Complex Population Models with Coalescent Simulations

by

Catherine Foo

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

Master of Engineering in ~~Computer Science and Engineering~~
Electrical Engineering and Computer Science
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2003

Author .....
Department of Electrical Engineering and Computer Science
May 21, 2003

Certified by .....
Mark J. Daly
Whitehead Fellow
Thesis Supervisor

Certified by ...
David C. Page
Professor
Thesis Supervisor

Accepted by ...
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Complex Population Models with Coalescent Simulations

by

Catherine Foo

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2003, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Simulated *single nucleotide polymorphism* data play an important role in human genetics. They provide a mechanism for analyzing genetic patterns on a genomic scale, examining theories of human evolution, inferring properties of recombination, and comparing efficiencies of medical genetic study designs. However, the simple population models used in most experiments pose a critical obstacle to drawing reliable inferences from these simulations.

This thesis introduces a powerful and flexible tool, CoSi, that generates haplotype data based on the *coalescent process*. CoSi includes key features for accurate simulation of empirical data: support for non-uniform recombination and the ability to model multiple populations and population events. Here we use CoSi to examine models of human recombination and demonstrate its capabilities by matching diverse empirical data.

Thesis Supervisor: Mark J. Daly
Title: Whitehead Fellow

Thesis Supervisor: David C. Page
Title: Professor

Acknowledgments

No thesis could be possible without the support of many others, and this one is no exception.

I would like to thank my lab mates – Claire Wade, Jeff Barrett, Shaun Purcell, and Andrew Kirby – with whom my time has been educational, entertaining, and at times absolutely confounding. Steve Schaffner and Nick Patterson spent hours brainstorming with me, and played a definitive role in designing this simulator. In addition, Steve taught me everything I know about coalescent simulations. David Page provided me with insights and context, and his interest in my work is greatly appreciated. Mark Daly is a one-of-a-kind advisor, whose enthusiasm and superlatives propelled me through my work. I could not have asked for better.

The genesis of this thesis began at the Fred Hutchinson Cancer Research Center, where Leonid Kruglyak and Mike Eberle gave me a crash course in computational biology in just three months. All of this is entirely due to the most fortuitous sequence of introductions. I owe a huge debt to my aunt and uncle, Luna Yu and Chen-Wei Lin, who introduced me to Leonid, who in turn introduced me to Mark.

I also need to thank the following people for making this work possible in many other ways: my parents for giving me constant support and guidance; Peip for reminding me what life was like three years ago; the one and only Megan Galbraith, for tolerating the unwashed dishes and the general disarray of my life; Dan Chak, my partner-in-crime; Aman Loomba, who gave me perspective through absurdity; Shawdee Eshghi, for suddenly appearing in my life; Deborah Ann White and Selam Daniel for their friendship and love over the past five years; Anna Levin and Elizabeth Hein whose presence in Boston made this year truly spectacular, along with all my girls from home; and Matt DeBergalis, who I can never thank enough for his unwavering belief in me.

I would like to dedicate this work to my grandmother, Wan-Chiao Chan Foo, who is an inspiration to all her grandchildren.

Contents

1 Introduction **13**

2 Background **15**
- 2.1 Single nucleotide polymorphisms (SNPs) . . . 15
- 2.2 Linkage disequilibrium and recombination . . . 16
- 2.3 Limitations of empirical data . . . 16
- 2.4 A brief review of population genetics . . . 18
  - 2.4.1 Effective population size . . . 18
  - 2.4.2 Out of Africa theory . . . 19

3 State of the art in population simulations **21**
- 3.1 The coalescent process . . . 22
- 3.2 Null model . . . 24
- 3.3 Non-uniform recombination . . . 26
- 3.4 Multiple populations . . . 26
- 3.5 Haplotypes . . . 27

4 Problem statement **29**
- 4.1 The need for an enhanced simulator . . . 29
- 4.2 Needed features . . . 30
  - 4.2.1 Non-uniform recombination rate . . . 30
  - 4.2.2 Support for multiple populations and population events . . . 31
  - 4.2.3 Haplotype output . . . 31
  - 4.2.4 History log . . . 31
- 4.3 Engineering constraints . . . 32
- 4.4 Advantages of a flexible simulator . . . 32

5 Implementation of CoSi **33**
- 5.1 Null model, no recombination . . . 33
- 5.2 Uniform recombination . . . 36
- 5.3 Non-uniform recombination . . . 37
- 5.4 Population events: single populations . . . 37
  - 5.4.1 Population size changes . . . 37
  - 5.4.2 Bottlenecks . . . 37

