Developing a Gene Model for Simulations that Incorporates Multi-Species Conservation

by

Brendan F. Liu

S.B., Massachusetts Institue of Technology(2013)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Master of Enginerring in Electrical Engineering and Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 15 2014
LIBRARIES

© Massachusetts Institute of Technology 2014. All rights reserved.

Signature redacted

Author ....
Department of Electrical Engineering and Computer Science
May 23,2014

Signature redacted

Certified by .....
David Altshuler
Professor of Biology(Adjunct Professor)
Thesis Supervisor

Signature redacted

Accepted by ....
Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Developing a Gene Model for Simulations that Incorporates Multi-Species Conservation

by

Brendan F. Liu

Submitted to the Department of Electrical Engineering and Computer Science
on May 23,2014, in partial fulfillment of the
requirements for the degree of
Master of Enginerring in Electrical Engineering and Computer Science and Engineering

Abstract

The genetic architecture, the number, frequency, and effect size of disease causing alleles for many common diseases including Type 2 Diabetes is not fully understood. Genetic simulations can be used to make predictions under specified genetic architecture models. Models whose predictions are inconsistent with empirical data can be rejected.

We extended a gene simulation model previously published by our lab. The distribution of number and length of coding and intron regions of each simulated gene was consistent with the distribution in the human genome. Selection pressure against mutations was modeled by utilizing the cross-species conservation of each region. The combined distribution of variants by their frequency over 500 genes was compared between the simulated genes and the corresponding empirical data. This distribution of variants between the simulated and empirical data was found to be consistent.

Thesis Supervisor: David Altshuler
Title: Professor of Biology(Adjunct Professor)

Acknowledgments

The completion of this thesis would not have been made possible without the support, mentorship, and encouragement of many individuals.

First and foremost, I would like to thank David Altshuler for allowing me into his lab and for his support and guidance throughout the project. I feel priviledged to be a part of his lab as the only Master's student. Through our meetings, he has given me so much advice that I wish I could write them down faster. Without his mentorship, I would not be where I am now.

I would like to thank my mentor Alisa Manning for all the time she has spent mentoring me. Even though she has many other projects that she is currently working on, she always tries to take the time to answer my questions, however dumb and frequent they are. Without her constant concern about the status of my project, there would be a good possibility that the project would not have been completed in a timely manner. I especially want to thank her for helping me write this thesis. Even though she was on vacation with her family in Disney World, she was willing to take some time to provide comments on this thesis. Her commitment to me as a mentor was one of the main reasons why my experience in the Altshuler Lab has been memorable.

I would also like to thank Vineeta Agarwala, the first person I met in this lab. I still remember that in our first meeting she was patient enough to spend two hours giving me an overview of population genetics. In addition, she was willing to meet with me for an hour for several months just to make sure that I would have the proper background in population genetics for this project. Her willingness to explain anything as well as her desire to make sure I understood everything really helped me get acclimated to this field. Even though she is currently in medical school, she still tries to find time to answer any questions I have. In addition, she managed to look over this thesis while being in the middle of medical school rotations.

Jason Flannick is the final member in the Altshuler lab that I would like to thank. Even though he may have been one of the busiest members of the Altshuler lab outside of David, he was still willing to answer questions whenever I had trouble using this pipeline that he

had developed. I also want to thank him for taking the time to look over my thesis.

I would not have been able to complete this thesis if it had not been for the support of my housemates. This past year, I feel like we have become more like brothers than housemates, supporting each other in times of hardship and celebrating during times of success. Especially these last few weeks, I have really felt your prayers and encouragement as I have been writing this thesis.

Finally I would like to thank my family for their support. For my parents who are always concerned about whether this project will be completed in a timely manner. I want to thank them for raising me, for providing me with an opportunity to go to an institution like MIT and for being with me every step of the way. Without them, it would have been exponentially harder to finish this project.

Contents

1 Introduction **15**

1.1 Human Disease Phenotypes are Inherited 15

1.2 Not all Traits Follow Mendelian Patterns of Inheritance 17

1.3 Common Diseases 18

1.3.1 Linkage Mapping fails for Complex traits 19

1.4 Genome-Wide Association Studies 19

1.4.1 The Relationship between Conservation and Selection 21

1.5 Simulations 21

1.6 Limitations of the Gene Model 24

1.6.1 The size of every gene is not constant 25

1.6.2 Causal Mutations in the non-coding regions 25

1.7 Roadmap of project 27

2 Reproducing the results of Agarwala et al. **29**

2.1 Overview 29

2.2 The Gene Model 29

2.3 ForSim Overview 30

2.4 ForSim Input 31

2.4.1 Calculating the Genetic Phenotype 32

2.5 Assigning Disease Status 33

2.6 Analysis of Output 34

3 Modification to the Gene Model **41**

3.1 Overview 41

3.2 Modeling Human Genes 42

3.3 Conservation and Selection 44

3.4 Comparisons with Empirical Data 49

3.5 Analysis of Small Sample with Approximate and Exact Models 50

3.6 Analysis on Large Sample with Approximate Model 57

3.7 Conclusion 64

4 Model Limitations and Future Steps **65**

4.1 Limitations 65

4.2 Future Steps and Implications 66

A References **69**

List of Figures

1-1 The results from Agarwala et al.[20] Each is divided into four parts, each representing one of the four tests. Arrows pointing up indicated that values for simulated data were higher than that of empirical while arrows pointing down indicated that values for simulated data were lower than that of empirical. Green boxes showed that results from all four tests for simulated population were consistent with that of the european populations in T2D. 24

1-2 The distribution of the total gene length of 500 random genes.. 26

2-1 The fitness in ForSim is calculated as a sum of the environmental phenotype plus the genetic phenotype. 31

2-2 Diagram mapping out how individuals are chosen for the next generation. Every individual is assigned a fitness score, which represents the probability that an individual gets put into the pool from which the next generation are drawn from. The individuals for the next generation are chosen randomly from this pool of possible individuals. 32

2-3 Figure showing how the disease status is assigned in the population. Note that this is if the population had a normal distribution. The important part is that individuals with the 8% highest Phenotype score, if the disease is Type 2 Diabetes, are cases. 35

2-4 Plots for this GWAS study as target size is constant at 50 and case/control sample size is at 2500. In **a)** is τ =0, in **b)** is $\tau = 0.5$ and in **c)** is $\tau = 1$. For each value of τ, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample. 37

2-5 Plots for this GWAS study as *tau* is constant at 0.5 and case/control sample size is at 2500. In **a)** is target size = 5 and in **b)** is target size = 50. For each value of Target Size, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample. 38

2-6 Plots for this GWAS study as *tau* is constant at 0.5 and target size is at 50. In **a)** is sample size = 500 and in **b)** is target size = 2500. For each Sample Size, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample. 39

3-1 The distribution of **a)** entire simulated region, **b)** number of exons, **c)** length of exons, and **d)** length of introns for the 500 simulated genes based off of 500 randomly chosen genes in the genome. 43

3-2 A flow chart of how the genes were chosen. Exon data for each gene was gathered from the NCBI gene database. Genes in regions with no conservation scores as well as Genes on the Y chromosome were not considered. 44

3-3 A plot of the LOWESS smoothing function applied to the conservation scores of an example gene. The top plot is a plot of the entire gene and the bottom plot is a zoomed in figure where only 1kb out of the 5kb flanking region is plotted on both sides. 46

3-4 Gene segment in the approximate model. Neutral and fitness impacting mutations occur throughout the segment. The probability that a mutation is fitness impacting is the proportion of subsegments that are conserved. 47

3-5 Gene segment in the exact model. In this model, each gene model is broken down into further conserved and non-conserved subsegments. Neutral mutations only occur in the non-conserved subsegments while fitness impacting mutations only occur in the conserved subsegments. 47

3-6 The distribution of selection coefficients as published in Kryukov et al. versus the distribution of selection coefficients for intron regions used in the gene model. 49

3-7 The distribution of selection coefficients as published in Kryukov et al. versus the distribution of selection coefficients for the flanking regions used in the gene model. 49

3-8 Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire simulated region, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 51

3-9 Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These counts came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 54

3-10 The site frequency spectrum for in the entire gene. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 55

3-11 The site frequency spectrum for **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These variants came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 56

3-12 Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire gene, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 59

3-13 Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire simulated region, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 61

3-14 The site frequency spectrum for the entire simulated region with a sample size of 379 individuals. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 62

3-15 The site frequency spectrum for **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These variants came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals. 63

List of Tables

3.1 Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in coding regions. 57

3.2 Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in intron regions. 58

3.3 Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in intron regions. 58

Chapter 1

Introduction

1.1 Human Disease Phenotypes are Inherited

It has been long recognized that physical characteristics can be passed on from generation to generation. The earliest theories of hereditary belonged to the Ancient Greeks. The first major theory of genetics was hypothesized by Hippocrates in the fifth century B.C. It is known as the "brick and mortar" theory[1]. The main idea was that hereditary material would be collected throughout the body and concentrated into the male semen, which developed into a human in the womb. Through this mechanism, Hippocrates believed that physical characteristics could be acquired. For example, a champion weight lifter who had developed massive biceps throughout his training would be able to pass his "big bicep" characteristic to his offspring through his sperm. Aristotle challenged this idea several decades later by pointing out that individuals with missing limbs often produced children with normal limbs. If the physical characteristics of a parent were passed to a child, how could the "limb" characteristic be passed on if an individual had no limbs in the first place[1]?

Two key independent discoveries in the late 19th century helped lay down the foundation of upon which modern genetics is based. The first was the publication of *The Origin of Species* by Charles Darwin in 1859[2], the publication where Darwin described his theory of evolution and natural selection. Natural selection states that genetic differences between individuals can make them more or less suited for certain environments. The individuals with the genetic material that resulted in advantageous traits were more likely to pass on their

genetic information. However, there was no mechanism to describe how this genetic information was passed on. In 1865, Gregor Mendel published *Experiments in Plant Hybridization*[3]. In his publication, he stated and discussed his observations from studying pea plants. There were seven phenotypes that were studied. A phenotype is a visible trait, such as the flower color, seed color, or stem length. Mendel observed that when plants with certain phenotypes were bred with each other, the ratio of the phenotypes was fairly constant. For example, if a yellow and a green seeded plant were bred together, this first generation of offspring would be all yellow seed plants. However, if this first generation bred with each other, the second generation of offspring would have a three to one ratio of yellow to green seeds.

Mendel came to three conclusions: inheritance of each trait is determined by "factors" that are passed on to descendants unchanged, an individual inherits one of these factors from each parent for each trait, and that a trait may not show up in an individual but can still be passsed on to the next generation. One phenotype appeared to be dominant over the other. In addition, these phenotypes had full penetrance meaning the presence of the factor of the dominant allele guaranteed the individual would have the dominant phenotype. In the yellow and green seed example, the two parents had two yellow and green factors respectively. Their children all had one green and one yellow allele, but they all were yellow because yellow was dominant over green. However, in the second generation of offspring, 1/4 of the plants had two yellow alleles, 1/4 had two green alleles, and 1/2 had one green and one yellow allele. This resulted in 1/4 of the plants being green and 3/4 being yellow. Plants with two of the same allele were called homozygous and those with two different alleles were heterozygous. These would be later known as the Mendelian Laws of Inheritance.

There have been many diseases that have been found to follow Mendelian Laws of Inheritance including Huntington's disease, Tay-sachs disease, Duchenne muscular dystrophy among others. The one common characteristic between all of these diseases is that they were single gene diseases[4].

1.2 Not all Traits Follow Mendelian Patterns of Inheritance

Carl Correns was one of the first to observe as early as 1900 that not all traits followed mendelian patterns of inheritance. He observed that there were certain traits that were more likely to be inherited with each other. He studied the plant *Mirabilis jalapa* and saw that the leaf color depended greatly on which parent had which trait. If green pollen fertilized a white stigma, the progeny were white, but if the sexes of the donors were reversed, the progeny were green. The phenotype seemed to depend on the identity of the parent which it came from and not on the actual phenotype.

In 1910, Thomas Hunt Morganwas able to combine the Boveri-Sutton chromosomal theory with Mendel's theory of inheritance to help explain what Correns had seen. The Boveri-Sutton chromosomal theory stated that the physical matter with which hereditary operated were chromosomes. Chromosomes came in pairs, one inherited from the mother and the other from the father. Morgan reported the sex-linked inheritance of white eyes in *Drosophila Melanogaster*, suggesting that the genes underlying these traits were physically coupled to the genes determining sex. The idea of "linkage groups" was developed to refer to the idea that genes on the same chromosome were more likely to be inherited together.

It was also discovered that recombination could occur between these linkage groups with the likelihood of recombination proportional to the distance between the two genes. Recombination is the event where homologous chromosomes, a set of one maternal and the corresponding paternal chromosome, exchange genetic information with each other resulting in a new combination of alleles. It occurs during meiosis, which is the process by which gametes(sperm and egg cells) are created. Before separating, there is an event of "crossing over" where each pair of homologous chromosomes exchange different segments of their genetic material to form recombinant chromosomes, neither of which is an exact copy of the original pair. In 1913, Alfred Sturtevant drew the first linkage map, a map showing the likelihood of two alleles being inherited together and thus the linear order of genes on a chromosome.

The linkage map is, with a few exceptions, a map of the distance between two alleles. If

the recombination rate is assumed to be constant over the entire chromosome, the closer two alleles are to each other in distance, the less likely recombination will occur between the two alleles. If recombination is less likely to occur between two alleles, there will be a greater the probability that they will be inherited together. Instances where this is not true are when there are recombination hotspots. Recombination hotspots are areas in the genome where the rate of recombination is elevated. However, they are spread sparsely with 25,000 hotspots in the entire human genome which has approximately 3 billion base pairs[5].

Linkage disequilibrium is the non-random association between two or more alleles. Alleles that are in the same LD block are inherited together because of their proximity on the chromosome. Therefore, LD blocks are entire regions of the chromosome that are likely inherited together because recombination rate of that region is low.

At every location, there are four possible bases: adenine(A), guanine(G), cytosine(C), and thymine(T). SNPs are specific locations in the genome that where two of these four bases are common in the population. The base that is more common is called the major allele and the base that is less common is called the minor allele.

In addition to mapping alleles that were inherited together, alleles could be mapped to diseases. By systematically correlating disease status with the transmission of particular alleles, it became possible to identify specific marker locations(and chromosomal regions) with which disease stauts was linked. This genetic mapping of alleles to disease in humans has resulted in the localization of genes underlying hundreds of 'Mendelian' disease phenotypes ranging from Huntington's Disease to Cystic Fibrosis[6].

1.3 Common Diseases

Most common diseases do not show Mendelian patterns of inheritance. For a disease to show Mendelian patterns of inheritance, it must be caused by single-gene defects[7]. The diseases that affect the largest number of people-Type 2 Diabetes(T2D), hypertension and others clearly have an inherited basis, but do not obey Mendelian properties and do not show patterns of recessive or dominant transmission in families.

The Biometrics movement in the 19th century viewed phenotypes as a continuously vary-

ing trait(such as height) rather than traits that showed discontinuous Mendelian inheritance. In 1918, Fisher resolved the controversy of how a disease trait should be viewed between the Mendelians and biometricians by pointing out that the variation of continuous traits could be explained by the combined action of a set of individual genes in his paper *The Correlation between Relatives on the Supposition of Mendelian Inheritance.*[8] He established that continuous phenotypes could result from the additive effects of many genetic factors(polygenic), each of which could be inherited in a Mendelian fashion and individually produce only a small effect on the total phenotype.

Common diseases are currently observed as a dichotomous trait. In 1965, D.S. Falconer suggested that dichotomous traits might be studied as if a continuously varying trait was underlying them; disease could be thought to result above a threshold on his continuous "liability" scale[9]. Many common diseases are already defined in this matter. For example, T2D is defined as having a Glycated hemoglobin level of above 6.5 percent in two independent tests. Glycated hemoglobin measures the percentage of blood sugar attached to hemoglobin[10].

1.3.1 Linkage Mapping fails for Complex traits

Linkage mapping, which had worked so well for rare Mendelian disease phenotypes, was only able to explain a small fraction of the total incidence of disease. This finding was consistent with the biometric hypothesis that common diseases may be polygenic. They may be caused by a large number of genetic mutations such that no individual mutation or marker linked to it shows any significant correlation with disease status.

1.4 Genome-Wide Association Studies

With linkage analysis unable to find the full set of causal gene, a new approach called genome-wide association studies(GWAS) was first used in 2005[11]. Instead of tracing the transmission of disease mutations through families, genome-wide association studies compared the frequencies of common polymorphisms across the genome for large numbers of affected and unaffected unrelated individuals.

