtDNA TMRCA, which have been otherwise uniformly higher than for Y chromosomal DNA (Table 1.1). This case serves as a reminder of the importance of deep and varied sampling when attempting to estimate the TMRCA of all living humans.

Although most estimates for the mtDNA and Y chromosomal TMRCA are between 50 and 300kya, autosomal TMRCA estimates are much larger. Table 1.2 briefly summarizes the extent of these estimates.

1.4 Thesis Outline

Despite the importance of derived variation, and in particular, shared derived rare variation, much work has yet to be done to integrate this information into existing population genetic methods. In the following two chapters, I set out to address the background that will make such integration possible. Namely, important questions

	MSY			mtDNA	
Paper	TMRCA (KYA)	Error (KYA)	Paper	TMRCA (KYA)	Error (KYA)
Pritchard (1999)(42)	46	(16-126)	Cann (1987)(8)	215*	(140-290)
Thomson (2000)(52)[1]	84	(55 -149)	Ingman (2000)(23)	171.5	(121 - 221)
Thomson (2000)(52)[2]	70	(11 - 129)	Tang (2002)(48)	238	(200-281)
Tang (2002) (48)	91	(60-130)			
Cruciani (2011),(12)	141.5	(126- 157)			
Hammer (2013)(35),	338	(237-581)			

Table 1.1: Summary table of a number of TMRCA findings for the Y chromosome and mtDNA Table is organized chronologically and separated for mtDNA and Y chromosomal analysis. In Thomson et al (52), two methods are explored, with [1] representing results using GENETREE and [2] using the divergence estimator. (*no value was given, so this is the center of the confidence interval)

about derived variation must be answered before the use of rare derived alleles can be used to extend such methods as the sequentially Markov coalescent.

In the second chapter, as I endeavor to verify that populations are evolving at the same rate, I question whether there are equal numbers of derived variants in populations across the globe. Although the literature has suggested variation in the number of deleterious alleles in different populations, we would still expect to see the same number of total derived alleles in all individuals since a MRCA. In addressing this question, I study departures from expectation of the derived site frequency spectrum in order to understand the types of mutation that might allow different population evolution rates.

To delve deeper into the causes of derived allele variability between populations, I also examine the derived site frequency spectrum in the context of different genomic backgrounds. Given Lohmueller et al's finding of varying extents of non-synonymous and deleterious mutations between populations (31), I ask these same questions specific to coding regions in order to understand departures from genome-wide patterns. In this chapter, I also explore the importance of variation at CpG sites and its potential role in the maintenance of derived variation in the population.

Last, I use the total number of derived alleles to estimate the TMRCA of all living humans, both as a verification that our analyses are sensible, and as a way to utilize

Paper	TMRCA (KYA)
Blum (2011) (6)	1500
Barreiro (2005) (3)	2800
Stefansson (2005) (45)	3000
Hayakawa (2006) (20)	2900

Table 1.2: Summary table of a number of TMRCA findings for autosomal DNA Table is organized chronologically and lists author, year of publication, reference and estimate.

the number of derived alleles to estimate the length of human history.

In my final chapter, I begin to address how these findings can influence methods for genetic analysis. In particular, I lay out the theoretical background behind the sequentially Markov coalescent model, and its pairwise equivalent. I then suggest, although do not implement, an important way in which derived alleles can be used to improve previously existing tests for pairwise coalescence between individuals across the genome.

2

Derived Allele Differences between Populations

2.1 Introduction

We hope to use derived variation to improve statistical methods to estimate genetic parameters such as the TMRCA. Before this is possible, we must ensure that we understand processes that might affect our use of this variation. For instance, if we intend to assume that the molecular clock applies uniformly globally, we must determine that populations have similar numbers of derived alleles. To aid us in understanding these processes, we will also examine the derived site frequency spectrum in order to elucidate patterns that may underlie differences in the total number of derived mutations.

2.2 Theoretical basis for expectations

Through coalescent theory, we can establish certain expectations for the derived allele frequency spectrum under neutrality. This will allow us to determine if exogenous factors are pulling the distribution away from this expectation.

An important first question asks, given a sample of sequences, what is the expected number of derived alleles at frequency i? We will follow the analysis of this question presented by Fu (16).

Consider a standard neutral coalescent with n variants at the branches numbered $1 \,..\, n$. Although the lineages are numbered separately, we note that nothing differs between the sequences, so they will be exchangeable for the purpose of our analysis.

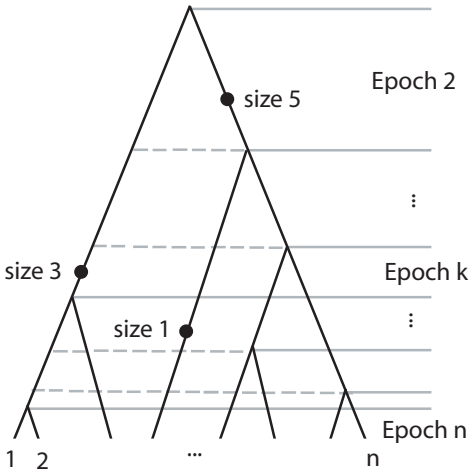


Figure 2.1: Visualization aid for terminology - Exchangeable branches 1 .. n are numbered at the branches of tree. Epochs 2 .. n are noted on the side, indicating the number of distinct lineages present at that time. Three sites are highlighted with sizes 1, 3 and 5, indicating the number of descendants of that branch in the most recent epoch n.

We can split this tree into $n-1$ epochs, such that in epoch k, there are exact k distinct variants. Therefore, just before the root of the tree, we fall in epoch 2. We will say that a variant in epoch k is of size i if it will be at frequency i in epoch n, which represents the current state of the population.

For this coalescent, we consider an infinite sites model in which each mutation occurs at a different locus, disallowing back-mutation. This assumption is later examined in order to understand the possible effects of back-mutation on our results. We also assume a constant mutation rate across the region.

We want to determine, under the conditions of the standard neutral coalescent, how many mutations are at a certain frequency i of n. In order to do this, we consider, for each epoch $k \in 2..n$, and for each lineage $l \in 1..k$ within the epoch, the number of mutations that fall on l within epoch k of size i.

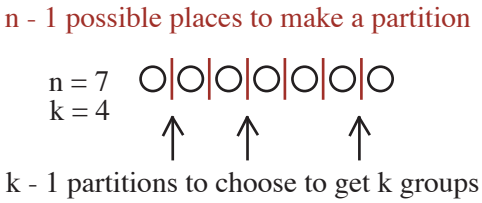
$$\xi_i = \sum_{k=2}^{n} \sum_{l=1}^{k} \epsilon_{k,l}(i)\chi_l \tag{2.1}$$

where χ_l represents the number of mutations that will fall on the length of l and

$$\epsilon_{k,l}(i) = \begin{cases} 1 \text{ if lineage } l \text{ in epoch } k \text{ is of size } i \\ 0 \text{ otherwise} \end{cases}$$

.

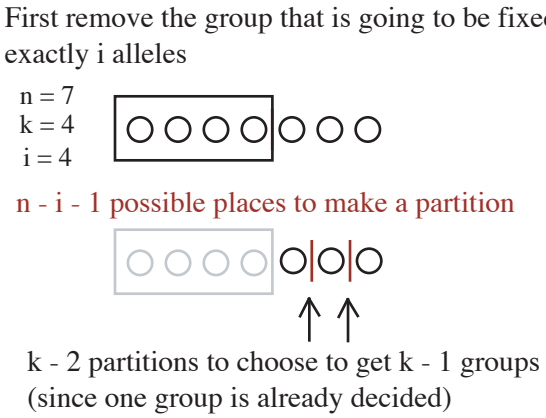
We first consider the proportion of possible lineages in epoch k that have size i. To do this, we will determine the total number of ways in which n lineages can coalesce into k lineages. This is the same as the number of ways one could throw n balls into k boxes, such that every box had at least one ball.



Therefore, the total number of configurations C_{tot} for coalescence is

$$C_{tot} = \binom{n-1}{k-1}. \tag{2.2}$$

To determine how many of these configurations allow for a given lineage to have size i, we observe that the number of configurations that allow for a lineage to have a group of i is:



Therefore, we calculate $C_{i|k}$, the number of configurations in which a derived allele has size i in epoch k, as follows

$$C_{i|k} = \binom{n-i-1}{k-2}. \tag{2.3}$$

In order to get the proportion, we examine the ratio of the configurations that allow i given k (2.3) as compared to the total possible number of configurations (2.2):

$$\begin{aligned} p(i|n,k) &= \binom{n-i-1}{k-2}\binom{n-1}{k-1}^{-1} \\ &= \frac{(n-i-1)!(k-1)!(n-1-k+1)!}{(k-2)!(n-i+1-k)!(n-1)!} \\ &= (k-1)\frac{(n-k)!}{(n-k-(i-1))!}\frac{(n-1-i)!}{(n-1)!} \\ &= (k-1)(i-1)!\binom{n-k}{i-1}\frac{1}{i!}\binom{n-1}{i}^{-1} \\ &= \frac{k-1}{i}\binom{n-k}{i-1}\binom{n-1}{i}^{-1} \end{aligned} \tag{2.4}$$

We now know the probability of observing a i-sized configuration in epoch k. Note that this probability is the expectation of the indicator variable $\epsilon_k(i)$. We must now determine the number of mutations χ_l that we expect to see on a lineage in this epoch. However, this is the simply the expected length of a lineage in epoch k and is given by

$$E(\chi_l) = \frac{\theta}{k(k-1)} \tag{2.5}$$

Here, θ represents the population scaled mutation rate, and is calculated by the product of the per base mutation rate μ, the effective population size N_e and a scaling factor ($\theta = 4N_e\mu$).

To determine $E(\xi_i)$, we can take the expectation of Equation (2.1).

$$E(\xi_i) = E(\sum_{k=2}^{n}\sum_{l=1}^{k}\epsilon_k(i)\chi_l) = \sum_{k=2}^{n}\sum_{l=1}^{k}E(\epsilon_k(i))E(\chi_l) \tag{2.6}$$

We can now apply the results of Equations (2.4) and (2.5) in Equation (2.6) to see that,

$$E(\xi_i) = \sum_{k=2}^{n}\sum_{l=1}^{k}\frac{\theta}{k(k-1)}\frac{k-1}{i}\binom{n-k}{i-1}\binom{n-1}{i}^{-1} \tag{2.7}$$

Since the ls are exchangeable, we can replace the sum over the ls with a k multiplier, and simplify:

$$E(\xi_i) = \sum_{k=2}^{n} \frac{\theta}{i} \binom{n-k}{i-1} \binom{n-1}{i}^{-1}$$

Pulling out terms not in the sum,

$$E(\xi_i) = \frac{\theta}{i} \binom{n-1}{i}^{-1} \sum_{k=2}^{n} \binom{n-k}{i-1}$$

Through repeated application of the recursive definition of the binomial coefficient,

$$\binom{a}{b} = \binom{a-1}{b-1} + \binom{a-1}{b}$$

We see that,

$$\sum_{k=2}^{n} \binom{n-k}{i-1} = \binom{n-1}{i}$$

And therefore,

$$E(\xi_i) = \frac{\theta}{i} \tag{2.8}$$

From this, we can see that because the density of variants at different frequencies is inversely proportional to the frequency, we expect, under neutrality, that if we chose a derived allele at random in the population, it would be equiprobable that it comes from any particular frequency class.

Importantly, these expectations only hold under a regime of constant population size with no selection, neither of which are necessarily good assumptions for human populations.

2.3 Data Background

Given our expectation for how the site frequency spectrum should look under neutrality, we can now examine how our observed data compare. There are a number of biological reasons that our data may not match expectations, but we must first discount possible sequencing related problems that might affect our results.

We underwent an extensive process to ensure that our derived site frequency spectra was based in biology and not sequencing artifacts.

2.3.1 Description

I used the 1000 Genomes variant call files from the Phase 1 release (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20101123/interim_phase1_release/). 1092 individuals are sequenced from across 14 populations with between 14 and 96 individuals per population. The color codes for these populations that will appear throughout the rest of the document are in Figure 2.2. Although more information can be found in the papers introducing the data, a short summary will be provided here.

The 1000 Genomes data used low coverage whole genome sequencing at 4-6x coverage in order to find variants outside the exome and to construct haplotypes. Within the exome, the project used targeted deep re-sequencing in order to accurately call alleles.

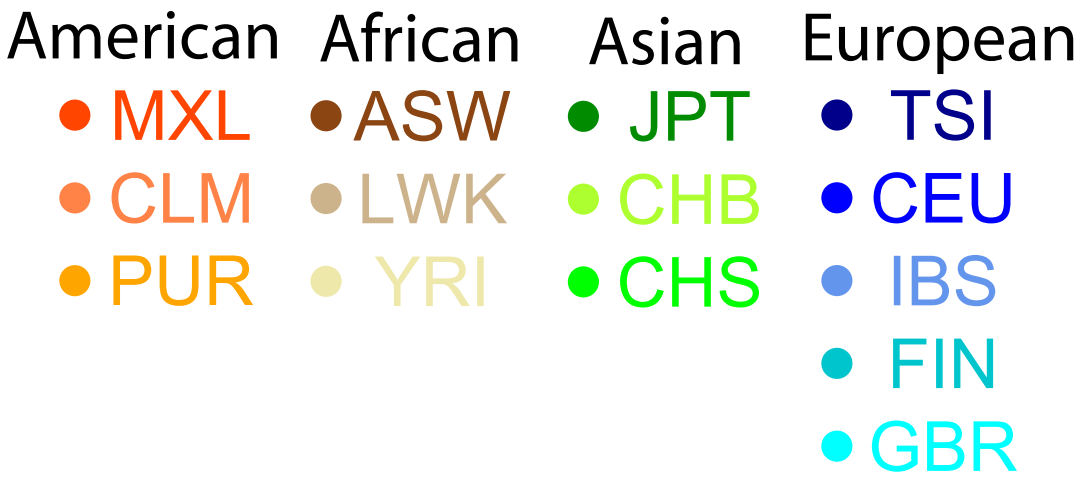


Figure 2.2: Population Codes - These codes describe the colors used in the rest of the analysis of the 1000 Genomes data for the remainder of the analysis and are divided by population. Population descriptions for each code are available in Table 2.1.

