Understanding the Patterns and Consequences of Single-Nucleotide Mutations in the Human Genome Using High-Throughput Sequencing

by

Valerie Marie Schaibley

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2013

Doctoral Committee:

Assistant Professor Jun Li, Chair
Professor David T. Burke
Associate Professor Julie A. Douglas
Associate Professor Donna M. Martin
Associate Professor Sebastian K. Zoellner

DEDICATION

To my family

ACKNOWLEDGEMENTS

I could not have done this without the never-ending support from those around me. First, I must thank my husband and my best friend, John. I could not have done any of this without his unwavering support. I also have to acknowledge my incredible parents and step-parents, Mom, Dad, Candis, and Dennis, for always believing that I could do anything. I also want to thank my siblings and step-siblings for always being there for me, and my nieces and nephews for always putting a smile on my face.

I also have to thank all of the friends I have made at the University of Michigan. First, I must thank Kaanan Shah and Michele Gornick for lunch breaks, coffee breaks, and everything in between. From our first year in PIBS together, Stephanie Coomes, Kadee Luderman, Heather McLaughlin, Cheryl Smith and Ilea Swinehart have been an amazing source of support.

One key person is my advisor, Jun Li, whom I must thank for his support and dedication to my education over these past several years. I must also thank the past and present members of the Li lab for their help throughout the years, specifically Jishu Xu and Bilge Ozel.

Finally, I would like to sincerely thank the members of my thesis committee, Dave Burke, Julie Douglas, Donna Martin and Sebastian Zoellner for providing me with excellent guidance on my dissertation projects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
BGC	Biased Gene Conversion
CNV	Copy Number Variant
DAF	Derived Allele Frequency
DHJ	Double-Holliday Junction
DSB	Double-Strand Break
DTH	Distance To Recombination Hotspot
EPO	Enredo, Pecan, Ortheus
ESP	Exome Sequencing Project
GWAS	Genome-Wide Association Study
HT	High-Throughput Sequencing
MPS	Martin-Probst Syndrome
MRCA	Most Recent Common Ancestor
OMIM	Online Mendelian Inheritance in Man
S	“Strong” G:C Base-Pairs
SFS	Site-Frequency Spectrum
SNP	Single-Nucleotide Polymorphism
SNV	Single-Nucleotide Variant
STR	Small-Tandem Repeat
Ti	Transition
Tv	Transversion
UCSC	University of California Santa Cruz
UTR	Untranslated Region
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
W	“Weak” A:T Base-pairs
XSS	X Chromosome Targeted Exome Sequencing

ABSTRACT

Recent advances in high-throughput genome sequencing technology have paved the way for the field to gain a better understanding of single-nucleotide mutations in the human genome. Until recently, analysis of rare single-nucleotide variants in humans was restricted by technology that limited the expansion to larger sample sizes and greater numbers of loci. The three projects presented here overcome these limitations, using data and results from high-throughput studies to understand the innate features of the genome that influence how frequently different types of mutations occur and identify those mutations that lead to human genetic disease.

First, I studied a rare Mendelian disorder, Martin-Probst Syndrome, which is characterized by sensorineural hearing loss and mental retardation. I used whole genome, whole exome, and X-specific exome sequencing across two affected male individuals from one family to identify mutations occurring in a previously identified X-chromosome haplotype block. After stringent filtering and validation steps, I identified two adjacent single-nucleotide mutations in the gene *RAB40AL*, likely leading to Martin-Probst Syndrome in this family.

The second project was aimed at understanding the degree to which innate features of the genome influence the spontaneous single-nucleotide mutation rate in humans and evolutionary processes that alter fixation rates of single-nucleotide variants. I used rare variants (derived allele frequency < 0.0001) to analyze mutation patterns, and common variants and substitutions to study fixation processes. I found

that GC content influences the mutation rate and fixation processes differently, especially with regard to distinct variant subtypes. Recombination rate, on the other hand, more strongly influences fixation, as evidenced by the stronger effect on common variants and substitutions than rare variants, consistent with biased gene conversion influencing variant patterns in humans.

Finally, I developed a forward genetic simulation program, SubSim, that models subtype-specific selection and mutation, along with base composition, recombination rate and biased gene conversion. Subtype-specific selection and altering the base compositions are two features unique to SubSim. These advances in the available simulation software will help the field gain a better understanding of the evolutionary forces that lead to patterns of single-nucleotide mutation events and fixation of variants in the human genome.

CHAPTER 1

Introduction

1.1 Mutation in the Human Genome

Mutations occur throughout the human genome. Depending on the location of the mutation, the time at which it occurs, and the resulting effect of that specific change to the DNA sequence, mutations can have beneficial, harmful, or negligible effects. Mutations have a bad reputation. They are thought of primarily in terms of cancer and other debilitating diseases. This gut-instinct reaction is not entirely incorrect. While the majority of mutations have no effect on cell function and health, those that do have a biological effect are often harmful and can lead to diseases such as inherited disorders and cancer. However, not all mutations are bad. Some, in fact, lead to more efficient or new cellular functions. These beneficial mutations are what evolution can act on, potentially changing the way an organism interacts with its environment or even leading the way for the evolution of a new species. We as a species would never have evolved without our mutations.

Mutations are classified into several major classes, depending on the changes that they make to the DNA sequence. Single-nucleotide mutations are the most prominent and most simplistic form of mutation. As the name suggests, they affect a single nucleotide in the DNA. A single-nucleotide mutation occurs when the nascent base pair is simply replaced by a different base pair. For example, if a DNA strand has

the sequence ATGTA, a single-nucleotide mutation could change the G base at the 3rd position to a T base, leading to the sequence ATTTA. The resulting mutation changes the sequence of As, Ts, Cs, and Gs, in the DNA sequence, but has no effect on the overall length of the DNA strand. Other mutations, however, give rise to small or quite drastic alterations in the length of the DNA. Small tandem repeats (STRs) are sets of repeated sequences in one to six nucleotide repeating units and are highly polymorphic in human populations (Sutherland and Richards 1995). The trinucleotide repeat subset of the STRs are prone to dynamic changes in their copy number due to polymerase slippage during replication that occurs at these repeat regions, leading to an increase in the number of repeats at that locus (Richards and Sutherland 1994). Small insertions and deletions (indels) affect 1 - 10,000 bp by removing (deletions) or adding (insertions) nucleotides at a specific locus (Mullaney et al. 2010). Copy number variants (CNVs) are similar to indels, in that they insert additional sequence or remove a specific set of base pairs from the genome. The difference, however, is the scale on which they act. CNVs affect much larger chromosomal regions, up to hundreds of millions of base pairs at once. The largest, and likely the most damaging forms of mutations, are large-scale chromosomal gains, losses, or rearrangements, known as cytogenetic abnormalities. These types of mutations result in an excess or lack of chromosome arms, as is seen in the fusion of chromosomes 9 and 22 in patients with chronic myelogenous leukemia (Rowley 1973), or entire chromosomes, such as an extra copy of chromosome 21 leading to Down Syndrome.

1.2 Discovery of Disease Genes in Rare Mendelian Disorders

Mendelian disorders are defined as diseases that follow Mendel's laws of inheritance. Disorders such as Huntington's disease and cystic fibrosis are both relatively common and therefore relatively well known. To date, there are > 21,000 recognized Mendelian phenotypes, of which only ~4,800 have a known molecular basis (Online Mendelian Inheritance in Man). Some of the first efforts to identify the genes for these diseases used a method known as positional cloning. Positional cloning is a technique that maps the locus of a specific mutation to a large region of the genome using segregation patterns of genetic markers, historically microsatellite markers, observed in affected and unaffected individuals. Once a large chromosomal location is identified, fine-scale mapping within that locus is performed to identify the narrow region of the genome containing the causative gene. PCR and sequencing of coding regions in in the interval is then performed to identify mutations in affected individuals. These techniques discovered mutations in genes that were previously unknown to lead to disease. Cystic fibrosis (Kerem et al. 1989; Riordan et al. 1989; Rommens et al. 1989), Huntington disease (1993), and Duchenne Muscular Dystrophy (Monaco et al. 1986) are just a few examples of the successful application of positional cloning to disease gene discovery.

Identification of genes that lead to Mendelian disease has several important implications. First, knowing what genes are affected in these disorders can help lead to potential treatment options. For example, the most common mutant allele in cystic fibrosis is a 3 bp deletion leading to an absence of the 508^{th} amino acid, phenylalanine, in the CFTR protein (Kerem et al. 1989; Riordan et al. 1989; Rommens et al. 1989). A

recent high-throughput (HT) screen identified several small molecules that can partially rescue CFTR function in individuals with this specific mutation (Pedemonte et al. 2005). Additionally, while these types of disorders are often extremely debilitating, researching them leads to a better understanding of biology and potential treatment options.

Positional cloning and other gene-mapping techniques have successfully identified genes causing many Mendelian disorders. However, there are still a large number of diseases where the underlying cause is unknown. This can be due to several factors. For one, not all Mendelian disorders are due to mutations in a single gene. These monogenic disorders are easily the most simplistic in terms of their segregation patterns. However, many disorders previously thought to be monogenic are in reality due to mutations in two or more genes. Fine-scale mapping is difficult, and many candidate genes are frequently identified in these regions. PCR and sequencing in large sample sizes or analyzing many genomic loci is difficult and time consuming. It is often unfeasible to apply these techniques on a large-scale. Overall, while these techniques have helped to significantly advance the field of human genetics, these inherent limitations necessitate development and application of new methodologies to further identify the root causes of many of these Mendelian disorders. Additional factors, however, can influence gene-mapping studies in Mendelian disorders. Gene-environment interactions, gene-gene interactions, allelic and locus heterogeneity all influence gene identification strategies, and new techniques and strategies aimed at studying these complex effects are necessary.

1.3 Quantification and Understanding the Single-Nucleotide Mutation Rate

Mutations not only cause disease, they also provide the raw material for evolution. Therefore, it is important to understand the frequency with which single-nucleotide mutations occur in the human genome. Historically, there are two main approaches to quantify the frequency of mutations in the human genome. The first is based on the frequency of dominant disorders and was pioneered by Haldane (Haldane 1935). Haldane's initial quantification using his method found the mutation rate to be on the order of 10^{-5} per-base pair per-generation in individuals with hemophilia (Haldane 1935). A number of other studies used the same or similar disease-based methods to quantify the rate of spontaneous mutation, estimating values of 3.6×10^{-9} (Sommer 1995), $\sim 11 \times 10^{-9}$ (Lynch 2010), and 1.8×10^{-8} per-base per-generation (Kondrashov 2003).

The second approach for estimating the single-nucleotide mutation rate is based on Kimura's theory of neutral evolution. This theory postulates that the majority of new mutations will be neutral and therefore the frequency of interspecies substitutions represents the frequency of those neutral mutations (Kimura 1983). The rate of divergent bases (substitution rate) between humans and closely related species, such as chimpanzee, has been used for these indirect estimates of the single-nucleotide mutation rate: $\sim 2.5 \times 10^{-8}$ (Nachman and Crowell 2000) and $\sim 1\text{-}2 \times 10^{-8}$ per-base per-generation (Kondrashov and Crow 1993; Drake et al. 1998).

For many years, it was thought that single-nucleotide mutations occurred randomly throughout the genome. Wolfe and colleagues (1989) were the first to show variability in the rate of mutations across different genes. This finding was the first to

suggest that in addition to pressure from the environment, innate genomic features could influence the types and frequency with which mutations occur. Other investigators observed similar variability in the neutral substitution rate and intraspecies diversity over both the total rate of mutations and also among different subtypes of single-nucleotide mutation (Nachman and Crowell 2000; Sachidanandam et al. 2001; Smith and Lercher 2002; Kondrashov 2003; Hodgkinson et al. 2009). Since these initial findings, a great deal of work has gone into understanding why the mutation rates appear to be variable from one region of the genome to another without any external stimulus.

Variability in the mutation rate is observed from large-scale genomic regions exhibiting different mutation patterns to adjoining bases having different rates of mutation. One of the most well-studied effects on the per-base mutation rate is the observation that C>T transitions at CpG dinucleotides occur at 10 - 40 times the rate of other mutations (Sommer 1995; Nachman and Crowell 2000; Kondrashov 2003; Hwang and Green 2004). The cytosine base at a CpG dinucleotide (a C base followed directly by a G base) is prone to methylation, forming 5-methylcytosine, which undergoes spontaneous deamination to produce thymine, leading to C>T transitions (G>A on the opposite strand) (Cooper and Youssoufian 1988; Cooper and Krawczak 1993).

Beyond the increased mutation rate at CpG sites, the nucleotides surrounding a base pair impact the type and frequency of mutations that occur. Hwang and Green found that the rate of specific single-nucleotide mutations depends on the two bases directly flanking a given nucleotide (Hwang and Green 2004). Beyond these immediately adjacent effects, the mutation rate appears to depend on the nucleotides

from 2 and up to 80 base pairs away (Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010; Nevarez et al. 2010).

On a larger scale, local base composition has been shown to play a significant role in the frequency and types of variants that exist in different regions of the genome. The base composition of the genome is defined as the relative proportion of A:T and G:C base pairs. GC content (the proportion of G:C bases) varies substantially across the genome (Lander et al. 2001), and there has been a large amount of work to understand what if any effect this has on the mutation rate. The combined results of the published studies, however, paint an unclear picture of what exactly is occurring in the genome. Many studies, though not all (Cai et al. 2009), show a positive correlation between GC content and both the rate of substitutions between humans and chimpanzee (Smith et al. 2002; Webster et al. 2003; Arndt and Hwa 2005; Duret and Arndt 2008) and diversity observed between humans (Sachidanandam et al. 2001; Hellmann et al. 2005). Individual variant subtypes, such as variants from an A base to a G base (A>G), show different patterns with regard to the local GC content, although there are major inconsistencies from one study to another (Lercher and Hurst 2002a; Lercher et al. 2002; Smith et al. 2002; Webster et al. 2003; Arndt and Hwa 2004; Duret and Arndt 2008).

Recombination rate also appears to influence diversity and substitution rates. Recombination rates vary widely across the genome (Kong et al. 2002) and many studies show a positive correlation between nucleotide diversity and recombination rate in humans (Nachman et al. 1998; Nachman 2001; Lercher and Hurst 2002b; Hellmann et al. 2005; Spencer et al. 2006; Cai et al. 2009; Lohmueller et al. 2011). A positive

correlation is also observed between interspecies divergence and recombination rate (Hellmann et al. 2003; Hellmann et al. 2005; Duret and Arndt 2008; Cai et al. 2009). Three separate theories, with varying degrees of scientific support, have been proposed to explain these findings: mutagenic recombination (Lercher and Hurst 2002b; Hellmann et al. 2003; Hellmann et al. 2005; Hellmann et al. 2008), selective-dependent processes (Charlesworth et al. 1993; Nachman 2001; Begun et al. 2007), and biased gene conversion (Meunier and Duret 2004; Duret and Arndt 2008; Berglund et al. 2009; Duret and Galtier 2009; Galtier et al. 2009). Although selection-dependent mechanisms, such as background selection and selective sweep, and biased gene conversion are largely favored compared to the hypothesis that recombination is mutagenic, none of the current studies have directly assayed the response of the mutation rate to the local recombination rate, but rather infer their findings based on common variant and divergence data.

In addition to the large-scale effects of base composition and recombination rate, several other genomic properties have been studied less extensively to explain the observed variability in the mutation rate. Several studies have shown a strand asymmetry in the frequency and types of mutations that occur in transcribed genes (Green et al. 2003; McVicker and Green 2010). Replication timing has also been suggested as having an impact on mutation rates in humans (Wolfe et al. 1989). Several recent studies report an increase in the neutral substitution rate and intra-species diversity in later-replicating regions of the human genome (Stamatoyannopoulos et al. 2009; Chen et al. 2010; Koren et al. 2012).

1.4 Discovery and Quantification of Mutations in the Genomics Era

The initial sequencing of the human genome in 2001 (Lander et al. 2001), catapulted the field of human genetics into the age of genomics, leading to HT SNP arrays and genome-wide association studies (GWAS) to identify loci for common traits. Later that same decade, advancement of DNA sequencing technologies further advanced the field from a focus on common polymorphic sites to an increasing wealth of information about rare variants, human demographic history, and disease-causing mutations.

In 2009, Ng and colleagues published the first application of HT sequencing to discover mutations in the gene *MYH3* causing the autosomal dominant disorder Friedman-Sheldon syndrome (Ng et al. 2009). This proof-of-concept paper established the use of targeted sequencing approaches to identify disease-causing mutations. In 2010, this same group published a finding of mutations in the gene *DHODH* in four unrelated individuals with the rare autosomal recessive disorder Miller syndrome, for which the underlying mutation was previously unknown (Ng et al. 2010b). As of November 2011, exome sequencing has been used to identify genes for 30 Mendelian disorders (Bamshad et al. 2011). That number has grown even higher since then and will continue to grow as more and more diseases are studied. In addition to these rare disorders, sequencing in more complex traits, for which GWAS and other common variant techniques have been unable to identify variants with large effect sizes, has started to identify potentially pathogenic mutations in autism (O'Roak et al. 2011; Neale et al. 2012; Sanders et al. 2012) and schizophrenia (Xu et al. 2012).

In addition to the advances in human medical genetics through the use of HT sequencing, these same technological advances have also led to discoveries about human evolutionary history and advances in the quantification of the single-nucleotide mutation rate. In 2010, Coventry et al. sequenced the exons of two genes in > 10,000 European individuals and found an excess of rare variants compared to expectations based on common polymorphism data (Coventry et al. 2010). Their results are consistent with the human population experiencing explosive growth sometime in the recent past, leading to an abundance of rare variants segregating in the population (Coventry et al. 2010). Following this initial report, two additional sequencing studies in large cohorts have been published: whole exome sequencing in 2,240 individuals (Tennessen et al. 2012) and sequencing of 202 drug target genes in > 14,000 individuals (Nelson et al. 2012). In addition to the similar finding of a large number of rare variants, these two studies were also able to analyze the degree to which individuals carry deleterious rare variants due to the larger number of loci sequenced in each study. They found that the individuals in their study harbor a large number of deleterious rare variants, without any overt effects on their overall health (Nelson et al. 2012; Tennessen et al. 2012).

HT sequencing has also revolutionized the way in which we can quantify the number of spontaneous mutations arising in the human genome. Sequencing of parent-offspring trios allows one to identify mutations present in the offspring that are not seen in either parent. These *de novo* mutations are important both in human disease and accurately measuring the human mutation rate. Recently, several groups applied this technique to quantify the spontaneous *de novo* mutation rate in humans, reporting an

average per-base per-generation mutation rate of 1.2 x 10^{-8} (Conrad et al. 2010; The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012).

1.5 Understanding the Factors Influencing Single-Nucleotide Mutations and Their Downstream Effects on Human Health using High-Throughput Sequencing

With the increasing availability of HT sequencing, I undertook three separate projects to further understand the role that single-nucleotide mutations play in disease, as well as the innate influence of the genome on the generation and proliferation of single-nucleotide mutations using HT sequencing technology. The application of this new technology allowed me to overcome the technological barrier in disease gene identification and the study of the spontaneous single-nucleotide mutation.

First, I used a combination of whole genome sequencing, whole exome sequencing, and X-chromosome targeted exome sequencing to identify a mutation with strong evidence of causation in a rare Mendelian form of mental retardation and deafness, Martin-Probst Syndrome (OMIM 300159). The technological details regarding the sequencing methodology and analysis are presented in Chapter 2, whereas a more thorough description of the functional analysis was previously published (Bedoyan et al. 2012). This application of HT sequencing techniques allowed me to identify the causative mutation for this disorder, which previous efforts failed to identify (Martin et al. 2000; Probst et al. 2004). The discovery that mutations in *RAB40AL* lead to mental retardation and sensorineural hearing loss (characteristics of Martin-Probst Syndrome) will help us to further understand the function of this gene.

In the next project, I used a unique data set derived from HT sequencing of a large cohort (Nelson et al. 2012) to understand how GC content and recombination rate impact single-nucleotide rare variants, common variants, and substitutions. My application of logistic regression on rare variants is the first study of its kind to assay how genomic context impacts patterns of rare variants, which are the result of recent mutation events and therefore more representative of the mutation rate. While previous studies analyzed the effect of GC content and recombination rate on common variants and substitutions, I observed patterns in rare variants that differed from those in common variants and substitutions. My results suggest that analysis of rare variants more accurately shows the true underlying effect of genomic context on the spontaneous mutation rate, not the effect of later acting evolutionary forces, such as selection and biased gene conversion.

Last, I developed a forward genetic simulation tool, SubSim, to jointly analyze how mutation and fixation bias impact patterns of observed variants in the human genome. SubSim fills a gap in the current selection of simulation algorithms. It has the ability to manipulate mutation bias and selection bias on specific variant subtypes and also alter the GC content of the simulated locus. These new advances open the door for future work to understand how the combined neutral effects of mutation and fixation bias in response to the local GC content and recombination rate can produce patterns of variants currently observed in human populations.

CHAPTER 2

High-Throughput Sequencing Identifies *RAB40AL* as the Gene Underlying Martin-Probst Syndrome

2.1 Introduction

Martin-Probst syndrome (MPS) (OMIM 300159) is an extremely rare genetic disorder. It was first described in 2000 by Martin and colleagues (Martin et al. 2000) in three related males. MPS is primarily characterized by congenital sensorineural hearing loss and mental retardation. Additional MPS phenotypes are variable and include short stature, congenital umbilical hernia, a variety of facial dysmorphisms, and abnormal teeth, among others (Martin et al. 2000). The inheritance pattern of MPS is consistent with an X-linked recessive form of inheritance (Figure 2.1). To date, MPS has only been observed in one family (Martin et al. 2000) with 3 clinically diagnosed male individuals (Figure 2.1).

Initial cytogenetic analysis of two affected males from this pedigree showed normal 46,XY karyotypes for each individual, eliminating any large-scale chromosomal abnormalities (Martin et al. 2000). Finer-scale haplotype mapping identified a shared haplotype on the X chromosome (Martin et al. 2000), including several previously candidate deafness genes: *POU3F4* (de Kok et al. 1995), *TIMM8A* (Jin et al. 1996), *COL4A5* (Jonsson et al. 1998), and *DIAPH2* (Lynch et al. 1997). Sequencing of *POU3F4*, *COL4A5*, and *DIAPH2* did not reveal deletions or point mutations in or around

these genes (Martin et al. 2000). The shared haplotype region was later refined using linkage to a 68 Mb region spanning the microsatellite markers *DXS1003* - *DXS1220* (Probst et al. 2004). This study also analyzed X-inactivation in female carriers in this pedigree. Typically, one X chromosome undergoes random X inactivation in females to compensate for the increased dosage of X chromosome genes. Instead of X-inactivation randomly inactivating one X chromosome, all females that were suspected of being carriers showed complete skewing of X-inactivation to one chromosome and a lack of inactivation on the other (Probst et al. 2004). Skewing of X-inactivation in other X-linked recessive disorders has been previously reported (Belmont 1996; Puck and Willard 1998) and was used to identify the gene *ATRX* in a different mental retardation syndrome (Gibbons et al. 1992). Together, the haplotype and linkage analysis, along with the X-inactivation data, strongly support the hypothesis that MPS is an X-linked recessive condition and further defines the genetic interval to a specific 68 Mb haplotype block on the X chromosome.

Recent advances in sequencing technology now allow for sequencing of whole genomes or specifically targeted genomic regions. In 2009, this technology was first put to use to identify a candidate locus for a Mendelian disease (Ng et al. 2009). Ng and colleagues developed a whole exome sequencing approach in which they specifically targeted the coding regions of genes in twelve individuals and then sequenced these targeted regions using high-throughput sequencing (Ng et al. 2009). In a subsequent publication, this same group was the first to utilize whole exome sequencing to identify the gene responsible for a rare recessive Mendelian disorder, Miller syndrome (Ng et al. 2010b). Using stringent filtering criteria to select for high-quality and potentially

damaging rare variants, this group identified mutations in the gene *DHODH* in two unrelated individuals affected with Miller syndrome (Ng et al. 2010b). To validate their finding, this group also sequenced the gene *DHODH* in additional unrelated affected individuals and found that these patients also carry mutations in this gene (Ng et al. 2010b). Since then, sequencing-based approaches have helped to identify candidate loci for a large number of Mendelian diseases, including deafness (Rehman et al. 2010; De Keulenaer et al. 2012; Diaz-Horta et al. 2012) and X-linked mental retardation (Hu et al. 2009; Tarpey et al. 2009; Jensen et al. 2011).

Here, we used a combination of whole genome sequencing (WGS), whole exome sequencing (WES) and X chromosome-specific exome sequencing (XSS) to identify the gene responsible for MPS. The combination of these three techniques allowed us to increase the amount of quality data we acquired by sequencing more loci than a typical whole-exome approach and increasing the overall depth of sequencing coverage. WGS sequences all bases, including exons and introns. WES and XSS are both target-based approaches, covering only the coding regions of the genome and the X-chromosome, respectively. Due to funding and sample availability, we performed WGS, WES, and XSS on one affected male, whereas only XSS was performed on another affected male individual from the same family. Because MPS is rare and exhibits a clear X-linked inheritance pattern, we hypothesized that with sufficient coverage in the previously defined haplotype region, sequencing would be an appropriate approach to identify a candidate locus.

The results from this study were published in 2012 in the Journal of Medical Genetics (Bedoyan et al. 2012), which focuses primarily on the functional analysis

performed by Dr. Jirair Bedoyan and colleagues in the laboratory of Dr. Donna Martin. In contrast, the work presented here provides a detailed description of the applied sequencing methods, analysis, and variant filtering necessary to identify mutations in the affected individuals.

2.2 Materials and Methods

2.2.1 Whole Genome Sequencing

We performed single-end and paired-end WGS on genomic DNA extracted from peripheral blood leukocytes from individual III-5 (Figure 2.1) across seven lanes on the Illumina Genome Analyzer IIx (Illumina; San Diego, CA). The University of Michigan DNA Sequencing Core (Ann Arbor, MI) performed the sequencing, generating 35, 39, and 79 bp reads.

The reads were aligned to the reference human genome (UCSC hg18) using BWA (Li and Durbin 2009). SAMTools (Li et al. 2009) was used to remove duplicate reads and call single nucleotide variants (SNVs) and indels.

2.2.2 Whole Exome Sequencing

Genomic DNA from peripheral blood leukocytes of individual III-5 (Figure 2.1) was extracted and used to generate a library for whole exome sequencing using the SureSelect^{TM} Human All Exon Kit (Agilent; Santa Clara, CA) based on CCDS2008 for exome target-capture. Weiping Peng in Jun Li's laboratory prepared the sample for sequencing. Paired-end sequencing was performed at Hudson Alpha Institute for Biotechnology (Huntsville, AL) on the Illumina Genome Analyzer II (Illumina; San Diego, CA), generating 75 bp reads.

I aligned the sequence reads to the reference human genome (UCSC hg18) using BWA (Li and Durbin 2009) and used SAMTools (Li et al. 2009) to remove duplicate reads and to call SNVs.

2.2.3 X Chromosome Targeted Resequencing

Samples from two affected males (III-5 and IV-1, Figure 2.1) were targeted using a NimbleGen custom capture array (NimbleGen; Madison, WI) followed by sequencing on the Illumina Genome Analyzer (Illumina; San Diego, CA) by collaborators at Emory University (Atlanta, GA). This generated 76 bp reads for each sample. Michael E. Zwick, along two members of his laboratory, Kajari Mondal and Amol C. Shetty, designed the custom capture array using the Microarray Oligonucleotide Probe Designer (Patel et al. 2010), prepared the samples, and generated the sequence data.

I aligned the sequence reads from both samples to the reference human genome (UCSC hg18) using BWA (Li and Durbin 2009) and used PICARD to remove duplicate sequences (http://picard.sourceforge.net/). I then recalibrated the base call quality scores and called variants using GATK (McKenna et al.).

2.2.4 Variant Filtering and Validation

Details regarding variant filtering are presented in Figure 2.2. In order to enrich the data set for high quality variants, I first removed SNVs with a quality score <30 and <4x depth of sequencing coverage. I filtered out all common polymorphisms present in dbSNP130 (based on hg18) and the 1000 genomes project (March 2010 release). I restricted further analysis to SNVs identified in the haplotype block covering 46,419,359 - 114,514,483 bp (Xp11.3 - Xq23) on chromosome X (Martin et al. 2000; Probst et al.

2004). In order to classify the functional effect of each variant, I used SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/) to annotate variants that met the filtering criteria thus far. Using these results, I included only potentially damaging variants, including missense, nonsense, splice site, and UTR variants. Next, I included only variants that were identified in both affected individuals. These individuals are first cousins once removed and the disease has been inherited across generations. Therefore, MPS is likely not caused by separate *de novo* mutation events in each individual and the two affected individuals should share the causative variant. I used a publically available gene expression database to include only variants identified in genes with known expression patterns during human fetal nervous system development (http://bgee.unil.ch/bgee/bgee). The final step in the variant filtering was to analyze the potential effect on protein function of each variant. I used prediction software to predict the effect of each variant on protein function, including PolyPhen (Ramensky et al. 2002), PolyPhen2 (Adzhubei et al. 2010), MuPro (Cheng et al. 2006), SIFT (Kumar et al. 2009), and AlignGVGD (Mathe et al. 2006; Tavtigian et al. 2006).

Jirair Bedoyan validated missense SNVs that passed the fetal nervous system expression filter. He used PCR followed by Sanger sequencing to validate these variants in both affected individuals. In addition, he tested additional members of this pedigree (Figure 2.1) to ensure that the variants segregated properly and to confirm carrier status of the suspected carrier females.

2.3 Results

In total, we sequenced two affected men from one multi-generation family. We performed WGS, WES, and XSS on individual III-5 (Figure 2.1) and XSS on individual IV-1 (Figure 2.1). Read counts, alignment scores, and duplication rates for each of the sequencing methods used in both individuals are presented in Table 2.1. Each sequencing method was of relatively high quality, as indicated by the high alignment rate of all reads and the low frequency of duplicate reads (Table 2.1).

After performing the initial alignment and quality control filters on the data, SNVs were called using two algorithms (Figure 2.2). In total, over 2 million SNVs were identified in the whole genome sequencing for individual III-5 (Table 2.2). WES and XSS in individual III-5 identified 45,182 and 1,718 SNVs, respectively (Table 2.2). XSS in individual IV-1 identified 1,197 variants (Table 2.2).

Because MPS is a rare genetic disorder, polymorphisms segregating in the general population are unlikely to lead to the disease. Therefore, I removed all SNVs identified in dbSNP130. Although dbSNP130 is a large database with 18,833,531 SNPs, it does not encompass all variants segregating in the population. I used variants identified in the 1000 Genomes Project to further enrich the data for rare variants that are not segregating in presumably healthy individuals (The 1000 Genomes Project Consortium 2010).

Previous work on MPS restricted the genetic interval containing the causative mutation to a 68 Mb window spanning the microsatellite markers DSX1003-DSX1220, which covers 46,419,359 - 114,514,483 bp (Xp11.3-Xq23) on chromosome X (Martin et

al. 2000; Probst et al. 2004). Therefore, I focused all subsequent analysis on variants identified in this locus.

SNV identification in high-throughput sequencing studies is often error-prone, but results tend to be more accurate with higher sequencing coverage (Li et al. 2011). For individual III-5, we obtained an average of 4.79x, 10.86x, and 86.89x depth of sequencing coverage in the exons for WGS, WES, and XSS, respectively, in the exons present in the haplotype block (Table 2.1). For individual III-5, this combined analysis covered 98.4% of all coding regions in the 68 Mb haplotype region with ≥ 4x average coverage (Figure 2.3). The average depth of coverage for individual IV-1 was 17.89x (Table 2.1). This extensive coverage of the haplotype region allowed me to generate accurate and reliable SNV calls.

The next step in the variant filtering process is to enrich for possibly damaging variants that could disrupt the function of the gene or resulting protein. To do this, I first analyzed all missense and nonsense variants, as well as variants disrupting a splice site or located in the untranslated region (UTR) of a gene. Nonsense and splice site variants have clear implications on protein function, by prematurely truncating the protein or disrupting the order of exons and introns, respectively. UTR variants have the potential to disrupt gene expression by altering transcription factor binding sites. Missense variants can alter protein function by changing the amino acid sequence of the protein. I identified a total of 50 nonsense, splice sites, UTR, or missense variants in individual III-5 (combining across WGS, WES, and XSS) and 18 in individual IV-1.

As the two sequenced individuals are affected by the same X-linked disease and are related on the maternal lineage (they are first cousins, once removed), they must have inherited the same mutation. Therefore, the next filtering step I employed was to only analyze variants observed in both individuals. There were a total of 15 variants shared across the two individuals (Table 2.3).

MPS is a neurological developmental disorder that presents early in life. Therefore, we hypothesized that the causative gene must be expressed during the development of the fetal nervous system. Using a publically available database of gene expression data, I filtered all genes that met the stringent criteria to those that were listed as having fetal nervous system expression. This particular analysis was performed on genes, whereas prior-filtering steps focused on each variant independently, without regard for the gene. Of the eleven genes with at least one identified SNV, I identified six genes that showed fetal nervous system expression (Table 2.2).

Although advances in alignment and base calling algorithms continue to improve the reliability of sequencing data, a separate validation step is required to ensure that the variants identified are true positives and not the result of sequencing artifacts or other technical issues. To validate the variants that met the previous filtering requirements, Jirair Bedoyan performed PCR followed by Sanger sequencing on the identified missense variants. Three variants identified in two genes were successfully validated.