The justification of this method was the common disease common variant hypothesis(CDCV)[12]. This hypothesis was ultimately grounded in two population genetic assumptions.

1. Human demographic history
2. Weak Natural Selection-causal alleles for common diseases do not have big effects on fitness and may not see a significant decrease in frequency over time.

The human population was known to have grown exponentially after a bottleneck[13]. When the population was small, every variant, even those with very few copies were considered common because of the small pool of total variants. When the population grew exponentially, if the selection against these variants was not strong enough, the frequency of the variants would not have decreased rapidly. The result was disease causing variants with small affect on overall fitness could appear at a common frequency in the current population.

The goal of GWAS was to find these common variants by looking across the entire genome for common variants and see if any are significantly associated with a disease. GWAS have only been made possible due to the rapid advances in technology in the early 2000s with the first human genome sequence completed in 2003[14]. For GWAS to work, millions of polymorphisms were identified across the genome. Single nucleotide polymorphisms(SNPs) are sites where 2 different alleles are both common in the population. The purpose of the International Hapmap project was to provide the data that could be used for GWAS studies[16]. The goal of the project was to provide a genetic map for SNPs that had at least a frequency of 1%. By 2007, the project had completed genetic maps of over 3 million SNPs in 270 individuals from four ethnically diverse populations.

The results of the first large-scale GWAS were published in 2007 for a large range of common human diseases traits[16]. Statistical standards were established and only variants with an association p-value of $< 5*10^{-8}$ were considered genome-wide significant after Bonferroni correction[15]. To increase the statistical power, larger numbers of unrelated samples were used[17]. The results of GWAS were fairly successful in finding numerous loci that were associated with common diseases with 114 being found for Type 2 Diabetes[18].

The translation of GWAS findings to actionable therapeutic and diagnostic insights has been challenging. This may occur for several reasons: the associated markers in most cases

are just located near the causal variation, the linkage blocks used in GWAS are often large and span multiple genes, and many variants are found in non-protein-coding regions with ambigious function. The total fraction of heritability explained by all the genome-wide significant loci discovered in GWAS has been limited for most common diseases, about 10% for Type 2 Diabetes[19].

1.4.1 The Relationship between Conservation and Selection

The two population genetic assumptions of CDCV were the human demographic history and causal mutations subjected to weak natural selection. Human demographic history is something that can be measured through fossil records and written records. Natural selection is the concept of mutations that have a negative impact on fitness will never reach high frequencies. It is difficult to measure how a mutation directly impacts the fitness of an individual so we sought other methods to help quantify natural selection against a mutation.

One possible solution to help quantify natural selection against a mutation is by looking at how well the base has been conserved between different species through evolution. Evolution of different species is a mechanism that occurs over a long period of time. As the two species split, some regions of the genome are changed while other regions are conserved. Natural selection determines which regions are conserved and which regions are not conserved by decreasing the fitness of individuals with mutations in the conserved regions. These conserved regions tend to have important functions in the body. If they didn't, mutations in these regions would not decrease the fitness of the individual. Therefore seeing how well a base has been conserved across several species is a good indication of how much negative selection there is against that base in the genome.

1.5 Simulations

The genetic architecture of a disease is the collection of the variants that contribute to the disease. Are these variants located in a few genes that each have a large effect size? Or are they located in many genes that each have a small effect size? Knowing the genetic architecture of human diseases has profound implications for the future of genetic research

and its impact on clinical medicine. For example, if a disease is caused by rare mutations of large effect, targeted diagnosis and therapeutics based on individual genome sequence will be much more successful.

In order to systematically evaluate which genetic architectures are plausible, it is necessary to compare the predictions of each model to empirical data from all available genetic studies in a unified framework. A paper from the Altshuler Lab, Agarwala et al. titled *To what extent can empirical data place bounds on the genetic architecture of complex human diseases?*[20] did exactly that. In this paper, experiments were performed to find models that were consistent with the cumulative results of studies already performed and which models could be excluded.

In order for the simulation to be accurate, the key forces of population genetics must be properly modeled. Mutations at some, but not all, loci across the genome have the potential to alter disease risk. Genetic drift, the random change in frequency of a variant in the population, and gene flow, the transfer of variants from one population to another, both influence the distribution of variants. Finally, natural selection results in the change in frequencies of variants that influence evolutionary 'fitness' or the composite of many traits that influence the chance of passing on the individual's genetic information to the next generation.

In the simulations done for the Agarwala et al., simple possible genetic architecture models were generated. These models considered only mutation, genetic drift, and purifying selection. If such simple models produced predictions inconsistent with empirical data, this does not imply that more complex models could not be consistent. However, if a simple model was consistent, then it can be concluded that its features are indeed plausible given current data. A three-stage framework was used: forward evolutionary simulation to generate multi-locus DNA sequence variation at large scale, mapping of genotype to phenotype under a range of disease models, and *in silico* prediction of genetic study results under each model.

Different genetic architectures were tested by varying 2 parameters, the total disease mutational target size T and a τ parameter. The number of disease variants carried by an individual was determined by T. Models of T ranging from 75kb to 3.75Mb were simulated. T was broken down into 'loci' that were each 2.4 kb. This size was chosen because it was the

'average' protein-coding gene from the RefSeq database[21] in terms on number of exons and introns and their size. 30, 100, 300, 500, 800, and 1500 loci were simulated. In the simulation, every variant has an effect on the overall 'fitness' of an individual, which is measured by a selection coefficient s and τ is how closely the value s for each variant is 'coupled' to that variant's contribution to the disease g as seen in Equation 1.1.

$$g = s^{\tau(1+e)} \tag{1.1}$$

where τ is the coupling parameter and e is drawn from a standard normal distribution. A τ value of 0 indicated that there is no correlation between the selection of variants with the variant effectson disease. A τ value of 1 indicated that variants with large effects on fitness have large effects on disease. Simulations were performed with τ values of 0, 0.1, 0.2, 0.3, 0.4, 0.5, and 1.

To define the set of genetic studies to simulate, results were collected from published genetic studies of T2D in European populations. These data included: epidemiological estimates of sibling relative risk, meta-analysis of linkage scans in 4,200 affected sibling pairs with T2D, discovery GWAS in 4,549 cases and 5,579 controls, replication of the top($p<0.0001$) signals from the discovery GWAS in an effective sample size of 55K, and *larger-scale meta-analysis* in 12,171 cases and 56,862 controls, followed by genotyping of top($p<0.005$) signals on the Metabochip genotyping array in 34K cases and 115K controls. The results are shown in Figure 1-1. The green boxes are the models that are consistent with all four tests and the red ones either have at least one study result that was inconsistent or had one study result that was excluded.

The results showed that no models with a τ value of 0 or 1 was a possible genetic architecture. This result is consistent with current knowledge of disease models because a τ of 0 would indicates that the frequency of variants are not correlated with the variant effects on disease and a τ of 1 indicates that each gene was tightly linked to the disease and would have been found through linkage mapping. In addition, we see that only models with at least 300 loci were consistent. This seems plausible if we look at our GWAS results. Having a minimum of 300 disease genes is realistic because the 114 known GWAS variants combined

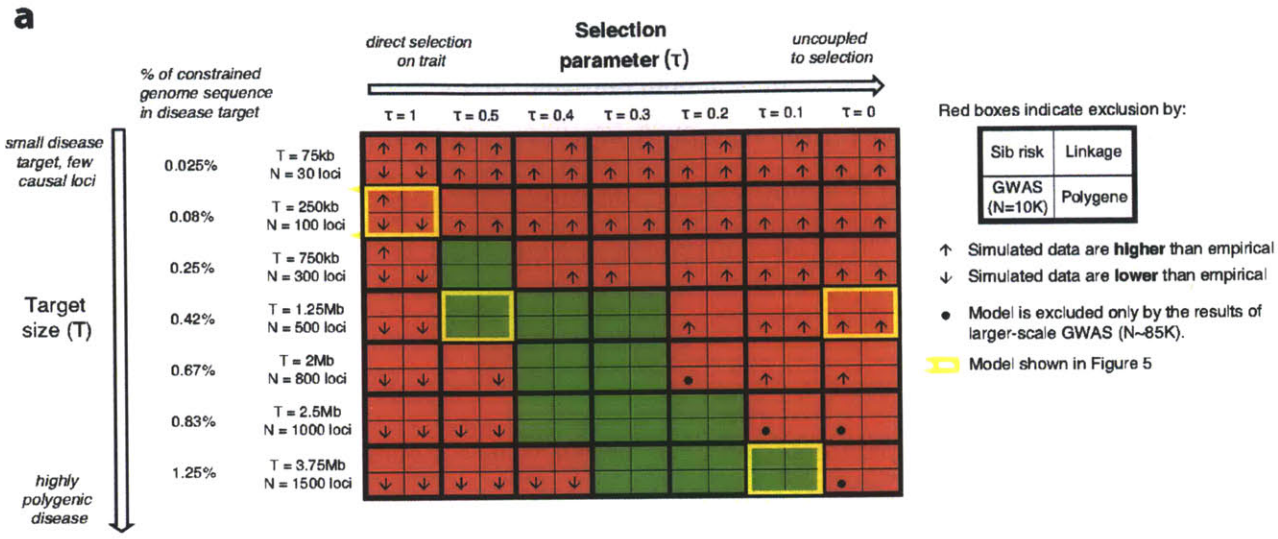


Figure 1-1: The results from Agarwala et al.[20] Each is divided into four parts, each representing one of the four tests. Arrows pointing up indicated that values for simulated data were higher than that of empirical while arrows pointing down indicated that values for simulated data were lower than that of empirical. Green boxes showed that results from all four tests for simulated population were consistent with that of the european populations in T2D.

only explain a fraction of the total heritability.

The main conclusion that can be drawn from the results of this paper are that there are some models that were consistent with empirical data and other models that were not. Some of the models that were found to be consistent contained genes that would not have been found through GWAS and linkage mapping.

1.6 Limitations of the Gene Model

The gene model that was used in Agarwala et al. was the same for every gene. But, in the human genome, gene lengths differ. The number of coding regions or length of each coding region also differs for every gene. In addition, the model allowed only mutations that occur in the coding regions to have the possibility of having a negative impact on fitness. This model is limited in two ways: one gene size cannot represent every gene and not all causal mutation are in the coding regions.

1.6.1 The size of every gene is not constant

There is a wide range of sizes of genes in the genome. The total gene length of 500 randomly selected genes are shown in Figure 1-2. The total gene length consists of all the coding regions and the non-coding regions in between the coding regions as well as 50kb flanking regions on each side. 50 kb regions were chosen because that is the general distance that influences the gene [22].

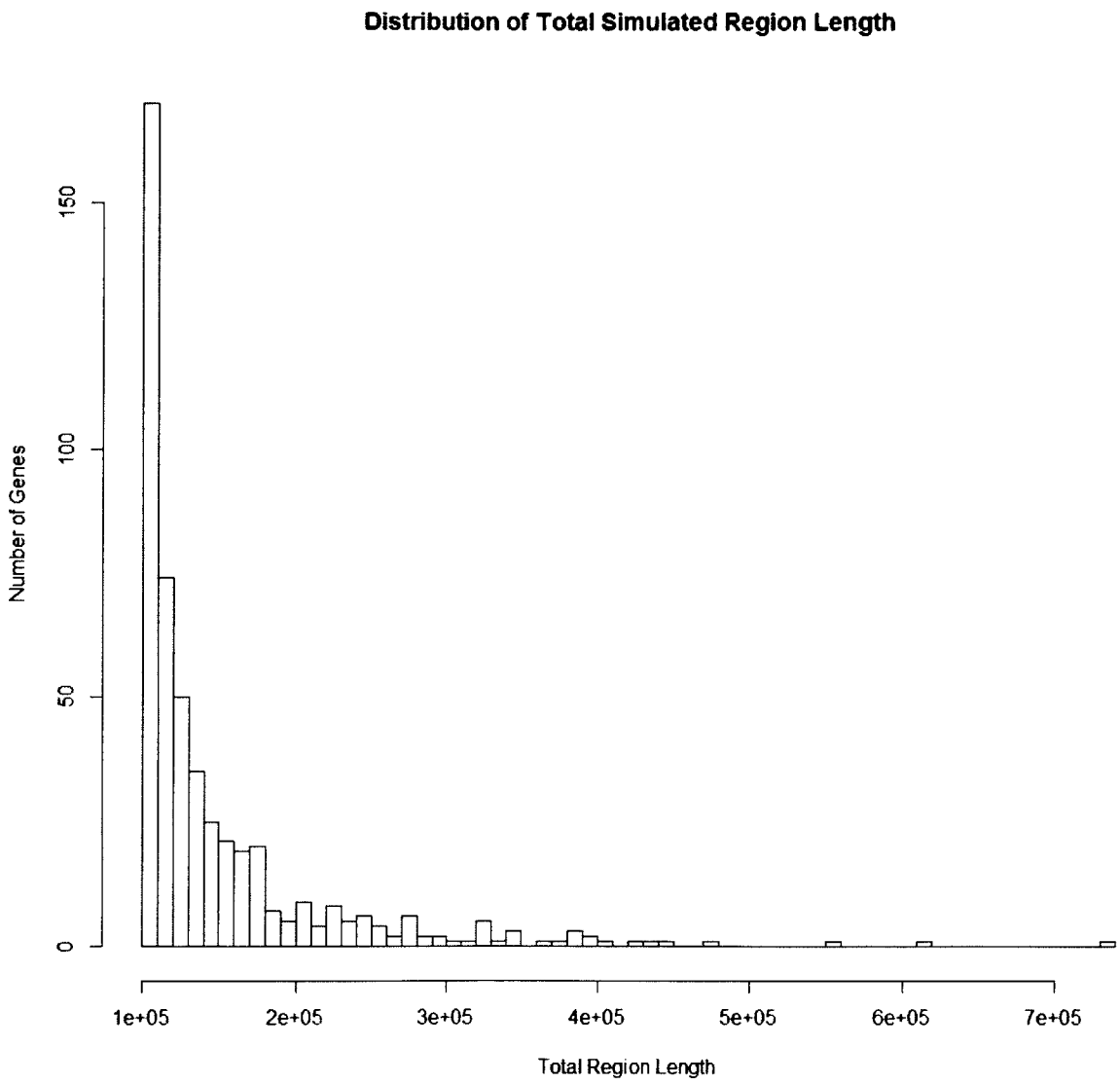


Figure 1-2: The distribution of the total gene length of 500 random genes..

The distribution of gene length does not follow a normal distribution. The distribution

has a one-sided tail. There is no simple conclusion on how to pick one gene length that would be representative of this distribution. If a gene with the median length was chosen, then the tail would simply be ignored. If a gene with mean length was chosen to symbolize all of the genes, then the genes in the genome that are in the tail would heavily skew the mean. By picking one gene length, the genes at the end of the tail will either be ignored or will have too much weight. Therefore the most accurate way to model this distribution is to have every gene be of different lengths and the distribution of different lengths model the distribution of different lengths of genes in the genome.

1.6.2 Causal Mutations in the non-coding regions

The gene model in Agarwala et al. required mutations that affect fitness to be in the coding regions. Currently, there is limited understanding of how non-coding regions affect biological processes and by extension the fitness of an individual.

One function of non-coding regions is to encode microRNA. Regular mRNA is the transcript from the coding region that is used to make the protein. MicroRNA however is a ncRNA(non-coding RNA). There have been studies shown that widespread disruption of microRNA has been seen in human cancer. In addition, there are many other types of ncRNA such as small nucleolar RNA, transcribed ultraconserved regions and large intergenice non-coding RNAs. Disregulation of these ncRNAs have been found in neurological, cardiovascular, developmental and other diseases [23].

Disruption of ncRNA is just one example of the affect of mutations in non-coding regions. Even though many of these pathways are not currently understood well, they are still important in the functionality of an individual. The absence of these mutations in the gene model used in Agarwala et al. is a limitation of that model.

1.7 Roadmap of project

The goal of the project described in this thesis is to extend the gene model used in Agarwala et al. in two ways. The first is modeling the distribution of number and length of coding and intron regions in the genome. This first extension will be referred to as

"modeling the distribution of gene length". The second extension is to model mutations in the non-coding regions that affect an individual's fitness in the simulations.

To be able to model the distribution of gene length in the genome, a large bank of genes will be built. This bank of genes must be large enough that the distribution of gene length will reflect the distribution of gene length in the genome.

Modeling mutations that affect fitness in non-coding regions is not straightforward because there is no direct way to measure how a mutation affects fitness. Instead, information on how well a base in the genome has been conserved was used to model these mutations. The more a mutation negatively impacted fitness, the stronger the selection is against that mutation. If there is strong selection against a mutation, it will be conserved over a long period of time.