5.5 Population events: multiple populations 38

6 Using CoSi **41**
6.1 Developing a recombination model 41
6.2 Comprehensive model 42
6.2.1 Allele frequency 44
6.2.2 LD versus physical distance 44
6.2.3 Fraction of ancestral chromosomes 44
6.2.4 Genetic distance between subpopulations 50
6.2.5 Haplotype blocks and block coverage 50

7 Conclusions **53**
7.1 Goals met 53
7.2 Uses and extensions of CoSi 54

A Additional background information **55**
A.1 Linkage disequilibrium 55
A.2 Finding haplotype blocks 56

B CoSi inputs **57**
B.1 Parameter file 57
B.2 Recombination input file 59

List of Figures

2-1 How SNPs are ascertained. 17
2-2 Average LD versus physical distance for African-American and non-African populations 19

3-1 Genetic history of a constant population size 10 23
3-2 A coalescent tree 24
3-3 Ancient versus recent mutations 25
3-4 Empirical versus simulated data. 26

5-1 Structure of CoSi 34
5-2 Distribution of coalescent times over 10^5 independent trials. 35
5-3 Minor allele frequency distribution for 10^2 independent trials. 36
5-4 Multiple populations and migration 39

6-1 LD versus physical distance for uniform recombination, H_0 42
6-2 LD patterns for regional uniform recombination rates chosen from a distribution, H_1 43
6-3 LD patterns for hotspot recombination model H_2 43
6-4 Population model used in Section 6.2. 45
6-5 Allele frequency of simulated European population 46
6-6 Allele frequency of simulated Asian population 46
6-7 Allele frequency of simulated African-American population 47
6-8 LD for African populations with uniform recombination. 47
6-9 LD for European and Asian populations with uniform recombination. 48
6-10 LD for African populations using the H_2 model of recombination. . . 48
6-11 LD for European populations using the H_2 model of recombination. . 49
6-12 Fraction of ancestral alleles in empirical data 49
6-13 Fraction of ancestral alleles in simulated data 50

List of Tables

4.1 Current simulators and available features 30

6.1 Block statistics for H_0,H_1, and H_2 recombination models. 42
6.2 F_{ST} values for simulated and empirical data 50
6.3 Haplotype block statistics for complex population model 51

Chapter 1

Introduction

Characterizing patterns of genetic variation in the human genome is a critical first step towards understanding how genotype and biomedical phenotype are related. This information can reveal genetic factors in disease that suggest possible treatments, or can help scientists understand the genetic basis behind a patient's response to certain drugs. Knowledge of these patterns has already enabled researchers to reduce the amount of time necessary to screen a region of the genome for involvement in disease [12, 24].

Simulated data is often used to formulate or test hypotheses about these patterns of variation. For example, models have been used to estimate the number of common single nucleotide polymorphisms in the human genome [16] or to suggest recombination patterns that lead to observed characteristics [29]. Simulations can also be used to further our understanding of human evolution. For example, by comparing empirical data to simulated data modeled upon an *out of Africa* hypothesis, we can make estimates as to when major movements of the human population occurred. And because simulated data contains complete information about the alleles and history of a hypothetical chromosome, we can analyze the effects of different methods of collecting empirical data on the completeness and accuracy of results.

Current models, however, are often limited in their flexibility. This leads to discrepancies between simulated and observed data. For example, recent empirical data have suggested much longer regions of high *linkage disequilibrium*, or correlation be-

tween variable sites, than observed in population simulations [3, 6, 23]. While simple models of population evolution and recombination are adequate for formulating basic hypotheses, more complex models must be available in order to make comparisons to specific features of empirical data. Many studies compare empirical data to simulated data to draw conclusions about selection events [4], but whether the observations could be explained simply by random drift in a more complex and accurate population model is still unclear. The recent rapid increase of empirical data provides us with the opportunity to create more complex models that can be matched to our emerging detailed understanding of genetic variation patterns.

This thesis describes the design and development of CoSi, a coalescent simulator with capabilities for multiple populations and non-uniform recombination. CoSi produces simulated data which has been tested to match the output of state of the art simulators in simple scenarios, but is also capable of simulating more complex models of human demography and recombination. These complex models, unlike those used frequently in population genetics literature [21, 29, 34], generate data which resemble empirical genetic variation patterns and can thus be used to draw more reliable inferences about human history and patterns of recombination.

Chapter 2

Background

To understand the need for simulations and the limitations of currently available tools, we first explain what we are simulating and the experimental reasons why we need simulations, then discuss some population genetics concepts.