2.3.2 Masking

I applied a strict mask to the 1000 Genomes VCF files separately for each chromosome. The masking file used (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/20120824_combined_mask.bed) was created by the 1000 Genomes Project Consortium by rejecting all reads without a coverage between between 0.5x and 2x of the mean read across all populations (2,566 - 7,698) inclusive, an average mapping quality below 56 and more than 0.1% of reads with a

Code	Population
MXL	Mexican Ancestry from Los Angeles USA
CLM	Colombians from Medellin, Colombia
PUR	Puerto Ricans from Puerto Rico
ASW	Americans of African Ancestry in SW USA
LWK	Luhya in Webuye, Kenya
YRI	Yoruba in Ibadan, Nigeria
JPT	Japanese in Tokyo, Japan
CHB	Han Chinese in Beijing, China
CHS	Southern Han Chinese
TSI	Toscani in Italia
CEU	Utah Residents with Northern and Western European Ancestry
IBS	Iberian population in Spain
FIN	Finnish in Finland
GBR	British in England and Scotland

Table 2.1: 1000 Genomes Population Codes - Presented here are the 1000 Genomes population codes and their population descriptions. They are grouped by continent of origin and can be used with the color codes in Figure 2.2 to determine populations in graphs throughout.

mapping quality of 0. After applying the mask, I was left with 28,254,834 sites, or 73.9% of the starting number. More information about the mask and its preparation can be found in section 10.4 of the supplementary information of 1000 Genomes Project paper (1). See Table 2.2.

2.3.3 SNP selection

Ancestral misidentification could cause a derived allele called at frequency x to actually be at frequency $1 - x$. If this is widespread, it could have potentially large effects on the derived allele frequency spectrum. Therefore, it is important to properly call ancestral alleles to ensure that the shape of the distribution is based on biology, and not misidentification.

To do this, I went through a rigorous validation scheme. The 1000 Genomes data lists a prediction for ancestral allele as extracted from an external database. Alignments

Chrom.	Unmasked	Strict Mask	Included
1	3007224	2216046	2066683
2	3307620	2537641	2390733
3	2763482	2108183	1989364
4	2736793	2057276	1932976
5	2530245	1920281	1805476
6	2424453	1805926	1680281
7	2215259	1584017	1472470
8	2183867	1679017	1570661
9	1652416	1229420	1156786
10	1882691	1401474	1309024
11	1894936	1385900	1281846
12	1828034	1325389	1229293
13	1373028	1062301	1000583
14	1258282	930035	862874
15	1130582	825732	766239
16	1210647	860052	799715
17	1046761	691900	632826
18	1088848	849800	803022
19	816143	425815	338833
20	855194	644399	604145
21	518993	386243	360807
22	494356	327987	294792
Total	38219854	28254834	26349429

Table 2.2: Masking of 1000 Genomes SNPs - Here, we report the total number of SNPs by chromosome in the 1000 Genomes Project autosomal data, and the number after applying a strict mask to the data. The third column represents the final number of SNPs that were included in our analysis.

are estimated for the ancestral state of the human-chimp ancestor and human-chimp-macaque. More details about this can be found on the 1000 Genomes website (54). If the human-chimp-macaque all agree, the ancestral allele is reported with high confidence. If human agrees with chimp or macaque but not both, the ancestral allele is reported, but with low confidence. If the human sequence does not agree with chimp

Chr.	All SNPs	Included	Excluded	Because of ...			
				1000G	dbSNP	Both	Disagree
1	2123316	2084479	38837	19980	11588	6540	729
2	2444093	2411815	32278	13569	10680	7321	708
3	2027547	2006970	20577	7735	7090	5285	467
4	1979772	1949993	29779	11670	11576	5834	699
5	1849414	1821282	28132	11070	8414	7950	698
6	1733578	1694654	38924	19624	8538	9832	930
7	1523574	1485319	38255	21192	8194	8092	777
8	1623184	1585629	37555	15668	7238	13819	830
9	1183983	1167084	16899	6583	5951	3774	591
10	1349334	1320319	29015	9494	12590	6442	489
11	1329085	1292975	36110	15909	9220	10462	519
12	1270169	1240043	30126	12865	6932	9673	656
13	1022619	1009179	13440	5573	5207	2377	283
14	892618	870473	22145	13402	4221	4212	310
15	792612	772983	19629	8276	3555	7496	302
16	826893	807434	19459	9245	5811	3939	464
17	655822	638241	17581	8598	5143	3014	826
18	819158	810186	8972	2775	3579	2243	375
19	395703	341856	53847	31107	10185	11680	875
20	618757	609298	9459	3905	3247	2032	275
21	370505	364085	6420	3175	1793	1234	218
22	316164	297343	18821	7668	4966	5997	190
Total	27147900	26581640	566260	259083	155718	139248	12211
(%)		(98%)	(.02%)				

Table 2.3: High quality ancestral allele filtering - Here we report the total number of SNPs included per chromosome after our ancestral allele (AA) filtering. The final four columns summarize the reasons for excluding a SNP because of insufficient ancestral allele information. They give the number of SNPs excluded because of a poor 1000G AA call, a poor dbSNP call, a poor call by both methods, or by disagreement between the two methods.

or macaque, the call is not made at all. In our first pass of analysis, I included all SNPs with either a low or high confidence ancestral allele call and an RSID for the purpose

of matching the SNPs to an external dataset.

I then looked at an independent dataset of SNP calls from dbSNP build 137 examining all RSIDs listed as having either low or high confidence ancestral alleles in the 1000 Genomes data. Those present in dbSNP were extracted. The ancestral alleles were extracted, and then polarized using a BED file from dbSNP which contained strand information (55). The polarized information at each RSID was then saved.

The two lists were then compared. Only SNPs at which we had a call from dbSNP and at least a low quality call from the 1000 Genomes data were included. Only SNPs where they agreed (even with a low confidence 1000 Genomes call) were included. All other SNPs were left out of the list of high quality SNPs.

This was the final list of high quality SNPs with complete ancestral allele information, although not all of these SNPs were used in our final analysis. A summary is provided in Table 2.3. The 1000 Genomes VCF lists a reference and an alternative allele. Sites at which the ancestral allele, listed as a consensus between the 1000 Genomes call and the dbSNP call, did not match either the reference or the ancestral allele were excluded. This represented a relatively small number of cases. Also excluded were SNPs where multiple nucleotides were called as either the ancestral or reference allele.

Using only this set of high confidence SNPs allows the best possible information with which to analyze the derived allele frequency spectrum. However, areas of the genome that are harder to sequence or that do not align well are more likely to be left out. This could potentially influence further analysis, which I note throughout.

2.3.4 Power Correction

A potential cause for concern is differential ability to detect alleles at different frequencies between populations. It may, for example, be possible that we do much better in detecting low frequency alleles in African populations than in others. In order to control for this, curves are fit using a separate dataset as an independent test of our ability to detect different alleles in different populations.

The OMNI (53) SNP genotypes was used as an external validation data set. The OMNI data has been used in many places throughout the 1000 Genomes project as a highly validated set of SNPs, acting as a verification for imputation analysis and testing exome calls (1). Following other 1000 Genomes analysis, the OMNI calls are taken to be the gold standard in terms of allele frequencies. Although we must be wary

of considering any data as the truth, it is important to try to adjust for the biases in the data, and the OMNI data allows an opportunity to do this.

Looking just at the same 1092 individuals sequenced as part of the 1000 Genomes project, our ability to rediscover OMNI derived alleles in the 1000 Genomes data $R_x(i)$ was computed. At the exact s sites at which there are i copies of derived alleles in population x of the OMNI data, I find v copies of derived alleles in the 1000 Genomes data:

$$R_x(i) = \frac{v}{s * i}$$

Plotting $R_x(i)$ reveals variable abilities to detect alleles especially at very low frequencies (Figure 2.3). In particular, the 1000 Genomes Project discovered an excess of rare alleles in Asian populations (JPT, CHS, CHB) as compared to the OMNI dataset, and a deficit of rare alleles in African populations (ASW, LWK, YRI). This is perhaps due to differences in how 1000 Genomes samples were processed. Also possible is variable coverage in the OMNI dataset by population.

At higher frequencies, the power to detect alleles was very high and uniform across populations. There was a slight increase in our ability to detect derived alleles in African populations at very high frequencies. This difference did not appear to majorly affect further analysis and results look very similar whether or not we apply the correction. The use of $R_x(i)$ as a power correction to control for differential abilities to detect alleles at different frequencies in further analyses becomes most important at low alleles frequencies.

It is worth noting that while the OMNI dataset was created independently, it did integrate SNP calls from the 1000 Genomes pilot study. Populations present in the pilot (CEU, YRI, CHB, JPT) may have more calls in the OMNI data. It is therefore unsurprising that there is an increase in the $R_x(i)$ value in these populations relative to their continental cohorts.

In a first attempt at creating this correction, I considered using the Axiom data as an external set, but I was unable to continue with the analysis because the Axiom data did not note RSIDs, which were necessary to determine the ancestral consensus alleles. The OMNI data had the desirable properties of containing sequencing information for all 1092 individuals and having RSID information with which to match up SNPs.

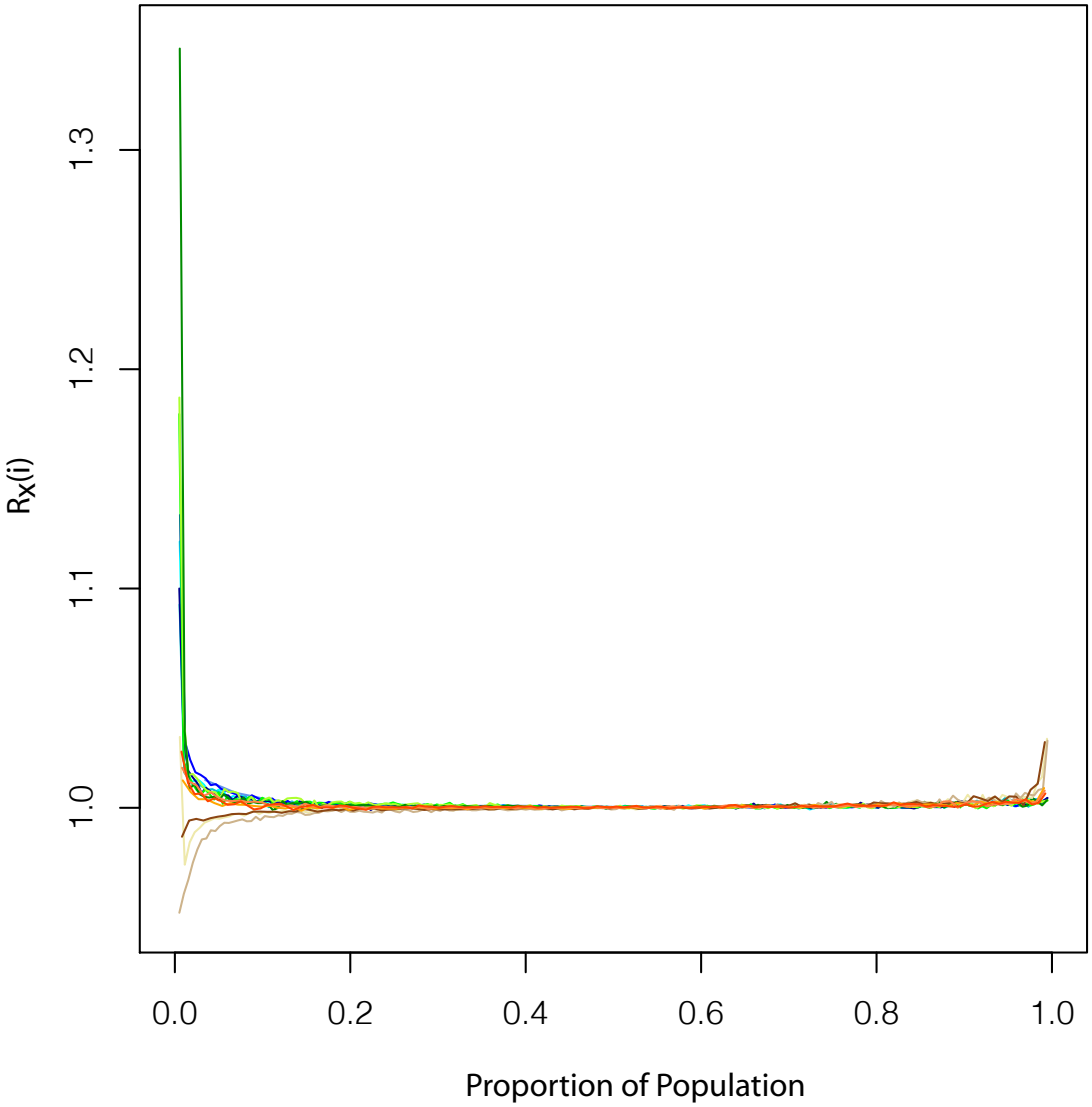


Figure 2.3: Power Analysis Curves - Shown is $R_x(i)$ describing our ability to detect alleles at different within-population frequencies across the populations of our sample as compared to the OMNI data.

2.4 Differences Between Populations in the Derived Allele Frequencies

2.4.0.1 Distortion in the Derived Allele Frequency Spectrum

The derived frequency spectrum was examined in relation to the findings in Section 2.2. The number of derived alleles at different frequencies should be a flat distribution (i.e., within a population, there should be n copies of an allele at frequency 1, $n/2$ copies at frequency 2, ... n/i copies at frequency i). The number of copies of alleles at frequency i was multiplied by i to get the total number of alleles of that frequency by population. In order to properly correct for power, all allele frequencies i in population x were divided by $R_x(i)$.

All populations show a clear departure from neutrality (Figure 2.4), with an excess of very rare and very common alleles. In addition, all the non-African populations show a deficit of low frequency alleles ($1\% < DAF < 40\%$) The three African populations (LWK, YRI, ASW) have an excess of low frequency alleles.

There are many possible reasons for these patterns to appear. The excess of high frequency alleles could be a result of positive selection pushing beneficial traits towards fixation in the population. The excess of very rare alleles is a good indicator of explosive population growth, as documented in Section 1.1.2. The deficiency of low frequency alleles could be a mark of a population bottleneck in non-African populations.