Not all missense mutations are deleterious to protein function. Some result in very similar amino acid substitutions, such as by those with similar sizes, structures, or hydrophilic or hydrophobic properties. There are a variety of available software packages available that predict the effect of missense variants on protein function using evolutionary conservation and other amino acid properties. I used PolyPhen (Ramensky et al. 2002), PolyPhen2 (Adzhubei et al.), MuPro (Cheng et al. 2006), SIFT (Kumar et al. 2009), and AlignGVGD (Mathe et al. 2006; Tavtigian et al. 2006) to analyze the potential impact of each missense mutation on protein function. Variants in the genes *ARHGEF9* and *RAB40AL* passed the prior filtering criteria and both contained missense mutations. Of these, the two variants at 102,079,078 and 102,079,079 in the gene *RAB40AL* were predicted to damage protein function.

The two mutations identified in *RAB40AL* are adjacent and lie within the same codon. They are both relatively well conserved on the nucleotide level, according to a GERP score (Cooper et al. 2005) (range: -11.6 to 5.82, 5.28 most conserved) of 0.77 and -0.99 and PhastCons scores (Siepel et al. 2005) (range: 0 to 1, 1 most conserved) of 0.78 and 0.46 for the 102,079,078 and 102,079,079 variants, respectively. Together, when both nucleotides are mutated, the resulting change to the DNA sequence is 102,079,078 – 102,079,079 AC → GC and the protein sequence is changed from an aspartic acid to a glycine at amino acid 59 (pD59G). The pD59G alteration to the amino acid sequence was predicted to be damaging to the function of RAB40AL using all of the above prediction software. Additionally, this amino acid is evolutionarily conserved from humans to invertebrates, indicating that it is extremely important for normal protein function (Figure 2.4).

I analyzed the presence of these variants in data from the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/). This database contains WES data from 3,510 European samples, and therefore can be used to remove potential non-disease causing polymorphisms from the analysis. Both variants that I identified in *RAB40AL* were not identified in the NHLBI database. Similarly to the 1000 Genomes Project analysis described above, the data for this analysis were acquired prior to completion of the project and a paper has recently been published describing their results (The 1000 Genomes Project Consortium 2010; Tennessen et al. 2012).

2.4 Discussion

We used a combination of WGS, WES, and XSS to identify a mutation in the gene *RAB40AL*, which is the likely cause of MPS. This unique combination of these three different sequencing methodologies resulted in high-quality data, covering more loci at a higher depth of sequencing coverage than any of them alone. In addition, the specific filtering criteria applied here, especially the previously identified haplotype block, the use of a database detailing genes expressed during fetal nervous system development, and the sequencing of two affected individuals, narrowed the large number of potential variants into a small handful.

Depth of coverage and the number of loci sequenced are important factors in any sequencing study. After the sequenced reads are aligned to the reference genome, sites where one or more alleles do not match the reference allele are called as variants by the variant calling algorithm (here, GATK or SAMTools). Sequencing errors can disrupt the calling algorithms, appearing as variant alleles. If there is insufficient

coverage at a locus, that site may be incorrectly called as a variant due to a single sequencing error in a single sequencing read. However, increasing the sequencing coverage increases the accuracy of calling a site as variant and decreases the rate of false positive variant calls, especially with low sample sizes (Li et al. 2011). The high depth of coverage at the majority of the loci analyzed ensures that the variant calls made in this study are highly accurate.

The different sequencing methods applied here target different regions of the genome. WGS is not a target-based sequencing approach. Instead, all bases in the genome are sequenced, although typically only low depth of coverage is feasible for WGS approaches. WES and XSS, on the other hand, are targeted sequencing methods. WES generates sequence of all coding regions in the genome, and the XSS targeted coding regions throughout the X chromosome. The WES and XSS techniques do not cover the exact same exons on the X chromosome. The XSS technique was developed to specifically target the entire X chromosome coding region and targets a larger proportion of the X chromosome than the WES. The use of both of these target-based approaches enabled us to sequence a larger number of exons to relatively high depth. Combining these two approaches with WGS gave a small amount of coverage on intronic and other noncoding regions of the X chromosome, where other causative mutations may reside.

RAB40AL is a member of the Rab40 family of small GTP-binding proteins. An inversion of the X chromosome disrupting the promoter region of *RAB40AL* was previously identified in a male individual diagnosed with Duchenne muscular dystrophy, who exhibited mental retardation, athetosis, nystagmus, and severe congenital

hypotonia (Saito-Ohara et al. 2002). Other Rab small GTPase proteins, along with proteins functioning in the same pathway as Rabs, have been associated with a number of disorders, including mental retardation, indicating that disruptions in the Rab pathway impact human health (Menasche et al. 2000; Seabra et al. 2002; Giannandrea et al. 2010).

Functional work by Jirair Bedoyan confirmed that *RAB40AL* is expressed in human brain tissue. Furthermore, he found that cells transfected with the p.D59G mutated form of *RAB40AL* showed decreased protein expression in a Western blot. Little is known regarding the function of *RAB40AL*. The Rab superfamily of proteins is involved in vesicular transport in the cell, trafficking organelles and intracellular vesicles to the extracellular membrane (Pereira-Leal and Seabra 2001). Different Rabs target different organelles, exhibiting specific subcellular localization patterns. *RAB40AL*, in particular, localizes to the mitochondria (Saito-Ohara et al. 2002). Dr. Bedoyan also demonstrated that the p.D59G mutation disrupts the subcellular localization of *RAB40AL*. While wild-type *RAB40AL* transfected in COS7 cells localized to the mitochondria, as expected, the mutated *RAB40AL* clustered in the nucleus, nucleolus, and/or perinuclear region of the cell. Together, this functional work provides additional evidence that the mutation we observed in the family with MPS leads to disruptions in the normal function of *RAB40AL*.

One difficulty that arose during this study was the lack of available patient samples to analyze. This disorder is extremely rare; only one family has been described exhibiting the unique combination of phenotypes that characterize MPS. Even though we sequenced two affected individuals, filtering for shared variants is somewhat

redundant since we already know they share a large region based on previous linkage and haplotype analysis (Martin et al. 2000; Probst et al. 2004). Without additional unrelated affected individuals, it is difficult to conclusively demonstrate that mutations in *RAB40AL* lead to MPS. The absence of any mutations in *RAB40AL* in the 1000 Genomes Project or the NHBLI Exome Sequencing Project, along with additional work performed by Jirair Bedoyan showing a lack of mutations in *RAB40AL* in 297 neurologically normal individuals (obtained from the Greenwood Genetic Center), demonstrate that mutations in *RAB40AL* are extremely rare and not present in individuals with normal hearing and cognitive function.

2.5 Conclusions

In conclusion, I showed that WGS, WES, and XSS can be effectively combined to identify causative mutations for rare Mendelian disorders. I found two mutations in the gene *RAB40AL* that segregate with MPS in the one family diagnosed with this disorder. The combination of these mutations is predicted to lead to damaging effects on protein function. There is growing evidence that humans can tolerate a large number of what appear to be severely damaging mutations without any noticeable effect on their health (Nelson et al. 2012; Tennessen et al. 2012). However, functional studies performed using the mutation we identified in these patients show a clear disruption in normal protein function. Overall, our work identified a likely candidate gene for sensorineural hearing loss and mental retardation in patients with MPS. Further work examining the role that *RAB40AL* plays in normal cognition and hearing will help us understand how mutations in *RAB40AL* lead to disease and potentially identify treatment options for these patients.

2.6 Figures

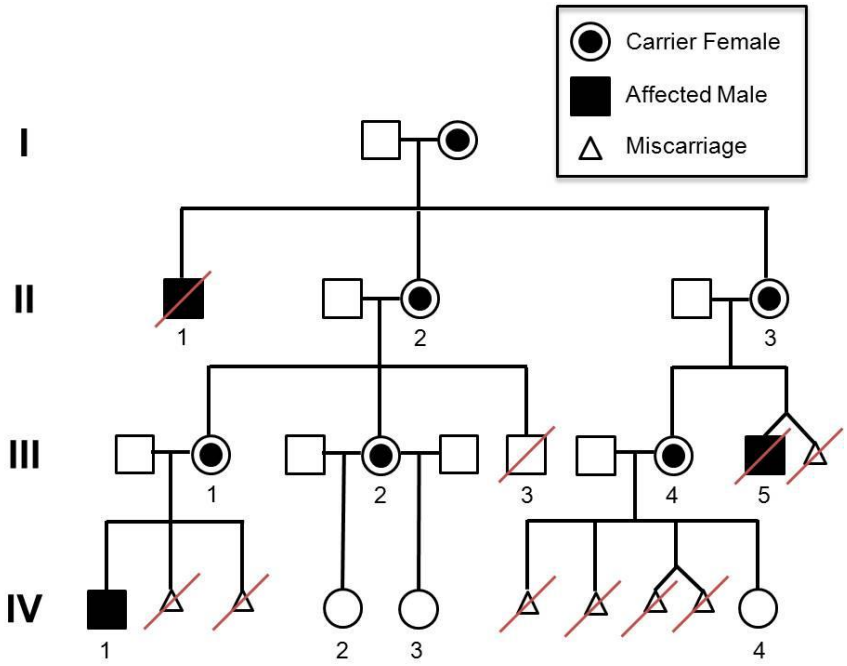

Figure 2.1: Pedigree for family affected with MPS

Affected individuals are shown in black and carrier females are indicated with a small black circle. Individual III-5 was sequenced using whole genome, whole exome and X-specific targeting platforms and individual IV-1 was sequenced using the X-specific sequencing. Verification of carrier status and validation of the p.D59G variant in RAB40AL was confirmed in all carrier females and affected males. The status of individuals IV-2, IV-3, and IV-4 was not determined. The red line through a symbol indicates a deceased individual.

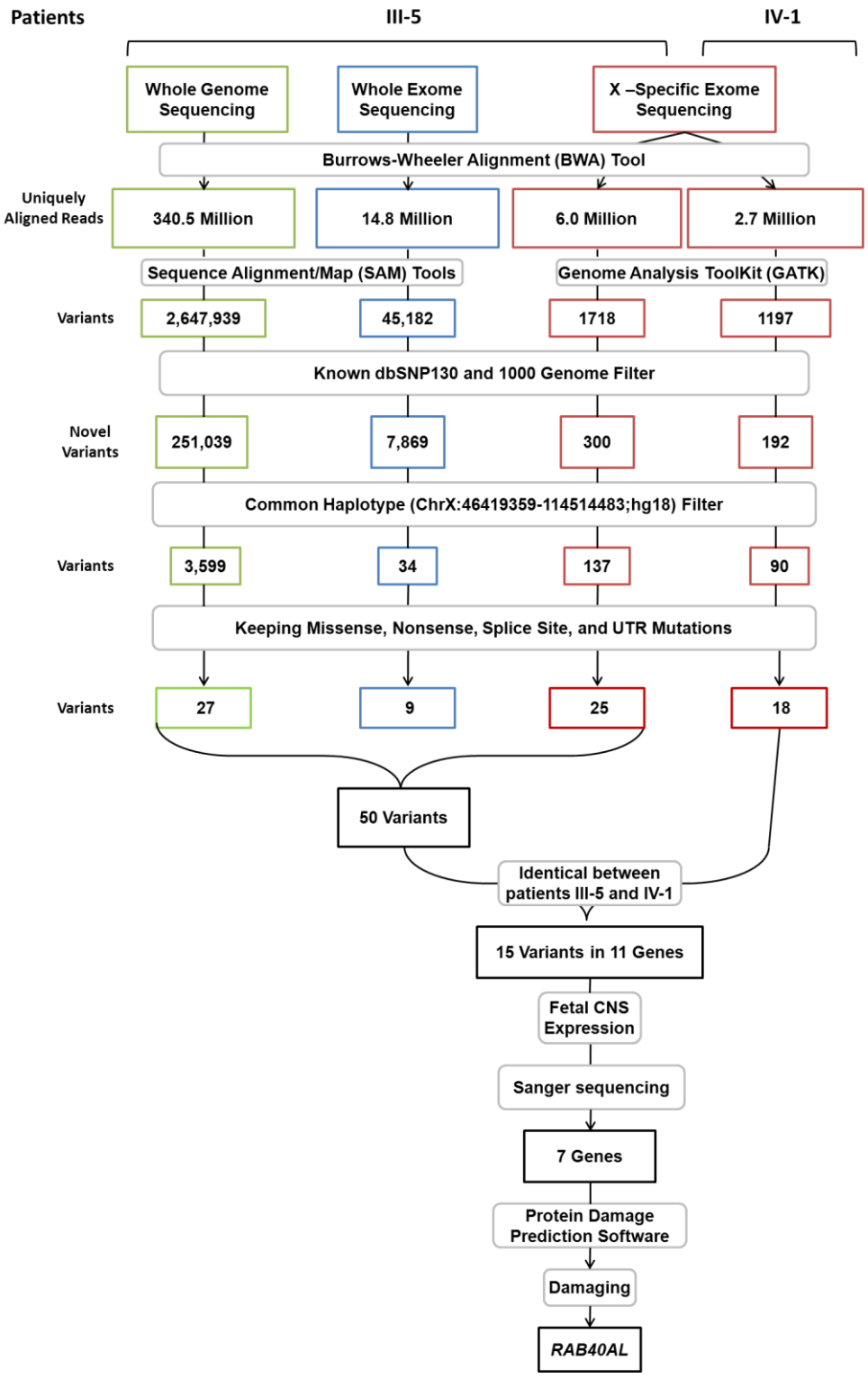

Figure 2.2: Data analysis pipeline for MPS WGS, WES, and XSS sequencing on individuals III-5 and IV-1

Sequence reads for WGS, WES and XSS were aligned using BWA (Li and Durbin 2009) and variants called using SAMTools (Li et al. 2009) or GATK . (McKenna et al.). Following removal of common polymorphisms identified in dbSNP or the 1000 Genomes Project, analysis focused on variants located in the previously identified haplotype block on the X-chromosome. Potentially functional variants that were shared between the two affected men were kept and validated. After filtering, we identified mutations in the gene *RAB40AL* as likely candidates for MPS.

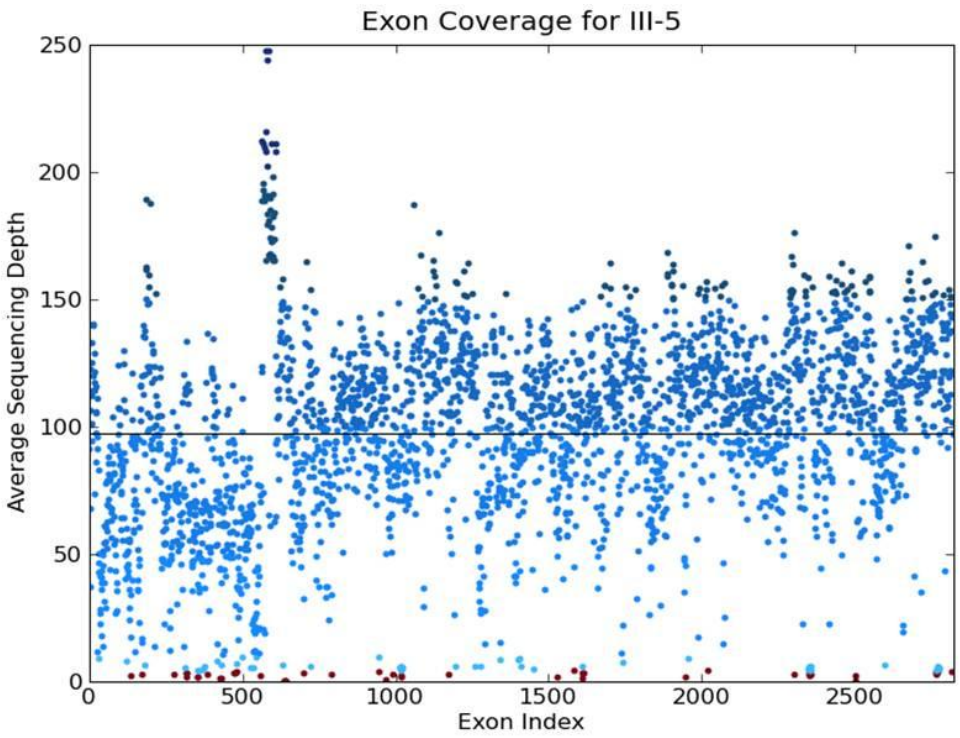

Figure 2.3: Extensive coverage of exons in the haplotype block for individual III-5

Sequencing read coverage depth plotted for each indexed exon in the haplotype block. Coverage depth was calculated as the average coverage across all exonic positions after combining reads from whole genome, whole exome, and X-specific targeted sequencing. Red points indicate those exons where the average coverage was less than 4x, the minimum sequencing depth required in my pipeline to call variants. The black line indicates the combined average coverage depth across all exons in the combined analysis, 97.55x

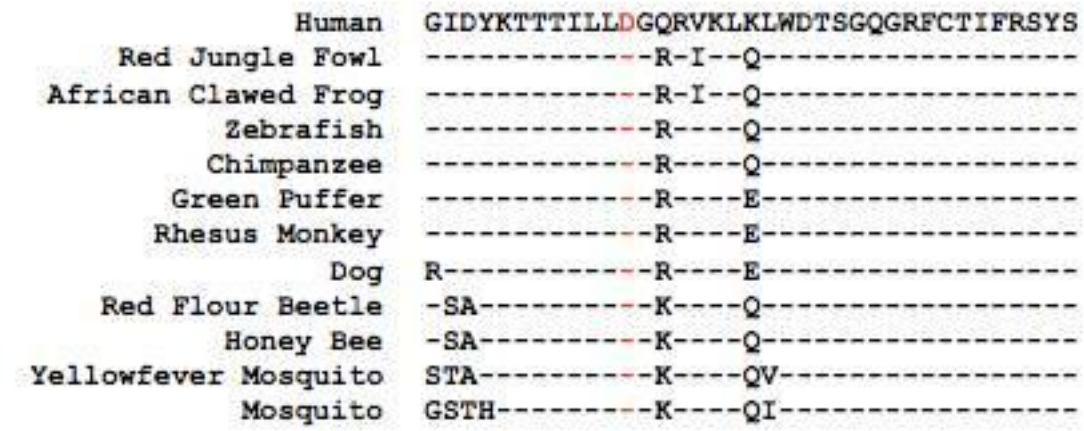

Figure 2.4: Evolutionary conservation of the p.D59G variant

The aspartic acid at amino acid 59 is conserved from humans through invertebrates, indicating that it is likely important for protein function.

2.7 Tables

Individual	Platform	Total Reads	Aligned Reads	Percent Aligned	Uniquely Aligned Reads	Percent Unique	Average Coverage
III-5	Whole Genome Sequencing	416,420,317	342,153,138	82.17%	340,502,582	99.52%	4.79
	Whole Exome Sequencing	17,612,511	17,366,720	98.60%	14,887,848	85.73%	10.86
	X-Exome Sequencing	25,355,543	23,869,724	94.14%	17,846,447	74.77%	86.89
IV-1	X-Exome Sequencing	19,944,310	17,368,317	87.08%	14,601,259	84.07%	17.89

Table 2.1: Alignment summary for whole genome, whole exome, and X-exome sequencing.

The total number of sequencing reads, total number of reads that aligned to one position in the genome, and the number of unique reads are shown here. The total number of unique reads describes the number of sequences where only one read mapped uniquely to a given locus. The average coverage is the average coverage of each sequencing methodology in the exons of the haplotype region.

Description of Variant Filter	III-5			IV-1	Intersect
	Whole Genome	Whole Exome	X Exome	X Exome	
Total Variants	2,647,939	45,182	1,718	1,197	--
Novel Variants	251,039	7,869	300	192	--
Variants in Haplotype Block	3,599	34	137	90	--
NS/MS/SS/UTR	27	9	25	18	--
Shared Variants	--	--	--	--	15
Fetal Nervous System Expression	--	--	--	--	6
Confirmed by Sanger Sequencing	--	--	--	--	2
Predicted Damaging					1

Table 2.2: Summary of variant filtering across whole genome, whole exome, and X-exome sequencing for individuals III-5 and IV-1.

The number of variants identified across the two individuals and across the three sequencing techniques used at each of the variant filtering steps. Nonsense (NS), missense (MS), splice site (SS), or untranslated region (UTR).

Chromosome	Position	Reference Allele	Alternative Allele	Functional Annotation	Amino Acid Position	Reference Amino Acid	Alternative Amino Acid	Gene
X	48128980	A	G	Missense	35	LYS	GLU	*SSX4*
X	48715512	G	A	3' UTR	--	--	--	*GRIPAP1*
X	49744485	A	G	3' UTR	--	--	--	*CLCN5*
X	55667068	C	G	Missense	67	PRO	ALA	*FOXR2*
X	55667843	C	G	3' UTR	--	--	--	*FOXR2*
X	57036943	A	C	3' UTR	--	--	--	*SPIN3*
X	62774487	C	G	3' UTR	--	--	--	*ARHGEF9*
X	62810649	G	C	Missense	306	ASP	GLU	*ARHGEF9*
X	63326117	G	A	3' UTR	--	--	--	*FAM123B*
X	66681507	C	A	5' UTR	--	--	--	*AR*
X	73873618	T	G	3' UTR	--	--	--	*KIAA2022*
X	73873874	C	T	3' UTR	--	--	--	*KIAA2022*
X	102079078	A	G	Missense	59	ASP	GLY	*RAB40AL*
X	102079079	C	A	Missense	59	ASP	GLU	*RAB40AL*
X	111211844	G	A	5' UTR	--	--	--	*TRPC5*

Table 2.3: Variants identified in individual III-5 and IV-1.

Variants identified after initial variant filtering observed in both affected sequenced individuals. I identified 15 variants in 11 genes that could potentially impact protein function (splice site, missense, nonsense, or UTR variants).

CHAPTER 3

The Influence of Genomic Context on Mutation and Fixation Patterns in the Human Genome Inferred from Rare Variants, Common Variants, and Substitutions[1]

3.1 Introduction

Mutation is one of the most fundamental processes in biology. It is the ultimate source of genetic variation and one of the driving forces of evolution. Mutation also plays a significant role in the etiology of human diseases. As such, there is considerable interest in understanding the underlying pattern and molecular spectrum of spontaneous mutations. Historically, two approaches were developed to estimate the single-nucleotide mutation rate in humans. The first analyzes divergent sites between humans and an ancestral species, typically chimpanzee. According to Kimura's neutral theory, the majority of substitutions are neutral and therefore the extent of between-species divergence can be used to estimate the neutral mutation rate (Kimura 1983). Many groups have applied this approach to estimate the spontaneous human mutation rate (Drake et al. 1998; Nachman and Crowell 2000; Kumar and Subramanian 2002; Silva and Kondrashov 2002). However, several forces, including natural selection, biased

[1] This work is currently under revision as: Valerie M. Schaibley, Matthew Zawistowski, Daniel Wegmann, Margaret G. Ehm, Matthew R. Nelson, Pamela L. St. Jean, Goncalo Abeçasis, John Novembre, Sebastian Zöllner, Jun Z. Li. 2013.
Rare Variants Reveal Patterns of Mutation in the Human Genome. *Genome Research*

gene conversion, and demographic history can alter fixation probabilities and reshape the spectrum and genomic distribution of between-species substitution patterns. A second, more direct approach, pioneered by Haldane (1935), uses incidence rates of dominant disorders in humans to estimate the mutation rate (Sommer 1995; Sommer and Ketterling 1996; Kondrashov 2003; Lynch 2010), although, this approach is limited by the fact that only a small subset of new mutations manifest as disease variants (Nachman 2004).

The mutation rates from these studies represent a genome-wide average. However, there is extensive variability in between-species divergence and within-species diversity among individual genes and genomic regions (Wolfe et al. 1989; Nachman and Crowell 2000; Sachidanandam et al. 2001; Smith and Lercher 2002; Kondrashov 2003; Hodgkinson et al. 2009). This suggests that spontaneous mutation rates are not constant throughout the genome, although the forces leading to this variability are unclear.

Local base composition has been shown to play a significant role in the frequency and types of variants that exist in any given region of the genome. The base composition is defined as the relative proportion of A/T and G/C base pairs. GC content (the proportion of G/C bases) varies substantially across the genome (Lander et al. 2001). There is a positive correlation between GC content and both the rate of substitutions between humans and closely related species, such as chimpanzee, (substitution rate) (Smith et al. 2002; Webster et al. 2003; Arndt and Hwa 2005; Duret and Arndt 2008) and diversity observed between humans (Sachidanandam et al. 2001; Hellmann et al. 2005).

Overall, there are twelve types of single-nucleotide mutations that can occur. Most of these alter the base composition of the site. For, example, a mutation from an A to G (A>G) would slightly increase the local GC content. An A>T mutation, however, maintains the same relative number of A/T and G/C bases and therefore has no effect on the GC or AT content of that region. Studies that analyze the correlation between these specific variant subtypes and the surrounding base composition support the hypothesis that a compositional stabilizing process occurs in the genome, maintaining regions of high and low GC content. For example, several studies report a negative correlation between the GC>AT substitution rate (G>A and C>T variants) and GC content (Arndt and Hwa 2004; Duret and Arndt 2008). Similarly, AT>GC SNPs segregate at a higher frequency in regions of the genome with high GC content (Webster et al. 2003) and there is an increased fixation bias toward GC bases in GC-rich regions (Lercher and Hurst 2002a; Lercher et al. 2002)**.** Together, these studies would suggest that GC-rich regions maintain their higher GC content through decreasing the rate of mutations that lower GC content (GC>AT and GC>TA mutations, jointly referred to as S>W) and increase the rate of fixation of GC-increasing variants (AT>GC and AT>CG, jointly referred to as W>S). However, analysis by Smith (Smith et al. 2002) and Webster (Webster et al. 2003) show that divergence rates for GC>AT substitutions increase with GC content. Over time, an excess of these variants would decrease GC content, bringing it closer to the genome-wide average of ~41% (Lander et al. 2001). Yet another study reported a contradicting finding, with a negative correlation between GC content and both intra-species diversity and inter-species divergence (Cai et al. 2009). These contradictory studies suggest that GC content does

influence the dynamics of mutations emerging in humans and also variants segregating in the population. Although these contradictory results could be due in large part to different studies focusing on a variety of datasets with differing allele frequencies, the true influence of GC content on variant patterns is still unclear.

Recombination rate has also been shown to influence diversity and substitution rates. Recombination rates vary widely across the genome (Kong et al. 2002) and many studies show a positive correlation between nucleotide diversity and recombination rate in humans (Nachman et al. 1998; Nachman 2001; Lercher and Hurst 2002b; Hellmann et al. 2005; Spencer et al. 2006; Cai et al. 2009; Lohmueller et al. 2011). A positive correlation is also observed between interspecies divergence and recombination rates (Hellmann et al. 2003; Hellmann et al. 2005; Duret and Arndt 2008; Cai et al. 2009). Three separate theories have been proposed to explain these findings: mutagenic recombination, selective-dependent processes, and biased gene conversion (BGC).

The hypothesis that recombination is directly mutagenic, leading to increased mutation rate in regions of high recombination and thus higher diversity, was initially proposed based on mutation studies in yeast (Magni and Von Borstel 1962; Magni 1964; Esposito and Bruschi 1993). The same conclusion was later reached in studies of human polymorphism, which observed a positive correlation between interspecies diversity and recombination rate (Lercher and Hurst 2002b; Hellmann et al. 2003; Hellmann et al. 2005; Hellmann et al. 2008). If recombination is inherently mutagenic, the same positive correlation that is seen in diversity levels should also be observed in intra-species divergence. However, there are conflicting reports in the literature, with some studies observing no correlation between recombination rate and divergence

(Begun and Aquadro 1992; Nachman 2001) and others reporting increased divergence in regions of high recombination rate (Hellmann et al. 2003; Hellmann et al. 2005; Duret and Arndt 2008; Hellmann et al. 2008; Cai et al. 2009).

The second theory proposed to explain the positive correlation between recombination and diversity deals with how natural selection alters patterns of variation in the human genome. Recombination generates new haplotypes and shuffles variants onto different backgrounds. This means that in regions of the genome where recombination rate is relatively high, there will be an increased number of distinct haplotypes. When natural selection acts on a beneficial or deleterious variant, it does not simply change the frequency of that specific locus. Rather, selection has a much longer ranging effect, changing the frequency of the entire haplotype in which a variant resides. Recombination diminishes the diversity-reducing effects of background selection, which uses purifying selection against deleterious variants, and selective sweep, which involves positive selection favoring beneficial variants (Smith and Haigh 1974; Kaplan et al. 1989; Charlesworth et al. 1993; Charlesworth et al. 1995; Hudson and Kaplan 1995; Nachman 2001). For example, background selection removes entire haplotypes containing a deleterious variant from the population. Recombination, however, increases the number of distinct haplotypes segregating in a population, and at the same time reduces the average length of the haplotypes. Therefore, if background selection occurs in the presence of high recombination, only haplotypes with the deleterious variant will be removed, while those without the deleterious variant would be kept. Recombination, therefore, promotes the maintenance of genetic diversity in the presence of selection. Studies in both humans and *Drosophila* have concluded

that background selection or genetic hitchhiking primarily drives the observed correlation between diversity and recombination (Begun et al. 2007; Cai et al. 2009; Keinan and Reich 2010; Lohmueller et al. 2011).

The third proposed mechanism leading to the positive correlation between recombination rate and genetic diversity is biased gene conversion (BGC). BGC is a recombination-associated process that preferentially converts base pairs at GC/AT mismatched sites generated during recombination into GC, leading to preferential fixation of GC alleles (Duret and Galtier 2009). Over time, the observed effect of BGC can mimic that of natural selection, leading to an excess of weak (W) A or T ancestral bases converted to strong (S) G or C base as if the latter were under positive selection (Berglund et al. 2009; Galtier et al. 2009). A major difference, however, between BGC and selection-dependent effects is that BGC acts on one nucleotide, without an effect on the surrounding loci. Background selection and selective sweeps, although they are driven by selection on one specific locus, are effects that alter the entire haplotype instead of just the single nucleotide.

These previous studies use common variants within humans and substitutions between humans and chimpanzees to model the effect of GC content and recombination rate on the mutation rate. However, these older forms of variation are effectively mutations accumulated over many generations. Their patterns, therefore, reflect the cumulative influence of many processes, including natural selection, population demographic history and BGC. A major challenge in the field is to elucidate the extent to which these forces have altered the distribution of variants over time and to distinguish their relative contributions. To minimize the effects of selection, many

studies restrict their analysis to non-coding regions of the genome. However, widespread signatures of recent positive selection, even within supposedly neutral regions (Williamson et al. 2007), suggest that even non-coding regions may also be influenced by selection.

Rare variants represent a newly available and expanding resource that can overcome some of these limitations. Rare variants are the result of recent mutation events and are relatively young compared to variants segregating at higher frequencies. Therefore, rare variants are typically less affected by population demographic history or natural selection (Messer 2009). Furthermore, BGC acts only on variants after they have arisen in the population (Duret and Galtier 2009), and does not influence innate mutation rates. As such, rare variants are an appropriate resource for studying the spectrum and genomic distribution of mutations while minimizing these potentially confounding influences.

While family-based whole-genome sequencing has begun to identify *de novo* mutations that provide more direct measures of mutation rates (Conrad et al. 2010; The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012), the mutations identified sparsely cover the genome. For example, if whole genome sequencing of each parent-offspring trio yields ~40 *de novo* mutations (Conrad et al. 2010), 500 such trios would need to be sequenced in order to accumulate roughly 20,000 mutations. These mutations, however, would occur roughly once per 150 kb, and the data would lack the spatial resolution necessary to detect the effect of local genomic context on a finer scale.

For this project, I wanted to understand the effect that GC content and recombination rate has on the spontaneous single-nucleotide mutation rate as well as fixation patterns of variants in the human genome. We hypothesized that rare variants would more accurately represent the underlying effects of GC content and recombination rate on mutation patterns in humans than common variants or substitutions. I studied a set of rare variants discovered using targeted re-sequencing of 202 genes in > 14,000 unrelated individuals. The 202 genes are drug targets relevant in 12 complex diseases. The 14,002 subjects were recruited for genetic association studies of these diseases (Nelson et al. 2012). I analyzed both the per-gene mutation rate as well as the probability of each site to contain a variant of a specific subtype relative to local GC content, recombination rate, and distance to recombination hotspot. In order to compare mutation rate inferences based on rare variants to those obtained by within- and between-species data, I compared rare variant patterns to common variant data from the 1000 Genomes Project and substitution sites between humans and chimpanzee. These three variant classes cover different evolutionary time scales, and the differences between them allow me to examine the distinct influence of genomic context on the initial mutation process, the subsequent rise of some mutations to become common variants, and eventual fixation.

3.2 Methods

3.2.1 Ethics Statement

All study participants in the component studies provided written informed consent for the use of their DNA in genetic studies. A careful review was conducted to verify that

the consents were consistent with the activities of this study. In selected instances further Institutional Review Board approval was sought and obtained where the appropriateness of the informed consent for the current study was not clear.

3.2.2 Data Sources and Processing

3.2.2.1 Rare Variants

I utilized single-nucleotide variants previously described in Nelson et al. (Nelson et al. 2012), which can be accessed at http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1056695. The variants were discovered from a targeted resequencing study of the exons of 202 potential drug target genes (including 50 bp flanking each exon). For this study, I analyzed 195 autosomal genes, and focused on variants identified in individuals of European descent (N = 12,515). I oriented all variants along the human ancestral lineage, as defined by the 1000 Genomes Project alignments (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_ancestor_GRCh37_e59.tar.bz2) (date accessed: January, 3, 2012) and defined rare variants as those with a derived allele frequency (DAF) $\leq 10^{-4}$. To minimize the potential confounding effects due to coverage and to enrich for high-quality variants, I selected variant and invariant sites with ≥ 10x coverage using per-site coverage data from a random sample of 500 individuals reported by Nelson et al. (2012).