2.1 Single nucleotide polymorphisms (SNPs)

Human DNA differs from individual to individual only at points where ancestral mutations have occurred. By far the most common type of mutation is one that occurs at a single site, referred to as a *single nucleotide polymorphism* (SNP)[1]. Other types of genetic markers exist as well but are not discussed here. SNPs, being most common, are likely responsible for most observed phenotypic variation, and are therefore the focus of many studies which seek out genetic factors in disease. SNPs can also serve as markers for other types of variation. Since mutations in DNA accumulate and are passed down to an individual's offspring, mutations that originate in ancient generations appear in the population with higher frequency than mutations from more recent generations. Based on an analysis of existing data, Kruglyak and Nickerson [17] predict about ten million SNPs in the human population with minor allele frequency $> 1\%$.

[1]**SNP**: a nucleotide base where some copies of a particular chromosome carry a particular base (A,C,G,T) and other copies carry a different base. In this document, the terms *marker* and *SNP* are used interchangeably.

Empirical data have revealed that groups of nearby SNPs are non-randomly associated with each other. This association, or *linkage disequilibrium* (LD), can be used to track and pinpoint genes associated with complex diseases, and suggests that genotypes can be determined using lower-resolution data and a good knowledge of LD patterns in the genome [6]. Some studies examine pairs of SNPs which are known to be linked, while others involve examining an entire region where limited *haplotypes*, or patterns of SNP alleles, are observed. For a review of the LD measure we use in this thesis, r^2, refer to Appendix A.1.

2.2 Linkage disequilibrium and recombination

The major force disrupting LD is recombination, because it promotes the creation of new haplotypes. When a mutation first appears on a chromosome, the mutation exhibits high LD with all neighboring markers, since it only appears in a particular haplotype. Recombination breaks up this correlation.

Because sites close to each other have a smaller chance of experiencing a recombination event between them, expected linkage disequilibrium will decrease with distance. Two SNPs tend towards *linkage equilibrium* at a rate dependent on the distance between the SNPs and the local recombination rate.

2.3 Limitations of empirical data

Error and bias is introduced into SNP studies through genotyping errors, sampling errors, and the ascertainment process. Since data are collected marker by marker, a *heterozygous individual*[2] is said to be of ambiguous *phase*: it is unclear from the observation which allele was paternally derived and which was maternally derived. Various methods can be employed to infer the haplotypes, but they introduce an element of uncertainty.

[2]**Heterozygous individual**: an individual who has two different alleles for a marker (one each on the two corresponding chromosomes).

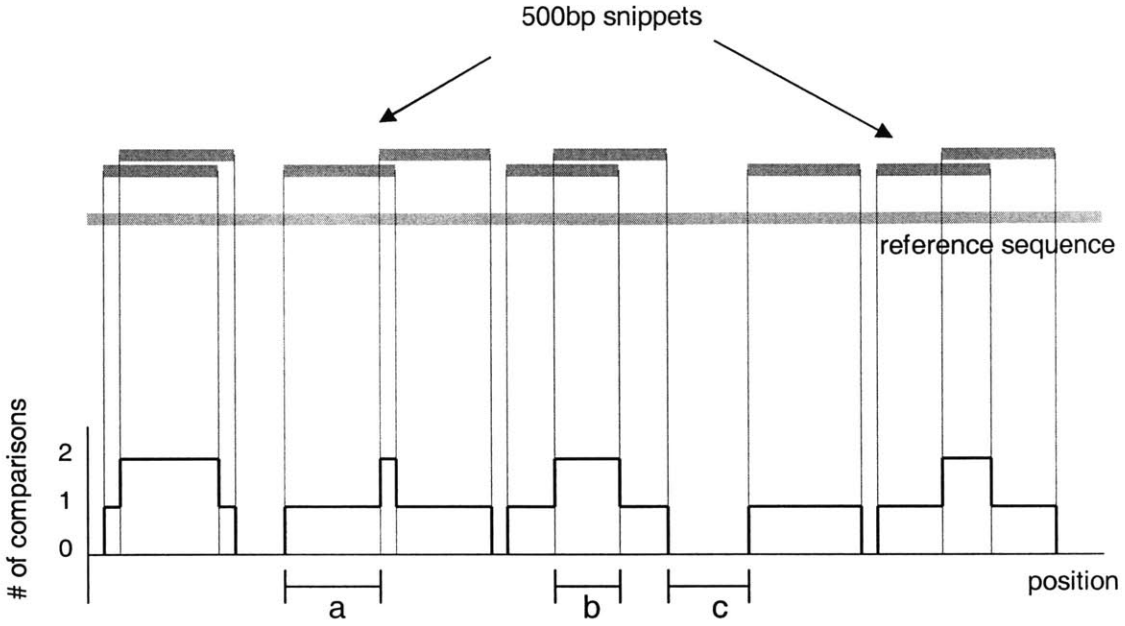

Figure 2-1: **How SNPs are ascertained.** Random shotgun snippets are lined up against a reference sequence (generally the Human Genome Project). The snippets are compared to the reference sequence and differences are noted as discovered SNPs. Some regions will be covered with more than one snippet, others will not be covered at all.