Ancestral bottlenecks could explain some of the divergent patterns between African, European and Asian populations. Evidence supports an out-of-Africa bottleneck in non-African populations (18), with a second smaller bottleneck occurring in Asian populations. These bottlenecks reduce the variation in the population and shrink the effective population size. In the coalescent framework, two branches coalesce with a rate inversely proportional to their effective population size. A historically larger African population will therefore have longer branch lengths with more mutations accruing than smaller Asian or European populations.

Differing mutation rates between populations could also explain the differences in the allele frequency spectrum. If African populations are accruing mutations more quickly than other populations, this could cause an inflation in derived alleles. These derived alleles may stay at low frequencies if they do not confer a selective advantage. There is some evidence that mutation rate could be related to temperature (30).

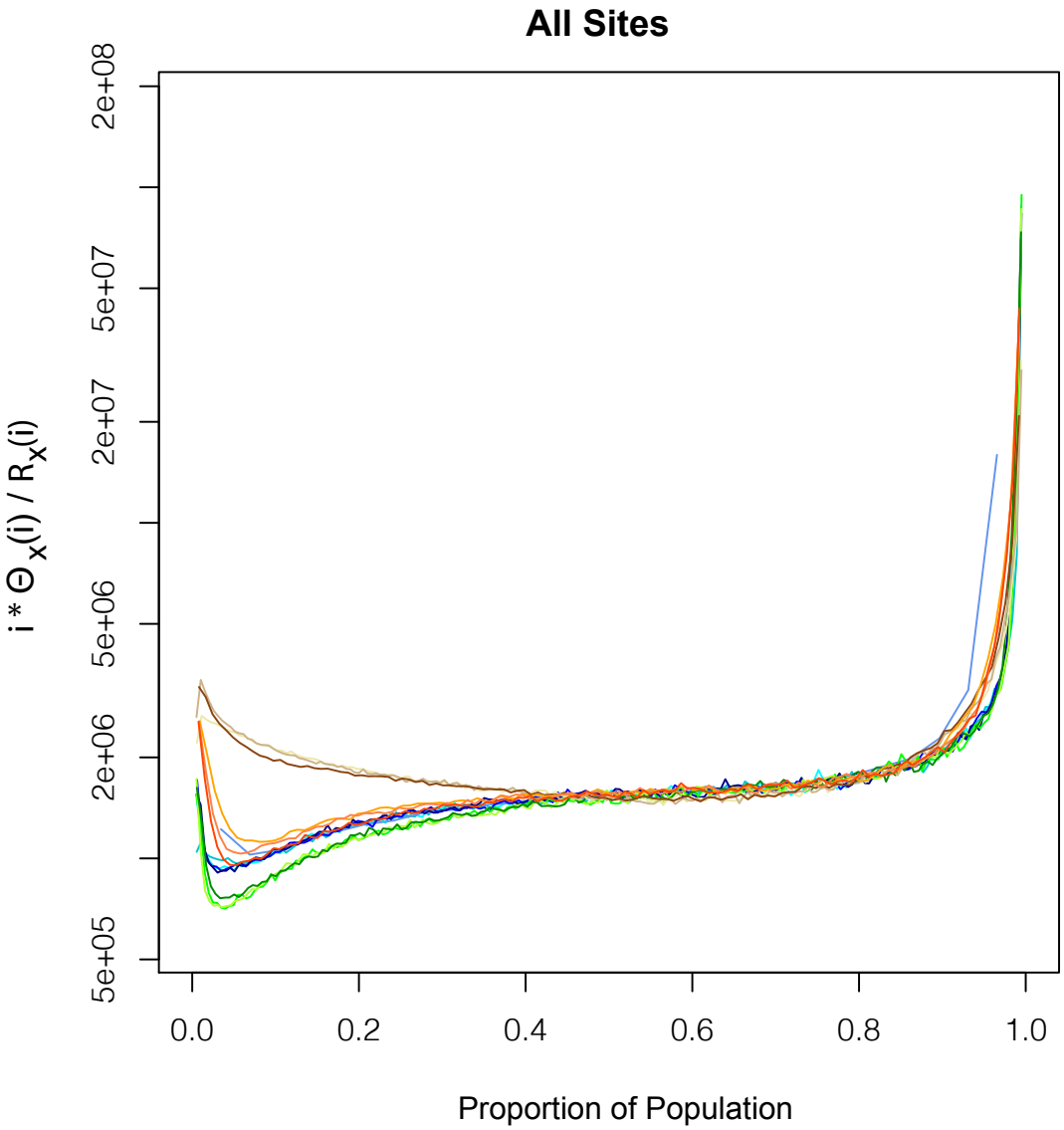


Figure 2.4: Derived Allele Density Curves by Population - Separated by population, the average number of derived alleles at each population proportion was plotted. This is computed by multiplying the number of sites present in i copies in population x $\theta_x(i)$ by the frequency i and dividing by $R_x(i)$, out ability to detect alleles of frequency i in population x.

The derived site frequency spectrum differs from neutrality, with an excess of high frequency alleles, and an enrichment of low frequency derived allele in African populations relative to others examined.

2.4.0.2 High frequency derived alleles

There are several reasons that we might see an abundance of high frequency alleles.

Positive selection acting on the genome can bring beneficial new mutations to near fixation. Standing variation in a bottlenecked population can allow neutral or near neutral alleles to rise quickly in frequency via drift, despite conferring a selective disadvantage.

Ancestral misidentification could play an important and non-biological role in causing the excess of very high frequency derived alleles. Despite our best efforts in using only alleles for which we have high confidence of the ancestral state, we observe an enrichment in low frequency rare alleles. Ancestral misidentification of low frequency alleles would increase the count of high frequency alleles. Given that these alleles are at such a high frequency in the population, their relative contribution to the density plots is larger than a misidentification resulting in a low frequency allele.

We must also consider the potential for back mutations. There are multiple ways that we might observe a derived allele at high frequency. First, the allele may have entered the population as a mutation and risen to high frequency because of selection or drift. This is illustrated in **A** of Figure 2.5. Alternatively, a high frequency derived allele might be a back mutation. If a mutation arises on the lineage after divergence from the outgroup, but before the most recent common ancestor, the allele will be fixed in the population. If the allele then mutates back to the ancestral allele on a branch ultimately ending up at frequency k in the current n sequences, our analysis will classify that as a derived allele at frequency $n - k$ (see **B** of Figure 2.5).

What is the probability that this will occur at an allele with apparent derived allele frequency k? The expected number of k-frequency variants A_k arising out of a mutation since the TMRCA is given by

$$E(A_k) = \frac{\theta}{k} \tag{2.9}$$

The expectation for the number of k-frequency variants arising out of back mutation is given by the probability of a mutation times the probability of a back mutation to

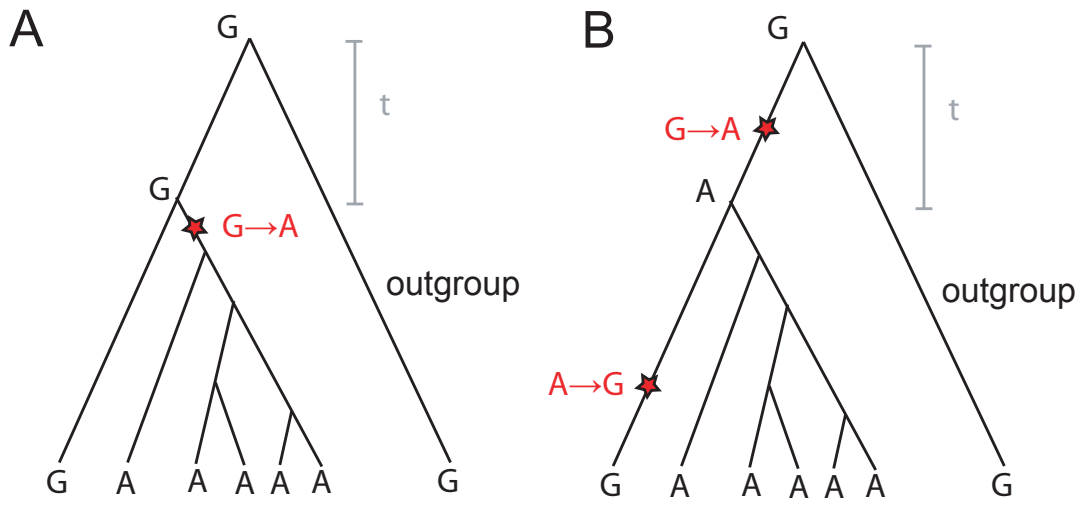


Figure 2.5: Scenarios leading to a high frequency derived mutation call - A) shows how a derived mutation can reach high frequency by entering into a population and then spreading widely. B) shows how a back mutation to an ancestral state can lead to a high frequency derived allele call. A derived allele fixes in a population before the MRCA of a set of sequences. A back mutation to the reference allele later enters the population. The number of non-ancestral alleles is high, leading to a high frequency derived allele call, even though the derived allele had previously fixed in the population.

the ancestral state. First, the probability of having a mutation on the tree section connecting the MRCA of the population and the root of the tree is μt, with mutation rate μ and t representing the time since divergence from the outgroup. The probability that we observe a back mutation to put the allele frequency at k is $\theta/(n-k)$. However, only in one third of these cases will the back mutation return the allele to its ancestral state. Therefore, the expected count B_k that we would see a back mutation at frequency k is

$$E(B_k) = \frac{\mu t}{3}\frac{\theta}{n-k} \tag{2.10}$$

Therefore, the probability that an allele called as derived at frequency k is actually derived is given by

$$p(k \text{ derived}|k \text{ called derived}) = \frac{\frac{\theta}{k}}{\frac{\theta}{k} + \frac{\mu t}{3}\frac{\theta}{n-k}} \tag{2.11}$$

$$p(k \text{ derived}|k \text{ called derived}) = \frac{3}{3 + \mu t\frac{k}{n-k}} \tag{2.12}$$

μt is approximately the divergence between the two sequences divided by 2 (since only mutations on one branch length affect the human sequences). Let us consider this result in the context of the 1000 Genomes data. The outgroup to humans has about 1% divergence, and $n = 2184$. Therefore, the probability that an allele called as derived is actually derived is given by

$$p(k \text{ derived}|k \text{ called derived}) = \frac{3}{3 + 0.005\frac{k}{n-k}} \tag{2.13}$$

Plotting Equation (2.13) for k in (1,2184), we see

We see that at very high derived allele frequencies, there is a substantial probability that the derived allele of frequency k is actually a back mutation of frequency $n-k$ (Figure 2.6). The considerable probability of a back mutation at high frequencies is driven by the length of the divergence between the outgroups and the MRCA of humans, relative to the human TMRCA. Resultantly, many more mutations will fix since human-chimp divergence and back mutate than will arise and spread to high frequency on the much shorter human lineage. While this explains the pattern we see in Figure 2.4, it does not help us distinguish high frequency derived alleles from back mutations.

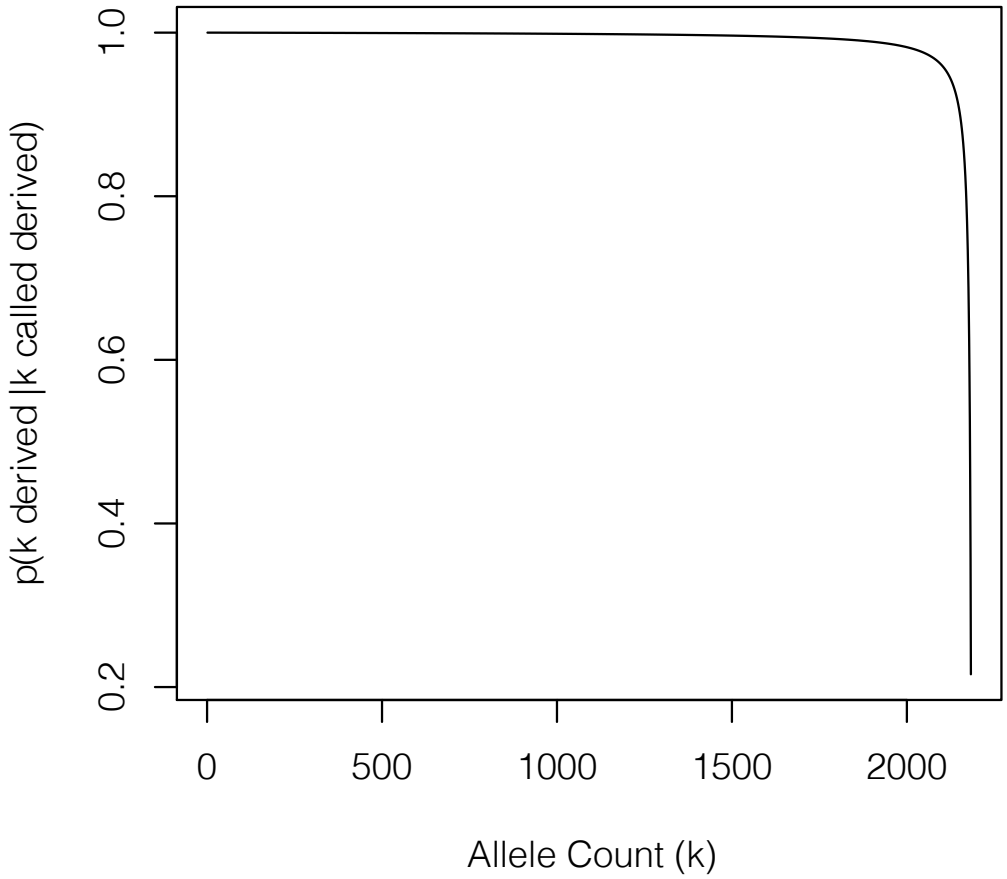


Figure 2.6: Probability of a k frequency allele being derived given that it was called as a k frequency derived allele - Equation (2.13) is plotted for k in 1..2183, as in the 1000 Genomes data.

2.5 Differing numbers of derived alleles in different populations

If we consider the most recent common ancestor of all human individuals, under strict neutrality, every modern human will have the same branch length back to this ancestor. Resultantly, regardless of demography or other changes discussed above that can explain the site frequency spectrum, each modern human should have approximately the same number of derived alleles.