There were two models constructed using this reasoning, an exact model and an approximate model. In the exact model, for every conserved base in the genome, there was one base in the simulated gene that would have selection against it. Mutations that occured in a base with selection against it would have a negative effect on fitness. For example, if there were 50 bases in a coding region that were conserved followed by 50 bases in a coding region that were not conserved, mutations that occured in the 50 bases that were conserved would have a negative impact on fitness while mutations that occured in the non-conserved segment would not have an impact on fitness. In the approximate model, for each segment, the percentage of 50 base subsegments that are conserved is the percent of mutations that have a negative impact on fitness.

Before the new model was implemented, the gene model of Agarwala et. al was first reproduced as a baseline for comparison. It was important to learn how this original gene model worked before it could be extended. Several genetic architectures were simulated and for each gene model, the results were what was expected from our knowledge of population genetics.

After the new gene model was implemented, it was first tested to see how well it was able to model a small sample of ten random genes. In this test, both models were simulated. Comparisons were made between the two models as well as empirical data of the ten genes and the model in Agarwala et al. Many annotations of regions were made including conserved

and non-conserved regions of coding, intron, and flanking regions. The purpose of this initial comparison was to test how well the model for fitness impacting mutations in non-coding regions worked. The comparisons between all four models showed that both the approximate and exact models were able to model non-coding fitness impacting mutations.

Next, a bigger sample of 500 genes, were simulated. The purpose of this bigger sample was to create a bank of genes where the distribution of gene length is consistent with that of the genome as well as confirm that the model for fitness impacting mutations in non-coding regions was still fairly accurate. Even though this larger sample only included approximately 2% of all the genes, the distribution of total gene length in this sample was representative of the distribution of total gene length in the genome. For this sample, only the approximate model was simulated because of computational performance considerations. Simulating the exact model took more than ten times the time of the approximate model. The comparisons between the three models showed that the approximate model was still able to model non-coding fitness impacting mutations.

Chapter 2

Reproducing the results of Agarwala et al.

2.1 Overview

Before the new gene model was implemented, it was first necessary to show that the results from Agarwala et al could be reproduced because this model would later serve as a baseline to which new models would be compared to. We describe the gene model and forsim, the forward evolutionary simulation software that is used. We simulates several genetic architectures and compared if the results to expectation.

2.2 The Gene Model

The gene model that was used in Agarwala et al. was designed to represent what an ' 'average" gene looked like in the genome. This was done by looking at the protein-coding genes from the RefSeq database[21]. The median number of exons, median total coding length and median total transcript length were used. The gene had the following characteristics.

1. 8 exons-each 300 bp long for a total coding length of 2.4k bp
2. 7 introns-each 3k bp for a total of 23.4 kb
3. 100 kb neutral flanking regions on both sides

4. Mutation rate constant across the gene.

In addition, only mutations in the exons could have a negative impact on the fitness of an individual. The synonymous and non-synonymous variants were modeled. 30% of the exonic variants are synonymous while 70% are non-synonymous. Synonymous variants are variants that do not change the protein sequence. This is because of the wobble effect, the concept where multiple sequences code for the same amino acid, and therefore have no effect on fitness. Approximately 80% of non-synonymous variants have an effect on fitness. The reason that there are some non-synonymous variants that do not effect fitness is that a change in amino acid sequence does not guarantee change in the protein structure. Therefore, 56% of mutations that occur in the exons will have a negative effect on fitness while the rest will have no effect.

The distribution of selection coefficients is a gamma distribution. The parameters for the gamma distribution were the set of parameters that resulted in the site frequency spectrum being the most consistent with empirical data. The site frequency spectrum is the distribution of the variants based on frequency. The empirical data used for these comparisons was the European population in T2D. The selection coefficient was the parameter that indicated how much of an effect a mutation had on the fitness of an individual. The more negative the selection coefficient, the greater the negative effect it would have on fitness. A shape parameter of 0.316 and a scale parameter of 0.01 was used. The mean for this gamma distribution was 0.00316 and the variance is 0.000032. Only mutations with negative impact on fitness were modeled so the selection coefficients that were drawn from the gamma distribution were multiplied by negative one.

2.3 ForSim Overview

ForSim is a forward evolutionary simulation system designed to be highly flexible. It takes in a list of parameters, including a gene model, mutation rate, population size among others and outputs several files. The version of ForSim used was developed by Brian Lambert and Ken Weiss when both were at Penn State University and modified by Vineeta Agarwala and Jason Flannick in the Altshuler Lab to decrease the runtime of the software.

Currently the software outputs two files, a ped file that contains a list of all the individuals in the final generation as well as all the minor alleles each possesses and a marker file that has a list of all the markers currently in the population as well as their frequency, location, and the identity of the minor and major allele are.

Analysis on the population can be performed by running tests of the ForSim output files. These tests include tests for the number of GWAS statistically significant variants.

2.4 ForSim Input

This section will provide detail on the parameters for the ForSim software and how ForSim creates a simulated population. In ForSim, every individual is assigned a fitness score. This fitness score corresponds to the fitness of an individual with a higher fitness score corresponding to a greater chance of survival. The score is a summation of the individual's genetic phenotype and the environmental phenotype as shown in Figure 2-1.

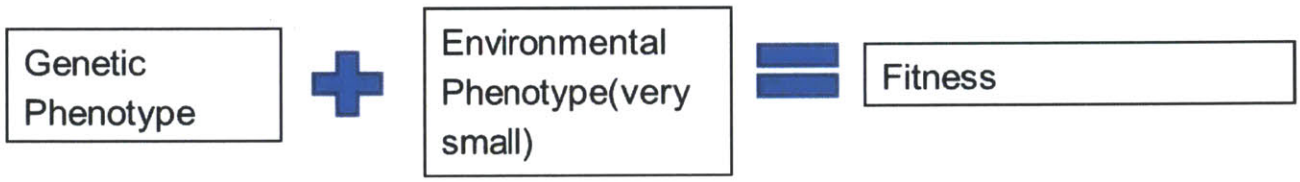


Figure 2-1: The fitness in ForSim is calculated as a sum of the environmental phenotype plus the genetic phenotype.

An individual's fitness is determined by both genetic and environmental factors. The environmental portion corresponds to factors such as diet and exercise. The environmental phenotype in this model is very small because the purpose was to model a population where the majority of the fitness was influenced by the genetic phenotype. The environmental phenotype was drawn from a normal distribution that had a mean of 0 and standard deviation of 0.0000001. For every individual, this fitness score will range from 0 to 1 and represent the probability of that individual gets into the pool from which the next generation is drawn from. For every individual, a random number is drawn from an uniform distribution from 0 to 1. If the fitness score is greater than the random number, then this individual will be considered for the next generation. The individuals for the next generation are drawn

randomly from this pool of possible individuals as shown in Figure 2-2. If an individual has a higher fitness, then it is more likely to survive to pass on its genetic information to the next generation.

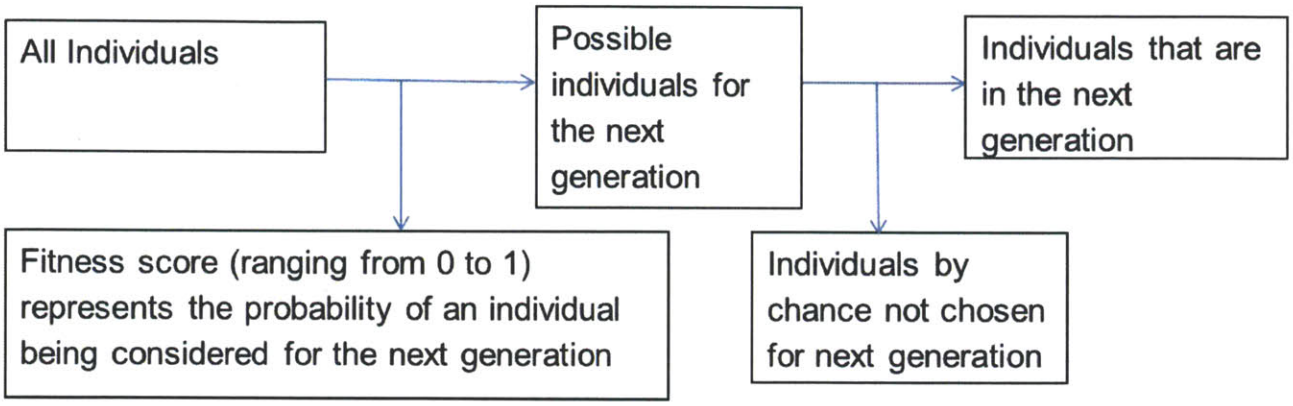


Figure 2-2: Diagram mapping out how individuals are chosen for the next generation. Every individual is assigned a fitness score, which represents the probability that an individual gets put into the pool from which the next generation are drawn from. The individuals for the next generation are chosen randomly from this pool of possible individuals.

The parameters of the population used in Agarwala et al. were tuned to the Northern European population. Initially, several previously published models of demographic history including those in Kryukov et al. and Gravel et al. were tested. These models were then modified until the site frequency spectrum of the simulated population was consistent with that of empirical data. A hybrid population was concluded to generate a simulated population that was the most consistent with empirical data. In this population, first 50,000 generations were simulated at a constant population size of 8100. This was followed by a bottleneck that reduced the population to 2000 and exponential growth for 370 generations to a size of 227,650.

2.4.1 Calculating the Genetic Phenotype

For each individual, the genetic phenotype starts at 1. For every fitness impacting mutation each individual has, the genetic phenotype decreases by the amount of that mutation's selection coefficient. The more negative the selection coefficient, the more it will affect the individual's genetic phenotype and ultimately the fitness of the individual. This is shown in

Equation 2.1.

$$1 + \Sigma s = GP \tag{2.1}$$

where s is the selection coefficient for every variant an individual has and GP is the genetic phenotype for that individual.

Additionaly, ForSim allows the user to set parameters that only apply to certain segments of the gene. These include the probability that a mutation that occurs has an impact on fitness and the distribution of selection coefficients. The different segments that are modeled are coding region, intron, and flanking, as described in Section 2.2.

2.5 Assigning Disease Status

There are several steps to determine diisease steps. The first step is to calculate each mutation's additive contribution to the disease risk score. This was calculated using Equation 2.2,

$$g = s^{\tau(1+\mathrm{e})} \tag{2.2}$$

where τ is one of the coupling parameters and e is drawn from a normal distribution with mean of 0 and standard deviation of 1. The second step was to calculate an individual's heritable phenotype G as shown in the following equation

$$G = \sum_{i=1}^{n} g_i \tag{2.3}$$

where g_{i} is the mutation additive effect for variant i and n is the total number of variants an individual has across all N target size genes. An individuals total Phenotype P is

$$P = z(G) + \sqrt{\frac{1-h}{h}} * E \tag{2.4}$$

where $z(G)$ is the z-score of G, E is the environmental phenotype drawn from a normal distribution with mean 0 and standard deviation 1, and h is the percent of variance that is due to heritability, which is 0.45 for Type 2 Diabetes. Disease status was calculated using a threshold derived from the prevalence of the disease. The threshold was calculated so that the percent of individuals with the disease in the simulated population would equal the prevelance of the disease in the real world. For example, if 8% of the population has the disease, then the 8% with the greatest P have disease status.

2.6 Analysis of Output

Data was generated for several genetic architectures by varying three parameters: coupling factor τ, the sample size of the study, and the target size. This data was then used to perform a GWAS and Manhattan and QQ plots were generated for analysis. τ values of 0, 0.5, and 1 were used. Target sizes of 5 and 50 were studied and the two sizes of GWAS studies that were done were 2500 cases and controls and 500 cases and controls.

500 genes using the gene model in Agarwala et al. were first simulated. The disease genes were chosen at random from the list of 500 genes. Next, the variant's additive effect was assigned based on the τ value. An individual's heritable phenotype as well as there total phenotype were assigned. The prevalence for type 2 diabetes was 8%. Disease status was then assigned and the case and controls were then drawn from the pool of diseased and non-diseased individuals at random.

A discovery GWAS study was then performed on the common variants with a frequency of greater than 5%. LD pruning was applied. LD pruning is randomly choosing one SNP to represent a group of highly correlated SNPs. A replication study was then performed in an independent sample of the variants that had a P-value < 0.0001, the replication threshold in the discovery sample. In this study, SNPs were declared significant if the P-value was less than 0.05 divided by the number of replication SNPs. This calculation of the P-value is from the Bonferroni adjustment where P-value equals 0.05 divided by number of independent tests. In this calculation, each SNP is treated as independent [15]. Manhattan and QQ plots were generated for the discovery sample and Manhattan plots were generated for the replication

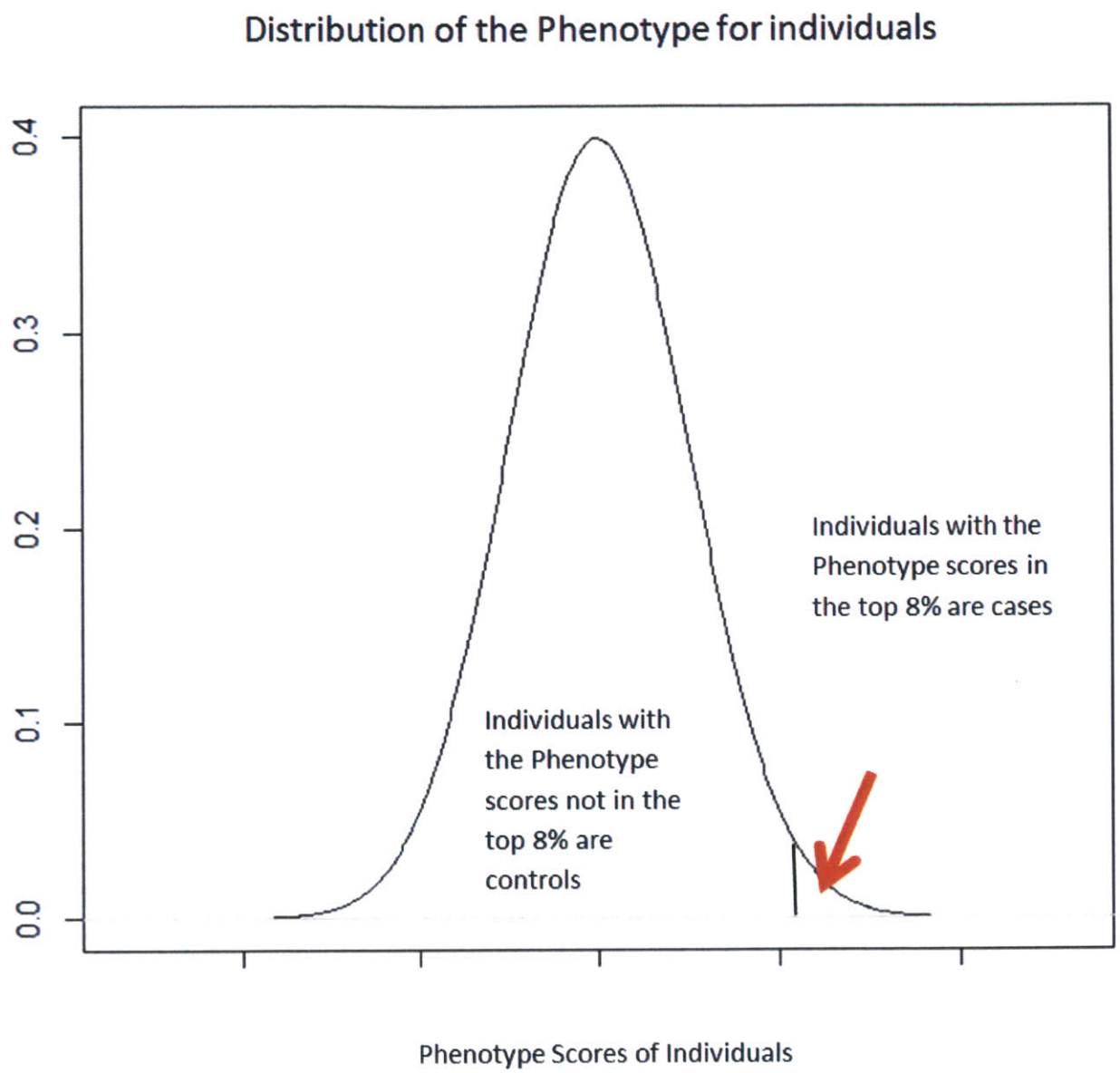


Figure 2-3: Figure showing how the disease status is assigned in the population. Note that this is if the population had a normal distribution. The important part is that individuals with the 8% highest Phenotype score, if the disease is Type 2 Diabetes, are cases.

sample. Manhattan Plots are plots that have every variant plotted according to base pair position on the x-axis and the y axis is the -log(p-value). This plot allows you to see those variants that have very small p-values easily as the smaller the p-value, the higher the point. For QQ plots, expected -log(p-value) is plotted against observed -log(p-value). The expected

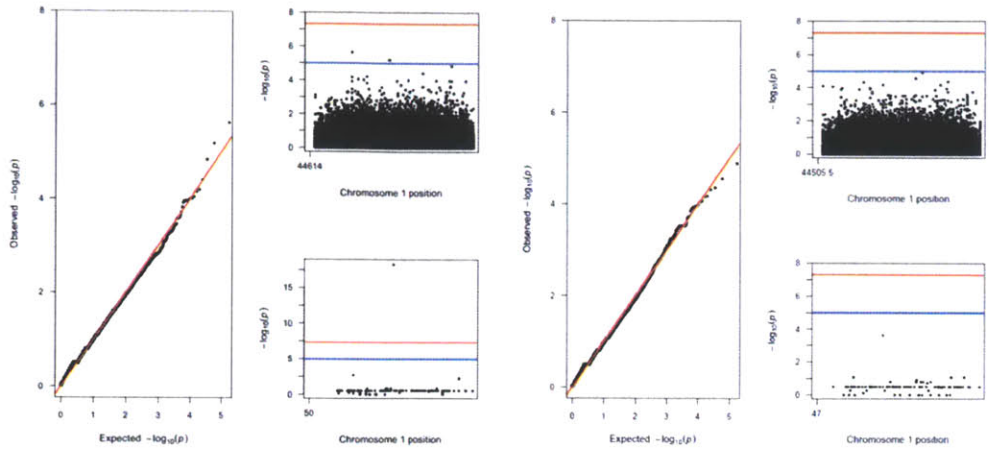
-log(p-value) is the distribution of -log(p-value) from a random distribution. If the plot starts to rise from a straight line(shown in red), then mutations that have a lower p-value then expected are present. Otherwise, the points will follow the straight line. The purpose of the Manhattan plot is to see if there are any variants that have a significantly low p-value and the purpose of the QQ plot is to see if there are any variants that have a lower p-value then expected from a random distribution.