Most SNPs are discovered by sequencing a small number of chromosomes before examining these SNPs in larger study samples. This ascertainment biases the collection of SNPs towards high frequency markers. For example, The SNP Consortium (TSC), the largest SNP discovery project to date, ascertains SNPs using a shotgun method (Figure 2-1). Most SNPs are detected using only one or two sequence comparisons, and some areas of the genome are not covered at all. A public database[3] contains all TSC SNPs along with a collection of SNPs discovered in other studies. Most genetic studies now base their experiments on the SNPs available in this database. Many studies such as as Phillips et al. [21] intentionally choose evenly spaced SNPs that are separated by several kilobases.

[3]dbSNP is available at http://www.ncbi.nlm.nih.gov/SNP

2.4 A brief review of population genetics

This section covers some basic concepts of population genetics that relate to this thesis.

2.4.1 Effective population size

Under neutral theory, where changes in variation are caused by genetic drift, population size determines the rate at which allele frequencies change. Most organisms, including humans, have frequent and significant changes of population size. The relevant population size for genetic studies is the size of the population which is currently breeding, which may be anywhere from one-third to one-quarter of the entire population [5, p.57]. Also, *panmictic*[4] populations exhibit lower genetic diversity than non-panmictic populations. Instead of using the actual population size, population geneticists use an *effective population size* N_e that is chosen to approximate the diversity of the actual population. Fluctuations in population size, non-uniform mating, and gender ratio can all affect N_e. Population size changes can be incorporated by taking the harmonic mean of the population sizes,

$$N_e \approx \left(\frac{1}{t} \sum_{i=0}^{t-1} \frac{1}{N_i} \right)^{-1}$$

where t is the total time, and N_i is the population size at time i. Thus, the time spent at a smaller sizes will have a greater impact on N_e than time spent at larger population sizes.

When simulating the human population as a constant size model, N_e is conventionally considered to be 10,000 individuals[5]. However, this estimate is not based on a large empirical analysis and, as we will note later, does not explain the full range of empirical data.

[4] **Panmictic**: breeding uniformly and randomly.

[5] See Takahata et al. [28], Yang [31], and references therein for a more detailed explanation.

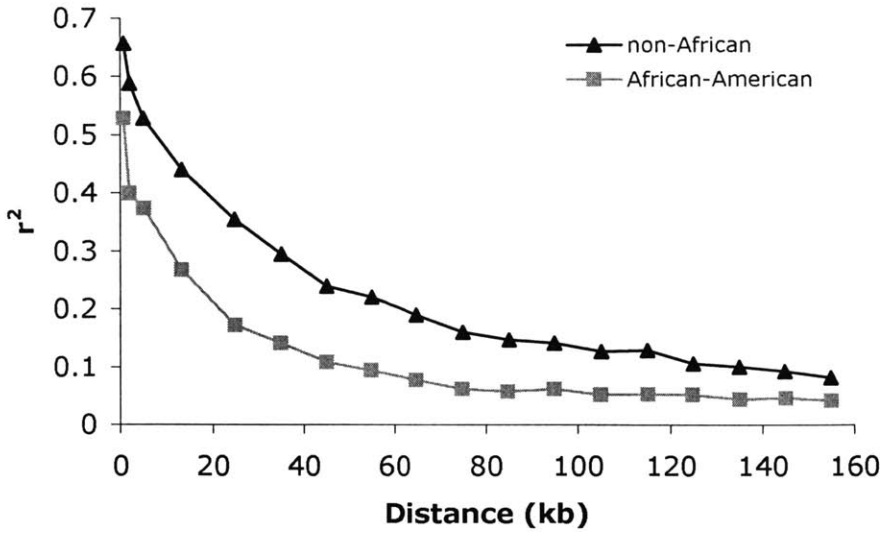

Figure 2-2: **Average LD versus physical distance for African-American and non-African populations.** Average LD is lower for African populations than non-African populations, indicating a longer genetic history for African sequences. Data from Gabriel et al. [6].