However, given the observed distortion in the site frequency spectrum, it is not apparent that this expectation will hold. The total number of derived alleles present in the genome was computed for each individual. Again, the value was corrected by dividing by $R_x(i)$, although the correction did not significantly change results. Appreciable differences between African and non-African populations were observed.

There are significantly more derived alleles in the African individuals, caused by the differences at low frequency alleles seen in Figure 2.4. As compared to all other populations, the African genomes have many more low frequency variants, which are responsible for driving the pattern. If all variants below frequency 0.4 are excluded, the population averages look far more similar, and Africans drop below Europeans in the number of derived variants. This demonstrates the amount of diversity harbored in a population by low frequency derived alleles.

Corresponding to the dip in low frequency derived alleles in the density curve for Asian populations, there are fewer derived alleles in Asian individuals. These differences are significant, with Asian populations possessing 0.2% fewer SNPs per genome as compared to European and American populations.

The patterns observed here, as explained above, cannot be generated from neutral evolution, as this would not change the branch length from a most recent common ancestor. Several non-demographic factors can affect how mutations fall on a coalescent tree and can potentially explain our observed patterns.

Population-specific selection can extend or shrink branch lengths in some parts of a tree but not others. This would allow relatively more or fewer mutations falling uniformly on a tree to affect a specific population or group of populations. Different generation times between populations can have a similar effect.

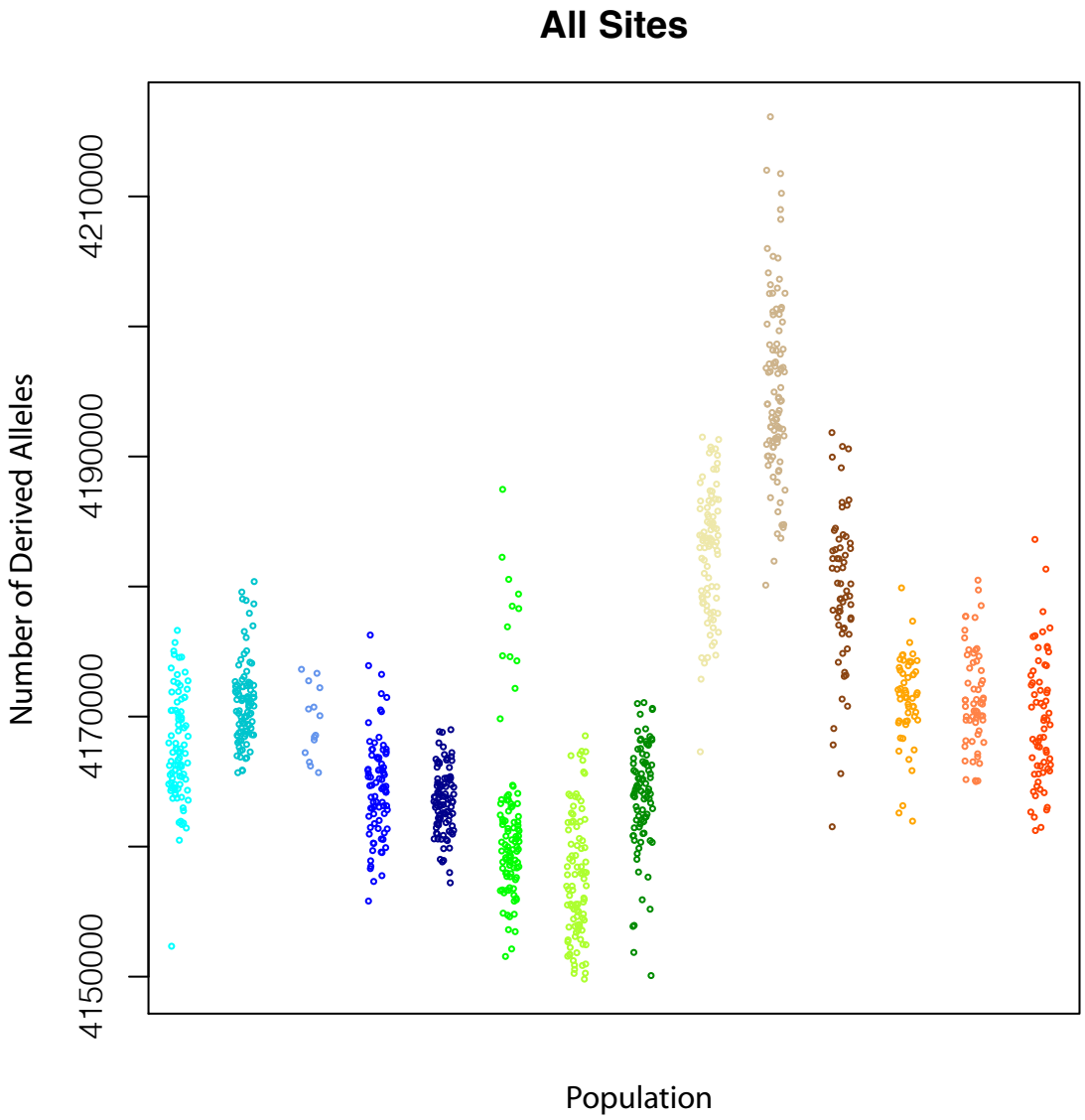


Figure 2.7: Total number of derived alleles per genome - Separated by population, the total number of derived alleles is plotted by individual. This number is power adjusted by $R_x(i)$ as described in 2.3.4.

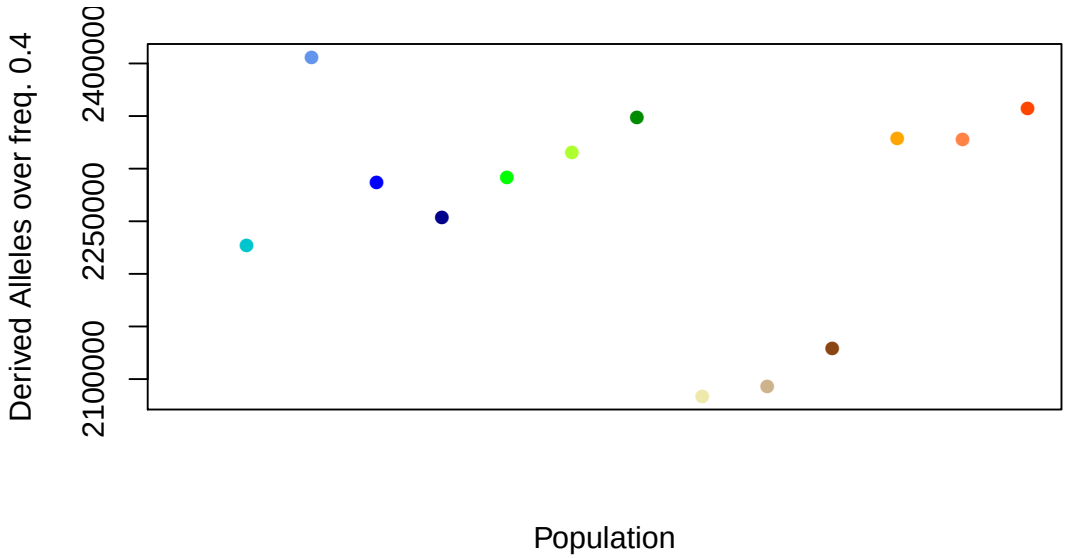


Figure 2.8: Total number of derived alleles per genome of frequency > 0.4 - The average number of derived mutations is plotted by population as computed from summing the allele frequencies greater than 0.4 in 2.4.

Alternatively, an increased mutation rate in some populations might mean that more mutations and hence more derived alleles can occur on branch lengths of the same size. There is some evidence that mutation rates vary based on temperature (30).

Differences in the number of derived alleles are observed between populations, with an excess of derived alleles in African populations and a deficit in Asian populations.

2.5.0.3 Time to Most Recent Common Ancestor

Having an estimate of the number of derived variants allows us to predict a time to most recent common ancestor for the autosomal DNA of human populations. Given a mutation rate of $1.5 * 10^{-8}$ per bp per generation and a genome of 1,949,370,478 autosomal bases in our strict mask, a genome is expected to accrue 29.24 mutations a generation. By this calculation, it would take 142,624 generations to accrue 4,170,450 derived variants, the mean number across all populations. At an estimated 25 years per generation, this suggests a common human autosomal ancestor around 3,565,622 years ago, which is in line with other predictions for the human autosomal TMRCA (Table 1.2).

Pop	All Derived	(SE)	CpG Derived	(SE)	Non-CpG Derived	(SE)
GBR	4167563	433.6	786356	111.4	3381207	342.3
FIN	4171119	327.8	787344	98.9	3383775	249.4
IBS	4169574	726.0	786995	198.9	3382578	619.8
CEU	4164547	408.5	785775	100.8	3378772	333.3
TSI	4163534	248.1	785656	70.0	3377878	209.8
CHS	4161616	666.4	785691	168.0	3375925	508.1
CHB	4157559	480.9	784887	118.7	3372672	375.5
JPT	4163845	438.0	786165	109.3	3377680	347.0
YRI	4182738	515.6	795293	139.1	3387444	399.3
LWK	4195285	705.5	797790	185.5	3397495	542.6
ASW	4179681	767.1	793659	239.3	3386022	560.8
PUR	4171386	453.4	788137	142.1	3383250	347.5
CLM	4171422	486.0	788050	139.8	3383372	382.0
MXL	4169619	594.8	787912	158.3	3381707	464.2

Table 2.4: Number of Derived Alleles by Population - Average number of derived alleles per individual by population among all sites and separated for CpG and non-CpG sites.

2.5.1 Derived Mutations Differences Driven by CpG Site Imbalance

One factor that could affect derived mutation accrual is whether or not a site is a member of a CpG dinucleotide.

CpG sites are those occurring as part of a C-G or a G-C dinucleotide combination. CpG sites are important for genomic regulation, as methylation at the cytosine nucleotide can act to downregulate nearby genes. Because of their functional importance, CpG sites are not randomly distributed throughout chromosomes, and tend to cluster in CpG islands with above average CG content (5). Despite their clustering, CpG sites are seen at only a fourth of their predicted presence in the genome indicating hypermutability in these regions (43).

Importantly, CpG sites have increased levels of mutability when compared to other sites. Specifically, there is evidence that the cytosine component of the dinucleotide mutates to a thymine more than any other nucleotide transition (24). Because of this hypermutability, alignment in CpG sites can be more difficult than in other parts of

the genome.

Ancestral CpG SNPs were called as those at which the ancestral allele call was either a C or a G, and the human reference genome for the 1000 Genomes data at flanking sites either created a CG or GC with the reference allele (as shown in Figure 2.9). About 6,752,818 (23.9%) of the SNPs were identified as ancestrally CpG.

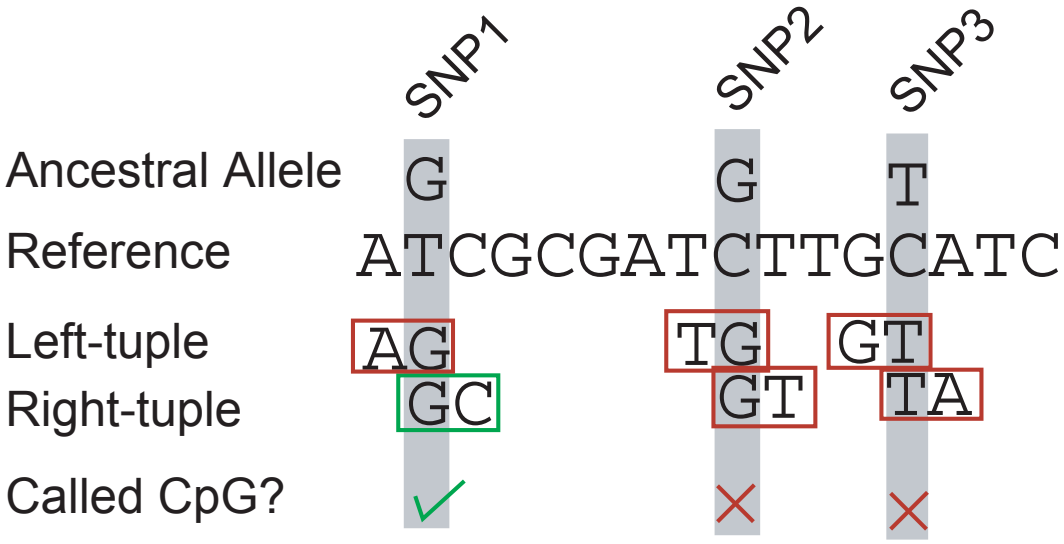


Figure 2.9: Process for calling Ancestral CG Sites - Ancestral allele information is recorded at polymorphic sites. Right and left flanking alleles are determined using the reference genome to form ancestral allele-neighbor tuples. If either the right or left tuple forms a C-G or G-C, the SNP is called as an ancestral CpG.

When dividing the genome this way, a significant enrichment of derived alleles was found within African populations at sites that were historically CpG.