The results are organized to see what changes are seen when one of the parameters has been modified. The first parameter that was modified is the τ value. Target size is kept constant at 50 while sample size for both the discovery and replication studies is 2500 cases and 2500 controls. Figure 2-4 shows the Manhattan and QQ plots for the discovery and Manhattan plot for the replication sample for when τ equals 0, 0.5, and 1. One observation seen is that as the τ value of 0 has more SNPs that are correlated with disease that tau values of 0.5 or 1 if a threshold of -$\log_{10}(5)$ is used. There are 3 SNPs in the discovery sample for tau equals 0, while there are only one when tau equals 0.5 or 1 in the discovery sample. This makes sense because when τ is high, variants with small selection coefficients are going to have the largest effects. However, these variants will not show up in the study because they are rare and only common variants were included in this study. If a larger sample size was used, there would be more statistical power resulting in the possibility of seeing more SNPs correlated with disease in the model where tau equals 0.5 compared to the model where tau equals 1.

The second parameter that was modified was the target size. The τ value was kept constant at 0.5 and the sample size was constant at 2500 cases and 2500 controls for both the discovery and replication studies. Figure 2-5 shows the Manhattan and QQ plots for the discovery and Manhattan plot for the replication sample for when target size equals 5 and 50. More associated variants are seen with the smaller target size. This is consistent with what was expected because with a smaller target size, there are fewer variants that contribute to the disease and thus every variant must have a larger effect and would have a smaller p-value.

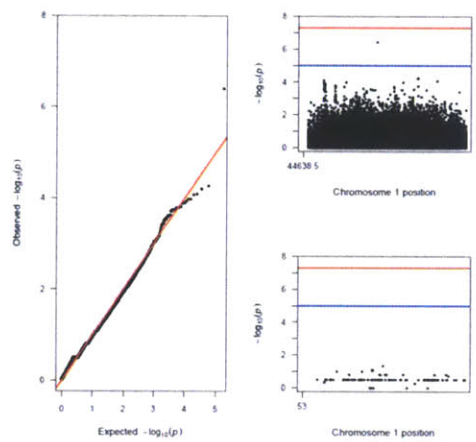
The third and final parameter that was modified is the sample size. The τ value was kept constant at 0.5 and the target size was constant at 50 genes. Figure 2-6 shows the

Manhattan and QQ Plots for studies where τ is varied



(a) $\tau = 0$.

(b) $\tau = 0.5$.

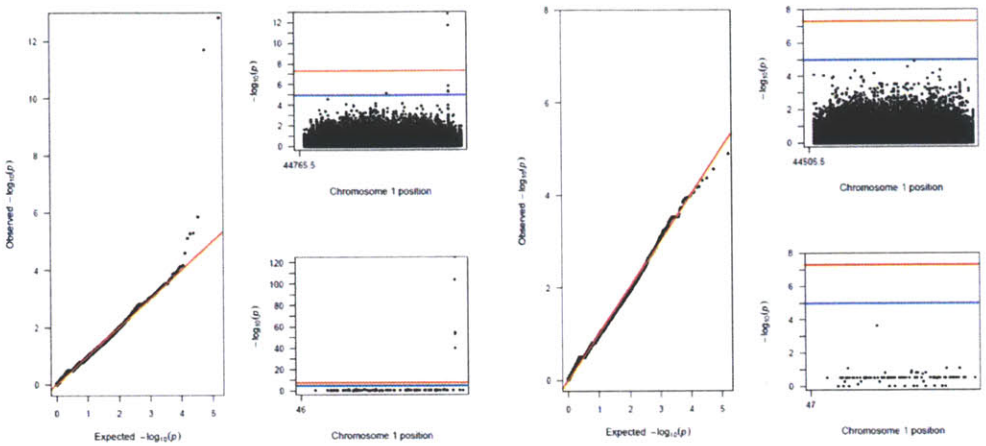


(c) $\tau = 1$.

Figure 2-4: Plots for this GWAS study as target size is constant at 50 and case/control sample size is at 2500. In **a)** is τ =0, in **b)** is $\tau = 0.5$ and in **c)** is $\tau = 1$. For each value of τ, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample.

Manhattan and QQ plots for the discovery and Manhattan plot for the replication sample for when sample size equals 500 and 2500. In the larger sample size, there are a limited number of variants that are seen to be significantly associated with the disease, but none are seen in the smaller sample size. This is consistent with what was expected because the small sample size did not provide the statistical power needed to be able to see variants that

Manhattan and QQ Plots for studies where target size is varied



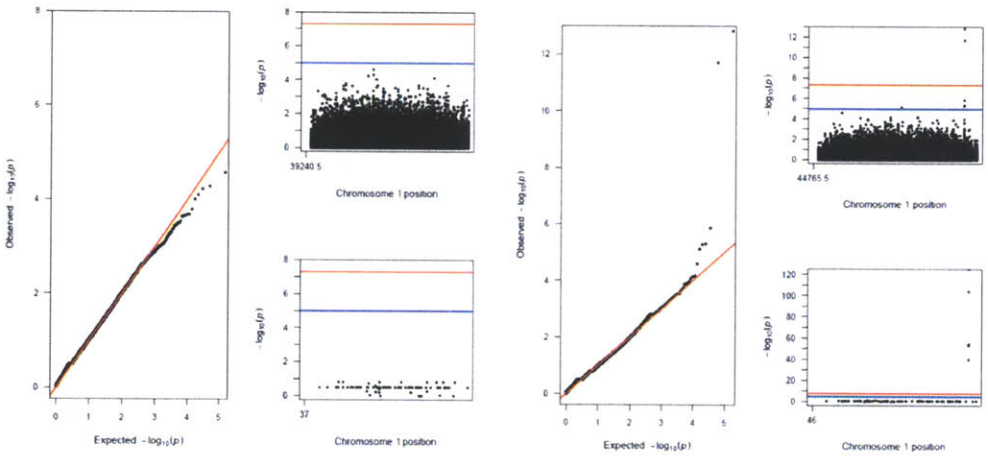
(a) target size=5. (b) target size=50.

Figure 2-5: Plots for this GWAS study as *tau* is constant at 0.5 and case/control sample size is at 2500. In **a)** is target size = 5 and in **b)** is target size = 50. For each value of Target Size, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample.

are associated with the disease and detectable using the statistical association test.

In conclusion, results using the gene model from Agarwala et al. were reproduced as the model was studied. Several studies were performed and three parameters were changed in the gene model, τ, target size, and sample size. The dependence of the GWAS results on the input parameters were consistent with what was expected. The model in Agarwala et al. will be used as a baseline in which all new models will be compared against.

Manhattan and QQ Plots for studies where sample size is varied



(a) Target Size=5. (b) Target Size=50.

Figure 2-6: Plots for this GWAS study as *tau* is constant at 0.5 and target size is at 50. In **a)** is sample size = 500 and in **b)** is target size = 2500. For each Sample Size, the plot on the left is the QQ plot for the discovery sample. The plot on the top right is the Manhattan plot for the discovery sample and the plot on the bottom right is the Manhattan plot for the replication sample.

Chapter 3

Modification to the Gene Model

3.1 Overview

The gene model that was used in Agarwala et al. was a model that represented what the "average" gene looked like in the genome in terms of protein coding exon and intron length and number as well as total transcript length. In addition, only mutations that occurred in protein-coding regions had a non-zero probability of having an impact on fitness. For the remainder of this chapter, this gene model will be referred to as the "static" model. The goal of this chapter is to improve the static model by applying two main modifications.

The two main modifications that will be applied to the gene model are as follows:

1. The number of protein-coding exons, intron, and their length come from a distribution that is representative of these characteristics in the genome.

2. Mutations that affect fitness in non-protein coding regions are included. The probability of these mutations will depend on how well the regions are conserved.

The purpose of these two modifications is to make the simulated genes more accurately represent the genes in the genome. The second modification addresses the fact that mutations in both the non-protein coding regions as well as protein-coding regions could impact fitness in a population.

Comparisons of the new model with empirical data will be performed on a smaller set of 10 random genes before building a bank of 500 genes. The purpose of the comparisons

with the smaller sample set is to test how well the fitness affecting mutations in the non-coding regions are being modeled. The purpose of the bigger sample set is to build a bank of genes where the distribution of number and length of coding and intron regions represent the distribution in the genome. The empirical data that is used for comparisons is the European population in the 1000 Genomes project. There will be two types of comparisons. The first comparison is the number of singletons, the number of variants that have a rare minor allele frequency(less than 1%) , the number of variants that have a low minor allele frequency(between 1% and 5%), and the number of variants that are common(frequency greater than 5%) between the simulated and empirical data. Singletons are variants that only show up once in the entire population. The minor allele frequency is the frequency of the minor allele. The second type of comparison will be comparing the site frequency spectrum of the simulated and empirical data. The site frequency spectrum plot has the minor allele count on the x-axis and number of variants on the y-axis. The minor allele count is the number of minor alleles in the population. The purpose of both of these comparisons is to compare the distribution of frequencies of the minor alleles in the population.

3.2 Modeling Human Genes

The goal of the first modification was to create a bank of genes that was representative of real human genes in terms of number and length of protein-coding regions and intronic regions. In this project, a bank of 500 genes was created. In Figure 3-1, the distribution of entire simulated region, number of exons, length of exons, and length of introns are shown. All of the distributions show a one-sided tail. By modeling each gene in the genome, genes that are located in the tail will be included in the simulation.

The protein-coding regions for every gene was obtained from the Consensus Coding Sequence Project(CCDS) [24]. CCDS Project is a collaboration between the National Center of Biotechnology Information, European Bioinformatice Institute, University of Santa Cruz, and Wellcome Trust Sanger Institute to agree upon a consistent set of protein-coding genes for humans. The latest release of CCDS, that was released 11/29/2013, was used with over 20,000 genes. NCBI base 37 base pair units were used.

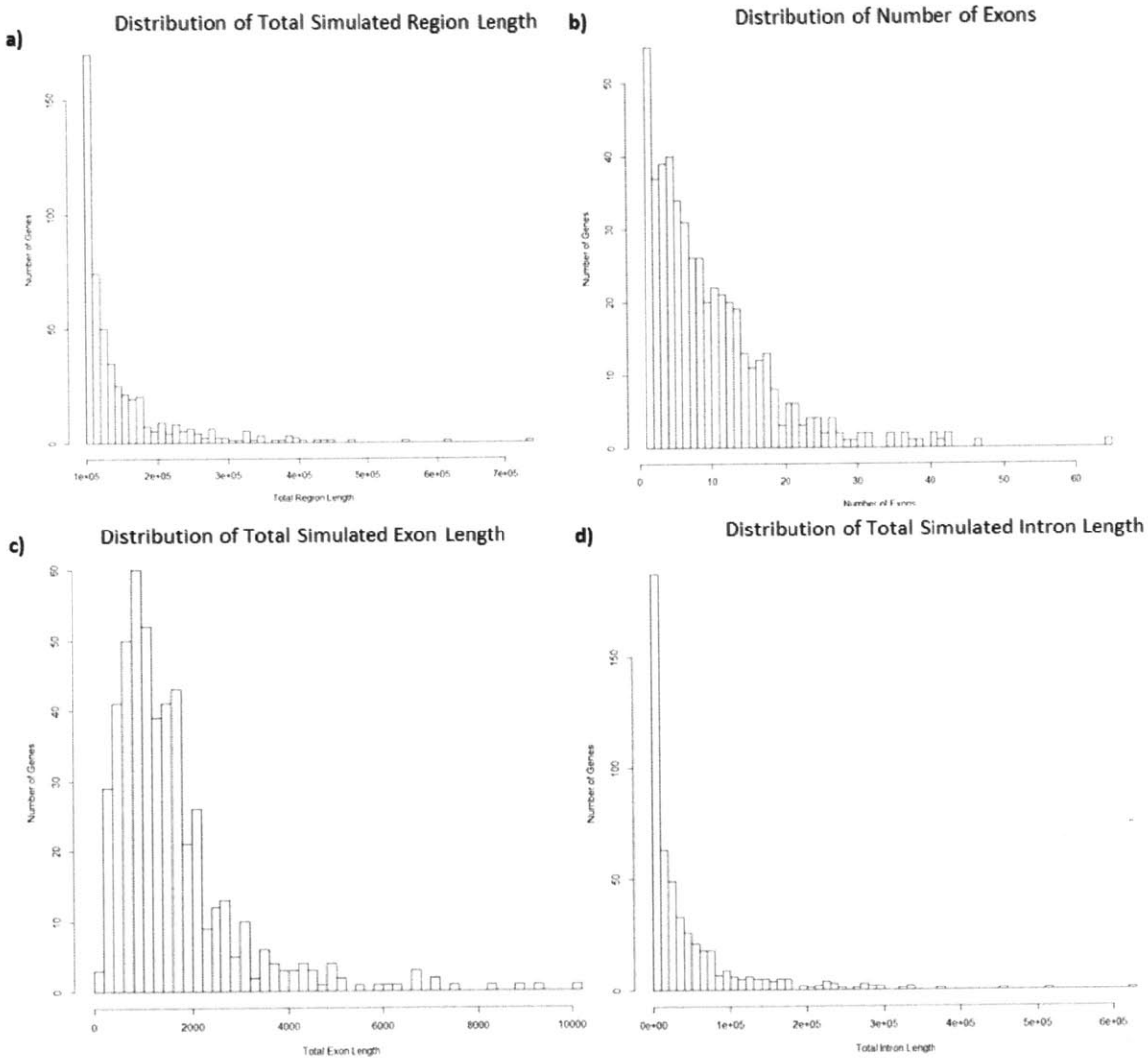


Figure 3-1: The distribution of **a)** entire simulated region, **b)** number of exons, **c)** length of exons, and **d)** length of introns for the 500 simulated genes based off of 500 randomly chosen genes in the genome.

In addition to simulating the protein coding regions and the intron regions that lay in between, 50 kb flanking regions were added on either side of the coding regions. 50 kb regions were chosen because that is the general distance that influences the gene [22]. Examples of how these flanking regions can influence the gene include coding for ncRNA or other molecules that can affect protein expression or protein structure.

There were several genes that were excluded in this project. CCDS genes that had the

status "Withdrawn" or "Review" were not considered. Because conservation scores will be needed to build the simulated genes, if the scores were not available for the bounds of the entire gene region including the 50 kb flanking regions, the gene was not considered. Genes with scores that were missing for small segments within the gene region were considered. Comparisons with empirical data will be made, which caused us to exclude genes on the Y chromosome because empirical data was not obtained for the Y chromosome.