I subdivided variants into 7 distinct subtypes based on the ancestral and derived alleles: AT>GC, GC>AT (non CpG), CpG GC>AT, AT>CG, GC>TA, AT>TA and GC>CG. GC>AT transitions that occurred at a CpG site (CpG GC>AT) (based on

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_ancestor_GRCh37_e59.tar.bz2) were analyzed separately because hypermethylation of the cytosine base at CpG dinucleotides leads to spontaneous deamination, resulting in C>T and G>A transitions that occur with substantially higher rates than other subtypes (Nachman and Crowell 2000; Kondrashov 2003; Hwang and Green 2004). In addition, previous studies found that substitution rates at CpG dinucleotides have a much stronger negative correlation with GC content and recombination rate than non-CpG-induced GC>AT transitions, suggesting that different molecular mechanisms may be involved (Arndt et al. 2005; Duret and Arndt 2008). The relationship between genomic context and GC>TA and GC>CG variants at CpG sites (which make up the eighth and ninth variant subtypes) was analyzed separately from non CpG-induced GC>TA and GC>CG variants. They were modeled in the multinomial logistic regressions with CpG as the ancestral base (see Section 3.2.3: Logistic Regression Analysis). As there were relatively few observed variants of these subtypes in my dataset (~200 each), it is difficult to accurately analyze mutation patterns and, therefore, I did not report these results. These variants, however, are included in the GC>TA and GC>CG variant subtypes presented in Table 3.1, and included when analyzing all variants subtypes combined.

3.2.2.2 Per-Gene Mutation Rates and Genomic Context

I analyzed mutation rates that were calculated for 193 of the 195 autosomal genes (two genes were excluded due to low numbers of variants), as described previously (Nelson et al. 2012). For each of the 193 genes, I calculated average GC content and sex-averaged pedigree-based recombination rates (Kong et al. 2002) within

the transcribed region of each gene based on definitions in RefGene. Linear regression was performed in R (R Development Core Team 2008).

3.2.2.3 Sampling of Intergenic Regions to Obtain Common Variants and Substitutions

To sample common variants and substitutions from random genomic intervals with the least selective pressure, I first defined intergenic regions by masking all genic regions ± 1 kb of the transcription start and end site of any gene based on RefGene in hg18. I then removed all regions that were not uniquely aligned in the 4-way Enredo, Pecan, Ortheus (EPO) alignments between human, chimpanzee and orangutan (ftp://ftp.ensembl.org/pub/release-54/emf/ensembl-compara/epo_4_catarrhini/) (date accessed: December 7, 2011). To match the distribution of genomic features with the rare variant data as closely as possible, I sampled 32,279 autosomal regions from all possible regions according to their genomic parameters. Specifically, I matched the size distribution as well as the joint distribution of GC content and recombination rate (Kong et al. 2002) of the selected regions to those of the target regions in the exome sequencing of the 202 genes. I used these regions to sample common variants. Because there were substantially more substitutions in these regions than common variants, I randomly subsampled 12,034 of the 32,279 regions to obtain substitutions. The median (± standard deviation) of GC content across the assayed regions was 0.493 ± 0.123, 0.469 ± 0.112, and 0.471 ± 0.116 for rare variants, common variants, and substitutions, respectively. The median and standard deviation for recombination rate (log transformed, in unit of cM/Mb) was 0.292 ± 0.167, 0.291 ± 0.165, 0.291 ± 0.166 for rare variants, common variants, and substitutions, respectively.

3.2.2.4 Common Variant Data

Single nucleotide variants from the interim phase 1 haplotype data from the 1000 Genomes Project were used to assemble a dataset of common variants. The frequency file for the European subset (N = 381) of the data was downloaded from (http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-PhaseI-Interim.html) (date accessed: December 20, 2011). All variants within the selected regions were oriented ancestral to derived, as described above. Successfully oriented variants with DAF > 0.05 were categorized into the seven variant subtypes and they form the common variant dataset.

3.2.2.5 Substitution Data

Substitutions between human and chimpanzee were obtained using the 4-way EPO alignments between human, chimpanzee, orangutan, and rhesus macaque (ftp://ftp.ensembl.org/pub/release-54/emf/ensembl-compara/epo_4_catarrhini/) (date accessed: December 7, 2011). To identify substitutions, only regions where there was a unique human, chimp, and orangutan alignment were used. Single-base human-chimpanzee differences were sampled from the 12,034 intergenic regions as described above. All sites were oriented along the ancestral lineage and categorized into the seven variant subtypes. Variant sites where the human lineage base represents the ancestral allele were excluded.

3.2.2.6 ESP Rare Variants

Variants from the NHLBI Exome Sequencing Project (ESP) from 5,400 individuals were downloaded from the Exome Variant Server (Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL:

http://evs.gs.washington.edu/EVS/) [date accessed: December 2, 2011]). I also utilized sequence coverage data downloaded from the Exome Variant Server (date accessed: December 2, 2011 and December 5, 2011) to restrict to sites with ≥ 10x coverage. Subsequent analysis focused on singleton variants (DAF = 1.4×10^{-4}) identified in Europeans (N = 3,510). Variants were oriented along the ancestral allele, as before.

3.2.3 Logistic Regression Analysis

I used a logistic regression framework to model the effect of GC content, recombination rate, and the absolute distance to the nearest recombination hotspot (DTH) on the occurrence of rare variants, common variants, and substitutions. I defined GC content at a given site as the percentage of GC bases in a 1 kb window surrounding the site (500 bp upstream, 500 bp downstream) based on the human genome reference sequence (hg18). Sex-averaged recombination rates from the deCODE project (Kong et al. 2002) were averaged in the same 1 kb window. The absolute distance to the center of the nearest recombination hotspot was calculated for each site using recombination hotspot coordinates from Phase II of the HapMap Project (McVean et al. 2004; Myers et al. 2005). I excluded sites if they were within repeats as defined previously by RepeatMasker. Recombination rates and DTH were log transformed to more closely resemble a normal distribution.

To examine the impact on total mutation (all subtypes combined), I regressed the logit of the probability of a site containing a rare variant of any subtype against GC content, recombination rate, or DTH using separate logistic regression models for each genomic context variable. Each logistic regression has the form, $\ln\left(\frac{p}{1-p}\right) = \alpha + \beta z$,

where p is the probability that the site contains a rare variant and z is either GC content, recombination rate, or DTH at that site. I assessed the significance of the regression using a Wald test on the β parameter. I also fit regression models for common variants and substitutions.

Next, to analyze the effect of genomic context on specific variant subtypes, I employed a multinomial logistic regression model that jointly analyzes the probability of all possible variant subtypes for a given ancestral state. The model treats the derived alleles at a site with a given ancestral allele state (AT, GC or CpG) as a multinomial random variable with four potential outcomes. Sites with an AT ancestral allele state, for example, can have one of four possible derived states: AT reference (invariant), GC, CG, or TA. I ran separate multinomial regressions for each ancestral allele state and set the invariant allele as the baseline outcome. From each of these regressions, I calculated unique slope and intercept parameters for each variant subtype. I let $X > Y_i$ denote a nucleotide site with ancestral allele X and derived allele Y_i. Then, the multinomial logistic regression for an ancestral allele state X to the derived state Y_i has the form,

$$\ln\left(\frac{\Pr(X > Y_i)}{\Pr(X > X)}\right) = \alpha_{X>Y_i} + \beta_{X>Y_i} z \qquad \textbf{3.1}$$

where $\Pr(X > Y_i)$ is the probability that a site with ancestral allele X is variant with derived allele Y_i, $\Pr(X > X)$ is the probability that a site with ancestral allele X is invariant, z is the GC content, log(recombination rate), or log(DTH) at a given site, and $\sum_i \Pr(X > Y_i) = 1$ for each nucleotide site. I used a Wald test on the β slope parameter to assess significance. I fit separate multinomial logistic regression models for each

ancestral allele state for rare variants, common variants, and substitutions in order to estimate the effect of genomic context on variant subtypes in these three distinct variant classes. All logistic regressions were performed in R (R Development Core Team 2008).

3.2.4 Analysis of Logistic Regression Robustness

I employed three strategies to assess the robustness of the logistic regression results on the rare variants: two to test the estimated coefficients and another to analyze statistical significance. First, I used a subsampling strategy in which I randomly sampled 2,000 sites (out of ~700 kb) and ran total logistic regression on these 2,000 sites. There are 2,126 exons in the target regions; therefore sampling 2,000 sites will generate ~1 site per exon, on average. This analysis was repeated 1,000 times. I also performed this analysis using multinomial regression separately on AT, GC, and CpG ancestral sites. To analyze the degree to which outliers could be driving the observed slope parameters from the logistic and multinomial regressions, I performed a bootstrapping analysis. I randomly re-sampled 195 autosomal genes with replacement, generating a dataset with the same number of data points to the full analysis but eliminating a random subset of genes in each run. I ran the bootstrapping for the logistic regression analysis on total rare variants and the multinomial logistic regression on AT, GC, and CpG ancestral sites 1,000 times. Finally, I used permutations to analyze the statistical significance of the observed coefficients. For the total logistic regression, I randomly distributed the variant and invariant sites across the entire ~700 kb target region and performed the regression. This analysis was repeated 1,000 times. I performed this same analysis for the multinomial regression separately on AT, GC, and CpG ancestral sites.

3.3 Results

3.3.1 Variant Counts and Densities among Rare Variants, Common Variants, and Substitutions

I obtained rare variants from a previously described sequencing study targeting the exons and flanking intronic regions of 202 genes in > 14,000 individuals to a median depth of 27x (Nelson et al. 2012). Several complimentary methods were used to assess the quality of rare variants in these data. Among singleton variants, the false positive rate was 2.0%, which was estimated by Sanger sequencing-based validation 240 out of 245 sampled singletons (Nelson et al. 2012). The false negative rate in singletons was estimated to be about 2.7%, based on call rates among sequenced regions (Nelson et al. 2012). Lower error rates were estimated for more common variants (Nelson et al. 2012). For this study I focused on the 195 autosomal genes, with ~700 kb targeted regions in ~2,000 targeted exons, which contained a total of 20,053 rare variants with a DAF ≤ 10^{-4} in the European subset (N = 12,515). Each variant was categorized into one of seven possible variant subtypes based on the ancestral and derived allele states: AT>GC, GC>AT, CpG GC>AT, AT>CG, GC>TA, AT>TA and GC>CG (Table 3.1). The notation of AT>GC indicates a site where the ancestral base A has a G as the derived allele or ancestral base T has derived allele C.

I summarized variant counts and conditional variant proportions by subtype (Table 3.1). Nearly 13% of CpG sites have a rare GC>AT variant, compared to only 1.71% of non-CpG GC bases, consistent with the known hypermutability of CpG dinucleotides (Nachman and Crowell 2000; Kondrashov 2003; Hwang and Green 2004). Among rare variants, there were nearly twice as many S>W variants (those converting a

G/C base pair into an A/T base pair) as the opposite W>S variants (Table 3.1). This mutational AT bias is consistent with previous observations (Lynch 2010), and can be partially explained by the relatively high frequency of GC>AT variants at CpG dinucleotides (Table 3.1).

For comparison, I also analyzed common variants and substitutions. I randomly sampled intergenic regions from the human genome to obtain common variants and substitutions for analysis while matching the genomic context of the rare variant dataset (see Section 3.2.2.3: Sampling of Intergenic Regions to Obtain Common Variants and Substitutions). Sampling intergenic regions allowed me to minimize selective effects. To achieve comparable statistical power, I sampled a similar number of common variants and substitutions as the rare variants. In all, I obtained 22,566 variants from the European subset of the 1000 Genomes Project with a DAF > 5% and 22,179 human-lineage-specific divergent sites between humans and chimpanzee (Table 3.1).

The relative proportion of variant subtypes differed among the three variant classes. Figure 3.1 shows the total variant proportion, defined as the number of variants of a given subtype over the total number of variants in that variant class. The relative proportion of AT>GC variants increased progressively from rare variants to substitutions, while CpG GC>AT transitions correspondingly decreased (Figure 3.1). Other variant subtypes showed little change across the three variant classes. As seen in Table 3.1, the ratio of W>S/S>W variants increased from 0.54 in rare variants, to 0.65 in common variants, and further to 0.75 in substitutions, resulting in a progressive increase in W>S variant "load" as the average frequency of the derived allele increased. In other

words, W>S variants were more likely to rise to high allele frequencies and more likely to become fixed in the human lineage than S>W variants.

The conditional variant proportion, defined as the number of a given variant subtype divided by the total number of bases that could produce the given subtype, was higher in all rare variant subtypes compared to common variants and substitutions (Table 3.1). The higher absolute conditional variant proportion in rare variants is expected, as the rare variants were discovered in > 14,000 individuals. Importantly, the relative magnitude of the rare variant subtypes is expected to more closely reflect the relative spontaneous mutation rate than common variants or substitutions. The results for rare variants in Table 3.1, therefore, provide more accurate estimates of the relative mutation rates among different mutation subtypes.

3.3.2 The Per-Gene Mutation Rate Was Influenced by GC Content but Not Recombination Rate

I analyzed the per-gene mutation rate for 193 genes (out of the 195 autosomal genes), calculated previously by Nelson et al. (2012), using the method described by Coventry et al. (2010) and Wakeley and Takahashi (2003). There was considerable fluctuation in the mutation rate across genes (Figure 3.2A). To assess the impact of genomic context on this variability, I calculated the average GC content and recombination rate within the transcribed region of each gene (Figure 3.2B and C, respectively). There was a weak but significant positive correlation between mutation rate and GC content (Pearson's $r = 0.22$, $p = 0.0031$, Figure 3.2D dashed line). Recombination rate, however, was not significantly correlated with mutation rate (Pearson's $r = 0.039$, $p = 0.60$, Figure 3.2E dashed line). To ensure that outliers did not

drive these results, I excluded genes that fell outside of two standard deviations from the mean GC content or mutation rate (n = 8) and for recombination rate (n = 10). There was a slight increase in the correlation with GC content and little in the correlation with recombination rate (dotted line in Figure 3.2D and E, respectively). As previously reported (Kong et al. 2002), GC content and recombination rate themselves are positively correlated (Pearson's r = 0.18, p = 0.017). Multiple linear regression including both GC content and recombination rate as covariates did not change the results from either regression alone, and recombination rate was still not significantly correlated with mutation rate (GC content p-value = 0.002, recombination rate p-value = 0.658).

3.3.3 Using Logistic Regression to Analyze Per-Site Variant Patterns

The per-gene mutation rates analyzed above were calculated using all variant subtypes in aggregate. Previous studies, however, suggest that GC content and recombination rate may have different effects on specific variant subtypes (Lercher and Hurst 2002a; Arndt et al. 2005; Duret and Arndt 2008; Berglund et al. 2009). Estimating subtype-specific mutation rates on a per-gene or per-exon basis lacks a sufficient number of sites, especially for subtypes with relatively few observed variants (such as transversions). Therefore, I combined the ~700 K targeted sites over all 195 genes, using a per-site logistic regression strategy to examine the effect of local GC content and recombination rate on the probability of observing a variant of a given subtype (see Section 3.2.3: Logistic Regression Analysis).

The dependent variable of the logistic regression was obtained by scoring each site as either variant or invariant. If the site was scored as variant, it was further categorized into one of seven variant subtypes based on the ancestral and derived

alleles. The log odds of a site being variant was regressed on GC content and recombination rate, calculated in 1 kb windows surrounding each individual site.

3.3.4 GC Content Affected Rare Variants Differently From Common Variants and Substitutions

Overall, the probability of observing any rare variant was positively influenced by GC content ($\beta = 0.68$, p-value $< 10^{-16}$). However, individual subtypes showed negative or relatively small positive effects of GC content on all other variant subtypes (Figure 3.3), including CpG GC>AT variants ($\beta = -2.64$, p-value $< 10^{-16}$). The observation that individual subtypes could show opposite regression results to all variants combined may seem counter-intuitive, but is an example of Simpson's Paradox, where trends observed in subsets of the data can be the opposite of what is observed in the entire dataset (Agresti 2002). CpG-induced GC>AT variants, one of the major variant subtypes, tended to lie in GC-rich regions (50-65% GC content), whereas AT>GC transitions tended to occur in GC-poor regions (30-45% GC content) (Figure 3.6). The unbalanced distribution of variant subtypes across GC content, combined with the much higher mutation rate at CpG dinucleotides, drove the observed positive slope for all variants combined (Figure 3.6). When all CpG sites (variant or invariant) were removed and the regression run, the relationship between total rare variants and GC content became negative and was no longer significant ($\beta = -0.17$, p-value = 0.028). A similar finding was noted previously for substitution data (Duret and Arndt 2008). These results highlight the importance of studying variant subtypes, as analysis of all variants in aggregate could miss the underlying pattern of individual variant subtypes.

Comparison of rare variants and common variants or substitutions revealed subtype-specific differences among the three variant classes (Figure 3.3). For AT>GC and AT>CG rare variants, there was a relatively strong negative relationship between variant proportions and GC content (Figure 3.3A and B). These same trends, however, were not observed in AT>GC and AT>CG common variants or substitutions, which were weaker, and sometimes positive (Figure 3.3A and B). In contrast, for the other four variant subtypes, there were relatively strong negative effects on common variants and substitutions, yet the effects on rare variants were smaller or absent (Figure 3.3C-E). Together, these results show that GC-rich regions tend to have fewer W>S rare variants and fewer S>W common variants or substitutions than GC-poor regions.

3.3.5 Recombination Affects Patterns of Common Variants and Substitutions, but Not Rare Variants

The influence of recombination rate on total rare variants (β = 0.15, p-value = 3.58 x 10^{-4}) and all variant subtypes, including CpG GC>AT variants (β = -0.13, p-value = 0.16), was relatively small compared to common variants and substitutions (Figure 3.4). In contrast, recombination rate had a much stronger effect on total common variants (β = 0.95, p-value <10^{-16}) and substitutions (β = 0.29, p-value = 2.03x10^{-13}) and all variant subtypes for both common variants and substitutions than rare variants (Figure 3.4). There was a relatively strong positive effect on W>S common variants and substitutions (Figure 3.4A and B), consistent with the expected impact of BGC on variant patterns in the human genome. For the other four subtypes (Figure 3.4C-F), the effect on common variants was positive but weaker than in the case of W>S subtypes. In contrast, the effect on substitutions was negative (Figure 3.4C-E) or slightly positive

(Figure 3.4F). While BGC could explain the trends seen in W>S subtypes, the positive effects observed on other common variant subtypes suggests that either selective sweep or background selection could also be acting on these variants. Importantly, the lack of an effect on rare variants suggests that mutation rates are not altered by local recombination rate.

3.3.6 Distance to Recombination Hotspot Negatively Influenced Common Variants, but Had Little Effect on Rare Variants or Substitutions

Previous studies suggested that recombination hotspots accounted for most of the observed correlation between nucleotide diversity and recombination rate (Spencer 2006; Spencer et al. 2006). To examine the effect that recombination hotspots have on variant patterns, I calculated the absolute distance between each site and the nearest recombination hotspot (DTH) as reported in the population-based estimates from the HapMap Project (Myers et al. 2005). Median per-site DTH was consistent across all variant classes (median and standard deviation of absolute log-transformed DTH for rare variants: 4.43 ± 0.65, common variants: 4.32 ± 0.60, and substitutions: 4.31 ± 0.60). The results, shown in Figure 3.5, were largely consistent with those for recombination rate (Figure 3.4). I observed relatively weak relationships between DTH and rare variants for total (β = -0.042, p-value = 1.61×10^{-4}) and all variant subtypes (Figure 3.5). The strongest of these, GC>AT rare variants, had a negative relationship with DTH, although, it was weaker than the relationship observed in common variants (Figure 3.5D). DTH had a negative effect on total common variants (β = -0.15, p-value < 10^{-16}) and for each of the six variant subtypes (Figure 3.5). For substitutions however,

the negative effects were either weaker than those observed for common variants (Figure 3.5A, B, D and F) or positive (Figure 3.5C and E).

3.3.7 Validation of Rare Variant Results in an Independent Dataset

The dataset used for rare variants involved 195 genes of pharmaceutical interest (Nelson et al. 2012), and therefore may not be representative of genome-wide patterns. To test this, I made use of a publically available dataset from the National Heart Lung and Blood Institute (NHLBI) Exome Sequencing Project (ESP). I applied logistic regression on 603,267 singletons in this dataset (DAF = 1.4×10^{-4}), limiting to sites with ≥ 10x depth of coverage. GC content and recombination rate were calculated as before in 1kb windows surrounding each site. The regression coefficients from the exome-wide rare variant data fell within the 99% confidence intervals of the coefficients estimated from the 195 gene data (Table 2), with the following exceptions. Recombination rate has a significantly larger effect on total variants in the ESP data (Table 3.2). Also, the proportion of CpG GC>AT transitions was positively influenced by recombination rate for ESP variants, but negatively for the previously described rare variants (although this negative influence was not statistically significant) (Table 3.2). Overall, these results show that for most variant subtypes, there was no significant difference in the way that GC content, recombination rate, or DTH influence rare variant patterns in the 195 gene dataset compared to a larger collection of genes. Therefore, I conclude that my analysis of rare variants in the 195 gene dataset is representative of a broader sampling of genes across the genome.

3.3.8 Robustness of the Logistic Regression

3.3.8.1 Comparison of Coding and Noncoding Rare Variants

A central premise of this study was that natural selection has limited effects on rare variants. The sequence data cover both targeted exons and 50 bp of flanking sequence, allowing me to compare between coding and intronic rare variants. Total and CpG GC>AT rare variants had a greater conditional variant proportion in coding compared to intronic regions, although the proportion for all other variant subtypes was greater in intronic regions (Table 3.3). While the differences in the conditional variant proportion between coding and noncoding sites were statistically significant for most subtypes, the magnitudes of the differences were small (average across subtypes: 0.27%). Thus, while purifying selection may have slightly reduced the absolute number of rare variants in coding regions, the relative proportion of individual variant subtypes was not substantially affected. Importantly, with regard to the main conclusions of this study, there was no significant difference (based on 99% confidence intervals) in the coefficients for GC content, recombination rate, or DTH regressions performed on coding, intronic, or the total dataset (Table 3.4).

3.3.8.2 No Difference in Regression Results across a Variety of Window Sizes for GC Content and Recombination Rate

The analysis presented above used GC content and recombination rate calculated in 1kb windows. To test the dependence of my results on window size, I extended the analysis for windows ranging from 100 bp – 10 kb. With the exception of CpG sites (Figure 3.7A), I observed no significant differences between regression

coefficients for any other variant subtype across the range of window sizes tested (based on 95% confidence intervals) (Figure 3.7).

3.3.8.3 Subsampling, Bootstrapping, and Permutation Analyses are Consistent with Logistic Regression Results in Rare Variants

The rare variants I analyzed were derived from exome sequencing and are distributed in tight clusters, corresponding to ~2,000 targeted exons in 195 autosomal genes. Genomic features of nearby sites are often not strictly independent. To evaluate the impact of spatial dependency on the regression results, I performed a subsampling analysis using 2,000 random sites (out of ~700K sites) in each run (see Section 3.2.4: Analysis of Logistic Regression Robustness). All observed coefficients in the original analysis fell within the 25th-75th percentile range of the coefficients from 1,000 subsampling runs for GC content (Figure 3.8A), recombination rate (Figure 3.8B), and DTH (Figure 3.8C). I also examined the potential impact of between-gene heterogeneity by performing a bootstrapping analysis involving random sampling of 195 genes with replacement. The distribution of the coefficients from 1,000 bootstrapping runs was symmetric around the original estimates for GC content (Figure 3.9A), recombination rate (Figure 3.9B), and DTH (Figure 3.9C), confirming that there was no systematic bias due to outlier genes driving the results of the regressions. In addition, as the p-values in the logistic regression were model-based, I assessed potential bias of the reported p-values by running 1,000 rounds of permutations of the variant and invariant status across sites, and found that the p-values calculated in the regressions were consistent for GC content, recombination rate, and DTH (Table 3.5).

3.3.8.4 Little Difference between Univariate and Multivariate Regression Results

GC content and recombination rate are positively correlated (Kong et al. 2002). To determine the extent to which the results for recombination rate and DTH could be driven by GC content, and vice versa, I performed multivariate logistic regression with two models, one using GC content and recombination rate as covariates and another using GC content and DTH as covariates. I did not observe a significant difference between the regression coefficients (based on 99% confidence intervals) estimated from the univariate (presented above) and the multivariate models for GC content, recombination rate, or DTH in rare variants (Table 3.6), common variants (Table 3.7), or substitutions (Table 3.8).

3.3.8.5 Coverage Does Not Alter Logistic Regression Results

Because GC content influences read depth in high-throughput sequencing studies, especially following target capture (Albert et al. 2007; Porreca et al. 2007), I verified that the observed influence of GC content on rare variants was not an artifact of sequencing depth. In addition to the 10x coverage filter imposed on all sites in the rare variant analysis (see Section 3.2.2.1: Rare Variants), I first performed logistic regression using per-base coverage as the explanatory variable. Total, AT>GC, CpG GC>AT, and GC>AT variants were significantly affected by coverage (Table 3.9). Including coverage as a covariate in the regression against GC content decreased the effect of GC content on CpG GC>AT transitions, but the coefficient was still negative (Table 3.9). The estimated coefficients for other variant subtypes were not affected by including coverage in the model (based on 99% confidence intervals). I concluded that coverage was not driving the results regarding the influence of GC content on rare variants.

3.4 Discussion

In this study, I examined mutation patterns of different variant subtypes in the human genome, using rare variants as a model for the *de novo* mutation rate. I also used common variants and human-chimpanzee substitutions to analyze the ongoing biases toward fixation present in the human genome, which can include natural selection and neutral evolutionary processes. The results suggest that mutation rates and fixation biases are affected by local GC content. However, my results suggest that fixation, and not mutation, is affected by the local recombination rate.

Using rare variants to analyze the spontaneous mutation rate was previously suggested in anticipation of the emergence of this new data type from the next-generation sequencing studies (Messer 2009). Rare variants, such as those used here, arose recently in the population. For example, the rare variants I analyzed, with a DAF ≤ 10^{-4}, arose an average of ~10 generations in the past, assuming a current population size of 50,000 individuals and a population growth rate of 0.001 (Slatkin 2000). Furthermore, for populations undergoing recent population expansion, such as the growth that humans have experienced in the recent past (Coventry et al. 2010), such low-frequency variants will be even younger. Because these rare variants are, on average, very young in the population, their patterns are primarily governed by random genetic drift. Therefore, unless the force of selection is strong, natural selection, along with population demographic history, and BGC, will not alter the observed patterns of rare variants.

Analysis of synonymous and nonsynonymous variants separately is a common approach for minimizing the effects of natural selection. However, the logistic regression

approach works on individual variant and invariant sites. It is difficult to analyze synonymous and nonsynonymous variants separately because each ancestral allele could mutate to three other nucleotides, and one needs to enumerate the potential synonymous and nonsynonymous variants that could occur at each site. Instead, I analyzed the effect that GC content, recombination rate, and DTH has on coding and noncoding rare variants in order to assess any potential confounding effect that natural selection may have on the results. I did not find any significant difference between these two functional classes of variants, consistent with theoretical analysis showing that the effect of selection is attenuated in rare variants (Messer 2009).

The average per-gene mutation rate, based on variants from 193 genes, was 1.02×10^{-8} per base pair per generation (Nelson et al. 2012), which is within the range of recently published estimates from family-based sequencing studies (The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012). I observed considerable variability in the mutation rate from gene to gene, consistent with previous work (Wolfe et al. 1989; Nachman and Crowell 2000; Kondrashov 2003; Hodgkinson et al. 2009).

Previous studies examining the effect of genomic context on the mutation rate relied on context measurements computed in fixed-length genomic windows. This window-based approach is difficult to implement in exome sequencing data because such data cover short intervals with variable length, representing targeted exons, separated by large gaps, representing introns. This leads to problems in defining window width and estimating average parameter values. In my study, most target regions are small (85% < 500 bp) and calculating rates for low frequency events, such

as transversions, in these windows could be highly inaccurate. I therefore adopted a logistic regression approach, using data for individual base positions and aggregating data across sites. This approach has several advantages. It eliminates the need to account for gaps in coverage from intronic and intergenic regions, and provides a sufficient number of sites to study the effect of genomic context on individual variant subtypes.

My results suggest that recombination rate has a relatively small effect on the mutation rate, but a significant impact on common variants in the population. First, I did not observe a correlation between the per-gene mutation rate and recombination rate. Second, the effect of recombination rate on rare variant subtypes was small, especially compared to the effect observed on common variants and substitutions. AT>GC and AT>CG common variants and substitutions, however, were both strongly affected by local recombination rate. Together, these results are consistent with BGC altering patterns of standing variation in the human genome. BGC has no effect on mutation rates, but over time, is expected to lead to a fixation bias toward GC bases at AT/GC polymorphic sites (Duret and Galtier 2009). A recent study reported a strong bias of W>S substitutions in Human Accelerated Regions that increased with increasing male recombination rate (Berglund et al. 2009)**.** Furthermore, BGC can drive deleterious GC alleles to fixation (Galtier et al. 2009) and is hypothesized to lead to the apparent increase in substitution rate with increasing recombination rate (Meunier and Duret 2004; Duret and Arndt 2008; Berglund et al. 2009; Galtier et al. 2009). While my results cannot completely rule out a mutagenic effect due to recombination, they suggest that if such an effect does exist, it is relatively small in comparison to the influence of BGC.

Background selection and selective sweeps can also drive positive correlations between diversity and recombination rate (Smith and Haigh 1974; Kaplan et al. 1989; Charlesworth et al. 1993; Hudson and Kaplan 1995; Cai et al. 2009; Lohmueller et al. 2011). These selection-dependent mechanisms are unlikely to affect rare variants because they are too young in the population to be substantially affected by selection. In addition to the relationship observed between recombination rate and AT>GC and AT>CG common variants and substitutions, I also saw relatively strong effects on other variant subtypes. Therefore, I cannot rule out the effect of these recombination-associated processes on patterns of variants in my dataset.

GC content varies throughout the genome, with long stretches of DNA exhibiting relatively stable GC content, known as isochores (Eyre-Walker and Hurst 2001). Previous studies propose that mutation bias or fixation bias toward or against GC bases drives the apparent regional variation in GC content and maintenance of isochores throughout the genome (Smith et al. 2002; Webster et al. 2003; Duret and Arndt 2008). My results are consistent with this hypothesis, suggesting that GC-rich regions of the genome may maintain base composition by simultaneously decreasing GC-enriching, W>S, mutations and reducing the fixation of GC-depleting, S>W, common variants.

The effect of genomic context on CpG dinucleotides has also been widely studied. The CpG dinucleotide is defined as a C base followed directly by a G base: 5'-CG-3'. The C at this specific base configuration is especially prone to methylation, and the resulting 5-methyl cytosine can undergo spontaneous deamination to produce T, generating a C>T mutation (G>A on the opposite strand) (Cooper and Youssoufian 1988; Cooper and Krawczak 1993). Due to this process, CpG dinucleotides mutate at

roughly 10 - 40 times the frequency of other nucleotides (Sommer 1995; Nachman and Crowell 2000; Kondrashov 2003; Hwang and Green 2004). The rate of CpG-induced GC>AT transitions has a strong negative correlation with both GC content and recombination rate (Fryxell and Zuckerkandl 2000; Arndt et al. 2005; Fryxell and Moon 2005; Duret and Arndt 2008) and it has been suggested that the increased thermal stability of GC-rich regions has a protective effect on the mutability of CpG dinucleotides (Fryxell and Zuckerkandl 2000; Fryxell and Moon 2005). My results for the effect of GC content and recombination rate on CpG-induced GC>AT transitions are consistent with previous studies.

Understanding the relationship between local genomic context and the mutation rate has several practical implications. More precise estimates of *de novo* mutation rates can improve genotype calling from short sequencing reads by providing better prior distributions for the spectrum of expected variation. Moreover, my results can help to identify potentially functional *de novo* mutations by highlighting new variants that are unlikely to arise spontaneously.

My study, however, has several limitations. I was not able to identify all potential mutations, as some will not be viable in humans. However, truly dominant lethal mutations are extremely rare and other approaches, including direct discovery of *de novo* variants via trio sequencing, will have similar limitations. Additionally, while rare variants are very young on the evolutionary time scale, they could still be influenced by the same confounding factors that affect common variants and substitutions, albeit to a lesser extent. At present, however, rare variants, especially the extremely rare variants I studied here, represent one of the most powerful datasets currently available for high-

sensitivity analysis of the rate and molecular spectrum of new mutations. Finally, the dataset I used involves only 195 genes, and could generate a biased representation of the genome. Indeed, these genes appear to be under stronger purifying selection than other genes (Nelson et al. 2012). Despite this caveat, I observed strong concordance between the results from the 195 genes and those from an exome-wide dataset, indicating that any selection acting on these genes does not influence the relationship with genomic context and that the results presented here are representative of the exome.