Because the differences between the CpG and non-CpG sites are not entirely apparent in this graph because of the log scale, side by side comparisons of curve densities are displayed for six populations 2.11. The rest of the populations can be found in Figure 2.12. The difference between African and non-African populations is particularly clear when the graphs are plotted separately for each population. The important thing to note is that at the upper end of the population frequencies, non-African populations have similar densities at both CpG and non-CpG sites. However, in African populations, there is a much greater density of non-CpG sites, indicating an enrichment of

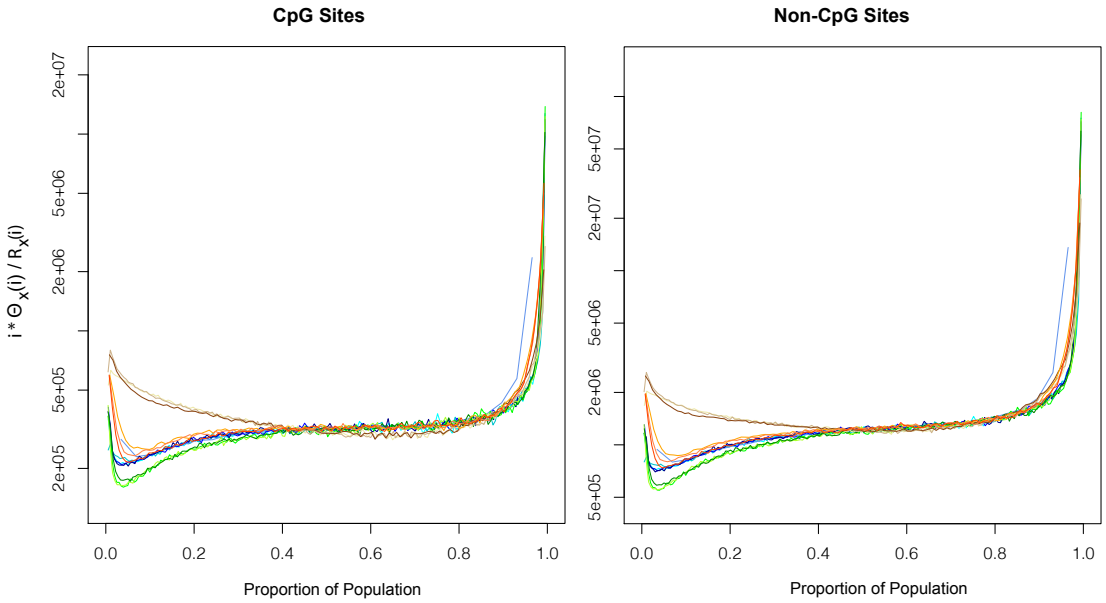


Figure 2.10: Comparison of derived site frequency spectra at CpG and non-CpG sites - In parts A and B, the number of derived variants present at different proportions in the population is plotted at CpG sites and non-CpG sites respectively. This number is power adjusted as described in 2.3.4, to account for differential ability to pick up low and high frequency mutations in different populations and plotted on a log scale. A legend for the different populations can be found in Figure 2.2.

CpG sites, at lower frequencies.

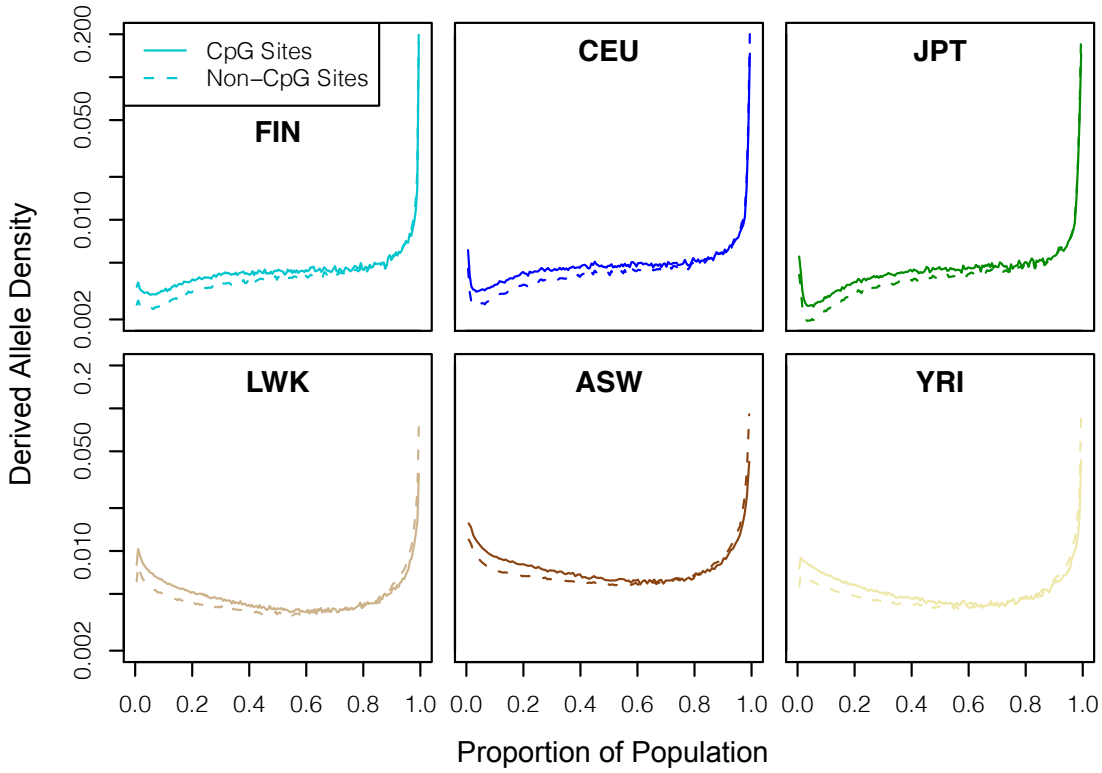


Figure 2.11: Derived site frequency spectra at CpG and non-CpG sites in six populations - Plotted are the density curves of derived variants at varying frequencies in the population at CpG sites (solid lines) and non-CpG sites (dashed) respectively in six populations. The top row represents two European populations and one Asian population. The bottom row shows the density curves for the three African populations. This number is power adjusted as described in section 2.3.4, to account for our differential ability to pick up low and high frequency mutations in different populations and plotted on a log scale.

When tabulating the number of derived allele differences within populations, this difference becomes even more apparent (Figure 2.13).

The average number of derived mutations in African populations is 1.2% higher than in non-African populations in CpG regions, as compared to just 0.35% higher in non-CpG regions, and 0.51% higher genome-wide. The number of derived alleles in Africans relative to non-Africans is significantly higher ($p < 2.2*10\text{-}16$).

Unlike some of the observations about diversity, there is no obvious explanation for this pattern based in demography. It is possible that this pattern is due to the role of

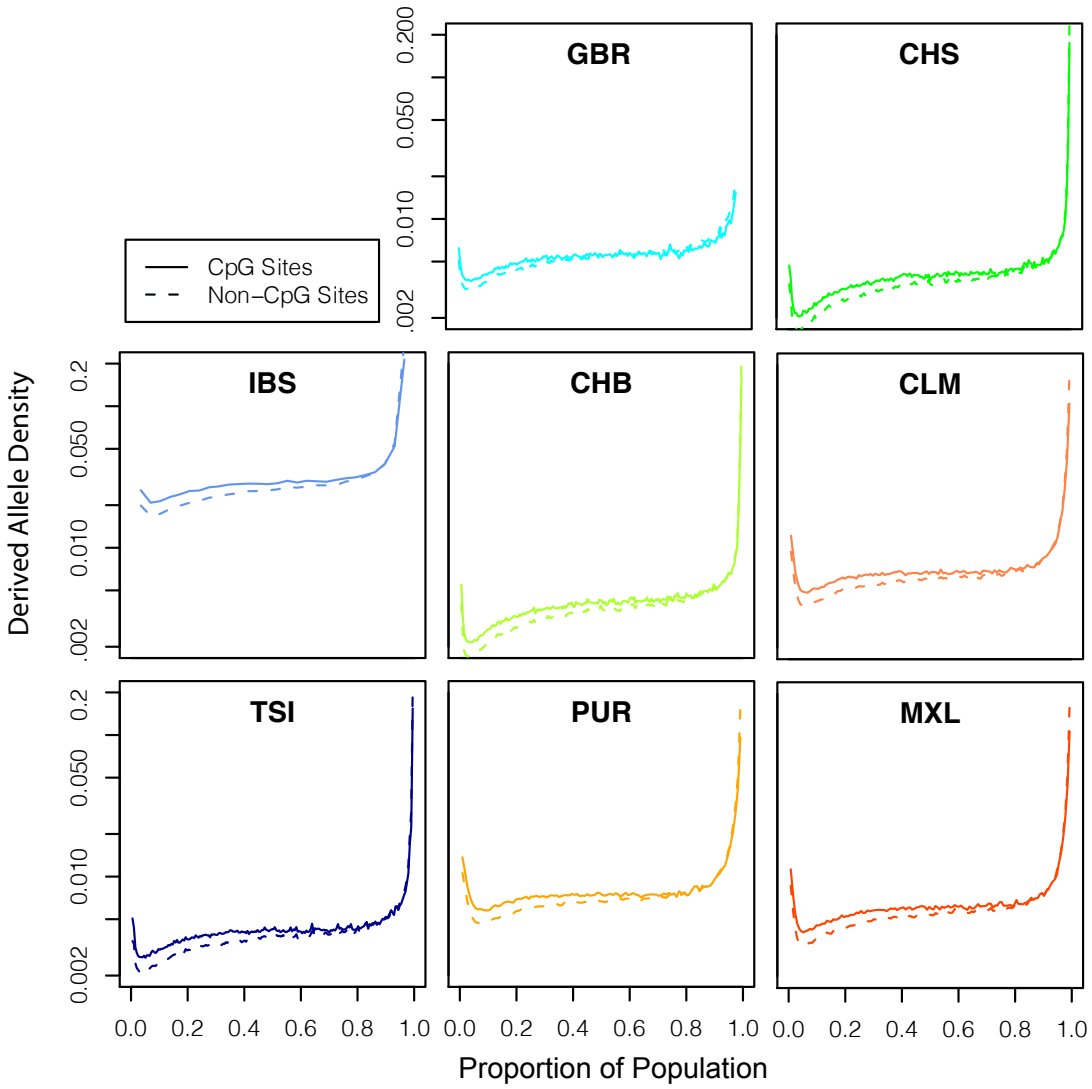


Figure 2.12: Derived site frequency spectra at CpG and non-CpG sites - Plotted are the density curves of derived variants at varying frequencies in the population at CpG sites (solid lines) and non-CpG sites (dashed) respectively separately by population. This number is power adjusted as described in section 2.3.4, to account for our differential ability to pick up low and high frequency mutations in different populations and plotted on a log scale. These populations show approximately the same patterns as the other non-African populations.

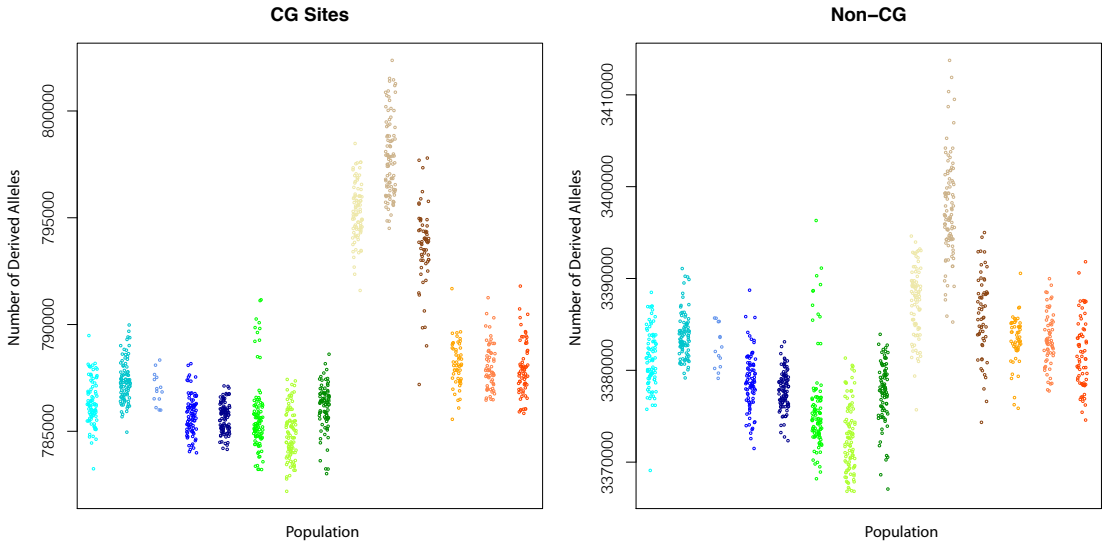


Figure 2.13: Comparison of derived alleles per individual at CpG and non-CpG sites - Separated by population, the total number of derived alleles per individual is plotted. This number is power adjusted as described in section 2.3.4.

methylation in the genome. It has been recently shown that methylation patterns vary widely across the world (36) and that methylation affects mutation rate (38). This suggests that there could be population-specific CpG site mutation rates that vary globally. If this is true, it suggests a possible explanation for the differences in the number of derived alleles between different populations.

If a mutation happens at a CpG site, in most cases it will render the site not a CpG site. Under assumptions of equal transition rates and nucleotide frequencies and the specifications for CpG sites (an ancestral C or G nucleotide adjacent to the corresponding C or G flanking nucleotide), the probability that a CpG site will remain CpG after a mutation is the probability that it mutates to the opposite CpG nucleotide (1/3) and that it is adjacent to a nucleotide of the ancestral state (1/4). Therefore, in 11/12 cases, the derived allele at a CpG site will make the site non-CpG. Since all samples share a common ancestor with a certain fixed number of CpG sites, if Africans really do show a greater level of derived mutations at CpG sites, we might expect that there are fewer CpG sites in total in African individuals.

This hypothesis was tested in pairs of individuals coming from African and European populations. Diploid genomes from 10 African individuals and 10 European individuals

were examined. For each pair of genomes, all sites at which either individual had a C or a G allele were located. The union of those two sets was taken to find a set of sites at which either individual potentially could have a CpG site. Among those sites, the human reference FASTA was used to determine whether the site was CpG or not for each genome tested. A site was called as CpG (as in Figure 2.9) if either the right or left flanking allele and allele at the locus in question resulted in a CG or GC tuple. At these sites, I computed the probability that if we sampled one allele from each individual, the alleles would be both CpG sites, both non-CpG sites, or CpG/non-CpG. This information was used to fill out the associated Table 2.5. The scoring matrix for any pair of individuals appears in Table 2.6. Indels and sites at which we did not have both genotypes in both individuals were excluded.

		Indiv 1	
		CpG	Non-CpG
	CpG	A	B
Indiv 2	Non-CpG	C	D

Table 2.5: Pairwise CpG Site Analysis Matrix Given a site that is C or G in either individual 1 or 2, the numbers A, B, C, and D respectively represent the number of times at which this SNP represents a CpG site in neither, one or both of the individuals.