2.4.2 Out of Africa theory

Based on archaeological evidence, and now substantiated by examination of the genome, scientists believe that a small group of modern *Homo sapiens* left Africa around 100,000 years ago and eventually founded the European and Asian populations. The founding group of these populations would have carried only a subset of the diversity present on the African continent at that time. The bottlenecks experienced during these migrations caused many alleles to fixate[6] or dramatically change in frequency. Thus, since the SNPs found in European and Asian populations are on average more recent, these populations should exhibit less diversity, fewer ancient mutations, and higher LD. Indeed, as shown in Figure 2-2, the empirical LD of non-African samples is higher than that of African-American samples.

[6]**Fixate**: To cease being a variation in a population, either by reaching 100% frequency or by disappearing entirely.

Chapter 3

State of the art in population simulations

When empirical data is limited, simulations are often used to generalize or confirm results of analyses. Gabriel et al. [6] used simulations to estimate the underlying structure that produced the given empirical results. Zhang et al. [34] used simulations to evaluate the effect of genetic drift on linkage disequilibrium.

The naïve method of generating a set of chromosomal samples is to start with a founding population or individual that replicates to produce the current day chromosomes. These chromosomes are then randomly sampled for analysis. This approach is computationally intensive, as the number of chromosomes increases to an unwieldy value. Moreover, at the end of the process, all the chromosomes which are not sampled represent wasted compute time and memory. Forward simulations are still used in cases where one needs to track the evolution of the entire population over time [18], but more efficient methods are now available.

Since we are only concerned with coalescent simulations in this document, we refer the reader to other sources for descriptions of approaches such as diffusion[1].

[1] A good overview of diffusion is given in Ewens [5].

3.1 The coalescent process

The coalescent process is commonly used to efficiently generate a population history, using a Monte Carlo method [7, 8, 14]. The simulation works *backwards*, starting with the final sampled chromosomes, and constructs their ancestral histories. Coalescence assumes the neutral theory, although selection has been considered in some models [9, 10, 19, 20]. In addition, each subpopulation is assumed to be panmictic.

As an example, Figure 3-1 shows a population of 10 chromosomes over several generations. Some chromosomes will have multiple descendents, while others will have no descendents. Eventually, looking back far enough will identify a common ancestor for all 10 current chromosomes in the population. As described later in Section 6.2.3, the *ancestral state* of a human polymorphism can be discovered by examining the corresponding site in non-human primates.

Coalescent simulations are based upon this assumption that all chromosomes are ultimately derived from a single ancestral chromosome. The lineages of chromosomes in the resulting sample are traced back until a common ancestor is found. From that point on backwards, these chromosomes share the same genetic history. The genetic history in Figure 3-1 can also be represented as the tree shown in Figure 3-2, which only shows the relevant nodes rather than the entire population.

A mutation that occurs in an ancestral chromosome also appears in all descendent chromosomes. After constructing the coalescent tree, mutations are placed on certain ancestral chromosomes, indicating that an individual during that branch of genetic history underwent a mutation. Since mutations only affect chromosomes that descend from the individual in which the mutation first occurred, recent mutations tend to have lower frequencies than older mutations (Figure 3-3).

The simulation then constructs the chromosomes using the ancestral trees and mutations. The simulated data gives complete sequence information for the population sample, where each mutated position is known, as are the identities of ancestral alleles.

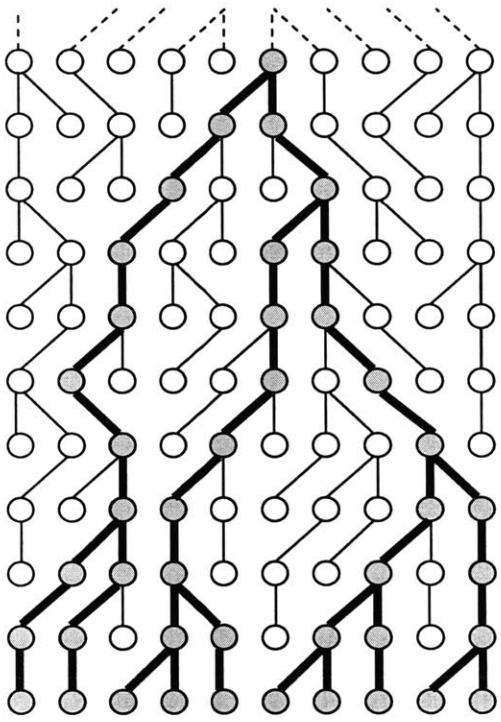

Figure 3-1: **Genetic history of a constant population size 10.** The oldest generation is at the top; the darkened nodes and lines represent the ancestors of the current population (bottom row).