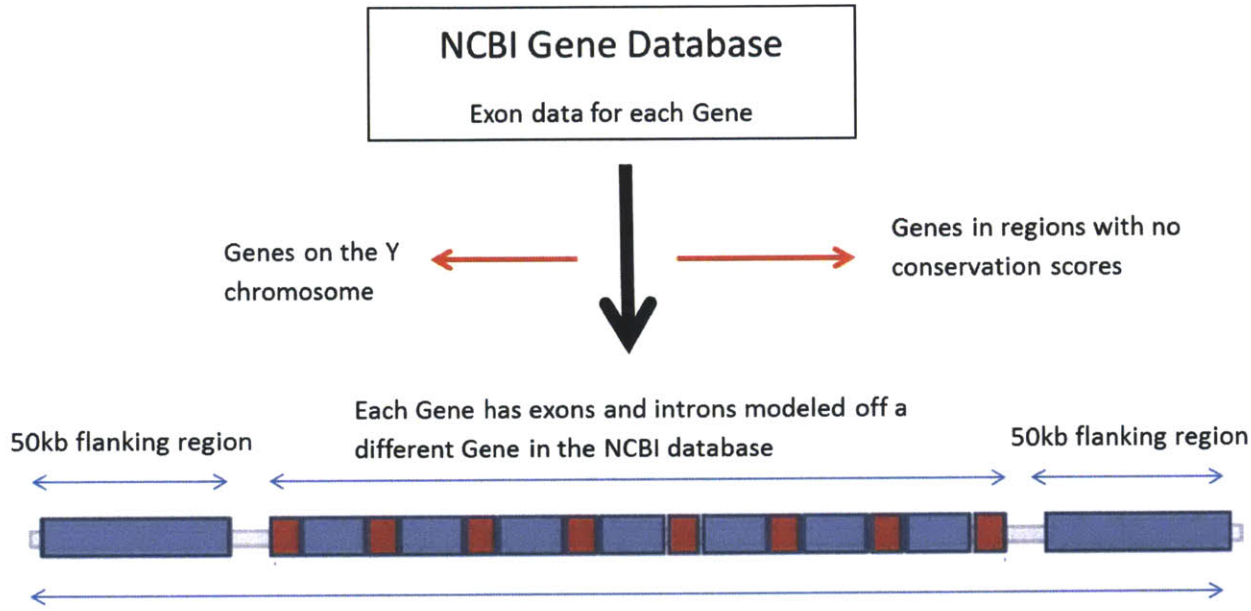


Figure 3-2: A flow chart of how the genes were chosen. Exon data for each gene was gathered from the NCBI gene database. Genes in regions with no conservation scores as well as Genes on the Y chromosome were not considered.

3.3 Conservation and Selection

The goal of the second modification was to incorporate mutations in the non-coding regions that impacted the fitness of an individual in the simulation. In the "static" model, the percent of mutations in the coding regions had been based off the percent of mutations that affected protein structure. Because non-coding regions of the genome do not have direct impact on the structure of a protein, a different approach will be taken.

The more negatively a mutation impacts the overall fitness of an individual, the greater the selection there is against that mutation. One way is to measure the selection pressure

against a base is to see how conserved the base is over time. If a particular base has strong negative selection, then a mutation at that base would be phased out over time. Therefore, a base that undergoes strong negative selection would have a higher probability of being passed down intact for many generations.

In this project, we calculated the selection pressure of a particular region by measuring how well that region of the genome was conserved across different species. By looking at how conserved a region in the genome is conserved over different mammalian species, we can observe how well that region has been conserved over millions of years of evolution. The conservation scores that were used were scores that looked at how well each base in the genome was conserved over 29 mammalian genomes[25]. The scores were downloaded from the UCSC genome browser and were split by chromosome[26]. The scores were available sequentially at every base pair.

The type of score that was used was the Phastcons score. This score is a number between 0 and 1 and represents the probability that the base in the genome is conserved. The score also takes into account how conserved the surrounding region is. Figure 3-3 shows a plot of the conservation scores of a randomly chosen gene. Each dot represents the average of a 50bp segment. The red dots are the variants in non-coding regions while the blue dots are variants that are in coding regions. There is also a line at 0.5 with variants above the line having more than a 50% chance of being conserved while those below the line have less than a 50% chance of being conserved. The majority of the coding variants are above the line while the majority of the non-coding variants are below the line, which is what we expected because most of the variants known to affect fitness are in coding regions. The boxes under the conservation scores are the coding regions.

The next step in building the simulated gene was to decide how to incorporate the conservation scores into the gene model. Each coding/non-coding section of each gene was broken up into 50 bp sub segments. 50 bp were chosen because that was the length of segment used in Kryukov et al.[25] when they were determining whether a region of the gene was conserved. A segment was considered conserved if the average conservation score was above 0.5. This cutoff was chosen because it indicated that the segment had more than a 50% chance of being conserved. If the coding/non-coding segment length was not an

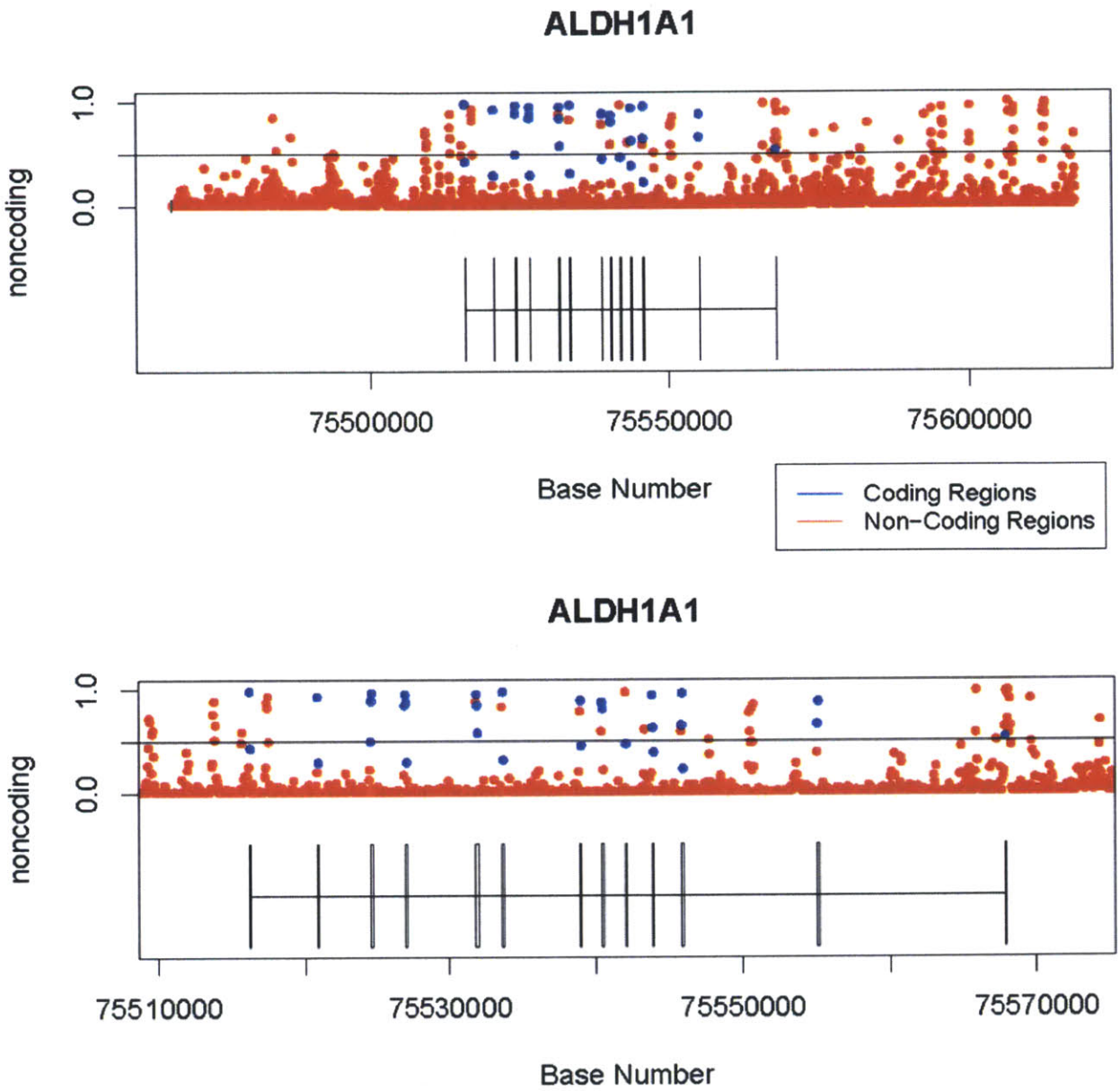


Figure 3-3: A plot of the LOWESS smoothing function applied to the conservation scores of an example gene. The top plot is a plot of the entire gene and the bottom plot is a zoomed in figure where only 1kb out of the 5kb flanking region is plotted on both sides.

exact multiple of 50, then the last sub segment would be the remaining scores and would be weighted accordingly when adding up the number of conserved and non-conserved sub segments in each coding/non-coding segment.

There are the two proposals of how to incorporate the conserved segments into the sim-

ulated gene model.

1. An "approximate" model where the gene is broken it coding, intron, and flanking segments and the percentage of 50 bp subsegments that were conserved in each segment would be the percentage of mutations in that segment that had a negative impact on fitness as shown in Figure 3-4. Intron regions are non-coding regions between 2 coding regions of the same gene. This model is named approximate because each section of the gene approximately models the distribution of conserved subsegments.

2. An "exact" model where each gene segment(coding, intron, and flanking) is broken further down into conserved and non-conserved subsegments. All mutations that occur in the conserved segments have a negative effect on fitness and all mutations that occur in the non-conserved segments will have no effect on fitness. This model is named exact because conserved and non-conserved segments are modeled exactly how they appear in the genome. Figure 3-5 shows a diagram of this.

Unbroken Gene Segment with both negative fitness impacting and neutral mutations throughout

Figure 3-4: Gene segment in the approximate model. Neutral and fitness impacting mutations occur throughout the segment. The probability that a mutation is fitness impacting is the proportion of subsegments that are conserved.

Broken Gene segment with regions of negative fitness impacting and regions of neutral mutations

Figure 3-5: Gene segment in the exact model. In this model, each gene model is broken down into further conserved and non-conserved subsegments. Neutral mutations only occur in the non-conserved subsegments while fitness impacting mutations only occur in the conserved subsegments.

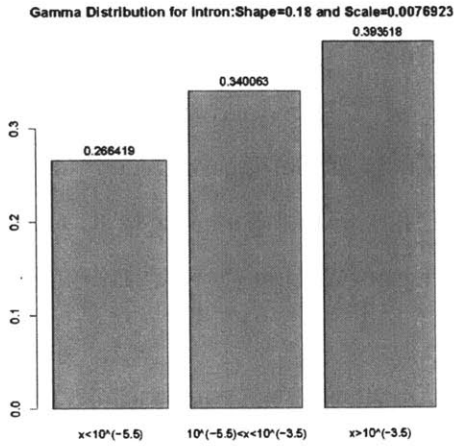
There are positive and negative aspects for both proposals. For the approximate model, the runtime for one gene is relatively fast, on average 4-6 hours. The downside for the approximate model is that the simulated gene may not accurately reflect the conserved

and non-conserved regions of the gene. For the exact model, the runtime for one gene is significantly slower, up to 50 hours. However, this model accurately reflect the conserved and non-conserved regions of the gene.

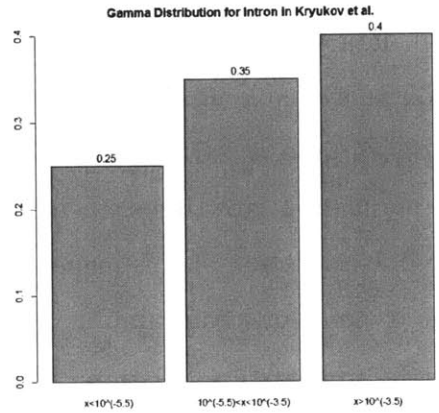
The two models dealt with the missing conservation scores that occurred in the middle of the gene differently. For the approximate model, the conservation of the first 50 bp with available conservation scores would be calculated followed by the next 50 bp with available scores until all segments were considered. The percentage of segments that were conserved was the percentage of mutations that would have an impact on fitness. This was based on the assumption that the missing scores would be consistent with the available scores. For the exact model, scores of 0.5 were added wherever scores were not available. Adding a score of 0.5 does not impact whether a region is conserved or not. In addition, if an entire 50 bp segment was missing, the segment was declared non-conserved because the majority of the genome is not conserved.

Ideally, both the approximate and the exact models would accurately simulate the real gene. In that situation, the approximate model would be chosen as the one to use for the larger sample size because of its faster runtime, but comparisons with empirical data must first be made before one model can be considered as the better one.

Before the model could be completed, the distribution of selection coefficients in the intronic and flanking regions needed to be established. The distribution of selection coefficients that was ultimately used was a gamma distribution that had been fitted to the distributions in Kryukov et al[25]. Different gamma distribution were tried until one with the same distribution as the distributions in Kryukov et al. was found for the intron and flanking regions in terms of the percent of the distribution that was less than $10^{-5.5}$, between $10^{-5.5}$ and $10^{-3.5}$ and greater than $10^{-3.5}$. For the intron regions, a shape parameter of 0.18 and scale parameter of 0.0076923 was the most consistent with the distribution. A comparison between the published distribution against the distribution used in the gene model can be seen in Figure 3-6. For the flanking regions, a shape parameter of 0.316228 and scale parameter of 0.0008 was the most consistent with the distribution. A comparison between the published distribution against the distribution used in the gene model can be seen in Figure 3-7.

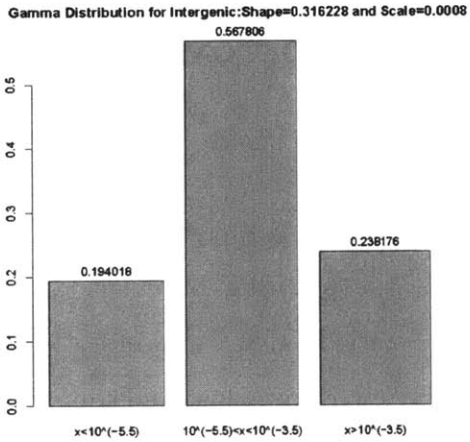


(a) The distribution of selection coefficients from a gamma distribution used for intron regions.

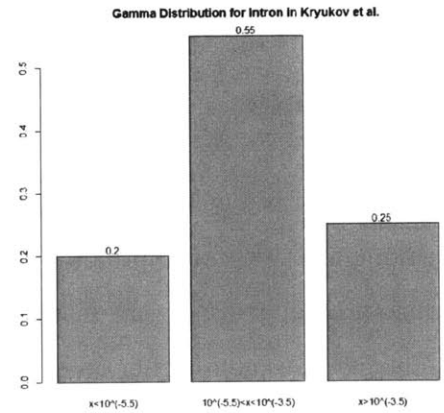


(b) The distribution of selection coefficients used for intron regions in Kryukov et al.

Figure 3-6: The distribution of selection coefficients as published in Kryukov et al. versus the distribution of selection coefficients for intron regions used in the gene model.



(a) The distribution of selection coefficients from a gamma distribution used for flanking regions.



(b) The distribution of selection coefficients used for flanking regions in Kryukov et al.

Figure 3-7: The distribution of selection coefficients as published in Kryukov et al. versus the distribution of selection coefficients for the flanking regions used in the gene model.

3.4 Comparisons with Empirical Data

Next, we asked how will it be known whether or not these two simulated models are consistent with empirical data? What should be compared between the simulated and empirical

data? We will compare the distribution of the frequencies.

Empirical data was obtained from the 1000 Genomes project[28]. Version 3 was used, which was made available April 30th, 2012. Out of 1092 individuals in the project, 379 were of European descent. Only those 379 individuals of European descent were included because the population growth parameters used in forsim had been tuned to the European population. One limitation of this model is that the population growth parameters are not generalizable to non-European samples.

To make the simulated population comparable, a subset of 379 individuals needed to be drawn from the simulated population. Out of the 227,650 individuals from the simulated population, a sub-population of 25,000 unrelated individuals was drawn. Fifty subsets of 379 were then drawn from this sub-population and averaged. By taking the mean value of 50 samples of the data we obtain more stable estimates of the desired statistics.