		Indiv 1		
		CpG/CpG	CpG/Non-CpG	Non-CpG/Non-CpG
	CpG/CpG	A = 1	A, B = .5	B = 1
Indiv 2	CpG/Non-CpG	A,C = .5	A, B, C, D = .25	B,D = .5
	Non-CpG/Non-CpG	C = 1	C, D = .5	D = 1

Table 2.6: Pairwise Genome CpG scoring matrix - Scoring matrix that determines the contribution to Table 2.5 for all examined SNPs. CpG/CpG and Non-CpG/Non-CpG refer to sites at which both chromosomes were called CpG or Non-CpG sites respectively in a given individual. CpG/Non-CpG classification refers to the case in which one allele is called CpG and the other is called non-CpG.

In this table, $B - C$ represents the difference between the CpG sites found just in individual 1 and those found just in individual 2, given sampling of one allele from

each individual. A positive value indicates that there are more individual specific CpG sites in 1 than 2. A negative score indicates an enrichment in CpG sites in individual 2 relative to individual 1. Within European populations or within African populations, this difference should be 0. However, if individual 1 is always chosen to be African, B-C is expected to be positive, since there should be more sites at which the African has mutated away from a CpG site than the European has, given a common ancestor.

The distributions of African-European, African-African and European-European CpG differences are plotted in Figure 2.14. Although all African-African and European-European pairs were tested among our 10 individuals, the choice of individual 1 was not randomized, so it is randomly selected whether I plot B-C (enrichment of individual 1 relative to individual 2) or C-B (enrichment of individual 2 relative to individual 1).

Contrary to our expectations, there is an enrichment of European specific non-CpG sites as compared to African specific non-CG sites. Indeed, it is the only distribution with mean significantly different from 0 (one sample t-test, $p = 0.01957$). Although this test is inconsistent with the hypothesis that there are more mutations in African populations at ancestral-CpG sites, it is worth noting that the variation in the values here is quite large, and the CpG enrichment pattern is slight.

The values of pair comparisons were examined in order to better elucidate this result (Figure 2.15). All values of B-C were normalized and their intensities were plotted for all pairs. There are a few notable individuals. The Luhya (LWK) samples have more CpG sites than any other population in this sample, including the other two African populations. The African-American (ASW) individuals do not appear to be less CpG enriched than the European populations. It can be seen from this figure that there is variation in the relative number of CpG sites within populations.

Importantly, this test examines our previous result without the use of ancestral alleles. Relative divergence from a common ancestor is determined without knowing the state of that common ancestor. Because our results do not agree, this may indicate problems with the selection of our ancestral allele.

I tried repeating the analysis in a separate dataset (Complete Genomics) and observed contradictory results. The patterns seemed to point to data processing problems, and we did not follow up on this line of investigation.

Summarizing, a greater enrichment of derived alleles in African genomes at ancestrally CpG sites was observed as compared to the rest of the genome. An additional test

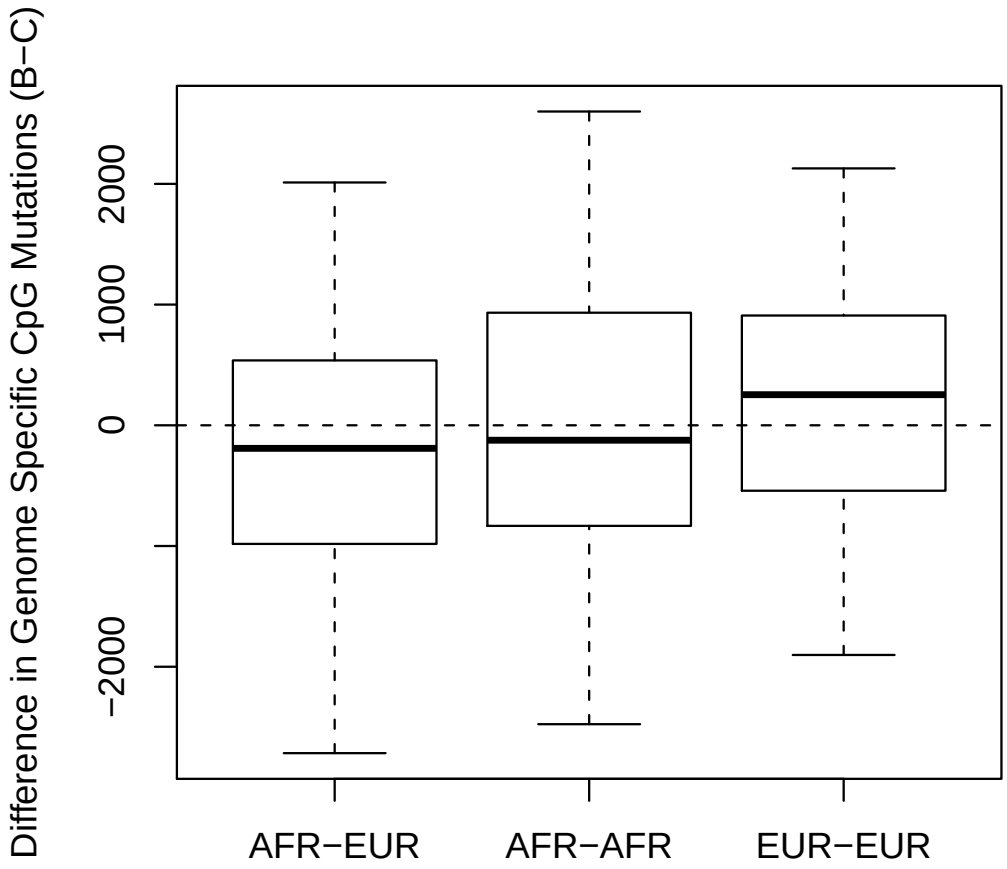


Figure 2.14: Comparison of AFR-EUR, AFR-AFR and EUR-EUR CpG differences - CpG comparisons (B-C) are plotted among African-European pairs (Number of CpG sites in Africans minus the number of CpG sites in Europeans), African-African pairs and European-European pairs.

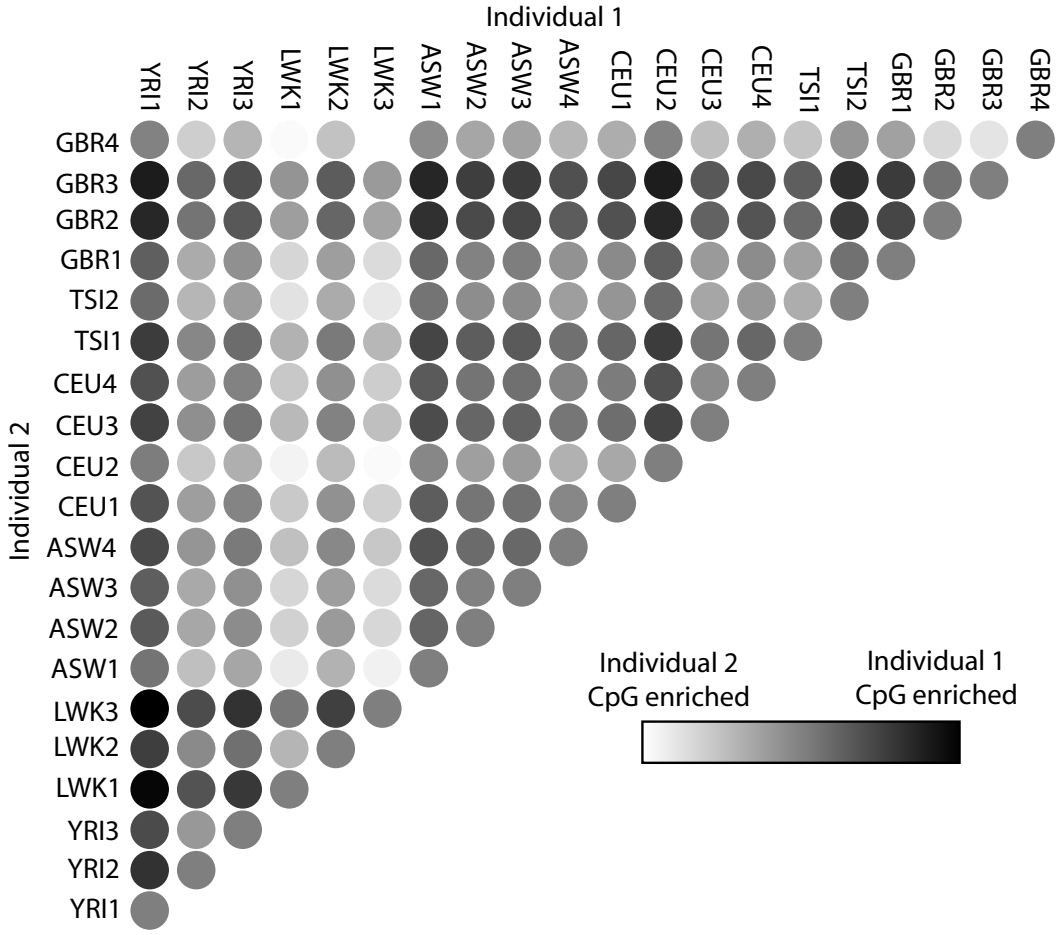


Figure 2.15: Comparison of CpG site load in pairwise comparisons between African and European individuals - Scores of CpG comparisons (B-C) are normalized and plotted, with darker colors indicating an enrichment of CpG sites in individual 1 and lighter colors indicating an enrichment of CpG sites in individual 2. The top right corner represents comparisons among Europeans and the bottom left represents comparisons among Africans.

showed no enrichment of CpG sites in non-African populations as might be expected, suggesting that the original finding is not robust.

2.5.2 Non-synonymous Load Similar in Different Populations

SNPs found in the exome have the potential to display very different types of patterns than those found in non-coding regions. While non-coding SNPs can play important regulatory roles, exome SNPs and, in particular, non-synonymous coding SNPs are under strong selective constraint. Although we might expect that there are the same number of derived alleles in the exome across all populations given the functional constraints, the literature points to an excess of non-synonymous and deleterious SNPs in European populations (see 1.2).

Our analysis was repeated restricted only to sites in the exome (for more details, see section 2.3.2). This was approximately 28,254,834 SNPs or 1.34% of all available SNPs.

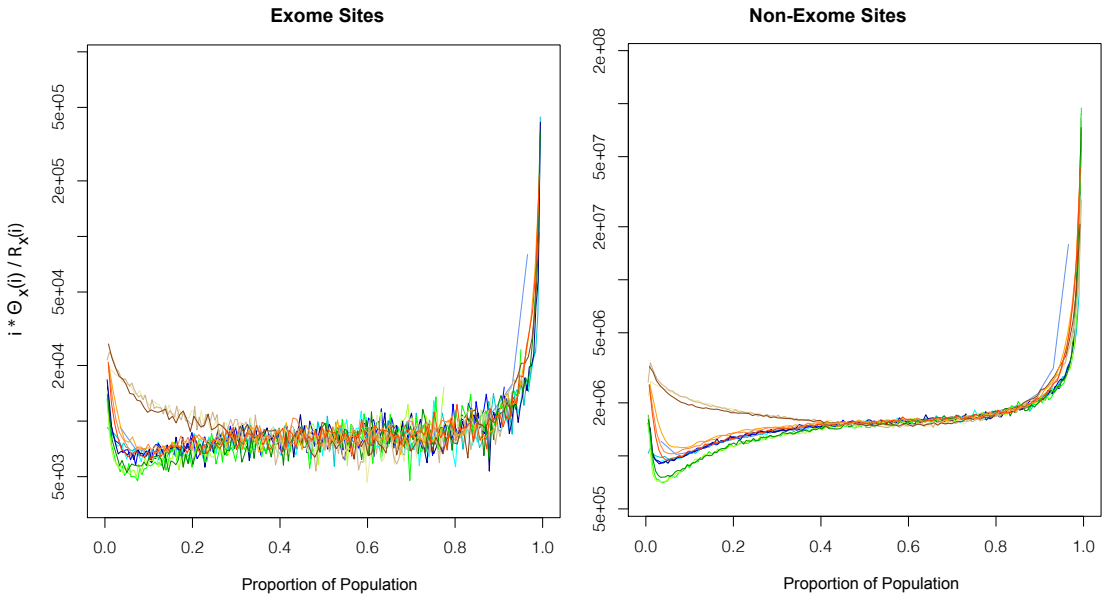


Figure 2.16: Comparison of derived site frequency spectra at exome and non coding sites - In parts A and B, the number of derived variants present at different proportions in the population is plotted in the exome and at non-coding sites. This number is power adjusted as described in 2.3.4, to account for differential ability to pick up low and high frequency mutations in different populations and plotted on a log scale. A legend for the different populations can be found in Figure 2.2.

Once again, a difference is observed in the number of low frequency derived variants between populations in both exome and non-exome sites (Figure 2.16). However, working with a relatively small number of sites in our exome analysis, the difference was obscured by stochasticity of the data. Also notable is the diminished dip for low frequency exome alleles in non-African populations, suggesting that the total number of derived alleles is more similar between the populations than in non-exome sites.

As before, the number of derived alleles in each individual was summed separately in the exome and the non-coding region. In a departure from our previous observations, the numbers of derived alleles among different populations at exome sites were much more similar compared with those falling outside the exome and across the genome as a whole. Although there was still a significant difference in the mean numbers of derived alleles between African and non-African populations for the exome, the difference was slight, with the 95% confidence interval of between 6 and 51 derived alleles per genome ($p = 0.0135$). This represented an enrichment in derived alleles among the African genomes of 0.1% in the exome, as compared to the enrichment of 0.5% outside of the exome.

Importantly, this suggests that individuals in all populations have approximately the same numbers of non-synonymous derived alleles. Not only are differences between populations small, but African individuals possess more derived alleles than European individuals, contrary to what might be suggested by the reported presence of more deleterious alleles in European populations, as the literature has described (Section 1.2).

If African populations accumulate mutations at a faster rate than the rest of the world, either because of a higher mutation rate or shorter generation times, it is possible that this increased rate of mutations would manifest itself genome-wide, except at sites where it was limited by selection. If functional constraint were responsible for the similarity in the number of exome derived alleles between populations, we might expect this difference to be even more pronounced at non-synonymous sites and less strong at synonymous sites under presumably weaker selection.