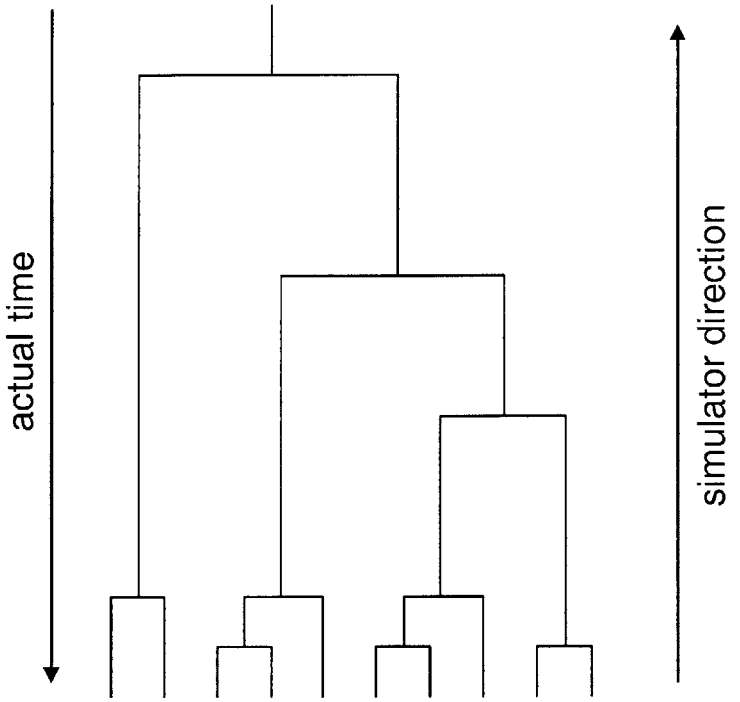

Figure 3-2: **A coalescent tree.** The same population as in Figure 3-1, but showing only the relationships among the chromosomes in the current generation.

3.2 Null model

The model most often used to generate simulated data produces data which differs from empirical data in many important respects. The typical model, hereafter referred to as the *null model*, includes the following parameters:

- constant population size $N_e = 10{,}000$
- constant mutation rate $\mu = 2 * 10^{-8}$ muts/base/gen
- uniform recombination rate $r = 1 * 10^{-8}$ recs/base/gen

As seen in Figure 3-4, LD patterns produced by this model do not match empirical data. The choice of constancy for the three parameters (N_e, μ, r) is not based on biological evidence (indeed, each is implausible), but is a computational convenience that produces ease of modeling. The null model provides a surprisingly good approximation for examining basic genetic observations, but in light of the increasingly detailed human genetic data available today, we need to go beyond these simple assumptions and build more realistic models.