3.5 Conclusions

In this study, I used a dataset of > 20,000 rare variants (DAF < 10^{-4}) as a new resource for studying patterns of single-nucleotide mutation in humans. Compared to common variants, the effects of confounding variables, including demography, selection and BGC, are reduced in these rare variants. This allows me to take a new step toward differentiating the initial mutation processes from the subsequent forces that act more gradually, affecting fixation processes of segregating variants. My results reveal a substantial difference in the relative abundance of variant subtypes between rare variants, common variants, and substitutions. GC content has a strong impact on all variant classes, although the effect is different both between variant classes and different variant subtypes. Recombination rate, on the other hand, has relatively little effect on rare variants, but a much stronger effect on AT>GC and AT>CG common variants and substitutions, consistent with BGC acting on existing variants. Furthermore, my results reveal a drastic difference between total mutations and individual mutation subtypes and this advocates for the importance of future research that focuses on

subtype-specific patterns in order to fully understand the effect of GC content and recombination rate on mutation and fixation rates in the human genome. Future research, aided by the ever-increasing deep sequencing data covering more genomic targets in larger population samples, is necessary to more precisely estimate these fundamental parameters.

Eventually, these studies will help unravel the relative contribution of diverse evolutionary forces acting over different time scales. Such an understanding will also provide the knowledge base necessary to study the allelic spectrum of inherited and somatic diseases, as well as the dynamics of human genome variation as it evolves under a variety of environmental and demographic conditions.

3.6 Figures

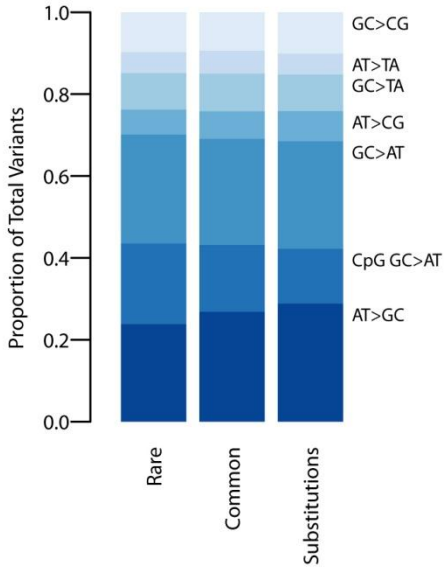

Figure 3.1: Comparison of total variant proportions of the seven variant subtypes across the three variant classes.

The total variant proportion is shown for each of the seven variant subtypes, defined as the number of variants of a given subtype over the total number of variants in that variant class, for each of the three variant classes analyzed: rare variants, common variants, and substitutions.

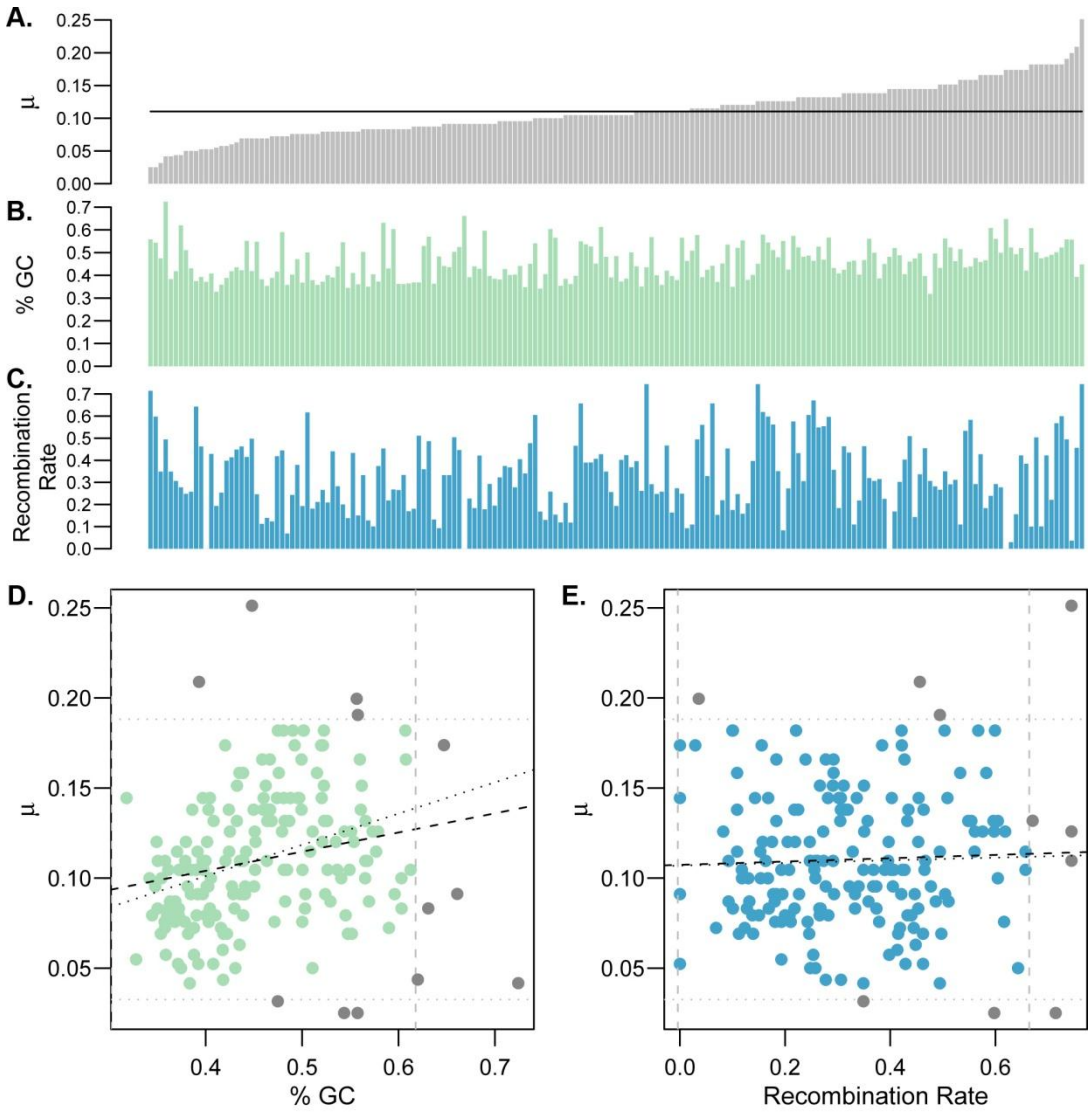

Figure 3.2: Variability of mutation rate across 193 genes and relationship with genomic context.

(A) Per-gene mutations rates (x 10^{-7} per base pair per generation) for 193 genes, estimated previously by coalescent modeling (Nelson et al. 2012), are shown ordered from lowest to highest. The black line indicates the average of 193 genes (1.02 x 10^{-8} per base pair per generation). (B) Per gene average GC content ordered as in A. (C) Per-gene average recombination rate ($\log_{10}$ cM/Mb) ordered as in A. (D) Relationship between GC content and mutation rate (x 10^{-7} per base pair per generation). The dashed line represents the linear regression fitting over all 193 genes. After removing outliers (grey filled points), the regression was recalculated (dotted line). (E) Relationship between recombination rate ($\log_{10}$ cM/Mb) and mutation rate (x 10^{-7} per base pair per generation). The dashed line represents the linear regression fitting over all 193 genes. Outliers were removed (grey filled points) and the regression was recalculated (dotted lines).

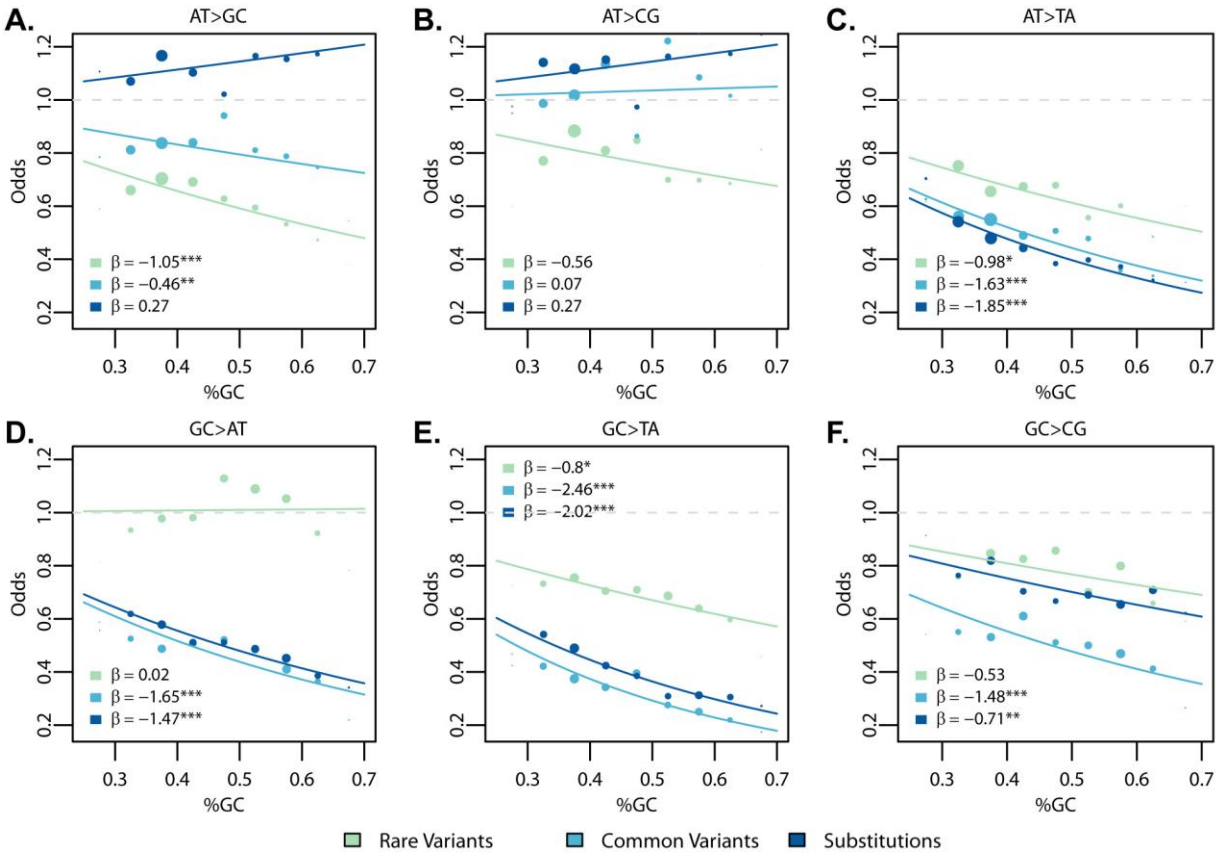

Figure 3.3: Regression results for GC content across variant subtypes for rare variants, common variants, and substitutions.

The relationship between local GC content and the observed conditional variant proportion for 6 variant subtypes: (A) AT>GC, (B) AT>CG, (C) AT>TA, (D) GC>AT, (E) GC>TA, and (F) GC>CG across observed GC content. Filled points show the conditional variant proportions in each GC content bin, scaled by the intercept of the logistic regression: $\frac{n_{X>Y,i}}{N_{x,i}} e^{\alpha}$ where α is the intercept calculated in the regression, $n_{X>Y}$ is the count of the given $X > Y$ variant subtype, and $N_{x,i}$ is the number of X ancestral invariant sites that could produce the given subtype in the i^{th} GC content bin. Symbol size represents the proportion of the given variant subtype falling into a given GC-content bin. The solid lines show the fitted logistic regression curve, where β is the slope fitted in the logistic regression and x_i is the GC content in the i^{th} bin. The grey dashed line represents the baseline of no effect, $\beta = 0$. Legends in each subplot show the regression slope calculated for each variant class and its significance. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

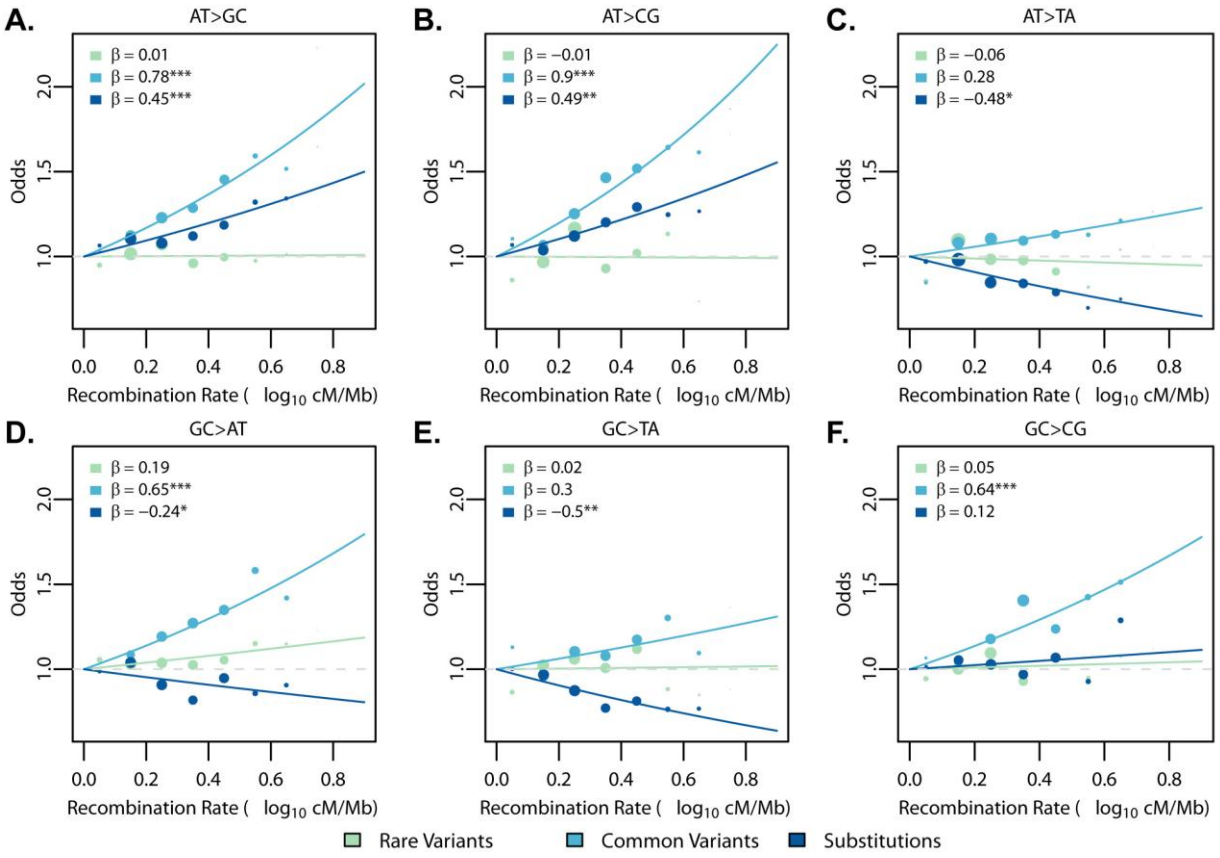

Figure 3.4: Regression results for recombination rate across variant subtype for rare variants, common variants, and substitutions.

The relationship between local recombination rate ($\log_{10}$ cM/Mb) and the observed conditional variant proportion for 6 variant subtypes: (A) AT>GC, (B) AT>CG, (C) AT>TA, (D) GC>AT, (E) GC>TA, and (F) GC>CG across observed recombination rate (plotted as in Figure 3.3). Filled points show the conditional variant proportions, scaled by the intercept of the logistic regression. Symbol size represents the proportion of the given variant subtype falling into a given recombination rate bin. The solid lines show the fitted logistic regression curve, where β is the slope fitted in the logistic regression and x_i is the recombination rate in the i^{th} bin. The grey dashed line represents the baseline of no effect, $\beta = 0$. Legends in each subplot show the regression slope calculated for each variant class and its significance. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

.

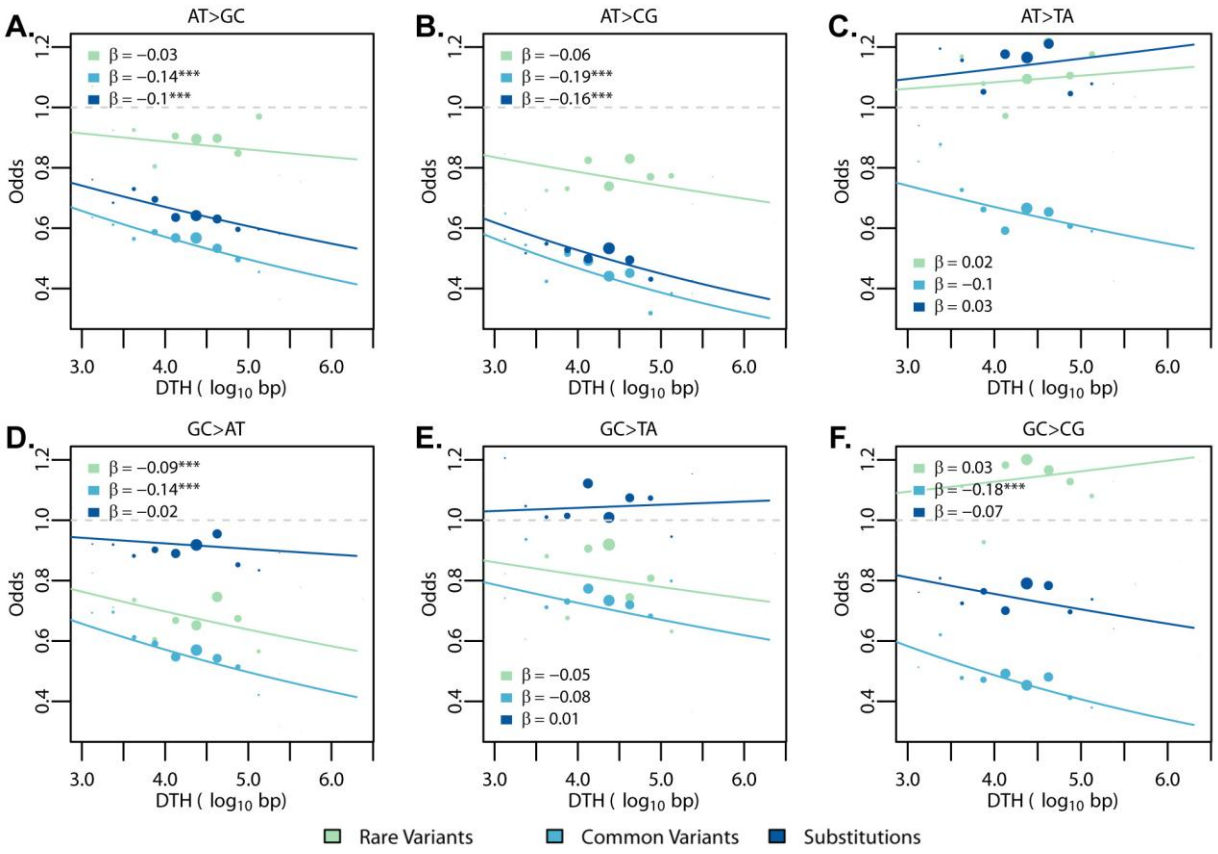

Figure 3.5: Regression results for DTH across variant subtypes for rare variants, common variants, and substitutions.

The relationship between DTH ($\log_{10}$ bp) and the 6 variant subtypes: (A) AT>GC, (B) AT>CG, (C) AT>TA, (D) GC>AT, (E) GC>TA, and (F) GC>CG across observed DTH (plotted as in Figure 3.3). Filled points show the conditional variant proportions, scaled by the intercept of the logistic regression. Symbol size represents the proportion of the given variant subtype falling into a given DTH bin. The solid lines show the fitted logistic regression curve, where β is the slope fitted in the logistic regression and x_i is the DTH in the i^{th} bin. The grey dashed line represents the baseline of no effect, $\beta = 0$. Legends in each subplot show the regression slope calculated for each variant class and its significance. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

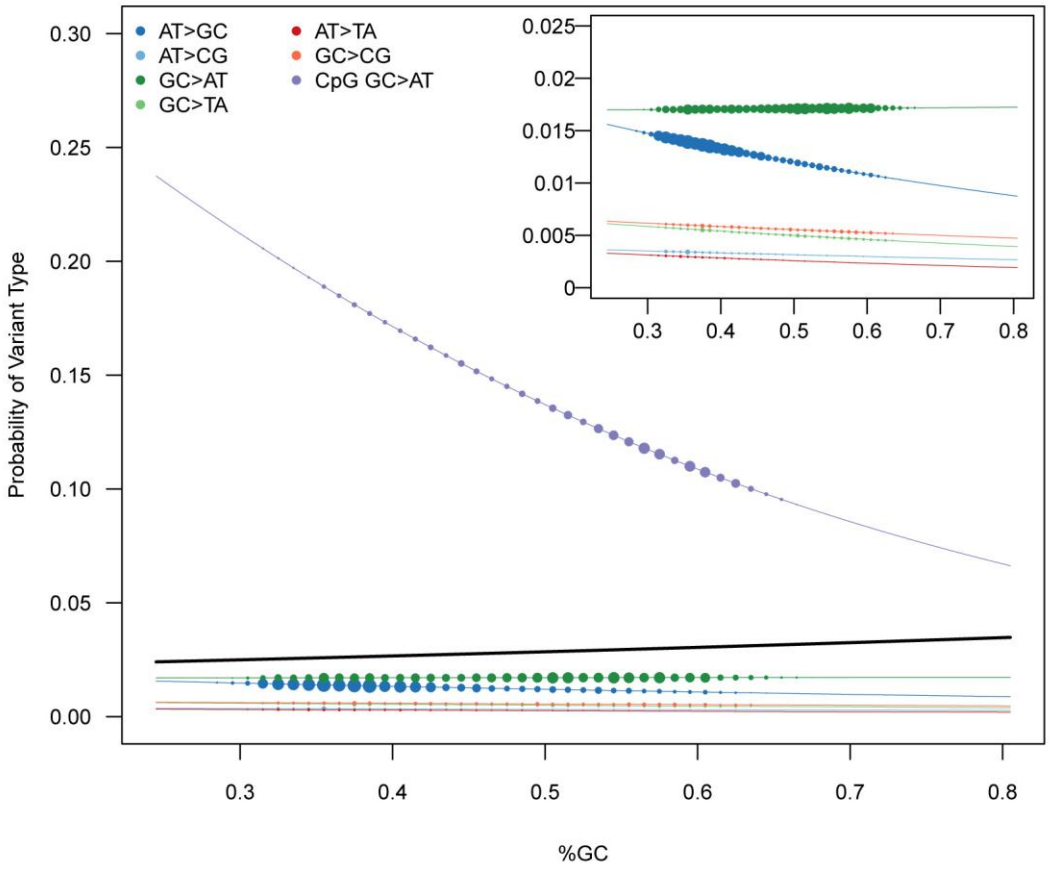

Figure 3.6: Difference in effect of GC content on rare variants between total variants and individual variant subtypes.

This plot shows the fitted logistic regression curves for a given variant subtype across observed GC content. The probability for total variants is shown in black. Point size corresponds to the proportion of the given variant subtype in each GC content bin. While most of the variant subtypes have a negative relationship between probability of occurrence and GC content, the trend between the overall probability of observing a rare variant and GC content is positive. This is driven by the increased mutation rate of CpG dinucleotides and the uneven distribution of CpG GC>AT and AT>GC variants across GC content. The inset shows the portion of the plot with variant probability ≤ 0.025 for all GC content bins to provide a better view of the probability across GC content for non-CpG-induced variants.

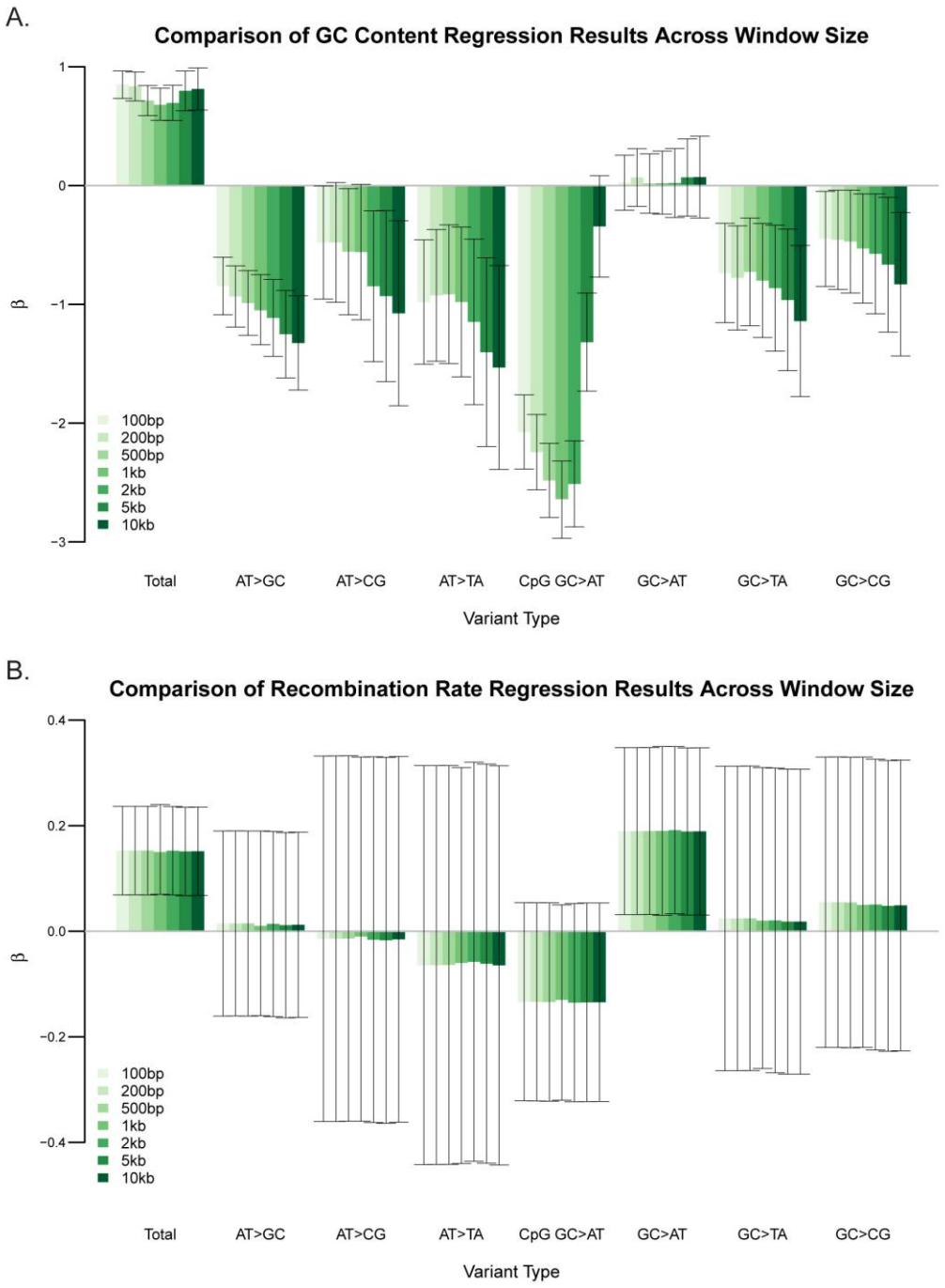

Figure 3.7: Sensitivity analysis for rare variants with varying GC content and recombination rate window sizes.

I compared regression analysis for GC content (A) and recombination rate (B) using window sizes of 100 bp, 200 bp, 500 bp, 2 kb, 5 kb, and 10 kb to the original 1 kb analysis. The barplots show the estimated regression coefficients for each of the window sizes including the 1kb described in the results. Error bars represent 95% confidence intervals for each regression coefficient.

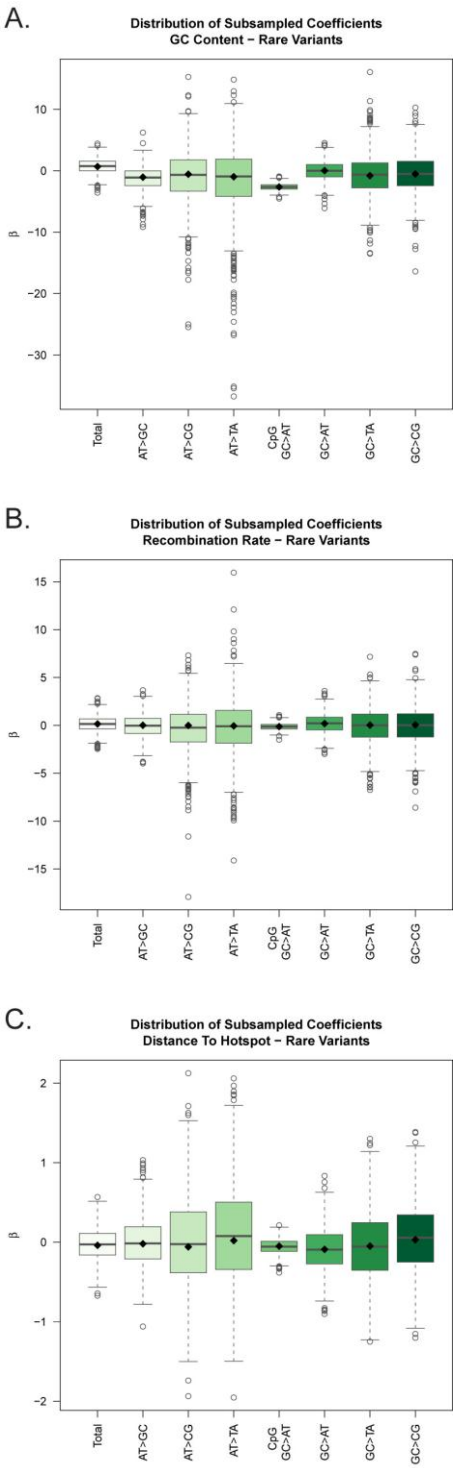

Figure 3.8: Distribution of estimated regression coefficients from subsampling analysis.

This plot shows the distribution of estimated regression coefficients from the 1,000 subsampling analyses for (A) GC content, (B) recombination rate, and (C) DTH for rare variants. Red diamonds indicate the coefficients obtained in the original analysis.

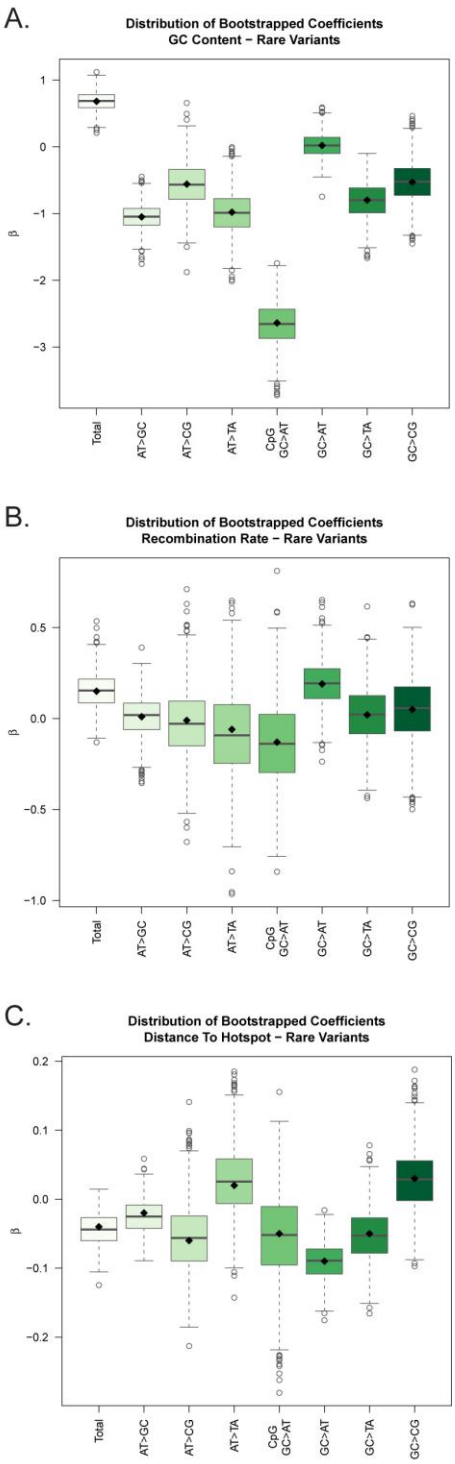

Figure 3.9: Distribution of estimated regression coefficients from bootstrapping analysis.

This plot shows the distribution of estimated regression coefficients over the 1,000 bootstrapping analyses for (A) GC content, (B) recombination rate, and (C) DTH for rare variants. Red diamonds show the coefficients obtained in the original analysis.

3.7 Tables

Variant Type	Transitions			Transversions				Total	Ti/Tv	W>S/S>W
	AT>GC	CpG GC>AT	GC>AT	AT>CG	GC>TA	AT>TA	GC>CG			
Rare Variants	4,778 (1.28%)	3,951 (12.76%)	5,338 (1.71%)	1,215 (0.32%)	1,796 (0.52%)	1,023 (0.27%)	1,952 (0.57%)	20,053 (2.79%)	2.35	0.54
1K Genomes	6,060 (0.10%)	3,684 (1.079%)	5,845 (0.11%)	1,519 (0.025%)	2,078 (0.038%)	1,261 (0.021%)	2,119 (0.038%)	22,566 (0.19%)	2.23	0.65
Substitutions	6,398 (0.28%)	2,971 (1.99%)	5,826 (0.29%)	1,633 (0.071%)	1,982 (0.10%)	1,135 (0.049%)	2,234 (0.11%)	22,179 (0.50%)	2.18	0.75

Table 3.1: Variant counts and conditional variant proportions across variant subtype for rare variants, common variants, and substitutions

Counts of all variant subtypes across rare variants, common variants, and substitutions are shown. Conditional variant proportion for each variant subtype, defined as the number of observed variants divided by the number of bases that could give rise to the given variant, are shown below in parenthesis. W>S/S>W was defined as the total number of weak to strong (W>S) variants divided by the total number of strong to weak (S>W) variants. Ti/Tv is the ratio of transitions to transversions.