The differences between the total numbers of derived non-synonymous and synonymous alleles are not significant between African and non-African populations. At synonymous sites, the number of African derived alleles represents only a 0.075% enrichment over the number of non-African derived alleles. At non-synonymous sites, this

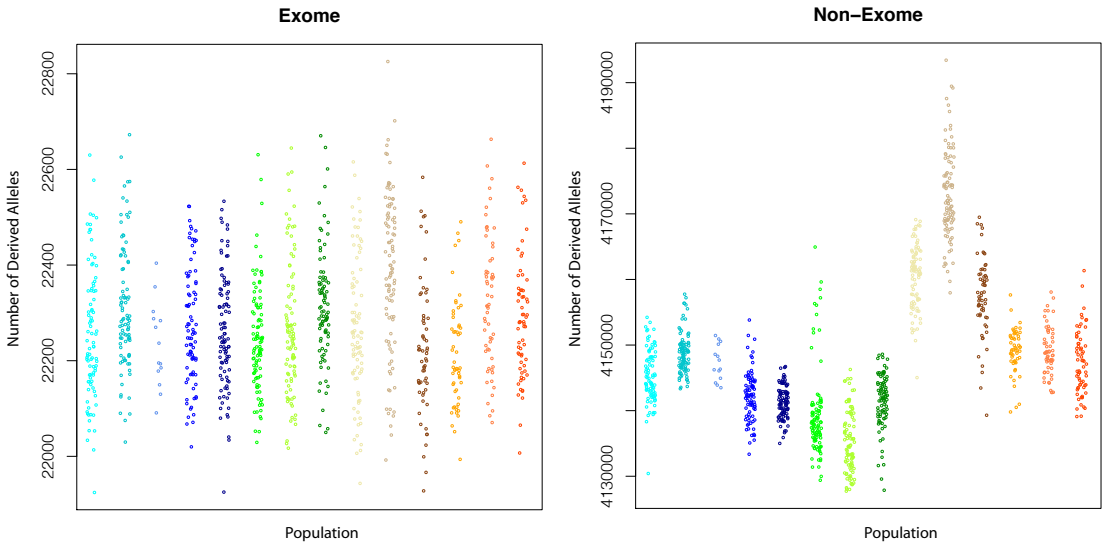


Figure 2.17: Comparison of derived alleles per individual at exome and non-coding sites - Separated by population, the total number of derived alleles per individual is plotted at exome SNPs and non-coding SNPs. The number of SNPs in the noncoding region is power adjusted as described in section 2.3.4.

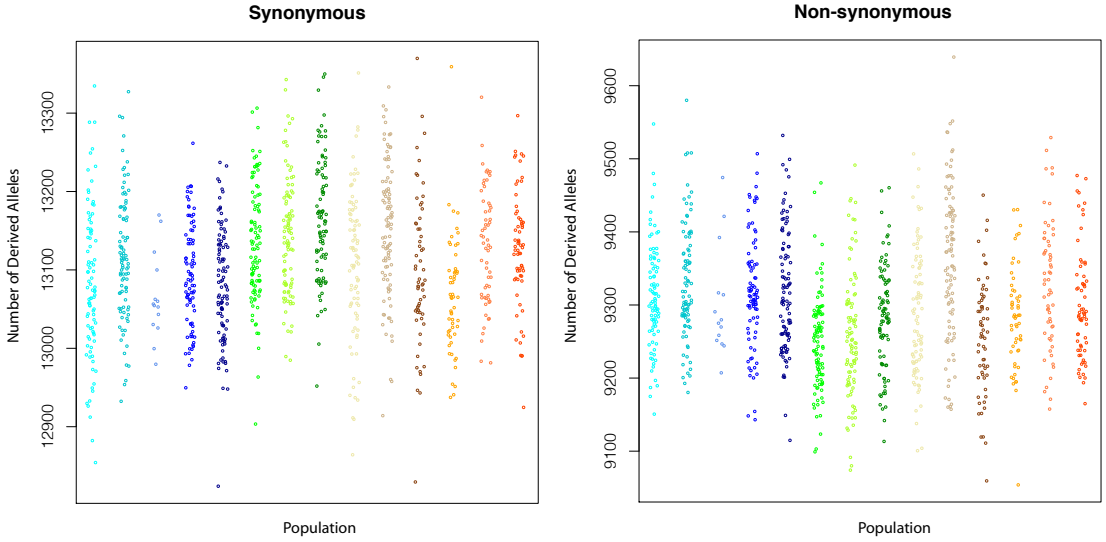


Figure 2.18: Comparison of derived alleles per individual at synonymous and non-synonymous sites - Separated by population, the total number of derived alleles per individual is plotted at synonymous SNPs and non-synonymous SNPs.

enrichment is 0.12%. Although I look only at the number of non-synonymous mutations and do not specifically categorize them as deleterious, these small differences do not support the hypothesis that there is a difference in the number of deleterious alleles between populations.

Pop	Exome	(SE)	Non-Exome	(SE)	Syn	(SE)	Non-Syn	(SE)
GBR	22257	14.7	4145306	428.0	13082	10.1	9312	7.5
FIN	22312	13.5	4148808	322.3	13116	8.2	9331	8.1
IBS	22233	23.7	4147341	728.1	13066	14.8	9303	20.3
CEU	22273	13.2	4142274	407.4	13093	7.2	9318	8.3
TSI	22262	12.1	4141272	244.1	13081	7.3	9320	8.3
CHS	22246	11.0	4139370	662.5	13133	7.4	9250	6.5
CHB	22266	13.9	4135292	474.4	13143	7.6	9257	8.6
JPT	22314	12.7	4141531	433.7	13166	8.2	9283	7.6
YRI	22267	15.4	4160471	508.0	13109	9.9	9287	9.2
LWK	22386	16.8	4172899	697.9	13153	8.5	9360	10.5
ASW	22225	17.6	4157456	761.1	13099	12.2	9257	10.1
PUR	22214	14.8	4149172	449.4	13066	9.9	9283	9.4
CLM	22324	18.1	4149098	474.7	13130	9.5	9328	11.3
MXL	22290	15.6	4147329	589.2	13127	9.0	9299	9.3

Table 2.7: Number of Exome Derived Alleles by Population - Average number of derived alleles per individual by population among at exome sites, non-exome sites, and synonymous and non-synonymous sites in the exome.

It is a good check to note that the number of non-synonymous derived alleles per diploid genome is 9,200-9,300. This is only slightly less than the theoretical predictions of 10,000 non-synonymous alleles by Sunyaev et al (47).

It would be useful in the future to attempt to categorize the functional result of each of the derived alleles and create our own estimates for the number of deleterious mutations. Since we don't have information on which of these non-synonymous alleles are deleterious, it is difficult to directly contradict the European deleterious load hypothesis. As noted, low frequency alleles tend to be private to populations. The non-synonymous SNPs in the European populations are not necessarily the same as the ones in the African populations, and the proportion of deleterious alleles could be

different. However, the similar numbers of non-synonymous mutations do suggest that functional constraint is keeping the number of derived alleles relatively similar between populations, regardless of demographic history.

2.6 Summary

This chapter suggests that the derived site frequency spectra in human populations differ both from neutrality and each other, particularly at low frequency alleles. While all populations show a strong excess of high frequency derived alleles, many of these calls are likely the result of ancestral misidentification.

African populations show an excess of derived alleles when compared to non-African populations. Some evidence suggests that this difference is more pronounced at sites that were ancestrally CpG. The between-population difference essentially disappears when looking only at sites that code for non-synonymous codon substitutions. This suggests that populations accrue mutations at different rates, but functional restraint keeps the number of derived alleles approximately equal across populations.

3

Incorporation of Derived Allele Information into Models of Coalescence Prediction

3.1 Introduction

As we have seen, understanding patterns of observed derived allele frequencies can allow us to make observations about demography, selection and relationships between populations. We would now like to leverage our observations about allele frequencies in order to make more informed statements about coalescent times between sequences.

While many of the methods introduced in section 1.3 seek to estimate the TMRCA of any set of sequences, most focus on mtDNA or DNA on the Y-chromosome which are special cases that do not require an understanding of recombination. When trying to construct ancestral lineages with recombination, the problem becomes much more difficult, as the space in which multiple sequences can recombine during coalescence is considerable.

The sequentially Markov coalescent (SMC) introduced in Cardin & McVean (34) seeks to address this problem by simplifying the search space. First, it disallows coalescence between any two sequences that do not share any common intervals, decreasing the time the algorithm must use in considering recombination events that have low probability. Second, it draws information about tree structure from neighboring sites, but only sites that are directly adjacent. This allows estimations to be informed by

other sites, but not be overwhelmed by having to review every possible combination of recombinations genome-wide. It represents a way to estimate the TMCRA between a group of sequences by simplifying the creation of ancestral recombination graphs.

The SMC was applied by Li and Durbin (28) to calculate pairwise coalescence times between two sequences, calling it the pairwise SMC (PSMC). In the current state of the PSMC, we know only if two sequences match or do not match. A long stretch of matching sequences might suggest recent coalescent time in that region. However, if we use the allele frequency spectrum, we can perhaps learn more. For instance, if we know that two sequences share a high frequency allele, it may be less informative than if they share a very rare one. Further, knowing that the allele that two sequences share is derived with respect to the human MRCA is probably indicative of an even more recent coalescent time.

3.1.1 Standard PSMC

In practice, the PSMC is a hidden Markov model (HMM) with hidden states representing the time of coalescence between the two sequences, and observed states of whether or not the two sequences match. As with other HMMs, the forward-backward algorithm can be applied to determine the most probable course through the hidden states given the emission states.

The first step in doing this is to determine the probability of transitioning between two adjacent states. Given that our coalescence time for a site $k-1$ is at time t_{k-1}, what is the probability density function for the coalescence time in an adjacent site k? We must consider all possible coalescence times x between the present and the earlier coalescence time of the previous state and the current state. At each of these times, we consider the probability that the first recombination event is at time x. This probability is given by the exponential distribution with parameter ρ, quantifying the degree of recombination.

In order to transition from coalescent time t_{k-1} to t_k, we must have a coalescence event at the minimum of the two possible coalescent events, and then also a recombination at some point more recently (i.e., sometime between the present and $\min(t_k, t_{k-1})$. We must integrate over all possible times for the recombination event on $[0, \min(t_{k-1}, t_k)]$. Recombination events are distributed as a Poisson process with parameter ρ. Once the recombination event time x is set, we must still include the probability

of a coalescent event at $\min(t_k, t_{k-1})$

$$d(t_k|t_{k-1}) = \int_0^{min(t_k,t_{k-1})} \rho e^{-\rho x} e^{-(min(t_{k-1},t_k)-x)} \mathrm{d}x$$

If no recombination occurs, then $t_k = t_{k-1}$. We will consider this case last. Let us consider the case with recombination first. Let $y = min(t_k, t_{k-1})$. Integrating we see that

$$d(t_k|t_{k-1}) = \frac{\rho e^{-y}}{\rho - 1}(1 - e^{-y(\rho-1)})$$

Since t_k is a continuous and infinite variable, we must perform a discretization in order to be able to define a transition density function at which any point contains non-zero probabilities. We want to map $[0, \infty)$ onto [0,1] and split it into n time subgroups within that domain. We define the intervals

$$q_i = -\log(1 - \frac{i}{n}) \qquad\qquad i = 0, 1, ..n$$

These intervals partition [0,1] and all times t_k map into one of the intervals. We can then allow τ_k to be the epoch at time t_k and τ_{k-1} to be t_{k-1}'s associated epoch. So now, we are interested in computing the transition probability between epochs $d(\tau_k|\tau_{k-1})$ instead of between time points. We delimit the interval τ_k with upper timepoint q_k^u and lower timepoint q_k^l. For now, we will assume that $\tau_k \leq \tau_{k-1}$, which is to say that we are moving from an older time epoch to a younger one. In order to compute this, we need double integrate over every possible point of coalescence in epoch $k-1$ to every possible coalescence time in the current epoch t_k.

So, from $d(t_k|t_{k-1})$, we get

$$\delta(\tau_k|\tau_{k-1}) = \int_{q_{k-1}^l}^{q_{k-1}^u} p(x) \int_{q_k^l}^{q_k^u} d(y|x) dy dx$$

We need to set a prior on x, the time in the previous epoch of τ_{k-1}. We will use an exponential distribution, so $p(x) = ne^{-x}$. Thus,

$$\delta(\tau_k|\tau_{k-1}) = \int_{q_{k-1}^l}^{q_{k-1}^u} ne^{-x} \int_{q_k^l}^{q_k^u} \frac{\rho e^{-y}}{\rho - 1}(1 - e^{-y(\rho-1)}) dy dx$$

Since y is the minimum of the coalescence times in the $(k-1)^{th}$ and k^{th} epochs, and because we have assumed that $t_k < t_{k-1}$, y is independent of x. Therefore, we can perform the integrations separately:

$$\delta(\tau_k|\tau_{k-1}) = \frac{\rho}{\rho-1}\int_{q_k^l}^{q_k^u} e^{-y}(1-e^{-y(\rho-1)})dy \int_{q_{k-1}^l}^{q_{k-1}^u} ne^{-x}dx$$

Because $e^{-q_{k-1}^u} - e^{q_{k-1}^l} = \frac{1}{n}$,

$$\int_{q_{k-1}^l}^{q_{k-1}^u} ne^{-x}dx = -1$$

Then,

$$\delta(\tau_k|\tau_{k-1}) = \frac{-\rho}{\rho-1}(-(e^{q_k^u} - e^{q_k^l}) + \frac{1}{\rho}(e^{-\rho q_k^u} - e^{-\rho q_k^l}))$$

And finally,

$$\delta(\tau_k|\tau_{k-1}) = \frac{1}{\rho-1}(\frac{\rho}{n} + e^{-\rho q_k^l} - e^{-\rho q_k^u})$$

Given that we now have a density function from state τ_{k-1} to τ_k, where $\tau_k < \tau_{k-1}$, we can determine the probability density when $\tau_k > \tau_{k-1}$ by calculating $\delta(\tau_{k-1}|\tau_k)$. The probability that $\tau_k = \tau_{k-1}$ (no recombination occurred, or recombination did not affect the epoch) can be determined by finding the remaining probability once the probability of both epoch transition probabilities have been determined.

We must now determine the emission probabilities by hidden state. Given that we are in state τ_k, what is the probability that our states match? The probability that the states match is given by the probability that no mutations have occurred since the coalescence time t. In the continuous case, this is just a Poisson distributed parameter, and

$$P_{\text{match}|t} = e^{-\theta t}$$

$$P_{\text{no match}|t} = 1 - e^{-\theta t}$$

From here, we can run the forward backward algorithm to derive an estimate to TMRCA at each site.