3.5 Analysis of Small Sample with Approximate and Exact Models

A small sample of ten genes based on ten random genes in the genome were first simulated to determine whether the approximate and exact models were consistent with empirical data. In addition, comparisons were made with the static model. Two kind of comparisons were made. The first was comparing the counts of singletons, rare, low, and common variants in the genes. The second was comparing the site frequency spectrum. The site frequency spectrum shows the distribution of variants based on their frequencies. Figure 3-8 shows the number of singletons, rare, low, and common frequency variants that there are in each region that was simulated. All of the regions in the first comparison have been normalized to number of variants per megabase.

Examining at the number of variants for the entire simulated region, the number of variants is fairly consistent across all four models, except for the number of singletons in the empirical data. Looking more closely at where this discrepancy comes from, it can be seen that both the intron and flanking regions have a low number of singletons. The 1000

Counts of Variants in Different Regions

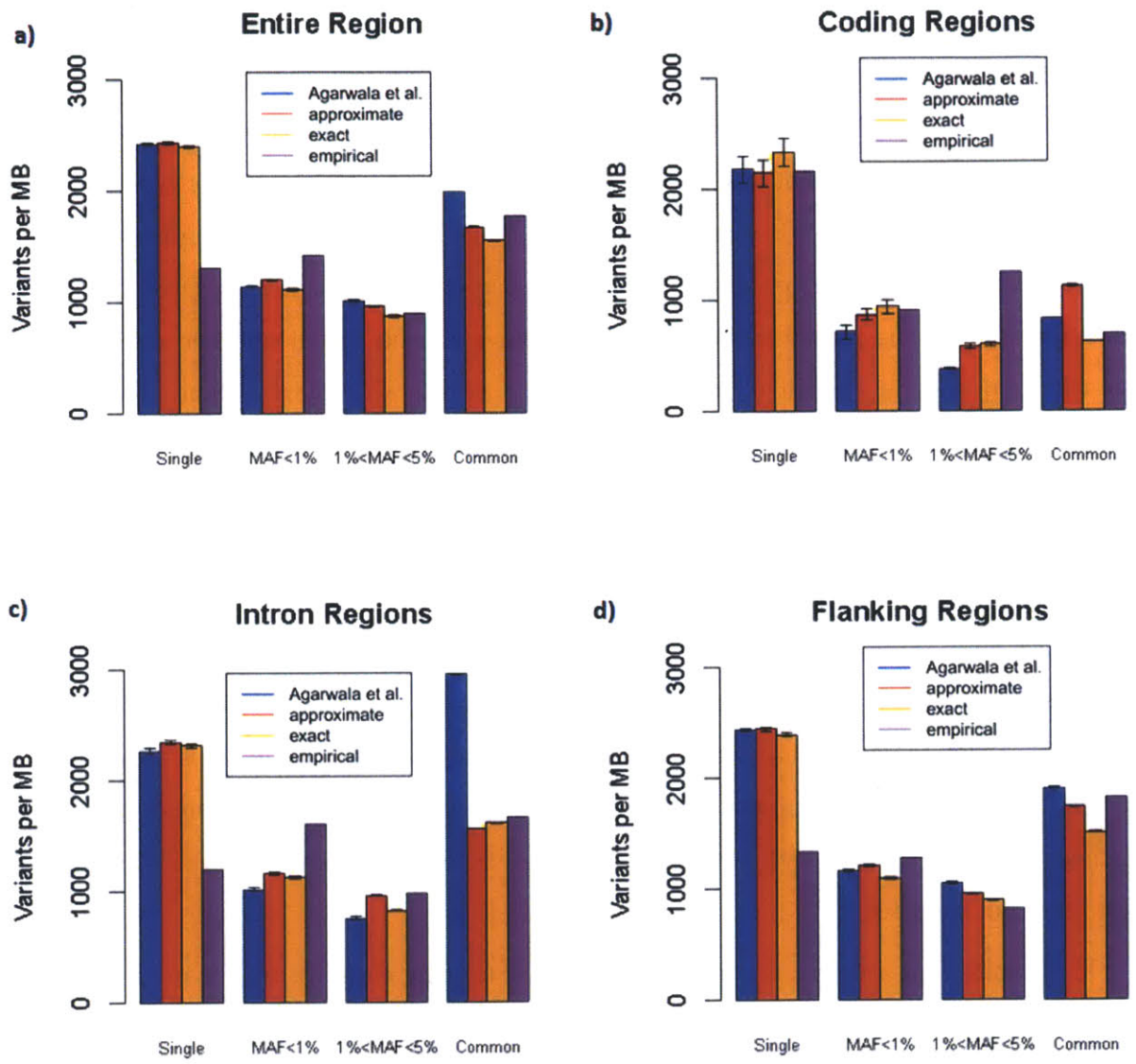


Figure 3-8: Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire simulated region, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

Genomes project was a combination of low coverage and exome whole genome sequence data. Exome whole genome sequence data is high coverage. The accuracy of the sequencing

depends of the level of coverage used by the sequencing technique. When DNA is sequenced, the common approach is to cut the segment into shorter DNA fragments and the cloned into a DNA vector and amplified in a bacterial host such as *Escherichia coli.* The short fragments are then purified from individual bacterial colonies, individually sequenced and assembled electronically into one long, continuous sequence. The higher the coverage, the more copies of each DNA fragment is present, the more accurate the DNA sequencing is. The type of variant that is most affected by this is rare frequency variants. During the sequencing process, there is always a chance for an error especially when the bacteria is amplifying the DNA fragments. Therefore, for rare frequency variants, it is sometimes difficult to tell whether a variant that has rare frequency is a result of sequencing error or if it is actually a variant. There was a study in Flannick et al[29]. on the effect of variants with a frequency on less than 1%. Even with the best low coverage techniques, the sensitivity is around 70%, while the majority of the techniques are below 50%. Sensitivity is the measurement the percent of real mutations were labeled as mutations. However, almost all the different techniques have a specificity of at least 99%, indicating that if in doubt, the variant is not reported. Specificity is the measurement of out of all the mutations that are labeled mutations, how many are actually mutations. This indicates that whenever there is a questionable variant, it is treated as an error. The original 1000 Genomes project had 25% power when detecting singletons in non-coding regions[28], a power level that indicated there were a significant number of singletons that were not detected.

The results that are seen are in line with the sequencing coverage that is used. Both the intron and flanking regions, that were a result of low coverage sequencing, have significantly fewer singletons in the empirical data, up to 50% less. There is one other significant pattern to be noted, in the old model, there are significantly more common variants than either the empirical or the two simulated models. This was an encouraging sign, indicating that the new model is modeling the gene more accurately. Because a random set of 10 genes is not very representative of the entire genome, analysis of all other differences between the four models would be made when comparisons of a larger subset of the genome was done if the differences still existed.

Comparisons were also made with conserved and non-conserved sub segments in each

type of segment(coding, intron, and flanking). In the simulated genes, conserved variants were those that had a negative impact on fitness. In the empirical gene, conserved variants were those variants that existed in the 50 bp sub segments that had an average conservation score of above 0.5. No comparisons were done with the conserved intron and conserved flanking regions for the static model because fitness affecting mutations in those regions were not modeled. The results can be seen in Figure 3-9.

In the coding variants, the approximate model looked closer to empirical data than the exact model. These results should not be weighted too heavily when considering how well each model performed because the total coding region that was modeled between all ten genes was 11.5 kb, approximately 1/5 of one flanking region of one gene. This is especially true in the non-conserved non-coding regions. There are an extremely high number of singleton for the exact model, however, the total region is only 2.7kb. In addition, the absolute counts are 14 versus 7 when comparing the exact versus the approximate models. These numbers are simply too small to have any significance. Another observation is the missing empirical singletons seen in the flanking and intron regions are mostly in non-conserved regions. This makes sense because when the genes were modeled, they did not take into consideration overlapping genes or those genes that existed in the flanking regions. Therefore, many of these conserved regions may be coding regions for other genes and being sequenced with high pass sequencing techniques.

The second type of comparison done was with the site frequency spectrum. The site frequency spectrum for the ten genes was calculated by adding the site frequency spectrum for each individual gene. There was no comparison with the static model because there was no trivial solution on how to normalize for the different gene lengths. The results for the entire simulated, coding, intron, and flanking regions can be seen in Figure 3-10, and for conserved and non-conserved segments can be seen in Figure 3-11.

The small number of variants in the coding regions indicates how small the coding region is compared to the other regions. In addition, the number of variants in the conserved intron and flanking regions is also fairly low compared to the non-conserved intron and flanking regions. These four regions have simulated and empirical data that follow the same general trend. In the non-conserved intron and flanking regions, there is significantly more variation

Counts of Variants in Conserved and Non-Conserved Regions

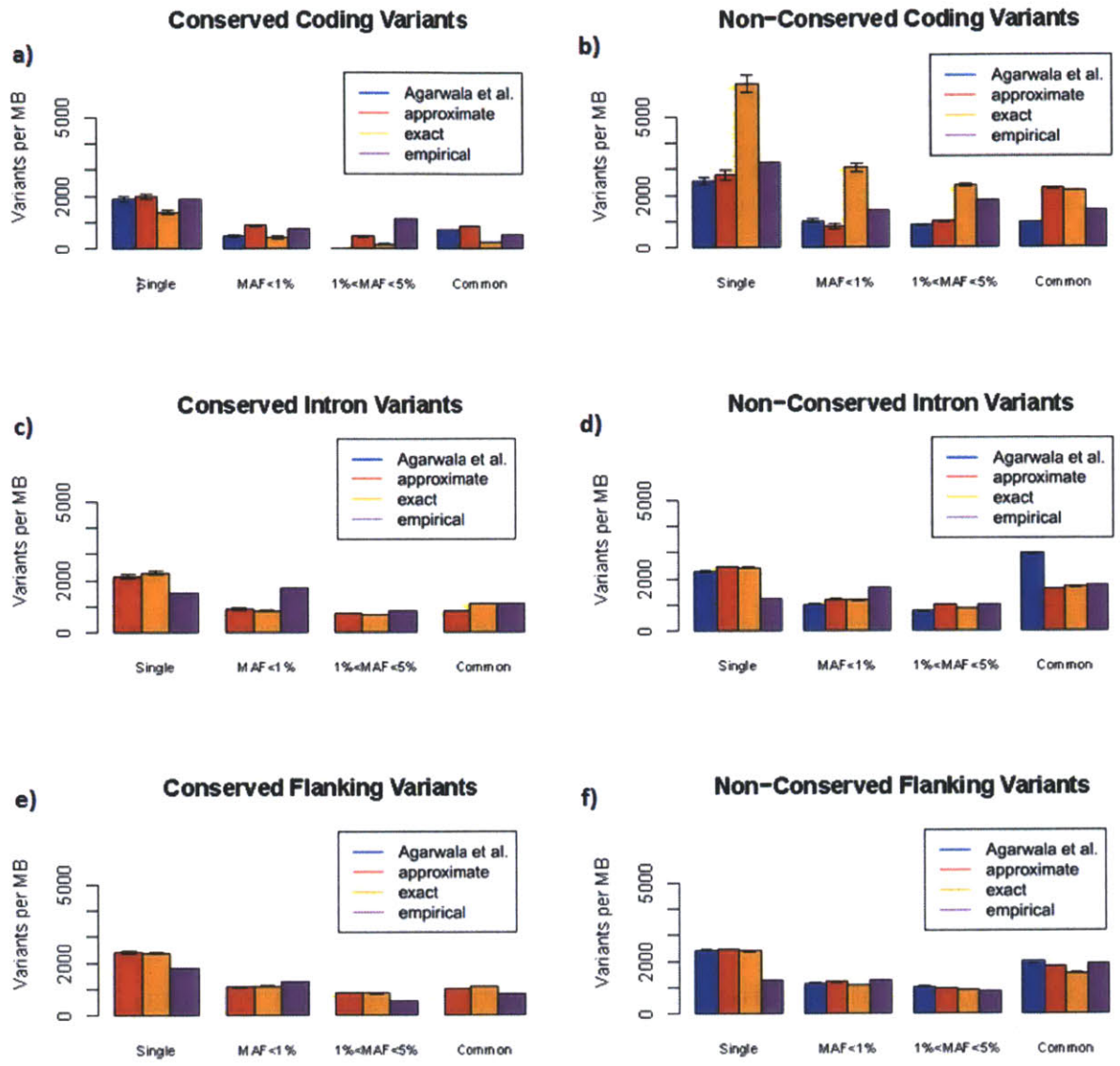


Figure 3-9: Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These counts came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

in the number of variants as the minor allele count increases in the empirical data. This is due to the simulated data being an average of fifty subsets of the population, thus decreasing

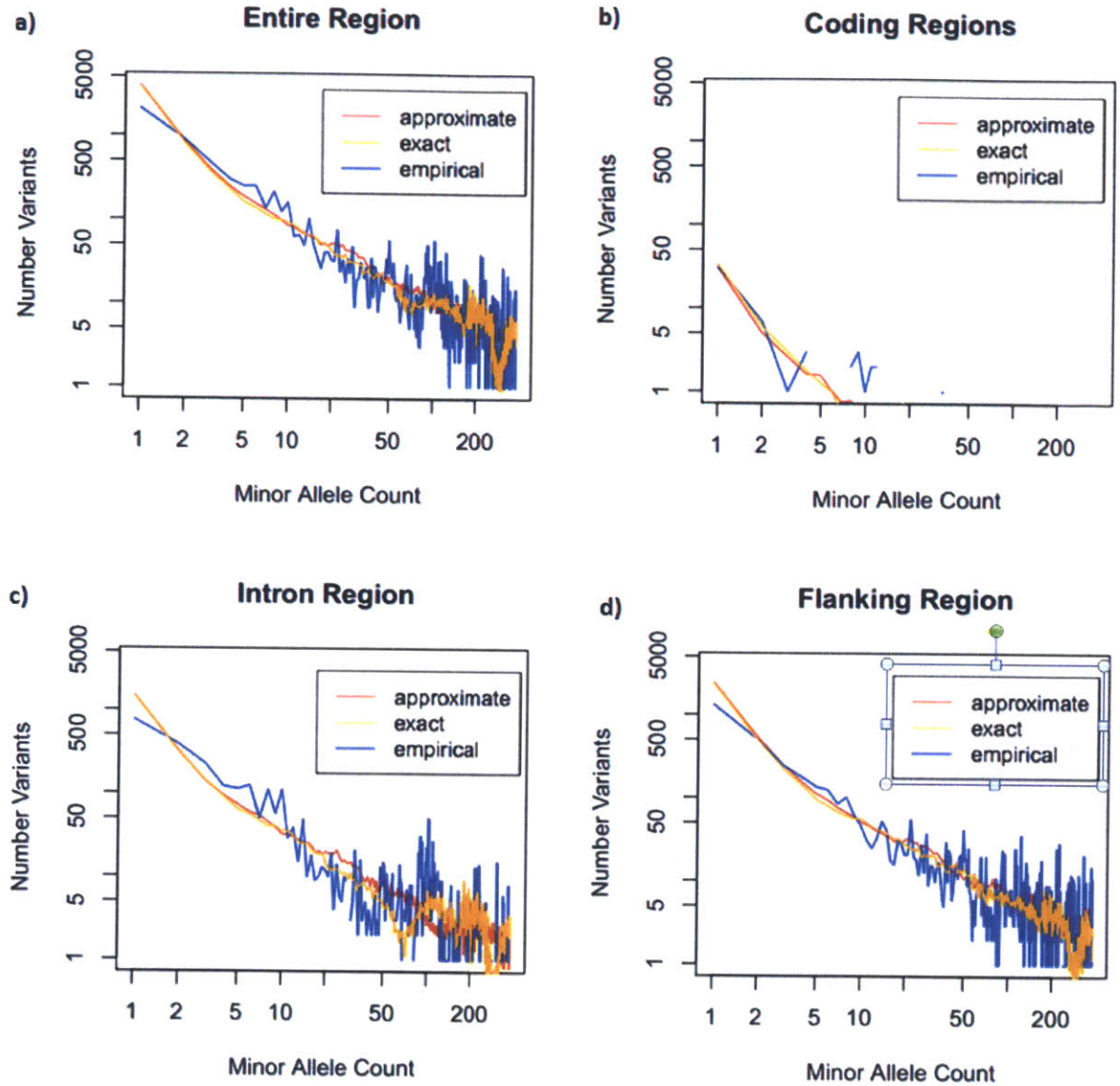


Figure 3-10: The site frequency spectrum for in the entire gene. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

the variation. The exact model tends to have a little more variation than the approximate model.