Variant Subtypes	GC Content				Recombination Rate				DTH			
	195 Genes		ESP		195 Genes		ESP		195 Genes		ESP	
Total	0.68	(0.069)	0.64	(0.012)	0.15	(0.043)	0.34	(0.0076)	-0.042	(0.011)	-0.057	(0.0020)
AT>GC	-1.05	(0.15)	-0.64	(0.028)	0.014	(0.089)	0.14	(0.017)	-0.025	(0.023)	-0.057	(0.0044)
AT>CG	-0.56	(0.29)	-0.17	(0.057)	-0.014	(0.18)	0.13	(0.034)	-0.060	(0.044)	-0.051	(0.0091)
AT>TA	-0.98	(0.32)	-0.21	(0.062)	-0.065	(0.19)	0.21	(0.037)	0.023	(0.049)	-0.034	(0.0099)
CpG GC>AT	-2.64	(0.17)	-3.072	(0.024)	-0.13	(0.10)	0.17	(0.014)	-0.047	(0.025)	-0.077	(0.0039)
GC>AT	0.024	(0.14)	-0.26	(0.025)	0.19	(0.081)	0.20	(0.015)	-0.089	(0.021)	-0.058	(0.0041)
GC>TA	-0.80	(0.25)	-0.91	(0.048)	0.024	(0.15)	0.15	(0.029)	-0.054	(0.039)	-0.061	(0.0077)
GC>CG	-0.53	(0.24)	-0.96	(0.043)	0.054	(0.14)	0.13	(0.026)	0.025	(0.037)	-0.055	(0.0069)

Table 3.2: Regression coefficients for rare variants in the 195 gene dataset compared to the ESP whole-exome dataset.

β coefficients and standard error (in parenthesis) for all variant subtypes from the original rare variant analysis in 195 genes compared to those from the ESP whole-exome sequencing data analysis. Values shown in bold indicate coefficients that are significantly different between the two datasets, based on 99% confidence intervals (not shown).

Variant Type	Coding		Intronic		p-value	
Total	8,738	(2.85%)	4,642	(2.70%)	0.0033	*
AT>GC	1,764	(1.19%)	1,245	(1.32%)	0.0028	*
AT>CG	398	(0.27%)	324	(0.34%)	0.00077	**
AT>TA	362	(0.24%)	249	(0.26%)	0.32	
CpG GC>AT	2,525	(13.28%)	551	(11.98%)	0.020	
GC>AT	2,147	(1.55%)	1,328	(1.81%)	0.0000037	***
GC>TA	746	(0.47%)	451	(0.58%)	0.00062	**
GC>CG	796	(0.50%)	494	(0.64%)	0.000056	***

Table 3.3: Comparison of Rare Variant Counts in Coding and Intronic Sequences

Counts of variants identified in coding and flanking intronic regions. Numbers in parenthesis show the conditional variant proportion of each variant subtype, defined as the number of variants of the subtype divided by the number of total sites that could produce the given variant. The p-values from a two-proportion t-test performed in conditional variant proportion are also presented.

Variant Subtype	Model								
	All Sites			Coding Sites			Intronic Sites		
	GC Content								
Total	0.68	(0.069)	***	0.61	(0.10)	***	0.74	(0.15)	***
AT>GC	-1.05	(0.15)	***	-1.14	(0.24)	***	-1.14	(0.31)	**
AT>CG	-0.56	(0.29)		-1.41	(0.50)	*	-0.19	(0.58)	
AT>TA	-0.98	(0.32)	*	-0.68	(0.51)		-0.82	(0.67)	
CpG GC>AT	-2.64	(0.17)	***	-2.62	(0.20)	***	-1.91	(0.48)	***
GC>AT	0.024	(0.14)		0.40	(0.21)		-0.39	(0.28)	
GC>TA	-0.80	(0.25)	*	-0.77	(0.38)		-0.22	(0.49)	
GC>CG	-0.53	(0.24)		-1.10	(0.37)	*	-0.45	(0.47)	
	Recombination Rate								
Total	0.15	(0.043)	**	0.15	(0.063)		0.29	(0.087)	**
AT>GC	0.014	(0.089)		-0.091	(0.14)		0.27	(0.17)	
AT>CG	-0.014	(0.18)		-0.55	(0.30)		0.15	(0.33)	
AT>TA	-0.065	(0.19)		-0.46	(0.32)		0.32	(0.38)	
CpG GC>AT	-0.13	(0.10)		-0.073	(0.12)		-0.12	(0.25)	
GC>AT	0.19	(0.081)		0.33	(0.12)	*	0.18	(0.16)	
GC>TA	0.024	(0.15)		-0.14	(0.23)		0.15	(0.28)	
GC>CG	0.054	(0.14)		0.16	(0.22)		0.14	(0.27)	
	DTH								
Total	-0.042	(0.011)	**	-0.069	(0.017)	***	-0.027	(0.022)	
AT>GC	-0.025	(0.023)		-0.0086	(0.038)		0.014	(0.044)	
AT>CG	-0.060	(0.044)		0.0076	(0.079)		-0.043	(0.086)	
AT>TA	0.023	(0.049)		0.093	(0.084)		-0.015	(0.10)	
CpG GC>AT	-0.047	(0.025)		-0.11	(0.031)	**	0.067	(0.062)	
GC>AT	-0.089	(0.021)	***	-0.085	(0.034)		-0.096	(0.040)	
GC>TA	-0.054	(0.039)		-0.088	(0.061)		-0.035	(0.072)	
GC>CG	0.025	(0.037)		0.012	(0.060)		0.0042	(0.068)	

Table 3.4: Comparison of Regression Results for Rare Variants In All Sites, Coding Sites, and Intronic Sites

β coefficients, standard error (in parenthesis), and significance from the regression on all sites, coding sites, and intronic sites. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

Variant Subtype	Model	
	Model-Based P-Value	Empirical (One-Sided) P-Value
	GC Content	
Total	$<2x10^{-16}$	$<1x10^{-3}$
AT>GC	$2.51x10^{-12}$	$<1x10^{-3}$
AT>CG	0.054	0.025
AT>TA	$2.28x10^{-3}$	0.001
CpG GC>AT	$<2x10^{-16}$	$<1x10^{-3}$
GC>AT	0.86	0.46
GC>TA	$1.15x10^{-3}$	0.001
GC>CG	0.024	0.009
	Recombination Rate	
Total	$3.58x10^{-4}$	0.001
AT>GC	0.87	0.47
AT>CG	0.94	0.49
AT>TA	0.74	0.38
CpG GC>AT	0.16	0.082
GC>AT	0.019	0.012
GC>TA	0.87	0.46
GC>CG	0.70	0.40
	DTH	
Total	$1.61x10^{-4}$	$<1x10^{-3}$
AT>GC	0.27	0.17
AT>CG	0.18	0.10
AT>TA	0.65	0.33
CpG GC>AT	0.059	0.028
GC>AT	$2.39x10^{-5}$	$<1x10^{-3}$
GC>TA	0.16	0.087
GC>CG	0.49	0.23

Table 3.5: Comparison of Model-Based and Empirical P-values calculated from 1000 Permutations of Variant and Invariant Sites

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	0.68	(0.069)	***	0.66	(0.070)	***	0.69	(0.069)	***
AT>GC	-1.05	(0.15)	***	-1.09	(0.15)	***	-1.05	(0.15)	***
AT>CG	-0.56	(0.29)		-0.58	(0.30)		-0.57	(0.29)	
AT>TA	-0.98	(0.32)	*	-0.99	(0.33)	*	-0.98	(0.32)	*
CpG GC>AT	-2.64	(0.17)	***	-2.64	(0.17)	***	-2.64	(0.17)	***
GC>AT	0.024	(0.14)		-0.014	(0.14)		0.027	(0.14)	
GC>TA	-0.80	(0.25)	*	-0.82	(0.25)	**	-0.80	(0.25)	*
GC>CG	-0.53	(0.24)		-0.55	(0.24)		-0.53	(0.23)	
	Recombination Rate								
Total	0.15	(0.043)	**	0.094	(0.043)		-	-	-
AT>GC	0.014	(0.089)		0.13	(0.092)		-	-	-
AT>CG	-0.014	(0.18)		0.048	(0.18)		-	-	-
AT>TA	-0.065	(0.19)		0.042	(0.20)		-	-	-
CpG GC>AT	-0.13	(0.010)		-0.044	(0.098)		-	-	-
GC>AT	0.19	(0.081)		0.19	(0.081)		-	-	-
GC>TA	0.024	(0.15)		0.086	(0.15)		-	-	-
GC>CG	0.054	(0.14)		0.095	(0.14)		-	-	-
	DTH								
Total	-0.042	(0.011)	**	-	-	-	-0.042	(0.011)	**
AT>GC	-0.025	(0.023)		-	-	-	-0.028	(0.023)	
AT>CG	-0.060	(0.044)		-	-	-	-0.061	(0.045)	
AT>TA	0.023	(0.049)		-	-	-	0.020	(0.049)	
CpG GC>AT	-0.047	(0.025)		-	-	-	-0.029	(0.026)	
GC>AT	-0.089	(0.021)	***	-	-	-	-0.089	(0.021)	***
GC>TA	-0.054	(0.039)		-	-	-	-0.054	(0.039)	
GC>CG	0.025	(0.037)		-	-	-	0.026	(0.037)	

Table 3.6: Comparison of Logistic Regression Results for Rare Variants between Univariate and Multivariate Models

β coefficients, standard error (in parenthesis), and significance for GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	-0.18	0.06	*	-0.77	0.064	***	-0.27	0.06	***
AT>GC	-0.46	0.12	**	-1.053	0.13	***	-0.56	0.12	***
AT>CG	0.070	0.24		-0.51	0.26		-0.053	0.24	
AT>TA	-1.63	0.28	***	-2.09	0.30	***	-1.73	0.28	***
CpG GC>AT	-3.82	0.15	***	-4.32	0.16	***	-3.88	0.15	***
GC>AT	-1.65	0.12	***	-2.21	0.13	***	-1.73	0.12	***
GC>TA	-2.46	0.22	***	-2.94	0.23	***	-2.52	0.22	***
GC>CG	-1.48	0.21	***	-2.0082	0.22	***	-1.58	0.21	***
	Recombination Rate								
Total	0.95	0.039	***	1.12	0.042	***	-	-	-
AT>GC	0.78	0.076	***	1.021	0.08	***	-	-	-
AT>CG	0.90	0.15	***	1.017	0.16	***	-	-	-
AT>TA	0.28	0.17		0.76	0.18	***	-	-	-
CpG GC>AT	0.30	0.10	*	1.036	0.11	***	-	-	-
GC>AT	0.65	0.077	***	1.11	0.082	***	-	-	-
GC>TA	0.30	0.14		0.92	0.15	***	-	-	-
GC>CG	0.64	0.14	***	1.049	0.14	***	-	-	-
	DTH								
Total	-0.15	0.011	***	-	-	-	-0.16	0.011	***
AT>GC	-0.14	0.021	***	-	-	-	-0.15	0.021	***
AT>CG	-0.19	0.041	***	-	-	-	-0.19	0.041	***
AT>TA	-0.10	0.046		-	-	-	-0.14	0.046	*
CpG GC>AT	-0.10	0.028	**	-	-	-	-0.13	0.027	***
GC>AT	-0.14	0.021	***	-	-	-	-0.17	0.021	***
GC>TA	-0.075	0.039		-	-	-	-0.11	0.039	*
GC>CG	-0.18	0.037	***	-	-	-	-0.20	0.037	***

Table 3.7: Comparison of Univariate and Multivariate Logistic Regression Results for Common Variants

β coefficients, standard error (in parenthesis), and significance GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

Variant Subtype	Model								
	Univariate			GC + Recombination			GC + DTH		
	GC Content								
Total	0.045	0.060		-0.12	0.064		0.011	0.060	
AT>GC	0.27	0.12		0.024	0.13		0.21	0.12	
AT>CG	0.27	0.23		-0.0077	0.25		0.17	0.23	
AT>TA	-1.85	0.29	***	-1.82	0.31	***	-1.86	0.29	***
CpG GC>AT	-4.18	0.17	***	-4.42	0.18	***	-4.19	0.17	***
GC>AT	-1.47	0.12	***	-1.50	0.13	***	-1.49	0.12	***
GC>TA	-2.016	0.22	***	-1.97	0.23	***	-2.03	0.22	***
GC>CG	-0.71	0.20	**	-0.85	0.22	***	-0.74	0.21	**
	Recombination Rate								
Total	0.29	0.040	***	0.32	0.042	***	-	-	-
AT>GC	0.45	0.075	***	0.44	0.080	***	-	-	-
AT>CG	0.49	0.15	**	0.49	0.16	*	-	-	-
AT>TA	-0.48	0.18	*	-0.054	0.20		-	-	-
CpG GC>AT	0.0034	0.11		0.61	0.11	***	-	-	-
GC>AT	-0.24	0.078	*	0.067	0.082		-	-	-
GC>TA	-0.50	0.15	**	-0.087	0.15		-	-	-
GC>CG	0.12	0.13		0.29	0.14		-	-	-
	DTH								
Total	-0.068	0.011	***	-	-	-	-0.068	0.011	***
AT>GC	-0.10	0.021	***	-	-	-	-0.098	0.021	***
AT>CG	-0.16	0.040	***	-	-	-	-0.16	0.040	**
AT>TA	0.029	0.050		-	-	-	-0.0096	0.051	
CpG GC>AT	-0.036	0.031		-	-	-	-0.052	0.031	
GC>AT	-0.024	0.022		-	-	-	-0.046	0.022	
GC>TA	0.014	0.041		-	-	-	-0.018	0.041	
GC>CG	-0.066	0.037		-	-	-	-0.076	0.037	

Table 3.8: Comparison of Univariate and Multivariate Logistic Regression Results for Substitutions

β coefficients, standard error (in parenthesis), and significance for GC content, recombination rate, and DTH. Results are shown for univariate and multivariate logistic regression models. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

Variant Type	Model					
	Univariate Model			Multivariate Model		
	GC Content					
Total	0.68	(0.069)	***	0.86	(0.072)	***
AT>GC	-1.05	(0.15)	***	-1.01	(0.15)	***
AT>CG	-0.56	(0.29)		-0.57	(0.29)	
AT>TA	-0.98	(0.32)	*	-0.98	(0.32)	*
CpG GC>AT	-2.64	(0.17)	***	-1.42	(0.20)	***
GC>AT	0.024	(0.14)		0.22	(0.14)	
GC>TA	-0.80	(0.25)	*	-0.81	(0.26)	*
GC>CG	-0.53	(0.24)		-0.44	(0.25)	
	Coverage					
Total	6.39E-03	(8.32E-04)	***	8.72E-03	(8.51E-04)	***
AT>GC	9.02E-03	(1.74E-03)	***	8.22E-03	(1.75E-03)	***
AT>CG	-2.88E-05	(3.41E-03)		-6.18E-04	(3.43E-03)	
AT>TA	1.58E-03	(3.72E-03)		6.91E-04	(3.75E-03)	
CpG GC>AT	3.41E-02	(1.82E-03)	***	2.65E-02	(2.11E-03)	***
GC>AT	6.31E-03	(1.58E-03)	***	7.08E-03	(1.66E-03)	***
GC>TA	2.42E-03	(2.87E-03)		-3.88E-04	(3.02E-03)	
GC>CG	5.07E-03	(2.74E-03)		3.54E-03	(2.88E-03)	

Table 3.9: Comparison of Rare Variant Regression Results for Univariate and Multivariate GC Content and Coverage Regressions

β coefficients, standard error (in parenthesis), and significance from the univariate regression models for GC content and coverage and multivariate model, using GC content and coverage as covariates in the regression model. ***p-value < 0.0001, **p-value < 0.001, *p-value < 0.01.

CHAPTER 4

SubSim: A Forward Genetic Simulation Program To Model Variant Subtype-Specific Mutation and Selection

4.1 Introduction

Simulations are used for a variety of purposes: to model the evolution and history of complex traits, to simulate populations under known conditions in order to test new statistical genetics methodologies, to estimate population genetic parameters, and to test theories relating to the complex interplay of different evolutionary forces in shaping patterns of genetic variation (Carvajal-Rodriguez 2008). Simulation-based approaches are useful for several reasons. First, it can be much more efficient to study evolution in simulated populations, especially in species with relatively long time spans between generations. Second, population-based data derived from simulations are useful for methodological development and testing because the conditions and parameters used to generate the data are known. Different simulation programs offer a variety of parameters that can be fine-tuned by users to suit their specific needs, with each offering its own unique combination of modeling capabilities. Two reviews of the available simulation software were recently published showing just how vast the pool of available tools is (Hoban et al. 2011; Yuan et al. 2012).

The two main methods for performing population genetic simulations are backward- and forward-in-time simulations. Backward-in-time simulations (called

coalescent simulations) are based on the coalescent theory and model the ancestry of a population backwards in time until the most recent common ancestor (MRCA) of all individuals in the population is identified (Kingman 1982). Commonly used coalescent simulation programs include ms (Hudson 2002) and SimCoal2 (Excoffier et al. 2000). Forward-in-time simulations take the opposite approach, simulating an initial population forward in time and tracking sequence variants as they rise and fall in frequency in the population (Hoban et al. 2011). Coalescent simulations only track the lineage of individuals present in the final population, and therefore can be run quickly with relatively small memory requirements. Forward simulations, however, track all individuals in every generation and require substantially more computational resources than coalescent simulations. Historically, coalescent simulations were favored due to their efficiency, although advances in computing technology have decreased the limitations that previously prevented advancement of forward simulation techniques. Over the past decade, many different forward simulation programs have been developed, including SFS_CODE (Hernandez 2008), GENOMEPOP (Carvajal-Rodriguez 2008), simuPOP (Peng and Kimmel 2005), and many more (Hoban et al. 2011).

Simulations can use a variety of mutation, selection, and demographic models. Mutation models, for example, can include the Jukes and Cantor model (Jukes and Cantor 1969), where all mutation types are equally likely, Kimura's 2 parameter model (Kimura 1980), which distinguishes between transition and transversion mutation rates, and Felsenstein's mutation model (Felsenstein 1981), which takes different base frequencies into account. Different simulation programs model selection acting on single

or multiple loci, typically modeling either deletions or beneficial variants in a population. Parameters can also often be adjusted, such as population growth and migration, to specify the demographic history of the resulting simulated population. Several forward simulation programs, such as FREGENE (Chadeau-Hyam et al. 2008) and GenomePop (Carvajal-Rodriguez 2008) use a resampling technique to improve efficiency. This scales the population size and the number of generations down by a specified degree, while scaling the mutation rate accordingly so that similar numbers of mutations are introduced into the population, even with reduced sample size and overall time.

In Chapter 3, I analyzed rare variants, common variants, and substitutions to determine the degree to which genomic context influences both mutation and fixation probabilities in the human genome. My results showed that GC content impacts rare variants, common variants, and substitutions (Figure 3.3), although the extent to which individual variant subtypes were affected among these three variant classes was different. This suggested that local nucleotide composition affects both mutation and fixation biases. Recombination rate, however, only altered patterns of common variants and substitutions, with much smaller effects on rare variants (Figure 3.4), suggesting that recombination primarily alters variant fixation rates. Specifically, AT>GC (an A ancestral base pair converted to a G, or the reciprocal T ancestral base to a C) and AT>CG common variants and substitutions were more strongly affected by recombination rate (Figure 3.4), which is consistent with biased gene conversion (BGC) increasing the fixation bias of these variants in regions of the genome with high recombination rates (Duret and Galtier 2009).

The effect of genomic context on patterns of variation in humans has been previously studied using computer simulation. Several *in silico* studies have shown that altering selection or BGC with recombination rate can mimic patterns of genetic variation observed in human populations (Charlesworth et al. 1995; Duret and Arndt 2008; Lohmueller et al. 2011). These studies, however, have not jointly modeled both mutation and fixation bias simultaneously. I sought to understand the degree to which mutation rates and fixation biases fluctuate in response to the local genomic context using forward simulations. Although there are a large number of coalescent and forward simulation programs available, none of them allows the user to define the base composition of the starting population, or subtype-specific selection coefficients. Compared to the coalescent, forward simulations are better suited to model selection. Therefore, I developed a forward simulation program that models these specific events. I had three goals for developing my simulation program. First, being able to model selection on specific variant subtypes was necessary for analyzing subtype-specific fixation biases. Second, I needed to be able to alter variant-subtype specific mutation bias in order to further understand the degree to which mutation bias on specific subtypes impacts variant patterns in humans. Finally, the ability to model the base composition of the ancestral chromosome, as well as the recombination rate, was important to understanding the effects that these features of the genome have on mutation and fixation rates. With these goals in mind, I developed a forward genetic simulation program, SubSim, which allows the user to alter GC content, recombination rate, subtype-specific mutation rates and selection coefficients, among other available parameter settings (Table 4.1). An overview of the basic simulation process is

presented in Figure 4.1. SubSim has many potential applications, including the study of mutation or selection bias in favor of or against specific variant subtypes, the effect of BGC on genome evolution, and the different effects that these processes have as GC content and recombination rate varies.

4.2 Methods and Implementation

4.2.1 Simulation Overview

SubSim has two basic steps: (1) generate a set of starting populations at mutation-drift equilibrium (the "burn-in" period) and (2) simulate populations forward-in-time from the starting population over a specified number of generations to output the final populations (Figure 4.1). The program simulates a modified version of a Wright-Fisher population, in which the number of individuals in each generation is constant, the individuals are diploid, and the generations are non-overlapping. Recent evidence in human population-based studies suggests that human populations have experienced exponential growth in recent history (Coventry et al. 2010). Although this is biologically more realistic, modeling growing populations is computationally inefficient and therefore, I modeled populations of constant size in order to improve efficiency.

The first step in the simulation is to generate a starting Wright-Fisher population under mutation-drift equilibrium. First, an identical chromosome is simulated for all N individuals (2N chromosomes). The user-defined GC content and the length of the chromosome, L, determine the sequence content of the starting chromosome. For each base, a uniform random variable is used to sample one of the four nucleotides, A, T, G or C, weighted by the user-defined GC content. This chromosome is copied 2N times

(for a diploid population) to generate the initial ancestral population at generation 0. After the initial chromosomes are simulated, the burn-in period consists of a user-defined number of generations to generate the staring population of chromosomes. The burn-in period in the simulations generates a realistic population containing genetic variation in a simulated chromosomal segment derived from a single original ancestral sequence.

After the burn-in period, the second step of the simulation is to simulate forward-in-time from the starting populations to generate the final population. These simulations are run using any user-defined changes to the simulation parameters. At the end of the user-specified number of generations, a final population is output, along with a file containing the location and the individuals with polymorphic sites, the ancestry information of all variant sites that existed in the population over the course of the simulation, and the chromosomes of each individual in the final population.

Each generation in the simulation (both the burn-in and simulating the final populations) goes through three steps. First, new individuals are formed based on a random selection of parents from the previous generation (section 4.2.2). Then, parental chromosomes are recombined and a single recombinant or non-recombinant parental chromosome is chosen from each parent to form the new individuals (section 4.2.3). Finally, single base pair mutations are introduced into the new generation (section 4.2.4).

The user can specify several parameters in either the burn-in or the test stage of SubSim, including the mutation rate, recombination rate, the rate of BGC, subtype-

specific selection coefficients, etc. For a more detailed description of the functions of the simulator, see below. For a list of all available parameter options, see Table 4.1. I wrote SubSim in a Linux/Unix environment in Python.

In order to efficiently test the functionality of SubSim, the simulations presented in the results section use the specified user-defined parameters described in each section throughout the entire simulation. For each test of the simulation functionality, each simulated population was run over the specified number of generations using the parameters described over the entire course of the simulation. Therefore, the populations in the results section are the result of running the burn-in simulations using the specified parameter settings, not subsequent test populations.

4.2.2 Parent Selection

To create each new generation, two parents are selected from the previous generation to obtain the chromosomes for each individual. Therefore, 2N total parents are drawn (with replacement) from the previous generation to generate N individuals. The two parents are independently drawn for each individual in the population. Each individual in the parental generation has a specific probability of being chosen as a parent, which is determined by the relative fitness of their genotypes. If each individual in the parental generation has fitness equal to 1, i.e. there is no selective effect (positive or negative) from any variant, then each parent is equally likely to be chosen, with probability $\frac{1}{N}$. If, however, selection is acting on the variants segregating in the population, then each individual has fitness,

$$f_i = \prod_{2l}(1 + s_l) \qquad \textbf{4.1}$$

where s is the selection coefficient across all l loci on each of the simulated chromosomes. In these simulations, selection is multiplicative, meaning that the fitness of the individual is the product of the fitness across all loci. When $s \neq 0$, the fitness of an individual in the population, i, is not necessarily the same as other individuals in the population. Sampling parents, therefore, is weighted based on their value of f_i. To perform this weighted sampling, I first calculate the relative fitness of each individual, defined as,

$$f_i' = \frac{f_i}{\sum_N f_N} \qquad \textbf{4.2}$$

where f_i' is the normalized fitness for individual i. The variable f_i' is calculated across all individuals, N, in the parental generation. I use this to take a weighted random sample (with replacement) of the parental generation with weight f_i' to sample parents for each individual in the offspring generation. When sampling parents for an individual, if the same individual is sampled for both parents, the second parent is resampled so that the two parents are different.

Sampling parents for many individuals requires iterative list searching, which can be extremely inefficient. For example, if $x = 0.99999$, $N = 10{,}000$, and $f_i = 1$ for all individuals, then the program must search 9,999 entries until it identifies individual 10,000 as the correct parent. To account for this, I applied a previously developed search algorithm, binary tree sampling (Gilberg and Forouzan 2001), to improve computational efficiency. The binary tree uses a series of if/else statements to identify the region of the list containing the appropriate element. In these simulations, the binary

tree sampling first determines whether the random variable is greater or less than a set number. For the first binary tree, this number is always 0.5. If $x > 0.5$, then the location of the appropriate index in the cumulative fitness vector will be in the latter half of the list. The next binary tree determines if $x > 0.75$ or if $x < 0.75$. Each branch of the binary tree essentially shrinks the indices in the cumulative fitness vector by half (Figure 4.2).

For simulations in which $s = 0.0$, 11 iterations of the binary tree led to the greatest improvement in efficiency (Table 4.2). After 11 iterations of the binary tree, the maximum number of list entries that need to be searched is 3. For a cumulative fitness vector with equally spaced entries (i.e. no selection occurring in the population), then the index identified by the binary tree method will be 1.5 entries away from the parental index, on average. To select 100,000 individuals, the starting index of the list search was on average 1.22 ± 0.81 entries from the appropriate individual index entry.

There is no family structure assumed in the simulations, and parents for each individual are chosen independently. If $s = 0.0$, then the probability of choosing any parental individual is $\frac{1}{N}$. Each individual is required to have two distinct parents. Therefore, the probability of choosing the second parent is $\frac{1}{N-1}$. The probability that an individual has parents i and j is $\frac{1}{N} \times \frac{1}{N-1}$. From this, the probability that two individuals in the simulation are full siblings is $\left(\frac{1}{N} \times \frac{1}{N-1}\right)^2$. When N is sufficiently large, this probability is low. For example, when $N = 10{,}000$, the probability that any two individuals are full siblings is 1.0002×10^{-16}.

4.2.3 Recombination

After each set of parents is selected to generate a new individual, the parental chromosomes are allowed to recombine to form the gametes that will eventually give rise to offspring. The number of recombination events that occur in the parents of the given individual follows a Poisson distribution. SubSim models uniform recombination, as opposed to recombination focused in hotspots. The mean number of recombination events occurring in the parents, then, is equal to the number of base pairs where recombination could occur multiplied by the recombination rate: $\lambda = 4Lr$, where r is the per-site per-generation recombination rate and L is the length of the simulated chromosome. By default, $r = 1 \times 10^{-8}$ per-base per-generation (Kong et al. 2002). The maximum number of recombinations that can occur is 2, one in each parent; the minimum is 0.

The first step in recombination is the formation of a double-strand break (DSB). The two parental chromosomes are randomly assigned as the acceptor and the donor chromosome (Figure 4.3A). The acceptor chromosome is the chromosome that is damaged by the DSB and the donor chromosome is used as the template to repair the DSB on the acceptor chromosome (Figure 4.3B). The nucleotide position of the double strand break in the paternal chromosome is determined by a random variable following a uniform distribution across the length of the simulated chromosome. Next, two random variables are drawn from a geometric distribution (used in SFS_CODE (Hernandez 2008)) to identify the length of the resection in the acceptor chromosome (Figure 4.3C). Following invasion of the 5’ strand of the acceptor chromosome (Figure 4.3D), the double Holliday Junction (DHJ) is formed (Figure 4.3E). Resolution of the DHJ can

occur in four ways, each of which is equally likely in the simulation (Figure 4.3F). Two of these resolutions result in a crossover and two result in a non-crossover event. Crossovers result in a more drastic alteration of the DNA sequence, as a larger amount of DNA is exchanged between the two chromosomes (Figure 4.3F).

An important feature of SubSim is the ability to determine how BGC alters the DNA. BGC is a recombination-associated process where mismatches formed during recombination are preferentially repaired to GC bases versus AT bases (Duret and Galtier 2009). At the beginning of the simulation, the user can specify the degree of bias in the repair process, B_{BGC}. For $B_{BGC} = 1$, the mismatch is always repaired to the GC base, if $B_{BGC} = 0$, the mismatch is always repaired to the AT base, and if $B_{BGC} = 0.5$ (default), repair to the GC or AT bases is equally likely. BGC occurs in either crossover or noncrossover events.

In order to improve computational efficiency, recombination only occurs between chromosomes with heterozygous loci. At homozygous sites, any exchange of DNA will not result in an altered genotype.

After recombination occurs in each of the parents, the gametes are chosen at random to produce the diploid offspring individual. One gamete from each parent is randomly chosen to make up the 2 chromosomes of the offspring individual.

4.2.4 Mutation

SubSim uses three different mutation models: equal mutation rates, where all mutation subtypes are equally likely (Table 4.3), transition-biased mutation rates, where

the transitions (Ti) to transversion (Tv) ratio (Ti/Tv) is 2.0 (Table 4.4), or a user-specified mutation rate option (Table 4.5). By default, mutation is Ti-biased.

For the equal and Ti-biased mutation models, the probability that a mutation originates at any base is equal (Table 4.3, Table 4.4). A Poisson distribution is used to determine the overall number of mutations for equal and Ti-biased mutation in each generation. The expected number of mutations in any generation is simply the number of possible base positions multiplied by the per-base per-generation mutation rate: $\lambda = 2NL\mu$, where μ is the per-site per-generation mutation rate. By default, $\mu = 1.2 \times 10^{-8}$, which has been reported in several recent publications using trio-based sequencing to identify *de novo* mutation in humans (Conrad et al. 2010; The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012). Mutations occur at random in the population. The chromosome and the nucleotide where the mutation occurs are also randomly chosen. Once the position of the mutation is identified, the mutation subtype is based on the nucleotide at that position and the relative probabilities of each mutation subtype that could occur from that allele.

For the user-specified mutation rate, the probability of each mutation subtype occurring is entered at the beginning of the simulation (Table 4.1, Table 4.5). The sum of these probabilities equals the total per-base per-generation mutation rate. Here, the probability of a mutation occurring at an AT or a GC base can be different, and therefore the number of AT mutations must be modeled separately from the number of GC mutations. The number of AT and GC mutations each follows a Poisson distribution with parameters, $\lambda_{AT} = L_{AT} \sum_{i=1}^{6} x_i$ and $\lambda_{GC} = L_{GC} \sum_{i=7}^{12} x_i$, where L_{AT} and L_{GC} are the number of AT or GC bases in the selected chromosome. The individual and chromosome in

which each mutation occurs is identified at random. The GC or AT base in the chromosome that experiences the mutation event is determined by a uniform random variable across all GC or AT bases, depending on the type of mutation that occurs.

4.2.5 Subtype-Specific Selection

Natural selection on a genetic variant changes the fitness of the individual, based on their genotype at that site. Individuals with variable fitness are more or less likely to contribute offspring to the subsequent generation. When fitness, f_i, is >1, individuals are more likely to have offspring and pass on their genetic material, whereas individuals with $f_i < 1$ have a lower chance of contributing offspring to the next generation. If $f_i = 1$ for all i, then every individual in the population has an equally likely chance of producing offspring. The variable f_i is calculated according to equation 3.1, where s_l is the selection coefficient (s) for the observed variant subtype at the locus, l. When $s < 0$, purifying selection is acting on the given variant and f_i will decrease. If $s > 0$, then the variant is under positive selection and f_i will increase. Here, the user can specify s at the beginning of the simulation. The simulation program processes subtype-specific s, allowing the user to set specific s for AT>GC, AT>CG, AT>TA, GC>AT, GC>TA, and GC>CG variants separately (Table 4.1). By default, $s = 0$ for all variant subtypes (Table 4.1).