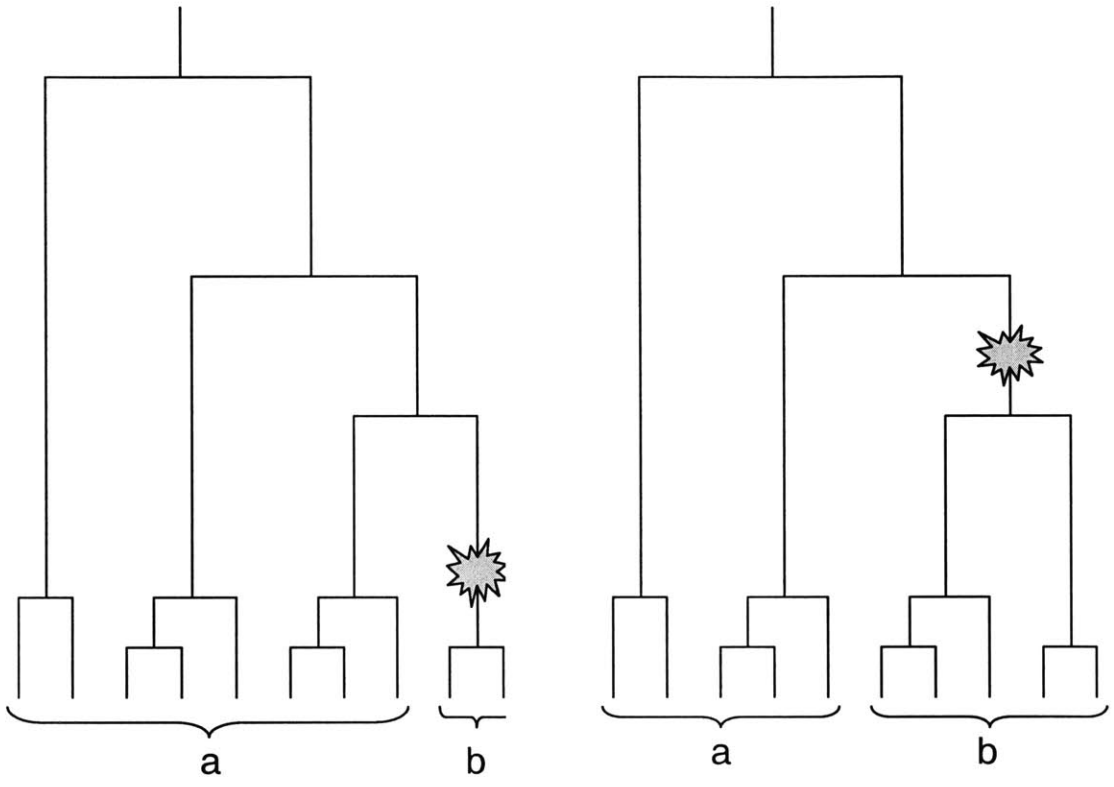

Figure 3-3: **Ancient versus recent mutations**. The figure on the left shows a recent mutation producing a low frequency marker, while the figure on the right shows an older mutation producing a higher frequency marker. The chromosomes in group a have the ancestral allele; the chromosomes in group b have the new mutation.

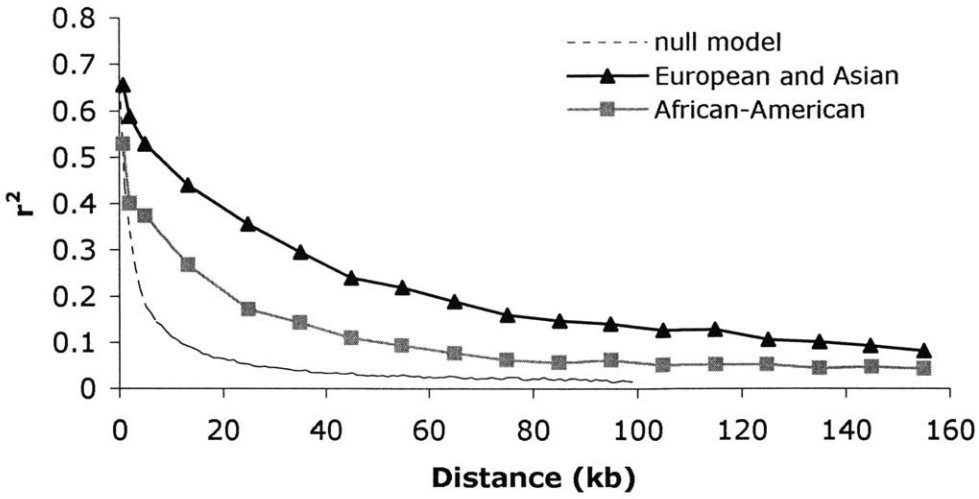

Figure 3-4: **Empirical versus simulated data.** Empirical data is from Gabriel et al. [6], simulated data is generated under the null model and ascertained with a TSC-like method described in Section 2.3.

3.3 Non-uniform recombination

Regional recombination rates (average rates over a few hundred kilobases) vary across the genome, from less than 0.5 cM/Mb to more than 3 cM/Mb [15]. In addition, Jeffreys et al. [11] and many other studies have shown evidence of recombination hotspots in humans and other organisms [2, 13, 27, 32, 33], implying some regional variation may be due to the variable distribution of these hotspots.

The ability to simulate these hotspots is key to obtaining an accurate picture of the LD patterns in the genome, since hotspots tend to preserve LD in areas between the hotspots. Without hotspots, the simulations produce data with a much shorter extent of LD than is observed in empirical data [34].

3.4 Multiple populations

Simulating multiple populations allows us to compare differences between populations. One such measure is F_{ST}, a measure of the genetic distance between subpopulations by comparing the allele frequencies for each marker. Here we use the Weir

and Cockerham formula [30].

Gabriel et al. [6] found that block boundaries are often preserved across populations. The ability to model multiple populations will allow us to simulate these observations.

3.5 Haplotypes

A simulator that outputs haplotypes rather than the relationship between only two markers has the advantage that patterns over several SNPs can be analyzed. Studies have investigated the extent and causes of regions of limited haplotype diversity, or *haplotype blocks* [3, 6, 12, 18, 21, 29, 34]. These regions often have only four or five common haplotypes throughout the population, even when 10 or more SNPs are examined. In this paper we use the haplotype block definition in Gabriel et al. [6], which searches for series of SNPs that have high pairwise LD. For a review of this definition, please refer to Appendix A.2.