3.2 Future Directions

Moving forward, I hope to take advantage of the understanding gained in the second chapter in order to build upon the PSMC. Although we have established that not all populations have the same number of derived alleles, the between-population differences are on the order of ¡1%. To integrate differential mutation rate information into a future model, we could only apply a correction after lineage differentiation. This represents a small proportion of the tree, and is unlikely to affect the TMRCA estimate. With this in mind, we can add the additional information provided by the polarization of these alleles to better inform the PSMC without needing to adjust for differential mutation rates.

The most natural first step would be an expansion of the emission states. If, for example, at each site, we knew not only if two sequences A and B agreed, but also if they agreed in the derived state or the ancestral state, this would already be additional information we could add to a model. As before, the probability that they do not match is the probability that there has been exactly one mutation since coalescence. However, if we attempt to categorize the mutational conditions that would result in A and B matching in the derived state, we see that considering only two sequences at a time is no longer possible. Any derived mutation present in both A and B will have arisen before the A and B's MRCA. Therefore, we must consider A and B in the context of a larger population, with a longer TMRCA. Given the TMRCA for the sample, and the pairwise coalescence time t, we can define the emission probabilities as follows:

$$P_{\text{no match}|t,T_{MRCA}} = 1 - e^{-\theta t}$$

$$P_{\text{match, derived state}|t,T_{MRCA}} = e^{-\theta t}[1 - e^{-\theta(T_{MRCA}-t)}]$$

$$P_{\text{match, ancestral state}|t,T_{MRCA}} = e^{-\theta t}[e^{-\theta(T_{MRCA}-t)}]$$

In the PSMC, the relevant information with respect to the hidden state at site i was the emission at i and the hidden state probability distributions of $i+1$ and $i-1$. If we augment this model with derived allele information, we must also simultaneously predict the TMRCA for the entire population at this site. Although we can define the emission probabilities given a population and pairwise TMRCA as above, producing good estimates for the population and pairwise TMRCA represents more of a challenge.

While a Gibbs sampler style algorithm could be used to determine maximally-likely values for the pairwise and population TMRCA simultaneously, this is computationally more intensive than the standard PSMC upon which the forward-backward algorithm can be easily applied.

While no implementation of this model has yet been developed, ensuring that there are a roughly equal number of derived alleles in different populations lays the groundwork upon which this SMC extension can be built.

References

[1] **An integrated map of genetic variation from 1,092 human genomes**. *Nature*, **491**(7422):56–65, 2012. 23, 26

[2] David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, SF Schaffner, F Yu, Penelope E Bonnen, PI De Bakker, Panos Deloukas, Stacey B Gabriel, et al. **Integrating common and rare genetic variation in diverse human populations**. *Nature*, **467**(7311):@articlealtshuler2010integrating, title=Integrating common and rare genetic variation in diverse human populations, author=Altshuler, David M and Gibbs, Richard A and Peltonen, Leena and Dermitzakis, Emmanouil and Schaffner, SF and Yu, F and Bonnen, Penelope E and De Bakker, PI and Deloukas, Panos and Gabriel, Stacey B and others, journal=Nature, volume=467, number=7311, pages=52, year=2010, publisher=Wellcome Trust Public Access , 2010. 5, 7

[3] Luis B Barreiro, Etienne Patin, Olivier Neyrolles, Howard M Cann, Brigitte Gicquel, and Lluís Quintana-Murci. **The Heritage of Pathogen Pressures and Ancient Demography in the Human Innate-Immunity¡ i¿ CD209/CD209L¡/i¿ Region**. *The American Journal of Human Genetics*, **77**(5):869–886, 2005. 15

[4] Doron M Behar, Mannis van Oven, Saharon Rosset, Mait Metspalu, Eva-Liis Loogväli, Nuno M Silva, Toomas Kivisild, Antonio Torroni, and Richard Villems. **A "Copernican" reassessment of the human mitochondrial DNA tree from its root**. *The American Journal of Human Genetics*, **90**(4):675–684, 2012. 6, 10

[5] Adrian Bird. **DNA methylation patterns and epigenetic memory**. *Genes & development*, **16**(1):6–21, 2002. 38

[6] Michael GB Blum and Mattias Jakobsson. **Deep divergences of human gene trees and models of human origins**. *Molecular biology and evolution*, **28**(2):889–898, 2011. 15

[7] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. **Assessing the evolutionary impact of amino acid mutations in the human genome**. *PLoS genetics*, **4**(5):e1000083, 2008. 8

[8] Rebecca L. Cann, Mark Stoneking, and Allan C. Wilson. **Mitochondrial DNA and human evolution**. *Nature*, **325**(6099):31–36, 1987. 11, 14

[9] Sung Chun and Justin C Fay. **Identification of deleterious mutations within three human genomes**. *Genome research*, **19**(9):1553–1561, 2009. 7, 8

[10] Sung Chun and Justin C Fay. **Evidence for hitchhiking of deleterious mutations within the human genome**. *PLoS genetics*, **7**(8):e1002240, 2011. 9

[11] Alex Coventry, Lara M Bull-Otterson, Xiaoming Liu, Andrew G Clark, Taylor J Maxwell, Jacy Crosby, James E Hixson, Thomas J Rea, Donna M Muzny, Lora R Lewis, et al. **Deep resequencing reveals excess rare recent variants consistent with explosive population growth**. *Nature communications*, **1**:131, 2010. 2, 3, 5

[12] Fulvio Cruciani, Beniamino Trombetta, Andrea Massaia, Giovanni Destro-Bisol, Daniele Sellitto, and Rosaria Scozzari. **A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa**. *The American Journal of Human Genetics*, **88**(6):814–818, 2011. 13, 14

[13] Adam Eyre-Walker. **Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies**. *Proceedings of the National Academy of Sciences*, **107**(suppl 1):1752–1756, 2010. 5

[14] Maroussia Favre and Didier Sornette. **Strong gender differences in reproductive success variance, and the times to the most recent common ancestors**. *Journal of Theoretical Biology*, 2012. 13

[15] Wenqing Fu, Timothy D O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A Nickerson, et al. **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants**. *Nature*, 2012. 6, 7, 8

[16] Yun-Xin Fu. **Statistical properties of segregating sites**. *Theoretical population biology*, **48**(2):172–197, 1995. 17

[17] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, David L Altshuler, et al. **Demographic history and rare allele sharing among human populations**. *Proceedings of the National Academy of Sciences*, **108**(29):11983–11988, 2011. 6

[18] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data**. *PLoS genetics*, **5**(10):e1000695, 2009. 3, 6, 29

[19] Marc K Halushka, Jian-Bing Fan, Kimberly Bentley, Linda Hsie, Naiping Shen, Alan Weder, Richard Cooper, Robert Lipshutz, and Aravinda Chakravarti. **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis**. *Nature genetics*, **22**(3):239–247, 1999. 9

[20] Toshiyuki Hayakawa, Ikuko Aki, Ajit Varki, Yoko Satta, and Naoyuki Takahata. **Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution**. *Genetics*, **172**(2):1139–1146, 2006. 15

[21] Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. **Context dependence, ancestral misidentification, and spurious signatures of natural selection**. *Molecular biology and evolution*, **24**(8):1792–1800, 2007. 6

[22] David A Hinds, Laura L Stuve, Geoffrey B Nilsen, Eran Halperin, Eleazar Eskin, Dennis G Ballinger, Kelly A Frazer, and David R Cox. **Whole-genome patterns of common DNA variation in three human populations**. *Science*, **307**(5712):1072–1079, 2005. 5

[23] Max Ingman, Henrik Kaessmann, Svante PaÈaÈbo, and Ulf Gyllensten. **Mitochondrial genome variation and the origin of modern humans**. *Nature*, **408**(6813):708–713, 2000. 12, 14

[24] Kamel Jabbari and Giorgio Bernardi. **Cytosine methylation and CpG, TpG (CpA) and TpA frequencies**. *Gene*, **333**:143–149, 2004. 38

[25] Alon Keinan and Andrew G Clark. **Recent explosive human population growth has resulted in an excess of rare genetic variants**. *science*, **336**(6082):740–743, 2012. 3, 4, 6

[26] Alon Keinan, James C Mullikin, Nick Patterson, and David Reich. **Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans**. *Nature genetics*, **39**(10):1251–1255, 2007. 6

[27] Motoo Kimura et al. **Evolutionary rate at the molecular level**. *Nature*, **217**(5129):624–626, 1968. 10

[28] Heng Li and Richard Durbin. **Inference of human population history from individual whole-genome sequences**. *Nature*, **475**(7357):493–496, 2011. 54

[29] Yingrui Li, Nicolas Vinckenbosch, Geng Tian, Emilia Huerta-Sanchez, Tao Jiang, Hui Jiang, Anders Albrechtsen, Gitte Andersen, Hongzhi Cao, Thorfinn Korneliussen, et al. **Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants**. *Nature genetics*, **42**(11):969–972, 2010. 2, 3, 6

[30] DAG LINDGREN. **The temperature influence on the spontaneous mutation rate**. *Hereditas*, **70**(2):165–177, 1972. 9, 29, 37

[31] Kirk E Lohmueller, Amit R Indap, Steffen Schmidt, Adam R Boyko, Ryan D Hernandez, Melissa J Hubisz, John J Sninsky, Thomas J White, Shamil R Sunyaev, Rasmus Nielsen, et al. **Proportionally more deleterious genetic variation in European than in African populations**. *Nature*, **451**(7181):994–997, 2008. 8, 14

[32] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. **Finding the missing heritability of complex diseases**. *Nature*, **461**(7265):747–753, 2009. 7

[33] Gabor Marth, Fuli Yu, Amit Indap, Kiran Garimella, Simon Gravel, Wen Leong, Chris Tyler-Smith, Matthew Bainbridge, Tom Blackwell, Xiangqun Zheng-Bradley, et al. **The functional spectrum of low-frequency coding variation**. *Genome biology*, **12**(9):R84, 2011. 2, 3

[34] Gilean AT McVean and Niall J Cardin. **Approximating the coalescent with recombination**. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459):1387–1393, 2005. 53

[35] Fernando L Mendez, Thomas Krahn, Bonnie Schrack, Astrid-Maria Krahn, Krishna R Veeramah, August E Woerner, Forka Leypey Mathew Fomine, Neil Bradman, Mark G Thomas, Tatiana M Karafet, et al. **An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree**. *The American Journal of Human Genetics*, 2013. 13, 14

[36] Erika L Moen, Xu Zhang, Wenbo Mu, Shannon M Delaney, Claudia Wing, Jennifer McQuade, Jamie Myers, Lucy A Godley, M Eileen Dolan, and Wei Zhang. **Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits**. *Genetics*, 2013. 43

[37] Newton E Morton, James F Crow, and HJ Muller. **An estimate of the mutational damage in man from data on consanguineous marriages**. *Proceedings of the National Academy of Sciences of the United States of America*, **42**(11):855, 1956. 7

[38] Carina F Mugal and Hans Ellegren. **Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content**. *Genome biology*, **12**(6):R58, 2011. 43

[39] Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. **An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people**. *Science*, **337**(6090):100–104, 2012. 2, 3, 6, 7

[40] Rasmus Nielsen, Melissa J Hubisz, Ines Hellmann, Dara Torgerson, Aida M Andrés, Anders Albrechtsen, Ryan Gutenkunst, Mark D Adams, Michele Cargill, Adam Boyko, et al. **Darwinian and demographic forces affecting human protein coding genes**. *Genome research*, **19**(5):838–849, 2009. 6

[41] Jonathan K Pritchard. **Are rare variants responsible for susceptibility to complex diseases?** *The American Journal of Human Genetics*, **69**(1):124–137, 2001. 5

[42] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. **Population growth of human Y chromosomes: a study of Y chromosome microsatellites.** *Molecular Biology and Evolution*, **16**(12):1791–1798, 1999. 13, 14

[43] WMD Rideout, Gerhard A Coetzee, Aria F Olumi, Peter A Jones, et al. **5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes**. *Science*, **249**(4974):1288–1290, 1990. 38

[44] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne,

Steven A McCarroll, Rachelle Gaudet, et al. **Genome-wide detection and characterization of positive selection in human populations**. *Nature*, **449**(7164):913–918, 2007. 9

[45] Hreinn Stefansson, Agnar Helgason, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Gisli Masson, John Barnard, Adam Baker, Aslaug Jonasdottir, Andres Ingason, Vala G Gudnadottir, et al. **A common inversion under selection in Europeans**. *Nature genetics*, **37**(2):129–137, 2005. 15

[46] Barbara E Stranger, Eli A Stahl, and Towfique Raj. **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics*, **187**(2):367–383, 2011. 5

[47] Shamil Sunyaev, Vasily Ramensky, Ina Koch, Warren Lathe III, Alexey S Kondrashov, and Peer Bork. **Prediction of deleterious human alleles**. *Human molecular genetics*, **10**(6):591–597, 2001. 7, 51

[48] Hua Tang, David O Siegmund, Peidong Shen, Peter J Oefner, and Marcus W Feldman. **Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition**. *Genetics*, **161**(1):447–459, 2002. 11, 12, 14

[49] Simon Tavare, David J Balding, RC Griffiths, and Peter Donnelly. **Inferring coalescence times from DNA sequence data**. *Genetics*, **145**(2):505–518, 1997. 12

[50] Alan R Templeton. **The "Eve" hypotheses: a genetic critique and reanalysis**. *American Anthropologist*, **95**(1):51–72, 1993. 11

[51] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. **Evolution and functional impact of rare coding variation from deep sequencing of human exomes**. *Science*, **337**(6090):64–69, 2012. 2, 3, 5

[52] Russell Thomson, Jonathan K Pritchard, Peidong Shen, Peter J Oefner, and Marcus W Feldman. **Recent common ancestry of human Y chromosomes: evidence from DNA sequence data**. *Proceedings of the National Academy of Sciences*, **97**(13):7360–7365, 2000. 11, 14

[53] http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37

[54] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/human_ancestor_GRCh37_e59.README

[55] http://www.ncbi.nlm.nih.gov/projects/SNP/