4.2.6 Testing Neutrality

I used several different measures to test that the simulations reached mutation-drift equilibrium. I calculated the number of observed segregating sites, S, the observed number of haplotypes segregating in the population, k, and the nucleotide diversity, π, given by the equation,

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij} \qquad \textbf{4.3}$$

(Nei and Li 1979). In equation 4.3, π_{ij}, is the number of sequence differences observed between sequences i and j and n is the total number of sequences in the sample. S and π both measure the amount of variation present in a population, although they are weighted toward variants at different allele frequencies. In a population without natural selection, S is driven entirely by the mutation rate and the population size, and captures the extent of DNA variation at all allele frequencies, although it is weighted more strongly toward rare variants. On the other hand, π is weighted more heavily toward variants present in many chromosomes in the population, which increases the number of pairwise sequence differences. Under mutation-drift equilibrium, the expected values of π and S are given by,

$$E[S] = \theta a_1 \qquad \textbf{4.4}$$

and

$$E[\pi] = \theta, \qquad \textbf{4.5}$$

where $\theta = 4N\mu L$ and $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$. The expected values of both S and π contain θ and can, therefore, be compared to test the assumption of neutrality using Tajima's D (Tajima 1989),

$$D = \frac{\pi - \frac{S}{a_1}}{\sqrt{Var\left[\pi - {}^{S}/_{a_1}\right]}} \qquad \textbf{4.6}$$

where $Var\left[\pi - {S}/{a_1}\right] = \frac{S}{a_1}\left(\frac{n+1}{3(n+1)} - \frac{1}{a_1}\right) + \frac{S(s-1)}{a_1+a_2}\left(\frac{2\left(n^2+n+3\right)}{9n(n-1)} - \frac{n+2}{na_1} + \frac{a_2}{a_1^2}\right)$, $a_1 = \sum_{i=1}^{n-1}\frac{1}{i}$, and $a_2 = \sum_{i=1}^{n-1}\frac{1}{i^2}$. A population is at neutrality when $D = 0$.

The expected number of haplotypes, k, in a population can be estimated using Ewen's Sampling Formula (Ewens 1972), given θ and the number of sequences in the sample, n,

$$E[k] = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1} \qquad \textbf{4.7}$$

(Hartl and Clark 1997). This equation shows that the number of haplotypes segregating in a population is determined primarily by the mutation rate and the population size. Large populations or large mutation rates will increase the expected number of haplotypes.

I also analyzed the observed site frequency spectrum (SFS) in the simulated populations. S and π are both weighted summaries of the SFS, with S being more heavily weighted toward rare variants and π weighted toward more common variants. The direct analysis of the SFS, however, describes the amount of variation in a population across all allele frequencies. Under neutrality, the number of alleles segregating on i sequences, ξ_i, is expected to be $E[\xi_i] = \frac{\theta}{i}$ (Wakeley 2009). As $E[\xi_i]$ is influenced by μ and N, higher values of μ or N will lead to more variants segregating at each allele frequency.

4.3 Results

4.3.1 Simulation Efficiency

In order to test efficiency, I ran the simulations using a range of sample sizes and a range of chromosome lengths (Table 4.6). One simulation was run for each combination of sample size and chromosome length over 1,000 generations. All other parameters were set to the default values (Table 4.1). The run-time increases with sample size faster than an exponential function. The same increase in time is seen as the length of the simulated chromosome increases.

4.3.2 Under Default Parameter Settings, the Variants Follow the Expected SFS and the Simulations Reach Expected Values of S, π, and k Under Neutrality

I ran 100 simulations with $N = 10{,}000$ and $L = 1kb$ under default parameter settings (Table 4.1) for 60,000 generations to evaluate the function of SubSim. In the final population (generation 60,000), the expected values of S, π, and k fell within the distribution of the observed values across the 100 simulated populations (Figure 4.4A-C). The expected values of S, π, and k were calculated according to Equations 4.4, 4.5, and 4.7, respectively. The observed value of π in each generation was calculated according to Equation 4.3. I also analyzed the average observed SFS across the 50 simulated populations and compared it to the expected SFS. The SFS analyzes the allele frequency of the derived alleles at polymorphic sites in the population. At all allele frequencies, there was no difference between the distribution of the observed values and those expected under neutrality (Figure 4.4D). After 60,000 generations, there was an average of 0.27 fixed derived alleles in each of the 100 simulated populations, likely due to the fact that these populations were simulated over 60,000 generations. On

average, the simulations reached expected values of S, π, k, and Tajima's D at roughly generation 20,000 (Figure 4.5). Comparing the SFS observed in generations 20,000 and 40,000 to the final generation (60,000) clearly shows that the number of fixed differences increases steadily, as expected, whereas other values in the SFS remain relatively unchanged (Figure 4.6).

4.3.3 Increasing the Recombination Rate Increases the Number of Haplotypes Segregating in Simulated Populations

In SubSim, recombination occurs between parental chromosomes according to the recombination rate, r. When recombination occurs between non-identical sequences, it increases the number of unique haplotypes segregating in the population. Ewen's Sampling Formula (Ewens 1972) can be used to estimate the expected number of haplotypes segregating in a population (Equation 4.7). This equation, however, does not take recombination into account. When r and μ are high, Equation 4.7 will underestimate the number of haplotypes. As there is no simple closed-form solution to estimate the expected number of haplotypes in a population with recombination, I employed a widely-used coalescent simulation program, ms (Hudson 2002), to estimate the number of haplotypes observed in populations with varying levels of recombination.

I ran three sets of simulations using SubSim: $r = 0.0$, $r = 1 \times 10^{-8}$, and $r = 1 \times 10^{-6}$ pe-base per-generation. I simulated 50 populations with 10,000 diploid individuals and 1 kb chromosomes for each recombination rate over 60,000 generations. All other parameters were set to default. 1,000 populations were simulated in ms (Hudson 2002) for each recombination rate using the same settings for the mutation rate, sample size, and chromosome length. The distribution of the haplotype

number between the populations simulated in ms and those simulated in SubSim overlap at all values of r (Figure 4.7). As r increases from 1×10^{-8} (Figure 4.7B) to 1×10^{-6} (Figure 4.7C), the range of observed haplotypes increases in both ms and SubSim, as expected.

4.3.4 Simulations with A Variety of Mutation Rates Result in Subsequent Changes in S, π, and k

I performed several tests to ensure that the simulation software would respond accurately to changes in the user-defined inputs. One of the goals in developing these simulations was to be able to manipulate the overall mutation rate in response to the local GC content and recombination rate. I ran simulations using three different mutation rates: $\mu = 1.2 \times 10^{-9}$, $\mu = 1.2 \times 10^{-8}$, and $\mu = 1.2 \times 10^{-7}$ per-base per-generation. Subtype-specific mutation rates were equal in each simulation. For each of the three mutation rates tested, I simulated 50 populations with $N = 10{,}000$ and $L = 1kb$ over 60,000 generations. All other parameters were set to the default values (Table 4.1).

I calculated S, π, and k as before in the final generation (60,000 generations total) in each of the 50 replicate populations across the three mutation rates. The expected values of S, π, and k all depend on μ, and therefore, their expectations will change with the changing mutation rate. For example, at a very low mutation rate ($\mu = 1.2 \times 10^{-9}$), $E[S] = 0.50$ for a 1 kb sequence in 10,000 diploid individuals. When the mutation rate increases to $\mu = 1.2 \times 10^{-8}$ or $\mu = 1.2 \times 10^{-7}$, $E[S] = 5.03$ or $E[S] = 50.3$, respectively. In the simulated populations, the number of segregating sites increases 10-fold with the increasing mutation rate: I observed an average of 0.28 ± 0.57 segregating sites across the 50 populations when $\mu = 1.2 \times 10^{-9}$ (Figure 4.8B),

4.92 ± 1.85 at $\mu = 1.2 \times 10^{-8}$ (Figure 4.9B), and 48.82 ± 8.26 when $\mu = 1.2 \times 10^{-7}$ (Figure 4.10B). There is also a corresponding increases in k and π with the increasing mutation rates (Figure 4.8 - Figure 4.10 A and C).

For each of the mutation rates, I analyzed the average of the SFS observed across the 50 simulations. The observed SFS for the simulations with $\mu = 1.2 \times 10^{-8}$ and $\mu = 1.2 \times 10^{-9}$ matched the expected values (Figure 4.9D and Figure 4.10D). The low mutation rate simulations, however, had fewer rare variants (derived allele frequency (DAF) ≤ 0.05) than expected. The mutation rate in these simulations is very low, and each iteration is only expected to have ~0.5 variant sites. In the low mutation rate simulations, only 11 iterations had at least one segregating site in the final population. These simulations likely do not match the expected SFS because there are too few to data points to obtain a precise estimate. Due to the time restraints in generating more populations, however, I did not pursue this further.

4.3.5 Introducing Subtype-Specific Mutation Bias Generates Expected Subtype SFS Patterns

Another goal in generating this simulation software was to be able to change the subtype-specific mutation rates in different simulated genomic contexts to understand the degree to which biasing mutation rates can lead to the observed patterns of rare variants observed previously (Chapter 3). I tested this using two separate simulations: one with mutation rates biased toward AT>GC and AT>CG (W>S) mutations and another with mutation rates biased toward GC>AT and GC>TA (S>W) mutations. For the W>S biased simulations, $\mu = 5 \times 10^{-9}$ per-base per-generation for AT>GC and AT>CG mutations and $\mu = 5 \times 10^{-10}$ for all other mutation subtypes. The S>W biased

simulations had $\mu = 5 \times 10^{-9}$ per-base per-generation for GC>AT and GC>TA mutations and $\mu = 5 \times 10^{-10}$ for all other mutation subtypes. All other parameters were set to the default values (Table 4.1). I simulated 98 populations with W>S mutation bias and 99 populations with S>W mutation bias. Two W>S and one S>W simulation failed due to server error. Each set of simulated populations had $N = 10{,}000$, $L = 1kb$ and was simulated over 60,000 generations.

Overall, the expected values of S, π, k, and the SFS fall within the observed distributions for both the W>S (Figure 4.11) and the S>W (Figure 4.12) mutation bias simulations. The mutation rate for the W>S mutations in the W>S biased simulations is 10x higher than the S>W mutation rate. As expected, S for the W>S sites is roughly 10x greater than S for the S>W sites: on average, there were 4.53 (± 2.05) W>S and 0.45 (± 0.66) S>W sites across the 98 W>S mutation bias populations. The opposite pattern was observed for the S>W biased simulations: there were 4.22 (± 2.15) S>W sites compared to 0.42 (± 0.65) W>S. For both sets of simulations, the observed SFS for both W>S and W>S variants roughly follows the expected SFS pattern (Figure 4.13 and Figure 4.14).

4.3.6 Subtype-Specific Selection Results in Subtype-Specific Deviations from Neutrality

My previous work on common variants and substitutions suggested that fixation patterns of variants are altered in response to the local genomic context (Chapter 3). As part of the simulation development, one of the major goals was to be able to alter subtype-specific selection coefficients in order to simulate an increase or decrease in fixation bias. To test the functionality of the subtype-specific selection function, I ran 2

sets of simulations: one used selection in favor of S>W variants and the other used selection in favor of W>S variants. The S>W biased selection was run with $s = 0.001$ for GC>AT and GC>TA variants, with $s = 0$ for all other variant subtypes. For W>S biased selection, $s = 0.001$ for AT>GC and AT>CG variants and $s = 0$ for all other subtypes. The value of $s = 0.001$ is within the range of selection values estimated in humans (Boyko et al. 2008). This relatively low selection coefficient allows me to see if the simulations elicit a response, even without a very large selective effect. I simulated 50 populations with $N = 10{,}000$ and $L = 1kb$ over 60,000 generations for each scenario.

The results from the S>W and W>S selection simulations are shown in Figure 4.15 and Figure 4.16, respectively. The simulations with positive selection for S>W variants shows an excess of S>W variants across nearly all allele frequencies > 0.05 (Figure 4.15), with a very large increase in the number of S>W fixed differences. W>S variants in these simulations were marginally affected, and show the pattern expected under neutrality (Figure 4.15). The opposite pattern was observed for W>S variants, with an excess of W>S variants at nearly all allele frequencies > 0.05 and the expected number of S>W variants under neutrality (Figure 4.16). Together, these results show that the simulation program responds by increases in the frequency and number of variants under positive selection, as expected, and will allow for modeling fixation biases of variant subtypes in future studies.

4.4 Discussion

Previously, I found that mutation and fixation bias are likely influenced by GC content and recombination rate, although I was unable estimate the extent of the bias

necessary to produce the patterns observed in rare variants, common variants, and substitutions (Chapter 3). As described in the introduction, forward simulation techniques are better suited to accurately model natural selection compared to coalescent methods. Therefore, I developed a flexible forward genetic simulation program to understand how genomic context influences mutation and fixation biases in the human genome.

My results show that the simulation program I developed works as expected, generating populations under neutral mutation-drift equilibrium. The simulations I presented using the default parameter settings generated values of S, π, and k that met expected values. The observed SFS matched the expected SFS for a population under mutation-drift equilibrium, indicating that overall, the simulations are functioning properly.

To test the goal of developing a forward simulation program that can alter variant subtype-specific mutation rates and selection coefficients, I ran a series of simulations analyzing how the program responds to changes in the different parameter settings. SubSim was able to generate populations that met expectation with changes in changes in the recombination rate, overall mutation rate, and subtype-specific mutation rates. These populations follow expectations from neutrality both in the overall degree of variant sites segregating in the final populations, but also followed expected values for the individual subtypes when mutation was biased toward a specific variant subtype. As expected, introducing subtype-specific selection resulted in populations that deviated from neutrality, since the expected values under neutrality are based on populations in the absence of selection. Because I simulated only short genomic segments (1 kb), I

cannot analyze any effect from background selection or selective sweep. Simulation of longer genomic segments is necessary to analyze the ability of the simulation tool to model these important evolutionary processes. Together, my results show that SubSim capable of producing populations under mutation-drift equilibrium, as expected, and can be used to further understand the degree to which subtype-specific mutation and fixation biases are present in the human genome.

Several recent review articles classify parameter options in available simulation software into distinct categories (Hoban et al. 2011; Yuan et al. 2012). Hoban and colleagues organized the modeling capabilities of both coalescent and forward simulations into 10 different groups: (1) spatially explicit considerations for population or individual modeling, (2) ways to model migration or dispersal, (3) mating system employed by the simulation tool, (4) fecundity, (5) life cycle, (6) population growth, (7) major events allowed to occur (colonization, extinction, population fission or fusion, etc.), (8) selection models used by software packages, (9) available mutation models, and (10) recombination models (Hoban et al. 2011). SubSim provides many of the capabilities, such as selection, recombination, and mutation. SubSim models individuals in a randomly mating isolated population. It does not model migration, dispersal, alternative mating strategies, fecundity, life cycle, model population growth or other major events. Mutation in SubSim can be user-specified, equal among all subtypes, or Ti-biased. While these options are available in other packages (Hoban et al. 2011), SubSim can model subtype-specific selection, a feature that is new to the body of available tools. Finally, SubSim models uniform recombination, given a user-defined recombination rate recombination, as opposed to recombination occurring in hotspots.

Although SubSim offers several options that previous programs did not, it still has several limitations. Because the hypotheses regarding genomic context and genome evolution do not deal specifically with population demography, SubSim does not model migration or changes in population size. Furthermore, SubSim tracks each nucleotide position in each diploid individual in the population over each generation, typically with very large sample sizes. Because of this, however, the simulations are somewhat slow, compared to other available forward simulation programs, such as SFS_CODE (Hernandez 2008) and simuPOP (Peng and Kimmel 2005). Future work to improve the computational efficiency would be necessary to rival the speed of other available programs.

SubSim can be used in the future to test specific questions regarding mutation and fixation bias of variant subtypes in response to GC content and recombination rate. There are, however, additional questions that could be asked using SubSim. In addition to modeling subtype-specific mutation and selection, SubSim can also model effects from BGC. Few available simulation programs model gene conversion events (Yuan et al. 2012)**.** SubSim therefore expands the available number of programs that can be used to examine the effect of BGC on genome evolution. Furthermore, BGC has been shown to mimic natural selection in empirical data (Berglund et al. 2009; Galtier et al. 2009). SubSim could be used to estimate the degree to which mismatch repair in gene conversion events must be biased toward GC bases in order to elicit a response in conventional tests for natural selection. SubSim models variant subtype-specific selection, as opposed to single or multi-locus selection. Therefore, it can also help to

understand selection on specific base pairs in the human genome, as has been suggested previously (Bernardi et al. 1985).

4.5 Conclusions

I developed a simulation program that allows users to define subtype-specific selection and mutation coefficients, alter rates of BGC, and set the base composition of the simulated sequence. The ability to manipulate these parameters will allow researchers to understand the extent to which bias in the mutation rate and selection leads to correlations between rare variants, common variants, and substitutions with GC content and recombination rate.

4.6 Figures

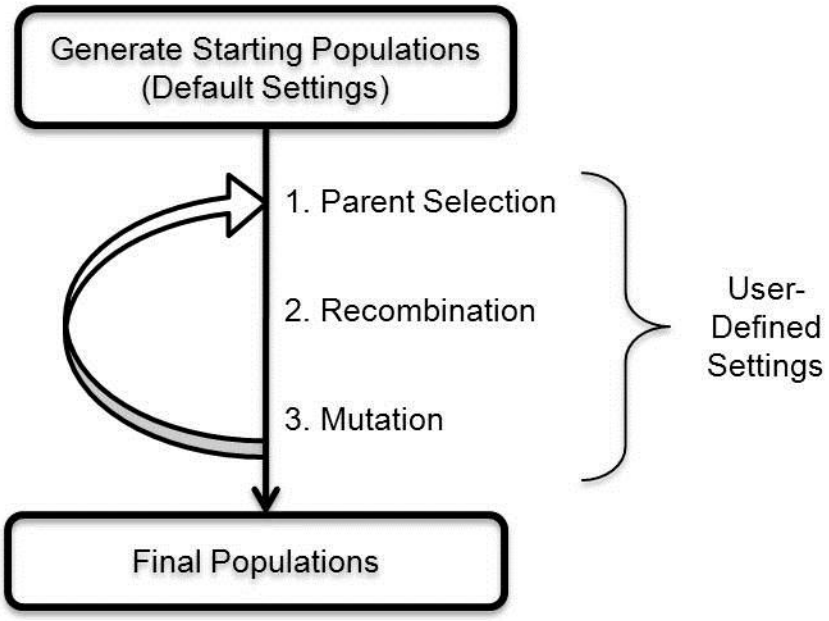

Figure 4.1: Simulation Overview

There are three main functions that occur in each generation of the simulation. First, parents are selected from the previous generation to provide the genetic material for the individuals in the generation. Next, recombination events occur randomly to the parental chromosomes and the individuals in the new generation are populated. Mutations occur randomly throughout the new generation. This occurs for a user-specified number of generations until the final population is produced.

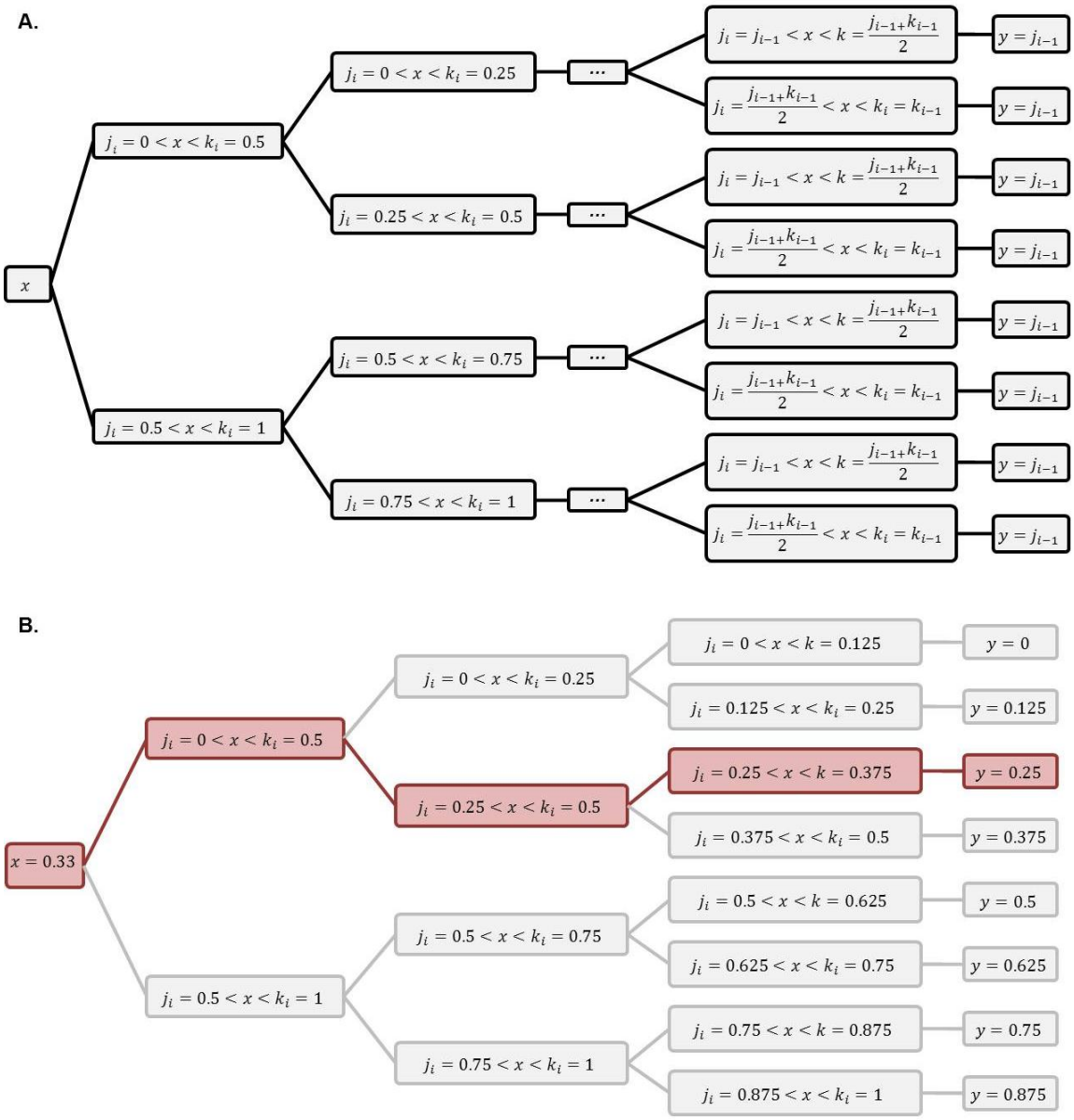

Figure 4.2: Binary Tree Scheme for Individual Selection

The binary tree sampling scheme used to select individuals for parent selection when generating offspring and for identifying individuals for mutation and recombination events. (A) Generic binary tree where x is the uniform random variable. Each level of the tree decreases the number of indices in the list that needs to be searched by half. The final variable, y, is the starting index for the subsequent list search. (B) An example of a binary tree with $i = 4$ levels. For $x = 0.33$, the starting index search point is $y = 0.25$ (red filled boxes).

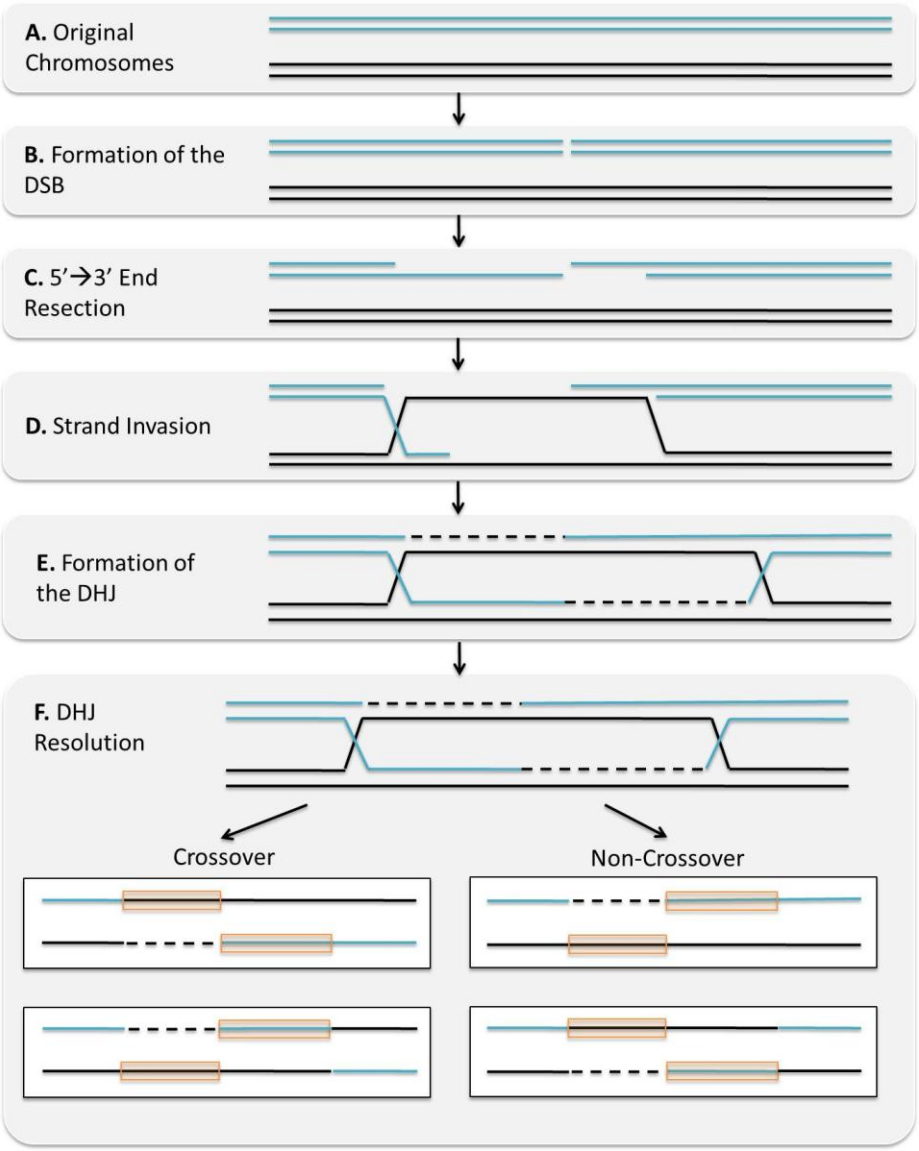

Figure 4.3: Homologous Recombination

Recombination schematic following the formation and resolution of a DHJ. (A) The acceptor chromosome, which is damaged by the DSB, is shown in blue and the donor chromosome, which is used to repair the DSB, is shown in black. (B) The position of the DSB is determined by a uniform random variable. (C) The length of the resection is determined by a Geometirc distribution. (D) The 5' end of the acecptor chromosome invades the donor chromosome and base pairs with the homologous sequence. (E) The DHJ is formed and the DNA lost in (C) is repaired using the homologous donor chromosome. (F) The DHJ is resolved either via crossover or non-crossover repair, resulting in 4 possible chromosomal configurations. The regions in orange are where mismatches can occur and be repaired, with or without BGC.

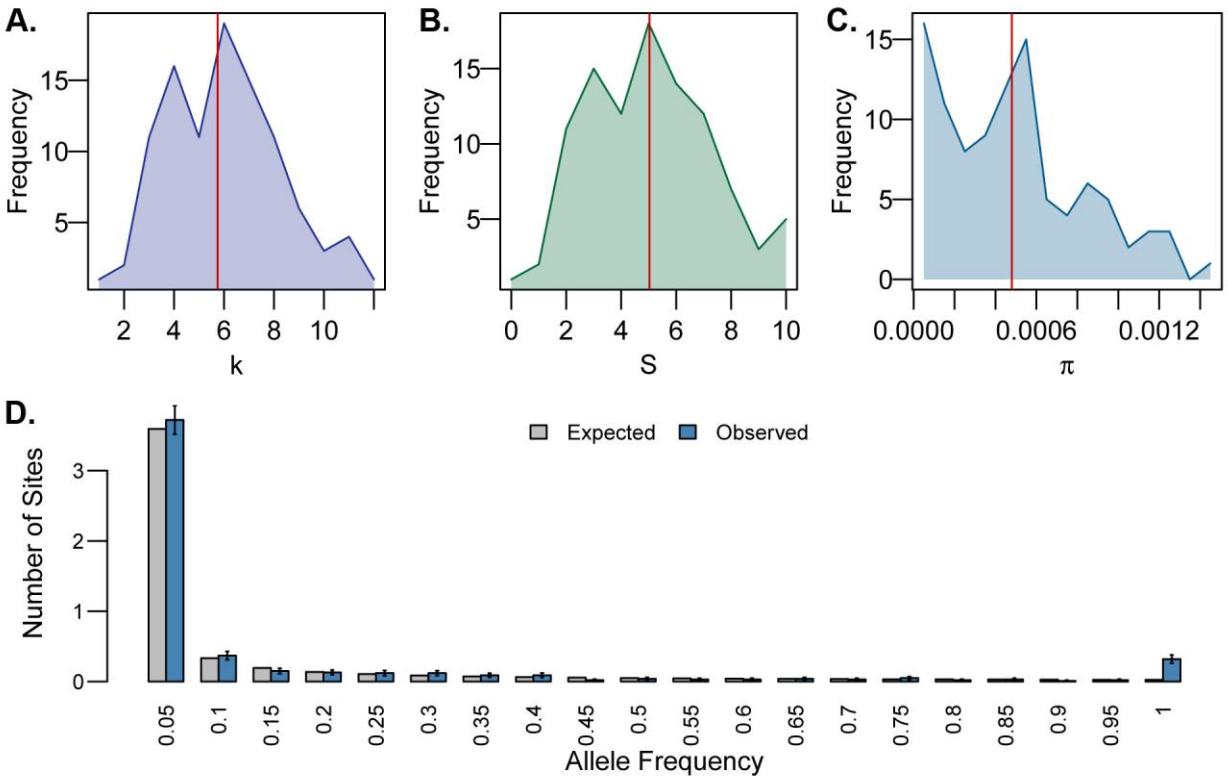

Figure 4.4: Final Population Summary Statistics for the Default Simulations

The distribution of k (A), S (B), and π (C) are shown for the final population across the 50 populations simulated under default settings. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 100 populations of the default simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. $0< x \leq 0.05$).

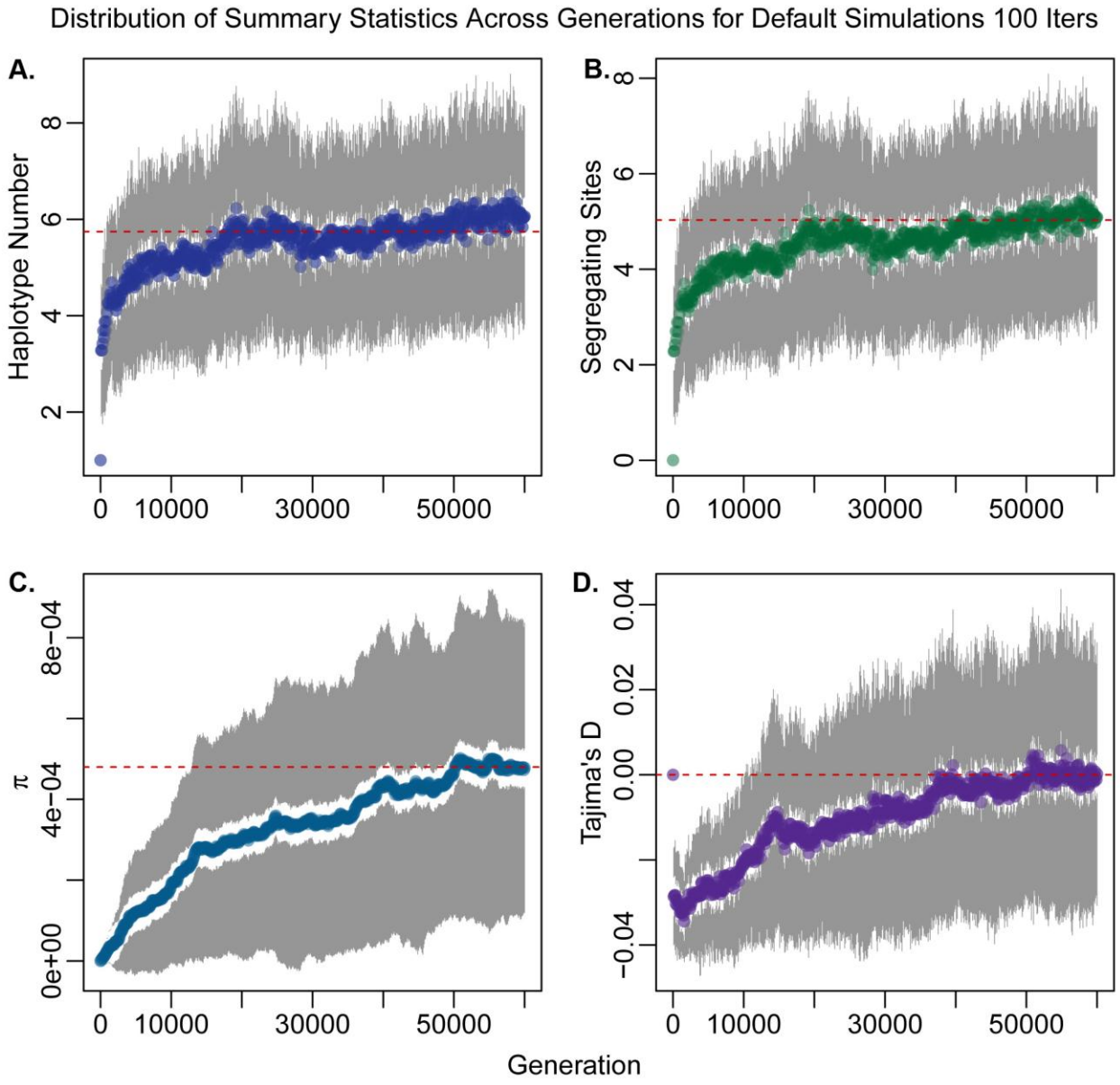

Figure 4.5: Distribution of S, π, k, and Tajima's D over Simulated Generations in the Default Simulations

The values of S, π, k, and Tajima's D were calculated every 100 generations over the course of the 60,000 generation default simulations. The average of k (A), S (B), π (C) and Tajima's D (D) across the 100 populations at each of the 100 generations is shown. The grey lines indicate the observed standard error across the 100 populations. Horizontal red dotted lines indicate the values expected under neutral mutation-drift equilibrium.