In conclusion, these comparisons show that all three models follow the same general trend that the number of variants decreases and the minor allele count increases. As the minor allele count increases, the variation in the number of variants increases in the empirical data

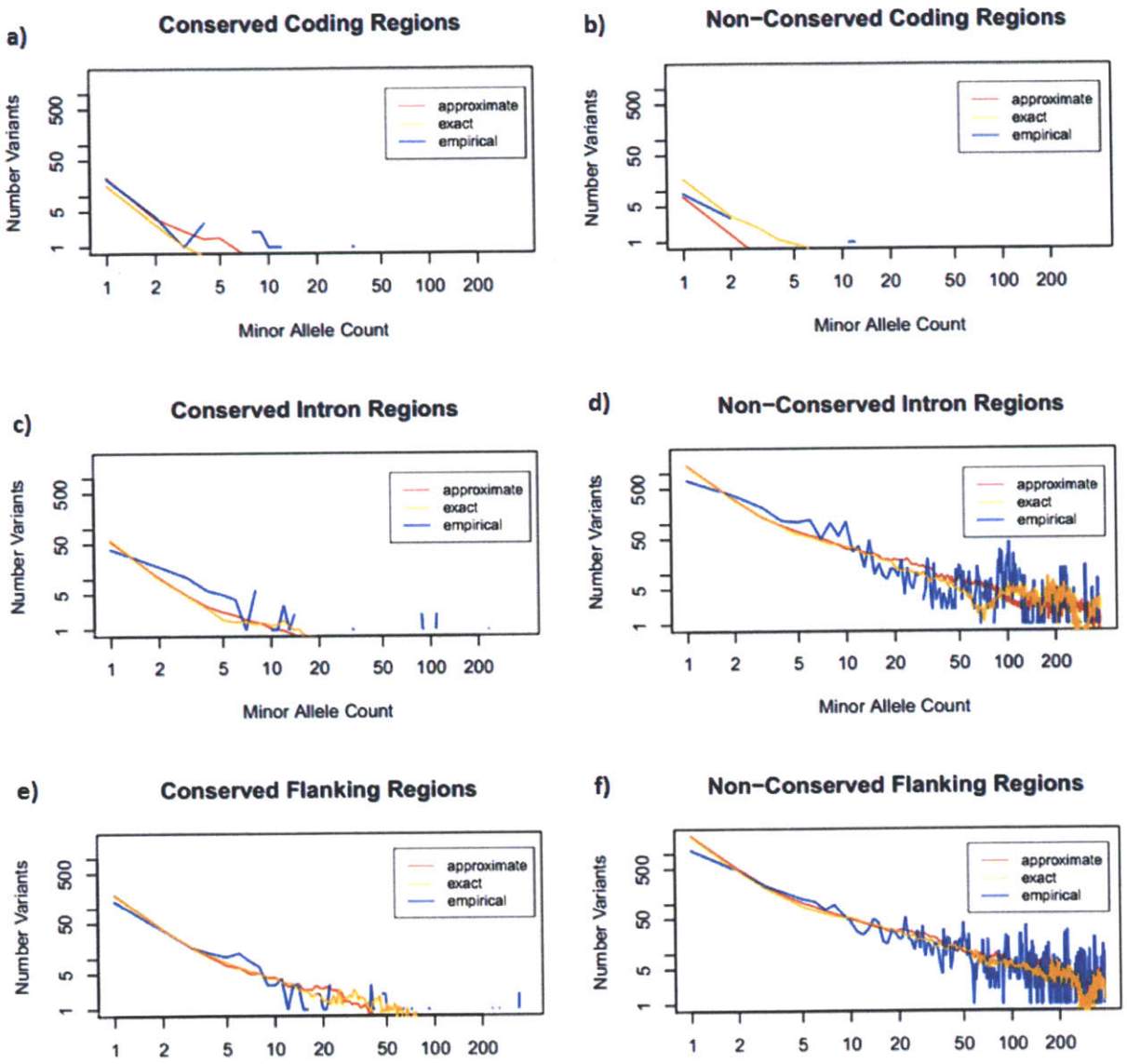


Figure 3-11: The site frequency spectrum for **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These variants came from simulating genes 10 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

compared to the simulated data. In addition, the exact model also has increased variation as the minor allele count increases. However, this increase in variation is insignificant compared to the increase in variation between the empirical data and the two simulated models.

The approximate and exact model are fairly consistent with each other and with empirical data. There were some instances where the approximate model was more consistent with empirical data like in the counts of non-conserved coding variants. There were some instances where the exact model was more consistent with empirical data such as having more variation as minor allele count increased in the site frequency spectrum for non-conserved introns and flanking regions. Because this is a small sample size, one more comparison was done to ensure that both models were consistent. A different random set of 10 genes were simulated and the total number of mutations as well as the number with no effect on fitness, a negative effective on fitness, and the average selection coefficient for coding, intron, and flanking regions for three of those genes were compared in **Tables 3.1-3.3**.

Comparing Counts of Total, Neutral, and Fitness Decreasing Mutations between the Approximate and Exact Models in Coding Regions

Coding Regions	CYP4B1 Approximate	CYPB1 Exact	DFFA Approximate	DFFA Exact	EPS15 Approximate	EPS15 Exact
Total Mutations	12	12	6	6	25	25
Neutral Mutations	10	11	3	2	4	1
Fitness Decreasing Mutations	2	1	3	4	21	24
Average Selection Coefficient	-1.17E-5	-2.84E-4	-1.78E-3	-3.51E-4	-2.68E-3	-2.39E-3

Table 3.1: Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in coding regions.

All three regions had fairly similar values across the board, leading us to choose the approximate model for the analysis on a larger subset of genes in the genome.

3.6 Analysis on Large Sample with Approximate Model

An analysis was performed on a subset of 500 random genes in the genome. Even though 500 genes is only approximately 2% of the entire genome, it can accuretely represent

Comparing Counts of Total, Neutral, and Fitness Decreasing Mutations between the Approximate and Exact Models in Intron Regions

Intron Regions	CYP4B1 Approximate	CYPB1 Exact	DFFA Approximate	DFFA Exact	EPS15 Approximate	EPS15 Exact
Total Mutations	938	975	496	543	8352	8524
Neutral Mutations	896	917	491	538	7929	8143
Fitness Decreasing Mutations	42	58	5	5	423	381
Average Selection Coefficient	-1.175E-3	-1.5E-3	-5.85E-4	-7.57E-4	-1.55E-3	-1.18E-3

Table 3.2: Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in intron regions.

Comparing Counts of Total, Neutral, and Fitness Decreasing Mutations between the Approximate and Exact Models in Flanking Regions

Flanking Regions	CYP4B1 Approximate	CYPB1 Exact	DFFA Approximate	DFFA Exact	EPS15 Approximate	EPS15 Exact
Total Mutations	5155	5069	5321	5230	5296	5312
Neutral Mutations	5034	4933	4925	4827	5013	5036
Fitness Decreasing Mutations	121	136	396	403	283	276
Average Selection Coefficient	-3.22E-4	-2.86E-4	-2.46E-4	-2.79E-4	-2.26E-4	-2.49E-4

Table 3.3: Comparisons of total mutations, neutral mutations, fitness decreasing mutations, and average selection coefficient for fitness decreasing mutations in intron regions.

the distribution of number and length of coding and intron regions of genes in the genome. In addition, this larger sample will be used to confirm the results of the smaller sample set.

The same two types of comparisons will be done on this larger sample as was done on the smaller sample. In this analysis, only the approximate model was simulated because the runtime for 500 exact model genes would have taken at least a few weeks. Figure 3-12

shows the results of the different number of counts for the different segments of the simulated regions.

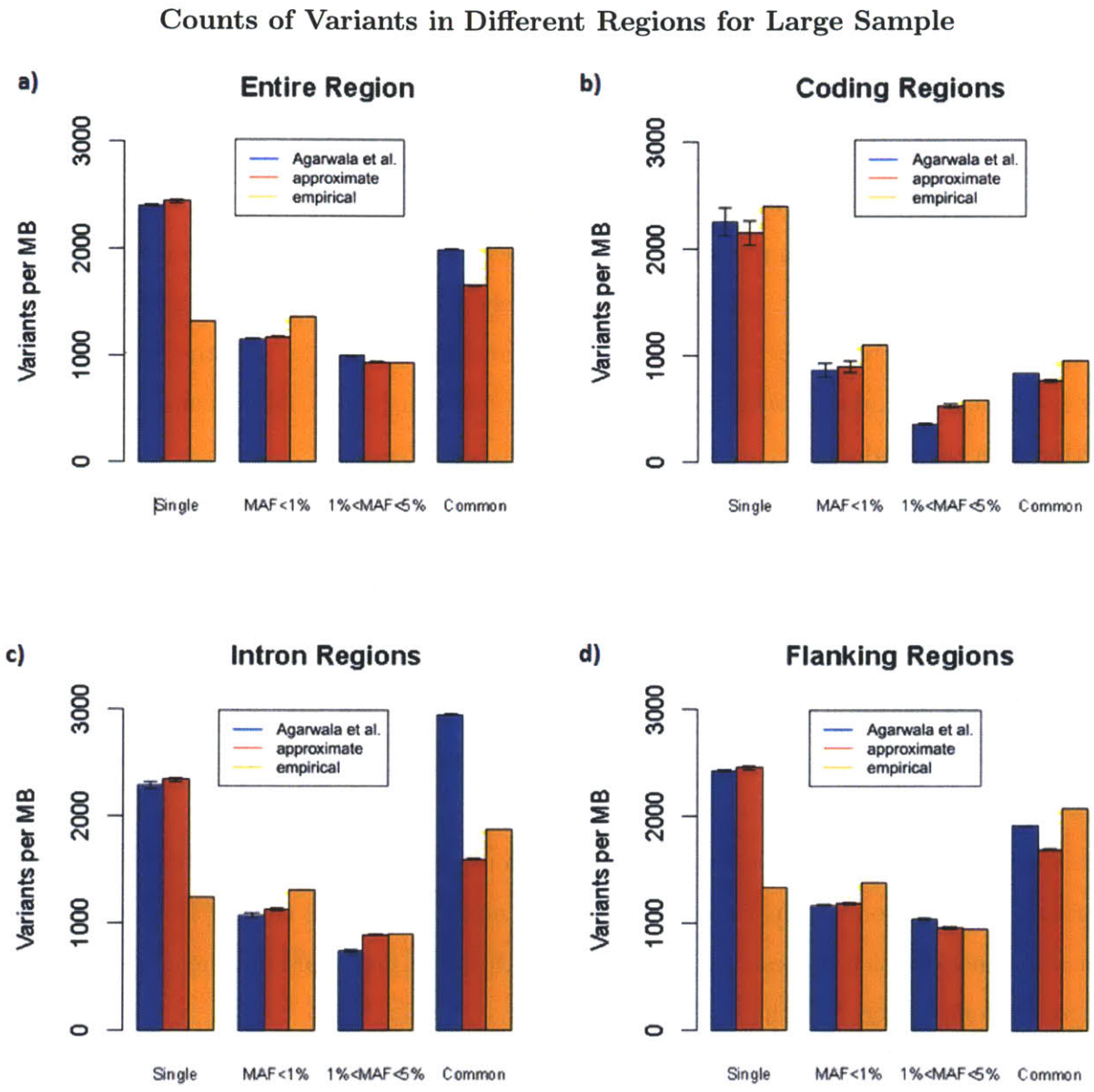


Figure 3-12: Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire gene, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

The two main trends that were seen in the smaller sample were also seen in this larger

sample. A large decrease of singleton's in the intron and flanking regions due to the low coverage in those regions as well as a sharp increase in the number of common variants in the introns. For the rest of the comparisons, the three models are fairly similar.

The results of the counts for different frequencies in conserved and non-conserved regions are shown in Figure 3-13.

In the conserved coding and non-coding regions, it was encouraging to see that these models were more consistent in this larger sample compared to the smaller sample. For example, in the smaller sample, there were significantly more variants between 1% and 5% while in this larger sample, there is approximately the same number of variants in that frequency range between the approximate model and empirical data. Another encouraging sign was the approximate non-conserved coding regions being able to model common variants better than the static model.

The comparisons of conserved intron and flanking regions in this larger sample set were fairly consistent with the comparisons made in the smaller sample set. The main difference is between the empirical data and approximate model is significantly more singletons for the approximate model, which makes sense because the 1000 Genomes project was not able to recover all the singletons in non-coding regions.

For the non-conserved regions for the intron and flanking regions, the result was what was expected. The number of empirical singletons is significantly smaller because of the low coverage sequencing techniques used in these regions. In addition, the number of common variants in the intron regions for the approximate model is closer to the empirical data then the static model which suggests that the approximate model may represent the non-conserved intron regions better than the static model.

Next, the site frequency spectrum were compared for the entire simulated region, coding, intron, and flanking regions in Figure 3-14. The site frequency spectrum is more consistent between the approximate and the empirical data in this larger sample than in the smaller sample. The main improvement comes from the reduced noise in the empirical data set. With 50 times as many genes, the number of variants at each minor allele count is much more stable. The site frequency spectrum for conserved and non-conserved segments is shown in Figure 3-15.

Counts of Variants in Conserved and Non-Conserved Regions for Large Sample

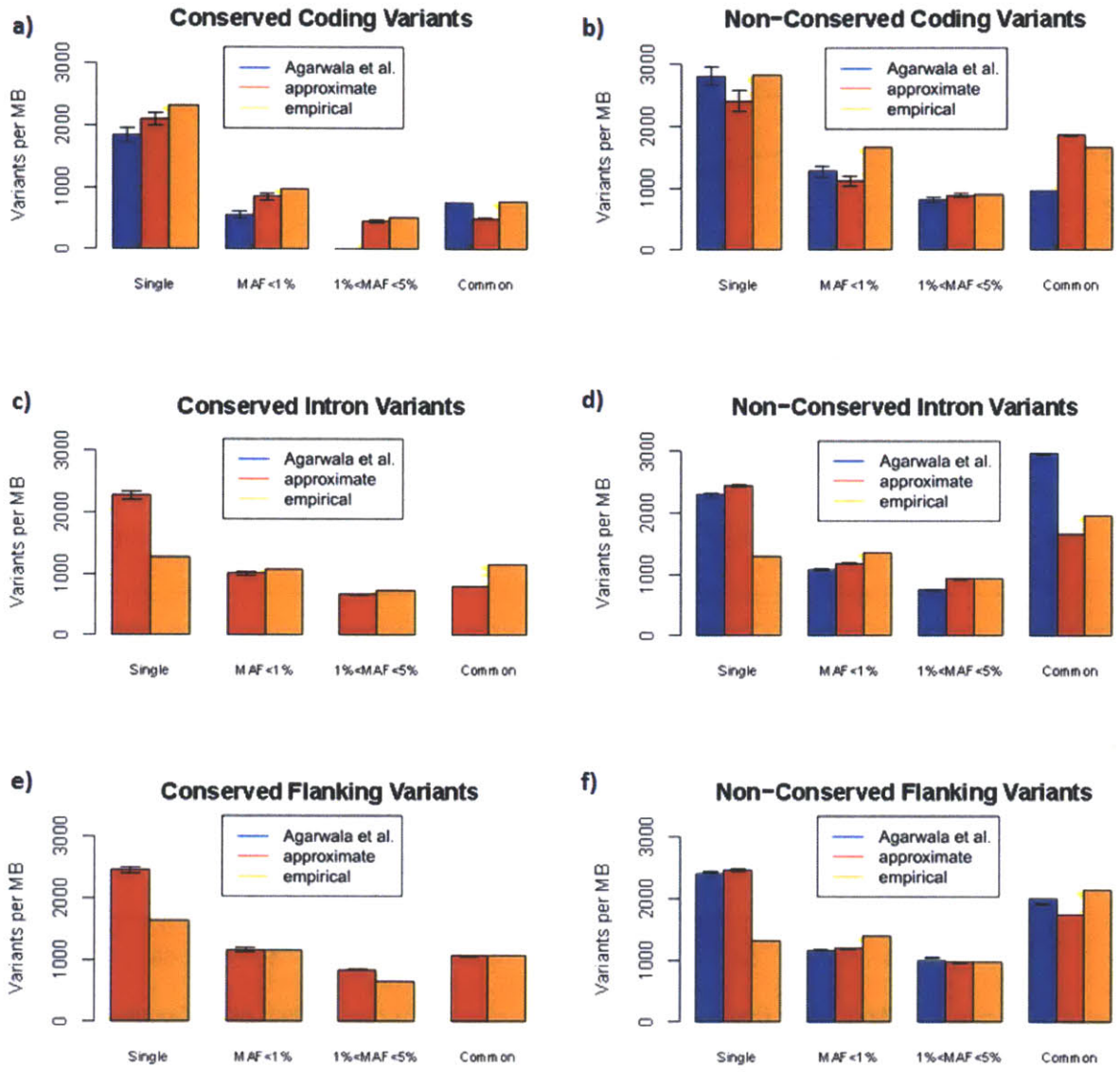


Figure 3-13: Number of singleton, rare(MAF<1%), intermediate frequency(1%<MAF<5%), and common(MAF>5%) in **a)** the entire simulated region, **b)** exon regions, **c)** intron regions, **d)** flanking regions. These counts came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