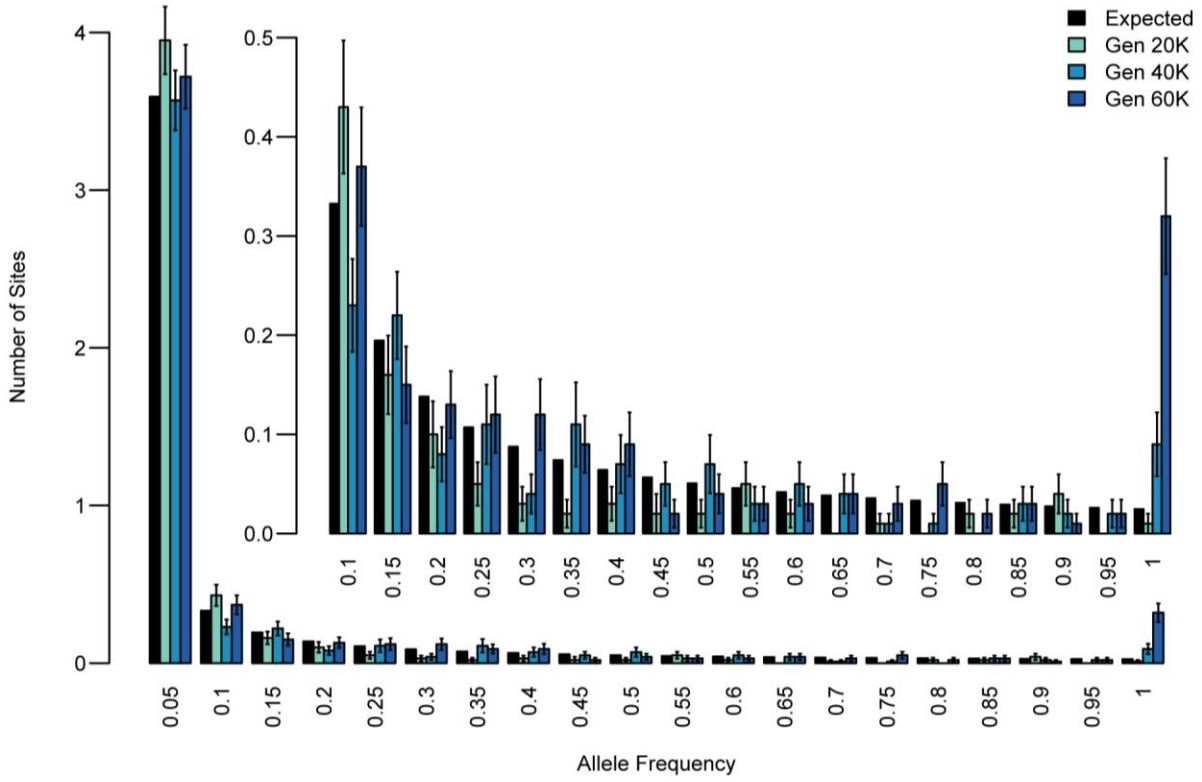

Figure 4.6: SFS for Default Simulations at Generation 20,000, 40,000, and 60,000

The average SFS observed at generation 20,000, 40,000, and the final 60,000th generation in the default simulations across the 100 populations. The inset shows the allele frequency subset where DAF > 0.05. Error bars represent standard error. The black bars indicate the number of segregating sites in each of the allele frequency bins expected under neutral mutation-drift equilibrium.

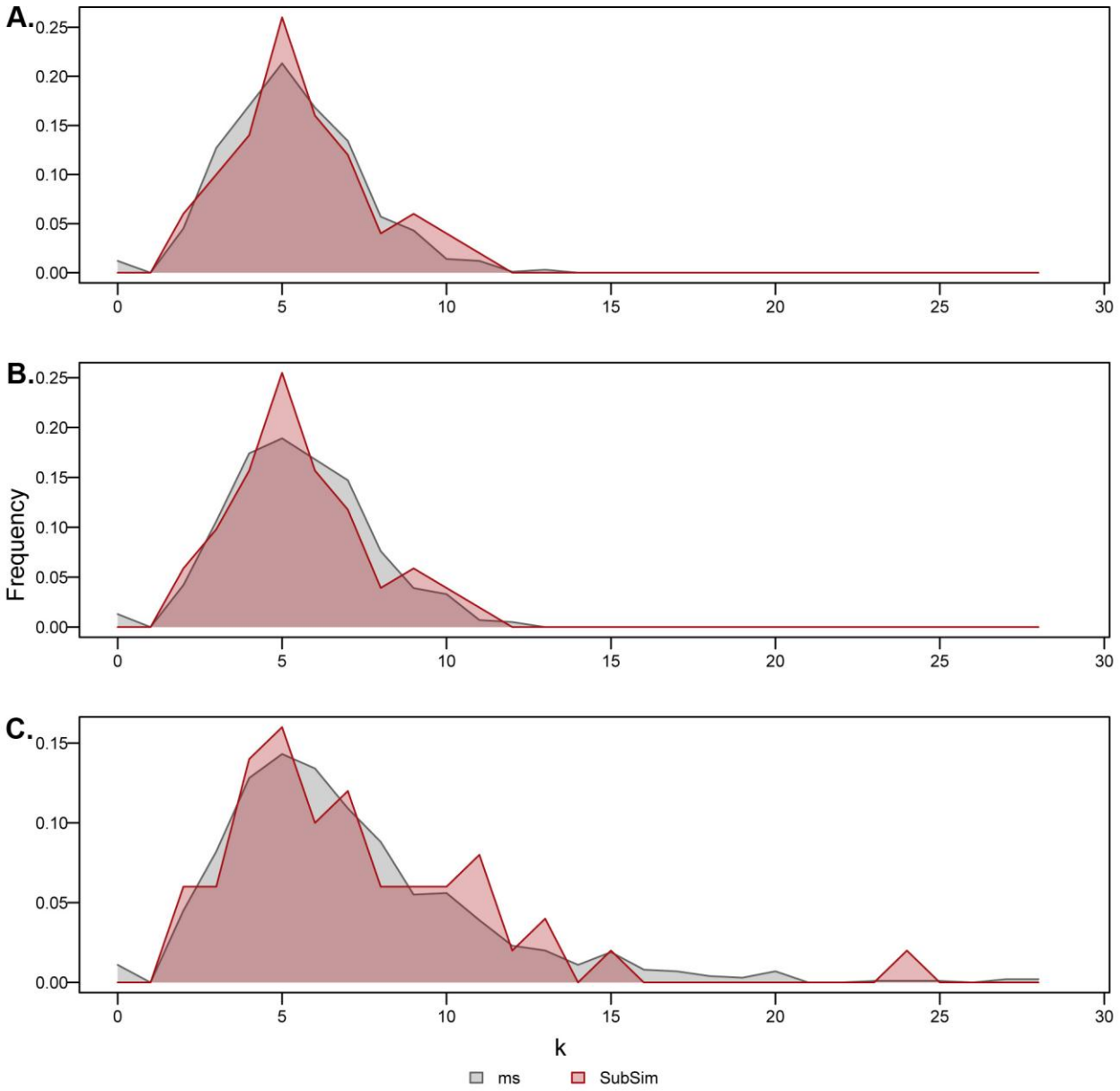

Figure 4.7: Comparison of Haplotype Number Between Simulated Populations in ms and SubSim

I ran simulations using three different recombination rates: $r = 0.0$ (A), $r = 1 \times 10^{-8}$ (B), and $r = 1 \times 10^{-6}$ (C). The distribution of the haplotype number across the 1000 populations simulated in ms (Hudson 2002) is shown in grey. The distribution of the haplotype number for the 50 simulated populations using SubSim is show in in red.

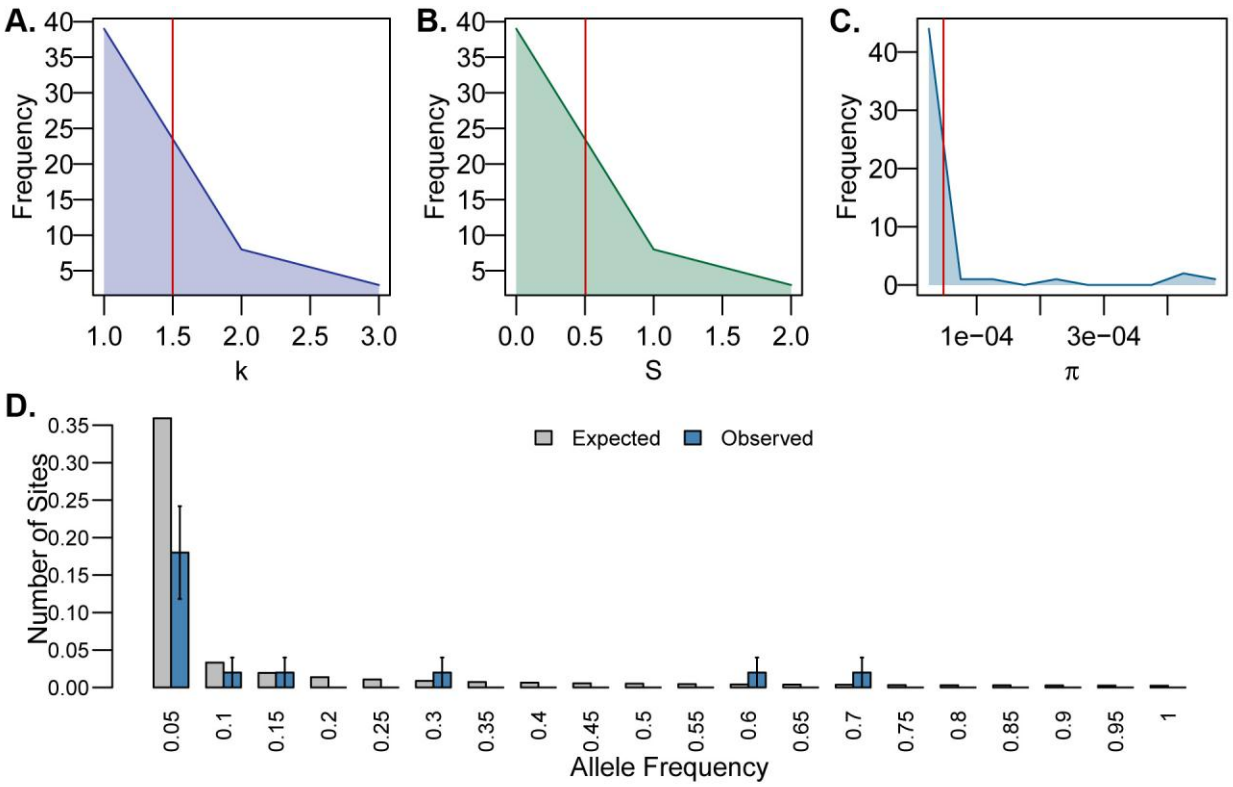

Figure 4.8: Final Population Summary Statistics for $\mu = 1.2x10^{-9}$

The distribution of k (A), S (B), and π (C) are shown for the final populations across the 50 populations simulated with $\mu = 1.2 \times 10^{-9}$. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 50 populations simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. 0<x≤0.05).

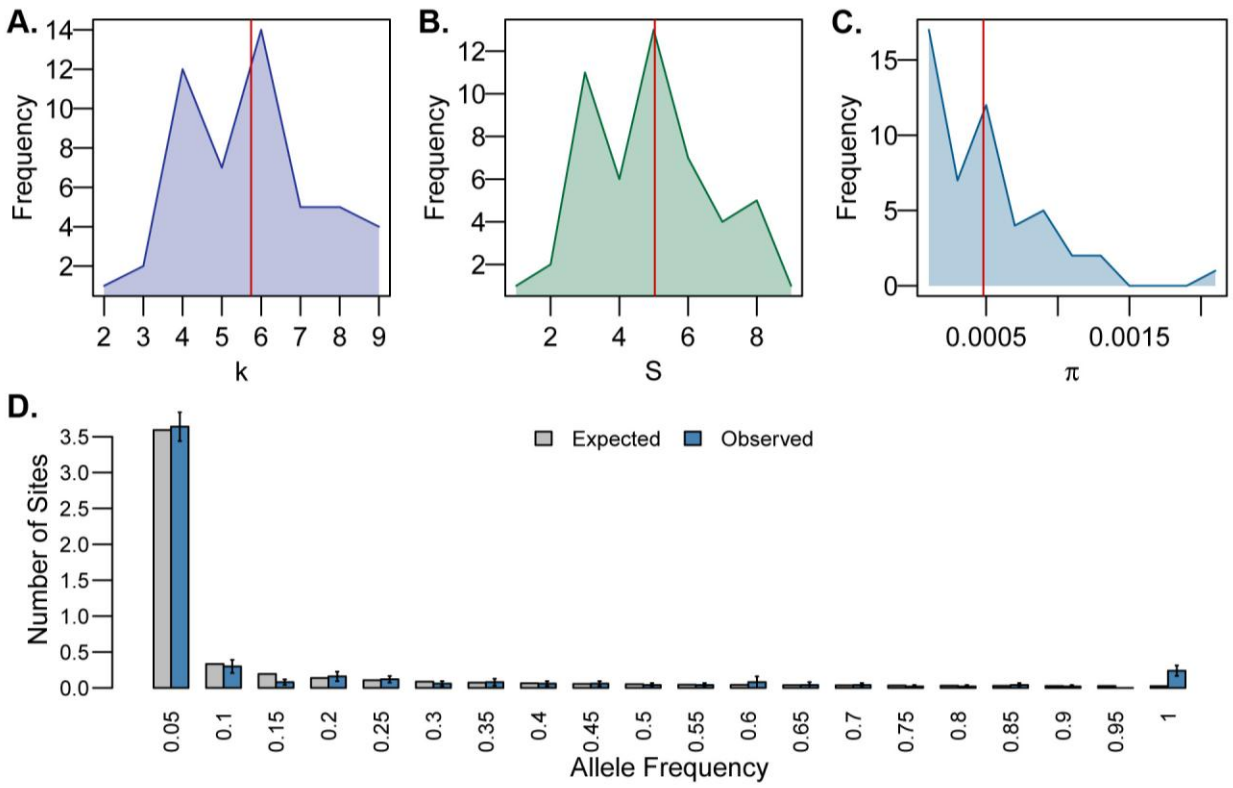

Figure 4.9: Final Population Summary Statistics for $\mu = 1.2x10^{-8}$

The distribution of k (A), S (B), and π (C) are shown for the final populations across the 50 populations simulated with $\mu = 1.2 \times 10^{-8}$. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 50 populations simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. 0<x≤0.05).

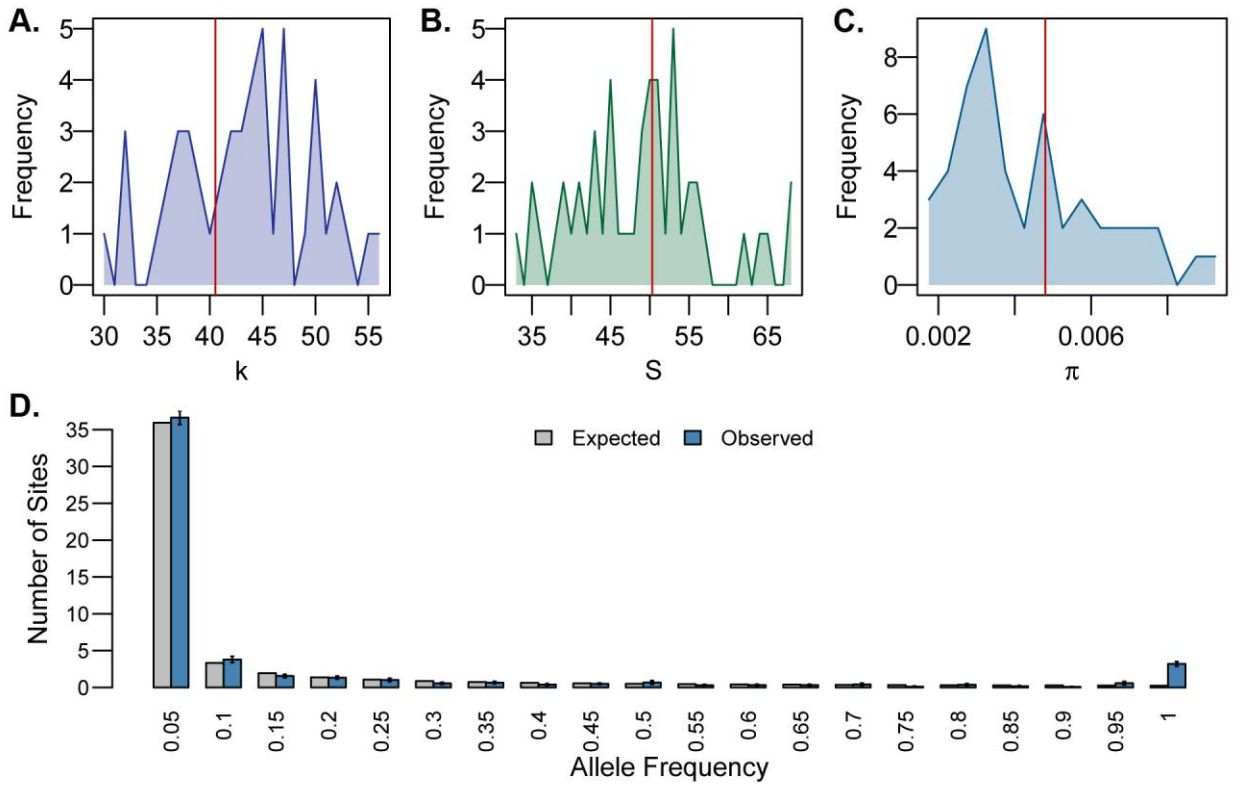

Figure 4.10: Final Population Summary Statistics for $\mu = 1.2x10^{-7}$

The distribution of k (A), S (B), and π (C) are shown for the final populations across the 50 populations simulated with $\mu = 1.2 \times 10^{-7}$. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 50 populations simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. 0 < x ≤ 0.05).

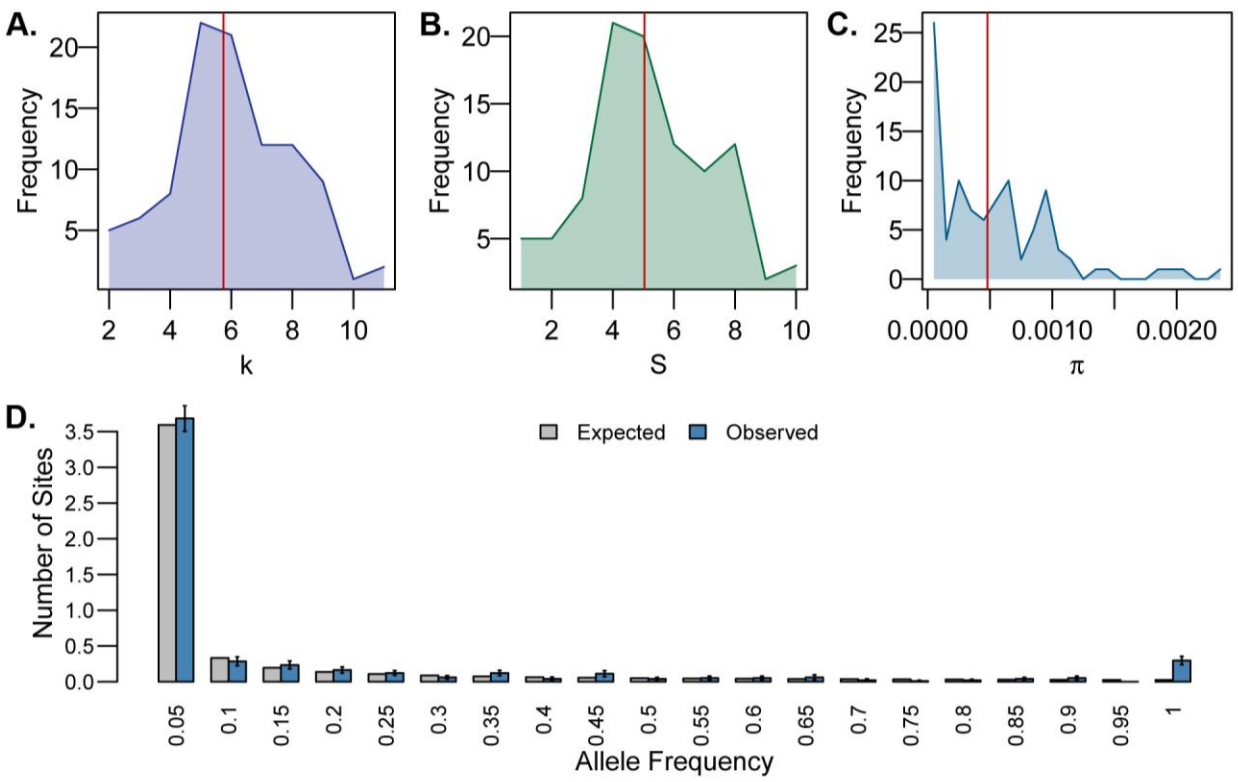

Figure 4.11: Summary Statistics for W>S Mutation Bias Simulations

The distribution of k (A), S (B), and π (C) are shown for the final populations across the 50 populations simulated W>S biased mutation. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 98 populations simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. $0 < x \leq 0.05$).

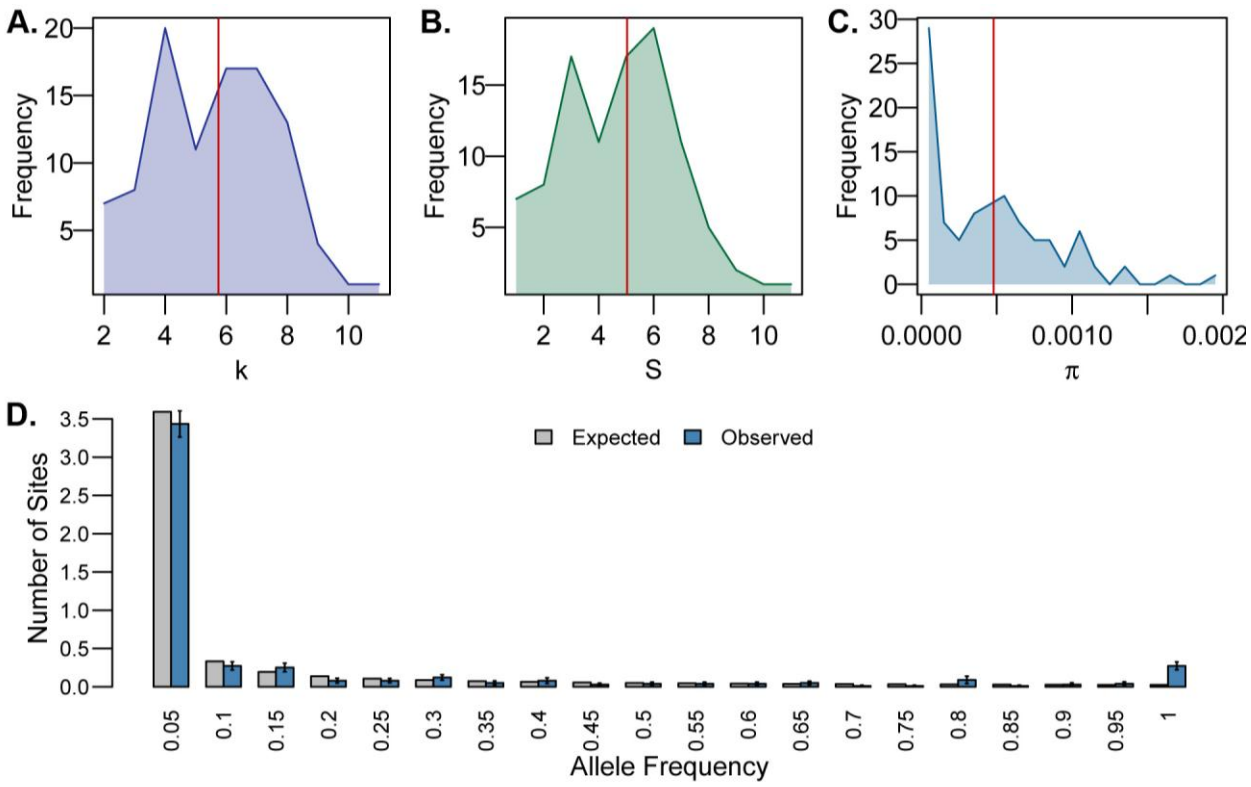

Figure 4.12: Summary Statistics for S>W Mutation Bias Simulations

The distribution of k (A), S (B), and π (C) are shown for the final populations across the 50 populations simulated with S>W biased mutation. Red vertical lines indicate values expected under neutral mutation-drift equilibrium. The average number of non-ancestral sites observed across all allele frequencies in the 99 populations simulations are shown in D. Error bars represent the standard error. Numbers on the x-axis are inclusive (i.e. $0 < x \leq 0.05$).

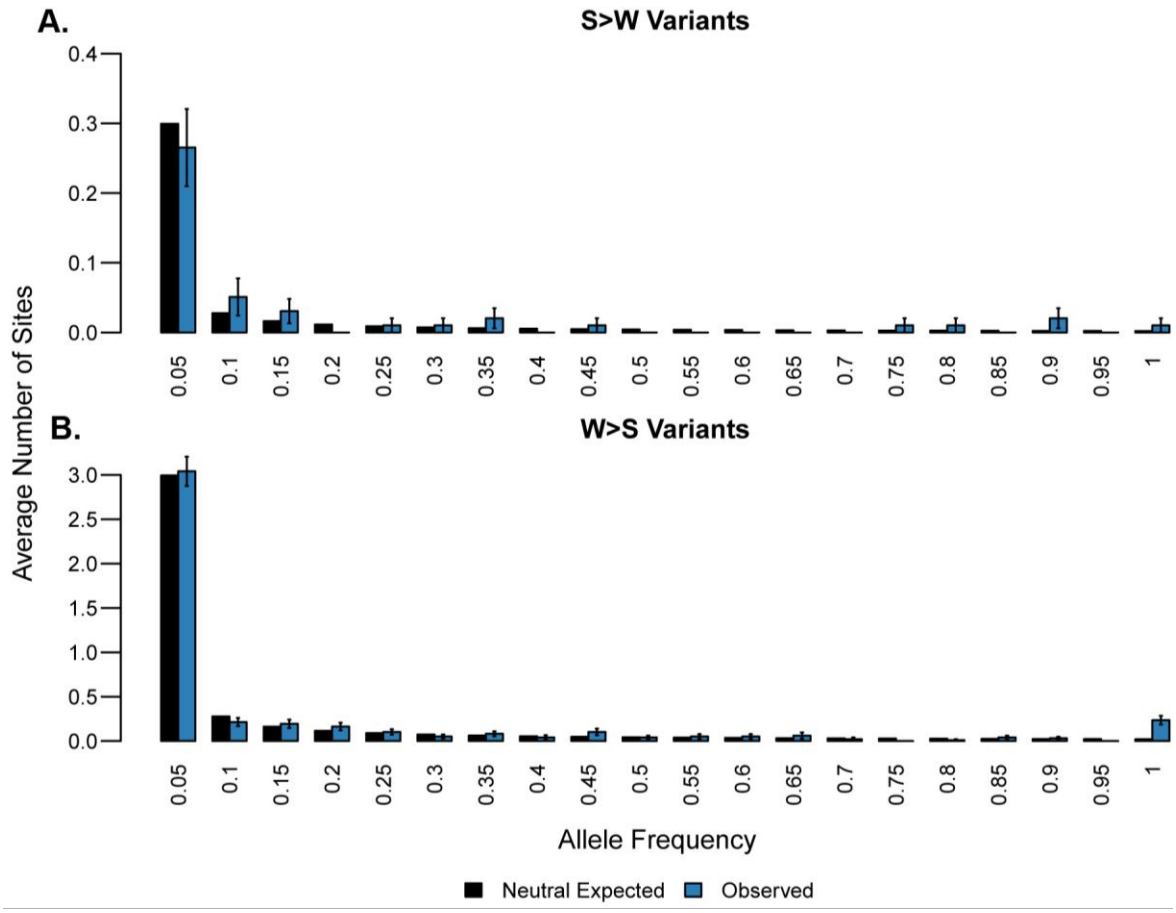

Figure 4.13: Comparison of S>W and W>S Variant SFS in W>S Mutation Bias Simulations

The average SFS observed in the final population for S>W (A) and W>S (B) variants in the W>S mutation bias simulations across the 98 populations. Error bars represent standard error.

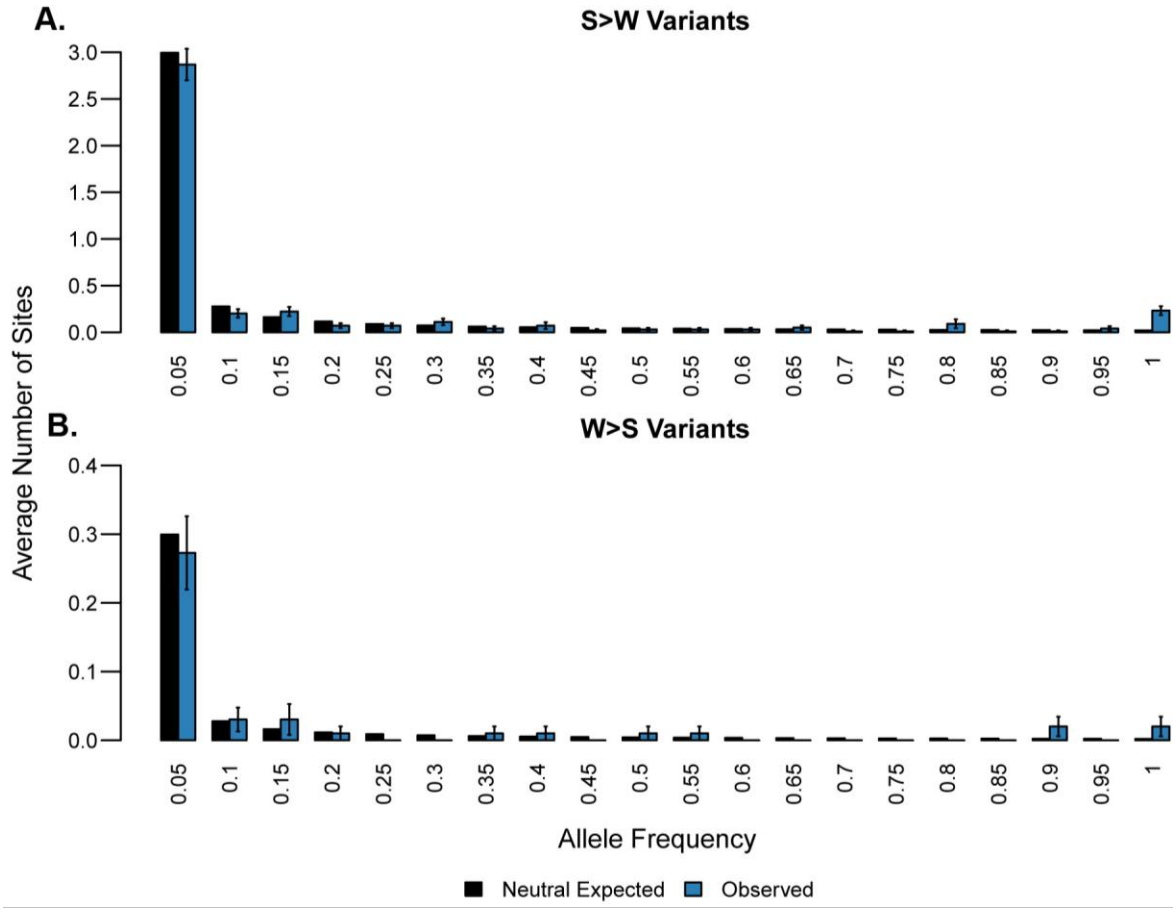

Figure 4.14: Comparison of S>W and W>S Variant SFS in S>W Mutation Bias Simulations

The average SFS observed in the final population for S>W (A) and W>S (B) variants in the S<W mutation bias simulations across the 99 populations. Error bars represent standard error.

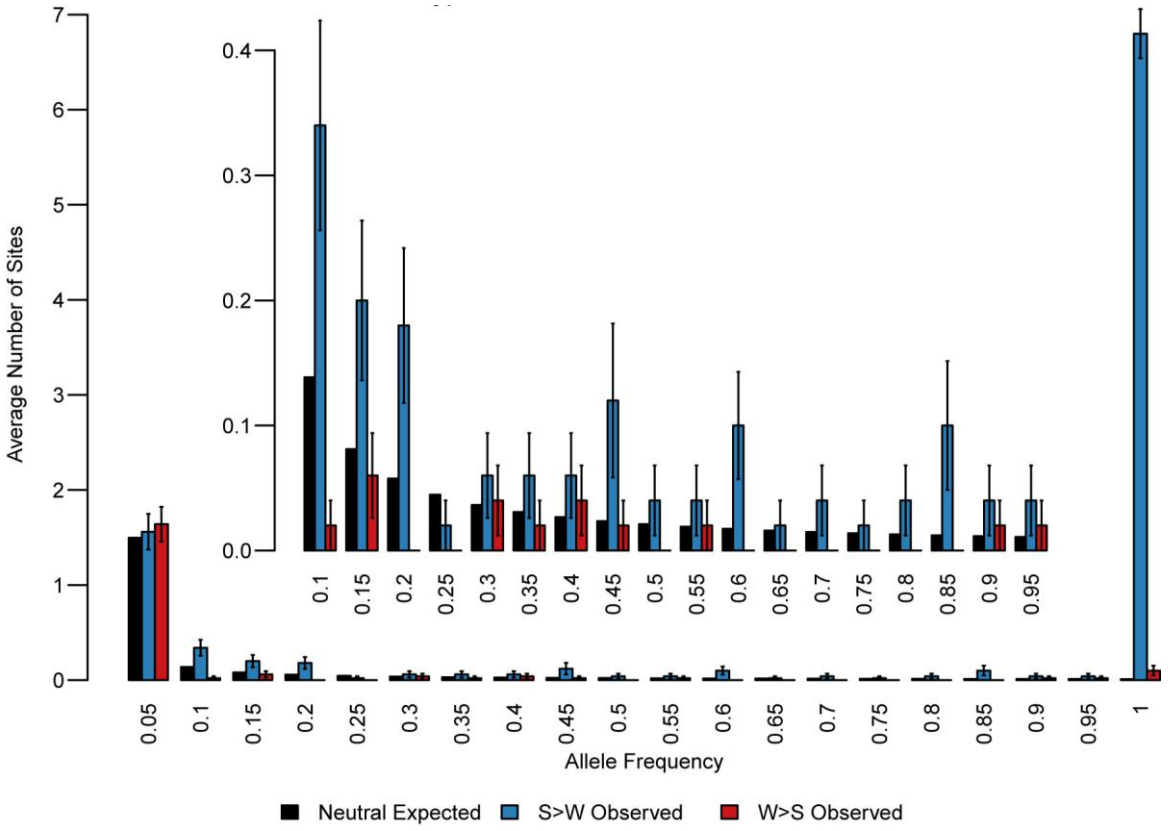

Figure 4.15: Final Population SFS for S>W Variant Selection Bias Simulations

The average SFS observed in the final population across the 50 S>W selection bias simulations. Error bars represent standard error. The inset shows the allele frequency subset from $0.05 < x \leq 0.95$.

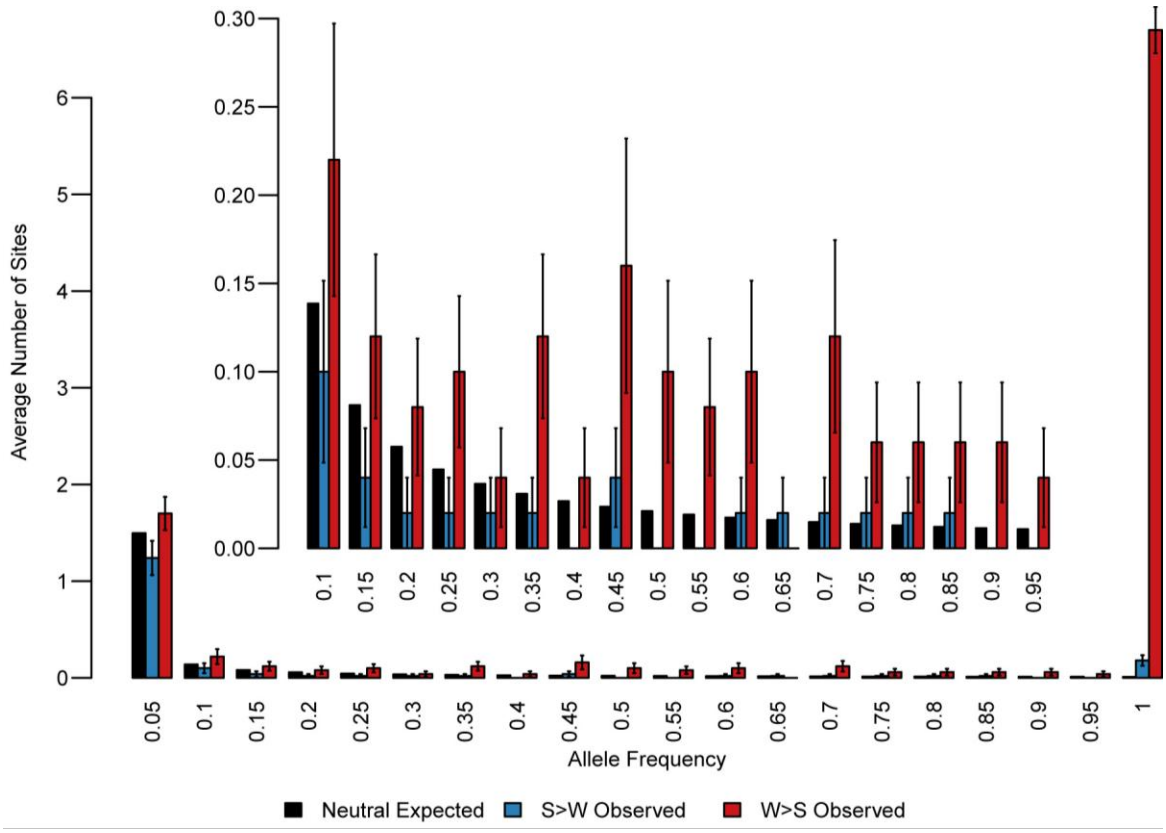

Figure 4.16: Final Population SFS for W>S Variant Selection Bias Simulations

The average SFS observed in the final population across the 50 W>S selection bias simulations. Error bars represent standard error. The inset shows the allele frequency subset from $0.05 < x \leq 0.95$.

4.7 Tables

Parameter	Shortcut	Arguments	Description
Mutation Rate	-m	equal	Equal mutation rates across variant subtypes
		titv (Default)	Transition-biased mutation rate
		user x_1 x_2 x_3 x_4 x_5 x_6	User specified mutation rates
Recombination Rate	-r	Float (Default = $1x10^{-8}$)	Per-site per-generation recombination rate
BGC	-bgc	Float (Default = 0.5)	Degree of GC-biased repair
Excision Mean	-ex	Integer (Default=100)	Mean length of excision in recombination
Selection	-sel	user x_1 x_2 x_3 x_4 x_5 x_6 (Default = 0)	Variant subtype specific selection coefficients
Ancestral Chromosome	-anc	File (Default = generation 0)	Specify the ancestral chromosome in file

Table 4.1 Available Parameter Options in the Simulation

Optional parameters that can be specified by the user along with the default values for each parameter.

Number of Binary Trees	Average Time Per Individual Selection (Seconds)	Average Time per Generation (Seconds)
0	0.00965	193.292
1	0.00452	90.653
2	0.00232	46.563
3	0.00116	23.352
4	0.000594	12.018
5	0.000304	6.217
6	0.000164	3.430
7	9.82E-05	2.109
8	6.07E-05	1.364
9	4.23E-05	0.990
10	4.39E-05	0.864
11	3.31E-05	0.805
12	3.29E-05	0.801
13	3.60E-05	0.865

Table 4.2: Improvements in Efficiency Using Binary Trees

Ten generations were simulated (N=10,000) and the time to sample each individual and the generation time was recorded. The average time per-generation is over 10 generations and the average time per individual selection is over 200,000 individual selections (10 generations x 10,000 individuals x 2 parents per individual).

Original Allele \ Mutated Allele	A	T	G	C
A	$1-\left(\frac{3}{12}\mu\right)$	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$
T	$\frac{1}{12}\mu$	$1-\left(\frac{3}{12}\mu\right)$	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$
G	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$	$1-\left(\frac{3}{12}\mu\right)$	$\frac{1}{12}\mu$
C	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$	$\frac{1}{12}\mu$	$1-\left(\frac{3}{12}\mu\right)$

Table 4.3: Equal Mutation Rates Across All Variant Subtypes

Relative mutation rates used when each mutation subtype is equally likely. The sum of each row adds up to 1, indicating the probability of each ancestral base to be each of the 4 possible derived states in the next generation.

Original Allele \ Mutated Allele	A	T	G	C
A	$1-\left(\frac{\mu}{6}+2\left(\frac{\mu}{24}\right)\right)$	$\frac{1}{24}\mu$	$\frac{1}{6}\mu$	$\frac{1}{24}\mu$
T	$\frac{1}{24}\mu$	$1-\left(\frac{\mu}{6}+2\left(\frac{\mu}{24}\right)\right)$	$\frac{1}{24}\mu$	$\frac{1}{6}\mu$
G	$\frac{1}{6}\mu$	$\frac{1}{24}\mu$	$1-\left(\frac{\mu}{6}+2\left(\frac{\mu}{24}\right)\right)$	$\frac{1}{24}\mu$
C	$\frac{1}{24}\mu$	$\frac{1}{6}\mu$	$\frac{1}{24}\mu$	$1-\left(\frac{\mu}{6}+2\left(\frac{\mu}{24}\right)\right)$

Table 4.4: Transition Biased Mutation Rate

Relative mutation rates used when transition mutations are 4 times more likely than transversion mutations. There are 4 potential transition mutation subtypes, A>G, G>A, C>T and T>C. The remaining 8 mutation subtypes are transversion mutations. Therefore, each transition mutation must be 4 times more likely to occur than each transversion mutation to equal the expected Ti/Tv ratio of 2.0. The mutation rate, μ, can be specified in the simulation or left to the default of $\mu = 1.2 \times 10^{-8}$.

Original Allele \ Mutated Allele	A	T	G	C
A	$1-(x_1+x_2+x_3)$	x_1	x_2	x_3
T	x_4	$1-(x_4+x_5+x_6)$	x_5	x_6
G	x_7	x_8	$1-(x_7+x_8+x_9)$	x_9
C	x_{10}	x_{11}	x_{12}	$1-(x_{10}+x_{11}+x_{12})$

Table 4.5: User-Defined Mutation Rates Across All Variant Subtypes

The user can specify mutation rates for the 12 possible mutation subtypes. The sum of each row adds up to 1, indicating the probability of each ancestral base to be each of the 4 possible derived states in the next generation.

N	L (kb)	Time
100	1	19.21 s
1,000	1	132.82 s
10,000	1	24.69 m
100,000	1	8.22 h
10,000	0.1	21.60 m
10,000	1	24.69 m
10,000	10	53.34 m
10,000	100	9.97 h

Table 4.6: Simulation Efficiency

Efficiency was tested for a variety of sample sizes and chromosome lengths. Each simulation was run over 1,000 generations using default parameters. The time reported is the time for each simulation to run over the simulations using each of the parameter settings over 1,000 generations.

CHAPTER 5

Conclusions and Future Directions

5.1 Technological Innovations in High-Throughput Sequencing Allow for a Better Understanding of Single-Nucleotide Mutation in the Human Genome

Recent advances in high-throughput (HT) sequencing have led to incredible steps forward in understanding single-nucleotide mutations in the human genome. Before the streamlined use of HT sequencing became a fairly routine practice, identification of mutations and rare variants in human samples was done using PCR followed by Sanger sequencing. While these techniques are extremely reliable, and are still an important step in properly validating mutations found via HT sequencing, expanding these assays to efficiently study a large number of loci or a large number of samples is difficult. HT-sequencing technology overcomes both of these major burdens, allowing researchers to rapidly sequence the entire exome or even genome of an individual and is feasible to scale for large sample sizes.

The three projects presented here are based on HT-sequencing data from human samples to understand the influence of innate genomic features on the frequency and types of mutations that occur in the human genome and identify those mutant sites to better understand disease. First, I uncovered a mutation in the gene *RAB40AL*, which likely leads to the rare Mendelian disorder, Martin-Probst Syndrome (MPS). Second, I leveraged rare variant data from an extremely large sequencing study

to understand how GC content and recombination rate influence patterns of mutations in humans. Finally, I developed a forward simulation program, which will pave the way for future work in understanding how the effects of GC content and recombination rate on subtype-specific mutation and fixation processes can influence variant patterns in humans.

5.2 Mutations in *RAB40AL* in Martin-Probst Syndrome

Martin-Probst Syndrome (MPS) is a rare X-linked recessive Mendelian disorder characterized by sensorineural hearing loss and mental retardation, among a constellation of other phenotypes (Martin et al. 2000). Previous work to identify the causative mutation in MPS identified a large haplotype block on the X-chromosome, although sequencing in this locus was unable to identify any causative mutations in potential candidate genes (Martin et al. 2000; Probst et al. 2004). With the increasing applicability of HT-sequencing techniques to find genes for these types of rare Mendelian disorders (Ng et al. 2009; Ng et al. 2010a), I undertook a multi-platform approach to find the gene underlying MPS. We applied whole-genome, whole-exome and X chromosome targeted exome sequencing in order to adequately cover the X-chromosome and enrich the dataset for high quality variants. After stringent quality control and filtering variants through a multi-stage protocol, I identified two adjacent single-nucleotide mutations in the gene *RAB40AL*. Together, these two mutations lead to a missense mutation in the amino acid sequence of the *RAB40AL* protein, changing an aspartic acid to a glycine at amino acid 59 (p.D59G). This alteration is expected to be damaging to protein function according to several prediction algorithms (Ramensky

et al. 2002; Cheng et al. 2006; Mathe et al. 2006; Tavtigian et al. 2006; Kumar et al. 2009; Adzhubei et al. 2010) and is highly conserved across the evolutionary lineage.

These findings now point to a potential biological mechanism underlying MPS, although, the function of the *RAB40AL* gene is poorly understood. *RAB40AL* belongs to a major of class of Rab small GTP-binding proteins that are responsible for intracellular organelle trafficking (Pereira-Leal and Seabra 2001). In 2002, a disruption to *RAB40AL* was identified in an individual with Duchenne Muscular Dystrophy and mental retardation (Saito-Ohara et al. 2002). While normal *RAB40AL* is located on the mitochondria (Saito-Ohara et al. 2002), *in* vitro functional evidence suggests that the mutated form of *RAB40AL* identified in MPS is unable to properly localize to the mitochondria and instead is found in the cell nucleus (Bedoyan et al. 2012).

Future functional analysis is necessary to fully understand the underlying etiology of MPS. Additional families exhibiting the same disorder would be ideal to conclusively demonstrate that mutations in *RAB40AL* can lead to this unique combination of phenotypes in affected individuals. Furthermore, there is a great deal to be learned regarding the function of *RAB40AL*. The preliminary functional studies by Bedoyan and colleagues (Bedoyan et al. 2012) show that the normal function of *RAB40AL* is disrupted by these mutations, however the exact deleterious effect that the p.D59G mutation has on protein function is as yet unknown. This mutation could potentially lead to a disruption in protein folding, signaling, transport within the cell to the mitochondria, or binding to the mitochondrial membrane. Further study is warranted to understand how these disruptions affect normal cellular functioning, leading to downstream hearing loss and impaired cognitive function.

5.3 The Influence of GC Content and Recombination Rate on Mutation and Fixation in the Human Genome

Researchers have studied mutation rates in humans for decades, and the widely-accepted estimate of the per-base per-generation mutation rate is 1.2×10^{-8} (The 1000 Genomes Project Consortium 2010; Campbell et al. 2012; Kong et al. 2012). Less understood, however, is how the mutation rate fluctuates from one locus to another and what influences those fluctuations (Wolfe et al. 1989; Nachman and Crowell 2000; Sachidanandam et al. 2001; Smith and Lercher 2002; Kondrashov 2003; Hodgkinson et al. 2009).

In this project, I used rare variants obtained from exome sequencing of 202 genes in >14,000 individuals. This extremely large dataset allowed me to assay very rare variants (derived allele frequency ≤ 10^{-4}). Because these rare variants are relatively young in the human lineage compared to variants with higher allele frequencies, their patterns are primarily governed by the spontaneous mutation rate and genetic drift, as opposed to other evolutionary forces. Analysis of rare variants can, therefore, be used to study the underlying rate of spontaneous mutation (Messer 2009). Application of rare variants to study mutation rate variability and the effect of GC content and recombination rate on the mutation rate had not previously been performed. Prior studies used common polymorphic sites segregating in humans or divergent sites between humans and chimpanzee to infer effects of genomic context on the mutation rate. These types of data, however, can be strongly influenced by population demographic history, natural selection, and biased gene conversion (BGC), and therefore might not be reflective of the actual underlying mutation rate. Additionally, this

unique application of HT-sequencing data was a creative extension of data that was generated for the purpose of disease gene mapping studies. These types of extraneous applications will be increasingly useful and important as more of these types of data are generated.

I found that GC content has a distinct effect on the probability of observing a rare variant of a specific subtype compared to both common variants and human-chimp substitutions. These results suggest that both the mutation rate as well as fixation may vary in different regions of the genome to maintain the base composition of that specific region. Recombination rate, on the other hand, had a much stronger influence on both common variants and substitutions, specifically those that converted an A:T base pair to a G:C base pair. These results are consistent with BGC, in which mismatches generated during recombination are preferentially repaired to a G or C base pair compared to an A or T (Meunier and Duret 2004; Duret and Arndt 2008; Duret and Galtier 2009).

The clear next step in this work is to determine if the joint effect of mutation and fixation bias on specific variant subtypes in response to the local genomic context can realistically lead to the effects observed in this study. Forward simulations are a clear choice for this analysis because of their effectiveness in modeling selection. Currently available simulation software, however, does not allow for modeling distinct selection coefficients on specific variant subtypes. For the final project presented here, we, therefore, developed a forward simulation program, which allows the user to define variant subtype-specific mutation and selection bias.

Additional genomic features have been shown to exert an effect on the spontaneous single-nucleotide mutation rate in humans. Strand-bias in transcribed genes (Green et al. 2003; McVicker and Green 2010), replication timing (Wolfe et al. 1989; Stamatoyannopoulos et al. 2009; Chen et al. 2010; Koren et al. 2012), and other more cryptic effects from neighboring nucleotides (Hwang and Green 2004; Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010; Nevarez et al. 2010) have all been shown to influence the mutation rate. Analysis of rare variants in these contexts will help to shed additional light on how these, and other, innate features of the genome influence mutation dynamics.

5.4 Forward Population Genetic Simulation Program

As the final part of this work, I developed a forward genetic simulation program, SubSim, to pursue additional research avenues generated from the project analyzing the effect of GC content and recombination rate on rare variants, common variants, and substitutions. Currently, there is a wide assortment of forward simulation programs to choose from, each offering its unique combination of parameters that can be modeled (Hoban et al. 2011; Yuan et al. 2012). Some of these include a variety of complicated mutation, selection and demographic models, each of which can be fine-tuned by the user to generate simulated data to test new statistical methods and algorithms, or to analyze population genetic and evolutionary questions. My unique contribution to this field is to offer a new program that models selection on specific variant subtypes, as opposed to selection on single-locus beneficial or deleterious mutations.

SubSim can be applied to answer a large number of biological questions, such as understanding the degree to which changes in mutation and fixation bias on specific variant subtypes lead to variability in mutation patterns across different regions of the genome. Additionally, SubSim allows the user to control the base composition of the starting chromosome, alter recombination rates, and model BGC. By manipulating these parameters, SubSim can be used to study the effects of BGC on genome evolution, understand how GC content and recombination rate influence variant patterns, study selection in favor of specific variants, and other important applications.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**(4): 248-249.

Agresti A. 2002. *Categorical data analysis*. Wiley-Interscience, New York.

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**(11): 903-905.

Arndt PF, Hwa T. 2004. Regional and time-resolved mutation patterns of the human genome. *Bioinformatics* **20**(10): 1482-1485.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**(10): 2322-2328.

Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol* **60**(6): 748-763.

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**(11): 745-755.

Bedoyan JK, Schaibley VM, Peng W, Bai Y, Mondal K, Shetty AC, Durham M, Micucci JA, Dhiraaj A, Skidmore JM et al. 2012. Disruption of RAB40AL function leads to Martin–Probst syndrome, a rare X-linked multisystem neurodevelopmental human disorder. *Journal of Medical Genetics* **49**(5): 332-340.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**(6369): 519-520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol* **5**(11): e310.

Belmont JW. 1996. Genetic control of X inactivation and processes leading to X-inactivation skewing. *Am J Hum Genet* **58**(6): 1101-1108.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol* **7**(1): e26.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228**(4702): 953-958.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**(5): e1000083.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet* **5**(1): e1000336.
Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ, Sudmant PH, Shendure J et al. 2012. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet* **44**(11): 1277-1281.
Carvajal-Rodriguez A. 2008. GENOMEPOP: a program to simulate genomes in populations. *BMC Bioinformatics* **9**: 223.
Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. 2008. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics* **9**: 364.
Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**(4): 1289-1303.
Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* **141**(4): 1619-1632.
Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**(4): 447-457.
Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**(4): 1125-1132.
Conrad DF, Keebler JE, Depristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV et al. 2010. Variation in genome-wide mutation rates within and between human families. *Nat Genet*.
Cooper DN, Krawczak M. 1993. *Human gene mutation*. Bios Scientific Publishers, Oxford, UK.
Cooper DN, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. *Hum Genet* **78**(2): 151-155.
Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**(7): 901-913.
Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**(8): 131.
De Keulenaer S, Hellemans J, Lefever S, Renard JP, De Schrijver J, Van de Voorde H, Tabatabaiefar MA, Van Nieuwerburgh F, Flamez D, Pattyn F et al. 2012. Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform. *BMC Med Genomics* **5**: 17.
de Kok YJ, van der Maarel SM, Bitner-Glindzicz M, Huber I, Monaco AP, Malcolm S, Pembrey ME, Ropers HH, Cremers FP. 1995. Association between X-linked mixed deafness and mutations in the POU domain gene POU3F4. *Science* **267**(5198): 685-688.
Diaz-Horta O, Duman D, Foster J, 2nd, Sirmaci A, Gonzalez M, Mahdieh N, Fotouhi N, Bonyadi M, Cengiz FB, Menendez I et al. 2012. Whole-exome sequencing efficiently detects rare mutations in autosomal recessive nonsyndromic hearing loss. *PLoS ONE* **7**(11): e50628.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* **148**(4): 1667-1686.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4**(5): e1000071.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285-311.
Esposito MS, Bruschi CV. 1993. Diploid yeast cells yield homozygous spontaneous mutations. *Curr Genet* **23**(5-6): 430-434.
Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3**(1): 87-112.
Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered* **91**(6): 506-509.
Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet* **2**(7): 549-555.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**(6): 368-376.
Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol* **22**(3): 650-658.
Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol* **17**(9): 1371-1383.
Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25**(1): 1-5.
Giannandrea M, Bianchi V, Mignogna ML, Sirri A, Carrabino S, D'Elia E, Vecellio M, Russo S, Cogliati F, Larizza L et al. 2010. Mutations in the small GTPase gene RAB39B are responsible for X-linked mental retardation associated with autism, epilepsy, and macrocephaly. *Am J Hum Genet* **86**(2): 185-195.
Gibbons RJ, Suthers GK, Wilkie AO, Buckle VJ, Higgs DR. 1992. X-linked alpha-thalassemia/mental retardation (ATR-X) syndrome: localization to Xq12-q21.31 by X inactivation and linkage analysis. *Am J Hum Genet* **51**(5): 1136-1149.
Gilberg RF, Forouzan BA. 2001. *Data structures : a pseudocode approach with C++*. Brooks/Cole, Pacific Grove, CA.
Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**(4): 514-517.
Haldane JB. 1935. The rate of spontaneous mutation of a human gene. 1935. *J Genet* **83**(3): 235-244.
Hartl DL, Clark AG. 1997. *Principles of population genetics*. Sinauer Associates, Sunderland, Mass.
Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**(6): 1527-1535.
Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**(7): 1020-1029.

Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res* **15**(9): 1222-1231.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**(23): 2786-2787.
Hoban S, Bertorelle G, Gaggiotti OE. 2011. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet* **13**(2): 110-122.
Hodgkinson A, Eyre-Walker A. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. *Genome Biol Evol* **2**: 547-557.
Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol* **7**(2): e1000027.
Hu H, Wrogemann K, Kalscheuer V, Tzschach A, Richard H, Haas SA, Menzel C, Bienek M, Froyen G, Raynaud M et al. 2009. Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. *Hugo J* **3**(1-4): 41-49.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2): 337-338.
Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* **141**(4): 1605-1617.
Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A* **101**(39): 13994-14001.
Jensen LR, Chen W, Moser B, Lipkowitz B, Schroeder C, Musante L, Tzschach A, Kalscheuer VM, Meloni I, Raynaud M et al. 2011. Hybridisation-based resequencing of 17 X-linked intellectual disability genes in 135 patients reveals novel mutations in ATRX, SLC6A8 and PQBP1. *Eur J Hum Genet* **19**(6): 717-720.
Jin H, May M, Tranebjaerg L, Kendall E, Fontan G, Jackson J, Subramony SH, Arena F, Lubs H, Smith S et al. 1996. A novel X-linked gene, DDP, shows mutations in families with deafness (DFN-1), dystonia, mental deficiency and blindness. *Nat Genet* **14**(2): 177-180.
Jonsson JJ, Renieri A, Gallagher PG, Kashtan CE, Cherniske EM, Bruttini M, Piccini M, Vitelli F, Ballabio A, Pober BR. 1998. Alport syndrome, mental retardation, midface hypoplasia, and elliptocytosis: a new X linked contiguous gene deletion syndrome? *J Med Genet* **35**(4): 273-278.
Jukes TH, Cantor CR, ed. 1969. *Evolution of protein molecules*. Academic Press, New York.
Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. *Genetics* **123**(4): 887-899.
Keinan A, Reich D. 2010. Human population differentiation is strongly correlated with local recombination rate. *PLoS Genet* **6**(3): e1000886.
Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**(4922): 1073-1080.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**(2): 111-120.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.
Kingman JFC. 1982. The coalescent. *Stochastic Processes and their Applications* **13**(3): 235-248.
Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* **21**(1): 12-27.
Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum Mutat* **2**(3): 229-234.
Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Wong WS et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**(7412): 471-475.
Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**(3): 241-247.
Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**(6): 1033-1040.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**(7): 1073-1081.
Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* **99**(2): 803-808.
Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.
Lercher MJ, Hurst LD. 2002a. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? *Gene* **300**(1-2): 53-58.
Lercher MJ, Hurst LD. 2002b. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**(7): 337-340.
Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162**(4): 1805-1810.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* **21**(6): 940-951.
Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**(10): e1002326.

Lynch ED, Lee MK, Morrow JE, Welcsh PL, Leon PE, King MC. 1997. Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the Drosophila gene diaphanous. *Science* **278**(5341): 1315-1318.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**(3): 961-968.
Magni GE. 1964. Origin and Nature of Spontaneous Mutations in Meiotic Organisms. *J Cell Physiol* **64**: SUPPL 1:165-171.
Magni GE, Von Borstel RC. 1962. Different Rates of Spontaneous Mutation during Mitosis and Meiosis in Yeast. *Genetics* **47**(8): 1097-1108.
Martin DM, Probst FJ, Camper SA, Petty EM. 2000. Characterisation and genetic mapping of a new X linked deafness syndrome. *J Med Genet* **37**(11): 836-841.
Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* **34**(5): 1317-1325.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**(9): 1297-1303.
McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**(5670): 581-584.
McVicker G, Green P. 2010. Genomic signatures of germline gene expression. *Genome Res* **20**(11): 1503-1511.
Menasche G, Pastural E, Feldmann J, Certain S, Ersoy F, Dupuis S, Wulffraat N, Bianchi D, Fischer A, Le Deist F et al. 2000. Mutations in RAB27A cause Griscelli syndrome associated with haemophagocytic syndrome. *Nat Genet* **25**(2): 173-176.
Messer PW. 2009. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182**(4): 1219-1232.
Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* **21**(6): 984-990.
Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM. 1986. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**(6089): 646-650.
Mullaney JM, Mills RE, Pittard WS, Devine SE. 2010. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* **19**(R2): R131-136.
Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746): 321-324.
Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet* **17**(9): 481-485.
Nachman MW. 2004. Haldane and the first estimates of the human mutation rate. *J Genet* **83**(3): 231-233.
Nachman MW, Bauer VL, Crowell SL, Aquadro CF. 1998. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**(3): 1133-1141.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**(1): 297-304.
Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**(7397): 242-245.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**(10): 5269-5273.
Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D et al. 2012. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**(6090): 100-104.
Nevarez PA, DeBoever CM, Freeland BJ, Quitt MA, Bush EC. 2010. Context dependent substitution biases vary within the human genome. *BMC Bioinformatics* **11**: 462.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC et al. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**(9): 790-793.
Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**(1): 30-35.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261): 272-276.
O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C et al. 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**(6): 585-589.
Online Mendelian Inheritance in Man O. Online Mendelian Inheritance in Man, OMIM. Vol 2013. McKusick-Nathans Institite of Genetic Medicine, Johns Hopkins University (Baltimore, MD).
Patel VC, Mondal K, Shetty AC, Horner VL, Bedoyan JK, Martin D, Caspary T, Cutler DJ, Zwick ME. 2010. Microarray oligonucleotide probe designer (MOPeD): A web service. *Open Access Bioinformatics* **2**(2010): 145-155.
Pedemonte N, Lukacs GL, Du K, Caci E, Zegarra-Moran O, Galietta LJ, Verkman AS. 2005. Small-molecule correctors of defective DeltaF508-CFTR cellular processing identified by high-throughput screening. *J Clin Invest* **115**(9): 2564-2571.
Peng B, Kimmel M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**(18): 3686-3687.
Pereira-Leal JB, Seabra MC. 2001. Evolution of the Rab family of small GTP-binding proteins. *J Mol Biol* **313**(4): 889-901.
Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F et al. 2007. Multiplex amplification of large sets of human exons. *Nat Methods* **4**(11): 931-936.
Probst FJ, Hedera P, Sclafani AM, Pomponi MG, Neri G, Tyson J, Douglas JA, Petty EM, Martin DM. 2004. Skewed X-inactivation in carriers establishes linkage in an

X-linked deafness-mental retardation syndrome. *Am J Med Genet A* **131**(2): 209-212.
Puck JM, Willard HF. 1998. X inactivation in females with X-linked disease. *N Engl J Med* **338**(5): 325-328.
R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**(17): 3894-3900.
Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, Shahzad M, Ahmed ZM, Riazuddin S, Khan SN, Friedman TB. 2010. Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet* **86**(3): 378-388.
Richards RI, Sutherland GR. 1994. Simple repeat DNA is not replicated simply. *Nat Genet* **6**(2): 114-116.
Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**(4922): 1066-1073.
Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N et al. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**(4922): 1059-1065.
Rowley JD. 1973. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**(5405): 290-293.
Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**(6822): 928-933.
Saito-Ohara F, Fukuda Y, Ito M, Agarwala KL, Hayashi M, Matsuo M, Imoto I, Yamakawa K, Nakamura Y, Inazawa J. 2002. The Xq22 inversion breakpoint interrupted a novel Ras-like GTPase gene in a patient with Duchenne muscular dystrophy and profound mental retardation. *Am J Hum Genet* **71**(3): 637-645.
Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL et al. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**(7397): 237-241.
Seabra MC, Mules EH, Hume AN. 2002. Rab GTPases, intracellular traffic and disease. *Trends Mol Med* **8**(1): 23-30.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8): 1034-1050.
Silva JC, Kondrashov AS. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet* **18**(11): 544-547.
Slatkin M. 2000. Allele age and a test for selection on rare alleles. *Philos Trans R Soc Lond B Biol Sci* **355**(1403): 1663-1668.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23**(1): 23-35.
Smith NG, Lercher MJ. 2002. Regional similarities in polymorphism in the human genome extend over many megabases. *Trends Genet* **18**(6): 281-283.
Smith NG, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res* **12**(9): 1350-1356.
Sommer SS. 1995. Recent human germ-line mutation: inferences from patients with hemophilia B. *Trends Genet* **11**(4): 141-147.
Sommer SS, Ketterling RP. 1996. The factor IX gene as a model for analysis of human germline mutations: an update. *Hum Mol Genet* **5 Spec No**: 1505-1514.
Spencer CC. 2006. Human polymorphism around recombination hotspots. *Biochem Soc Trans* **34**(Pt 4): 535-536.
Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet* **2**(9): e148.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**(4): 393-395.
Sutherland GR, Richards RI. 1995. Simple tandem DNA repeats and human genetic disease. *Proc Natl Acad Sci U S A* **92**(9): 3636-3641.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3): 585-595.
Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A et al. 2009. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* **41**(5): 535-543.
Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* **43**(4): 295-305.
Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al. 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science*.
The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.
The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**(6): 971-983.
Wakeley J. 2009. *Coalescent theory : an introduction*. Roberts & Co. Publishers, Greenwood Village, Colo.
Wakeley J, Takahashi T. 2003. Gene genealogies when the sample size exceeds the effective size of the population. *Mol Biol Evol* **20**(2): 208-213.
Webster MT, Smith NG, Ellegren H. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol Biol Evol* **20**(2): 278-286.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**(6): e90.
Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**(6204): 283-285.
Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos JA, Karayiorgou M. 2012. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet* **44**(12): 1365-1369.
Yuan X, Miller DJ, Zhang J, Herrington D, Wang Y. 2012. An overview of population genetic data simulation. *J Comput Biol* **19**(1): 42-54.