For all comparisons, the general trend between the approximate simulated model and empirical data was consistent. The one comparison where the two diverged the most was the non-conserved coding regions, which was because the regions of non-conserved coding

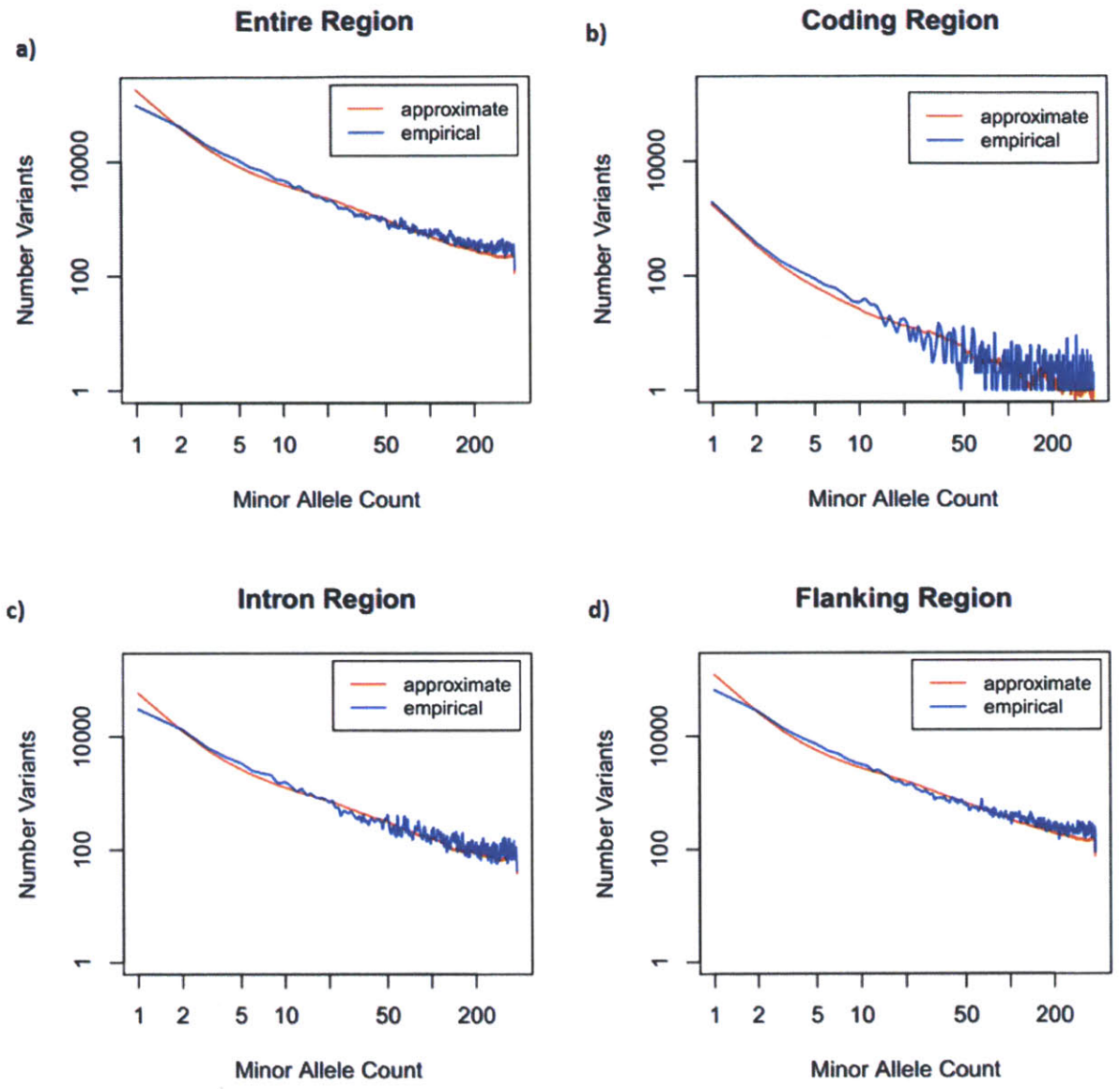


Figure 3-14: The site frequency spectrum for the entire simulated region with a sample size of 379 individuals. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

regions was the smallest. For the conserved and non-conserved coding regions, conserved intron and flanking regions, the same observation of an increase in variation as the minor allele count increases seen in the small sample is also seen in the large sample. An interesting observation is that this noise is not seen in the non-conserved intron and flanking regions. In those two regions, the two models are very consistent. The reason for the lack of noise is

Allele Frequencies in Conserved and Non-Conserved Regions

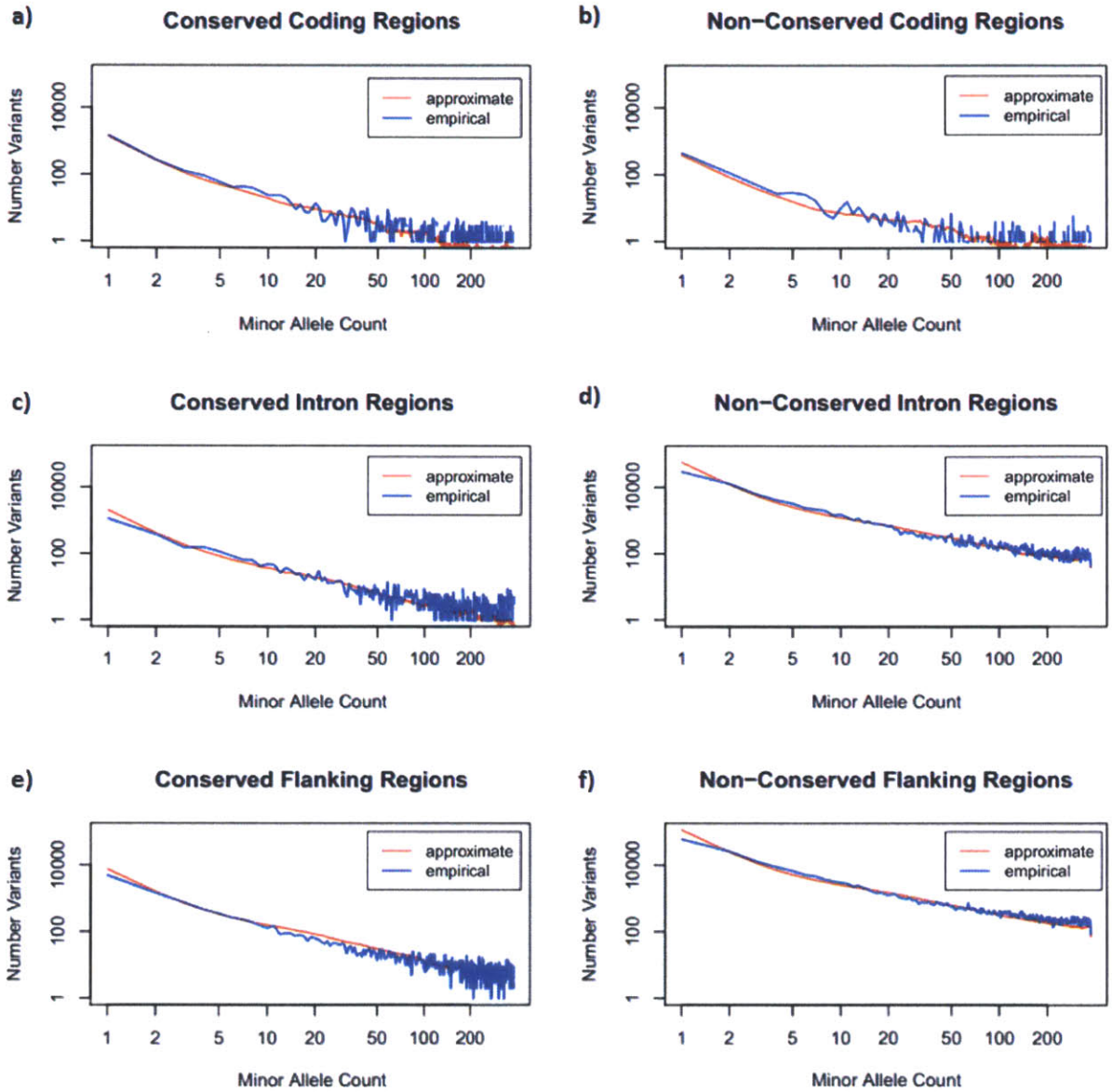


Figure 3-15: The site frequency spectrum for **a)** conserved coding regions, **b)** non-conserved coding regions, **c)** conserved intron regions, **d)** non-conserved intron regions, **e** conserved flanking regions, and **f** non-conserved flanking regions. These variants came from simulating genes 500 random genes, adding them up and normalizing by length. A sample of 379 individuals were used. The empirical data was from the 1000 Genomes project. The simulated data was an average of 50 subsets of 379 individuals.

that in these two regions cover a distance in the genome. There are 47 Mb of non-conserved flanking and 23 Mb of non-conserved intron compared to 3 Mb of conserved flanking and

903 kb of conserved intron. By covering a significantly larger larger, the variation decreased because the sample distance was larger.

3.7 Conclusion

The goal of this project was to extend the static gene model in two ways.

1. The distribution of number and length of coding and intron regions reflect this distribution in the genome.

2. Include mutations that affect fitness in non-coding regions.

The first modification was successfully implemented by creating a bank of 500 genes whose distribution of gene length reflected that of the genome. The second modification was accomplished by using conservation scores to measure selection pressure of mutations in the gene.

Comparisons were done between the simulated and empirical data in both a small subset of ten random genes as well as a larger subset of five hundred random genes. The counts of singletons, rare frequency, low frequency, and common variants were compared along with the site frequency spectrum. In conclusion, the simulated data based off the new gene model produced results that were fairly consistent with the empirical data except for the number of singletons in non-coding regions. This was caused by the inability of the low-pass empirical data to catch all of the singletons in non-coding regions.

Chapter 4

Model Limitations and Future Steps

4.1 Limitations

Even though this project was able to address two of limitations in the gene model used in Agarwala et al., there are still many limitations, both ones that existed before and new ones that were introduced.

One of the limitations that is still present from the previous model is the absence of recombination hotspots. The recombination rate across the genome is not constant, there are certain areas with elevated levels of recombination. If these hotspots occur in the middle of a gene, it will cause pairs of variants that are normally highly associated due to their close proximity, to no longer be as strongly correlated with each other.

Another limitation of the model is that each gene is simulated independently and the effects of every variant are then added together. In reality, genes do not act independently. A mutation in one gene may have significant implications on many other genes. These gene to gene interaction is called epistasis. By not modeling epistasis, an important function of biological pathways are ignored.

One of the main limitations with the new model is the difficulty to label synonymous and non-synonymous variants in the coding regions. The source of this problem is the 50 bp segments that were either labeled conserved or non-conserved. The majority of the exons are less thatn 200 bp or 4 50 bp segments. Because the majority of the exon is conserved, all 4 50 bp segments can be conserved. Therefore, this entire region would be conserved and

only non-synonymous mutations would occur because only fitness affecting mutations would occur in this region. However, in the real world, this is not the case. Just because a region contain 50 bp segments that on average have more than a 50% chance of being conserved does not imply that every mutation that occurs in that region is non-synonymous.

The solution of this limitation is to increase the resolution when counting conservation. One model that could easily be implemented is to completely disregard the 50 bp segments and for each coding/intron/flanking segment, the percent of mutations that have a negative impact on fitness is the percent of bases that have greater than a 50% chance of being conserved.

4.2 Future Steps and Implications

There are many more features that could be added to the model. One that was considered would have been to implement recombination hotspots. There is a new version of forsim software that does allow for recombination hotspots to be modeled.

Another extension to the gene model that could be implemented is non-coding exons, namely the 5' and 3' UTR's. Currently, only coding regions, introns, and flanking regions are modeled. The UTR's are very important in that they regulate the translation of the protein. The main difficulty with modeling the UTR's is that it have been difficult to find annotations of the locations of these UTR's.

In this project, comparisons were only done with data from the 1000 Genomes project. Additional comparisons can be done with other sets of empirical data to confirm the accuracy of the comparisons in this project.

What are some of the new questions that can be asked with this extended model? One question is how does this extension impact the bounds of possible disease architectures of common diseases. Are there disease architecture that were possible architectures under the previous gene model that no longer are possible and vice versa?

Now that the gene model includes mutations that affect fitness in both the coding and non-coding regions, questions can be asked about the role of variants in both coding and non-coding regions. What possible genetic architectures are there for diseases that have

causal genes whose fitness affecting mutations are located in mainly non-coding regions? For diseases that have causal genes whose fitness affecting mutations are located in mainly coding regions? a mixture of the two?

There is currently significantly less understanding of the role of variants in non-coding regions compared to the role of variants in coding regions. By extending the gene model to model fitness affecting mutations in the non-coding regions, we will be able to increase our understanding of the role of these variants by observing how they affect possible genetic architectures of diseases.

Appendix A

References

1. Foley, Mackenzie, Genetics: Past, Present, and Future. *Dartmouth Undergraduate Journal of Science* Spring 2013

2. Darwin,C. On the Origin of Species 1859

3. O'Neil D. Mendel's Genetics May 2013 http://anthro.palomar.edu/mendel/mendel_1.htm

4. Genetic Alliance; District of Columbia Department of Health. Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals. Washington (DC): Genetic Alliance; 2010 Feb 17. Appendix B, Classic Mendelian Genetics (Patterns of Inheritance)

5. Myers S. et al. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* **310**, 321 2005

6. Altshuler D. et al. Genetic mapping in human disease *Science* **322**,22 881-8(2008)

7. Kumar V, Abbas A. et al. Mendelian Disorders: Diseases Caused by Single-Gene Defects.*Robbins Basic Pathology* 9th edition

8. Fisher RA The Correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc* (Edinburgh) **52** :399-433,1918

9. Falconer, D.D. The inheritance of liability to certain diseases, estimated from the incidence among relation. *Annals of Human Genetics* **29** 51-76 (1966)

10. Tests and Diagnosis, Type 2 Diabetes *Mayo Clinic* http://www.mayoclinic.org/ May 2014

11. Klein RJ et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration Science 308 (5720): 385âĂŞ9 April 2005

12. Hemminki, K. The 'Common Disease-Common Variant' Hypothesis and Familiial Risks *PLOS ONE* June 18,2008

13. Chakravarti, A. Population genetics-making sense out of sequence. *Nature Reviews Genetics* **21**. (1999)

14. http://www.genome.gov/11006943

15. Johnson, R. Accounting for multiple comparisons in a genome-wide association study(GWAS) *BMC Genomics* 2010, Dec 22, 2010

16. International HapMap Project May 2014 http://hapmap.ncbi.nlm.nih.gov/

17. Spencer, C et al. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip *PLOS Genetics* May 15 2009

18. A Catalog of Published Genome-Wide Association Studies, *National Human Genome Research Institute* www.genome.gov May 2014

19. Voight, B.F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**, 579-89 (2010)

20. Agarwala et al. To what extent can empirical data place bounds on the genetic architecture of complex human diseases? *Nature Genetics* 2013

21. RefSeq Database http://www.ncbi.nlm.nih.gov/refseq/

22. Ayallet S. et al. Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLOS Genetics* August 2010 Vol 6 Issue 8

23. Esteller M. Non-coding RNAs in human disease *Nature Reviews Genetics* **12**, 861-874 December 2011

24. Pruitt, K. The consensus coding sequence(CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 2009 Jul; 19(7):1316-23

25. Lin, M. et al. Locating protein-coding sequences uner selection for additional, overlapping function in 29 mammalian genomes. *Genome Research* 2011 21:1916-1928

26. UCSC Genome Browser http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/placent

27. Kryukov, G et al. Small fitness effect of mutations in highly conserved on-coding regions. *Human Molecular Genetics* 2005 Vol. 14, No. 15

28. The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing *Nature* 28 October 2010, Vol 467

29. Flannick J. et al. Efficiency and Power as a Function of Sequence Coverage, SNP Array Density, and Imputation. *PLOS Computational Biology* July 2012 Vol 8 Issue 7