Chapter 4

Problem statement

4.1 The need for an enhanced simulator

As seen in Figure 3-4, the null model does not fit empirical data well for characteristics such as LD. To draw accurate conclusions from simulated data, we need to be able to create data which looks realistic.

Specifically, we would like to match the following characteristics:

Allele frequency

Allele frequency is the distribution of frequency of alleles at variable sites.

LD patterns

As described in Section 2.1, LD is the measure of non-random association between two markers. We examine the relationship between LD and physical distance.

Fraction of ancestral alleles

This is the probability that an allele of a given frequency is ancestral.

Genetic distance between subpopulations

As defined in Section 3.4, F_{ST} is a measure of genetic distance between subpopulations.

	variable recombination	multiple populations	haplotypes
Hudson	N	Y	Y
Schaffner	Y	Y	N
Posada	Y	N	Y

Table 4.1: **Current simulators and available features** The three simulators compared are Hudson's *make samples* simulator[8], Schaffner's two-locus simulator[26], and Posada's SNPsim program[22].

Length of haplotype blocks

For each population, we examine the average length of haplotype blocks for three groups of blocks: all blocks, blocks longer than 20kb and blocks longer than 10kb. As noted in Section 3.5, the definition used in this paper is from Gabriel et al. [6].

Haplotype block genome coverage

After calculating the length of haplotype blocks, we are also interested in how much of the genome falls within blocks.

Of the capabilities described in Sections 3.3-3.5, Table 4.1 shows that of the three coalescent simulators surveyed, none of them has all the desired features. The goal of this thesis is to create a tool to fulfill that need.

4.2 Needed features

This section describes the needed features for the simulator.

4.2.1 Non-uniform recombination rate

We need the flexibility to model non-uniform recombination rates. The user should have complete freedom to specify any recombination model.

4.2.2 Support for multiple populations and population events

The user should be able to specify a wide variety of population events involving one or more populations. These events include:

Population size changes
Populations size changes will include both instantaneous and exponential size changes.

Bottlenecks
A bottleneck represents a brief period with a very small population size, or any event which produces a similar effect.

Expansions
An expansion is an exponential growth of a population, such as the population growth following the agricultural revolution.

Admixtures
Admixtures occur when two populations mix and breed with each other.

Splits
Splits occur when one part of a population leaves to form, or to join, another population.

4.2.3 Haplotype output

The output should be simulated haplotypes, in order to examine relationships between a sequence of SNPs, rather than just pairwise SNP comparisons. This allows us to study characteristics such as haplotype blocks.

4.2.4 History log

After drawing conclusions from the simulated data, we can look at the various events and correlate the characteristics of the haplotypes to the underlying historical events. For example, we may be interested in determining the contribution of recombination

events to haplotype block boundaries. We can generate our sequences, find the haplotype blocks, and return to the history log to determine where the recombination events occurred.

4.3 Engineering constraints

In addition we also need to observe the following constraints.

Programming in ANSI C Current software in population genetics is often written in C. Because we are creating a tool that population geneticists will need to be able to alter and add functions to, we need to use a commonly used language.

Modularity The program must be as modular as possible, both for ease of development but also ease of customization by others. We may want to add features as new datasets become available.

Ease of use Since we are providing the ability to use complex models, we need to ensure that the input to the program is easily read by the user.

4.4 Advantages of a flexible simulator

The data generated by this simulator will better resemble empirical data. This flexibility will let researchers tailor the simulated output to their relevant data sets. They can then analyze available data in the context of more accurate simulations.

This simulator can also be used to examine theories of human migration. The out of Africa theory has been shown to contribute to the reduced diversity in non-African populations. One can estimate the likelihood of a proposed event by examining the event's effects on the characteristics of the simulated sequence, and comparing those characteristics to the same measures derived from empirical sequence.

Chapter 5

Implementation of CoSi

The options when approaching this project were to modify an existing program by adding necessary features, or to write a simulator from scratch. After investigating existing programs, we decided to write an original program, in order to meet our desired constraints specified in the last chapter.

Figure 5-1 shows the structure of the program, CoSi (*Co*alescent *Si*mulator). The simulator engine executes the simulation by manipulating the data representation maintained by the demography manager, through Poisson and historical events. After constructing the complete coalescent tree, the engine calls the mutation subroutine to place mutations and output haplotypes.

5.1 Null model, no recombination

Initially we implemented a simple coalescent simulator and ensured that it met theoretical values for the time to the most recent common ancestor. The time to reduce the population from n to $n-1$ chromosomes is T_n,

$$E[T_n] = \frac{4N}{n(n-1)}$$

where N is the population size. For k chromosomes sampled from a population of N diploid individuals, the time to coalesce to one ancestor has an expected mean of

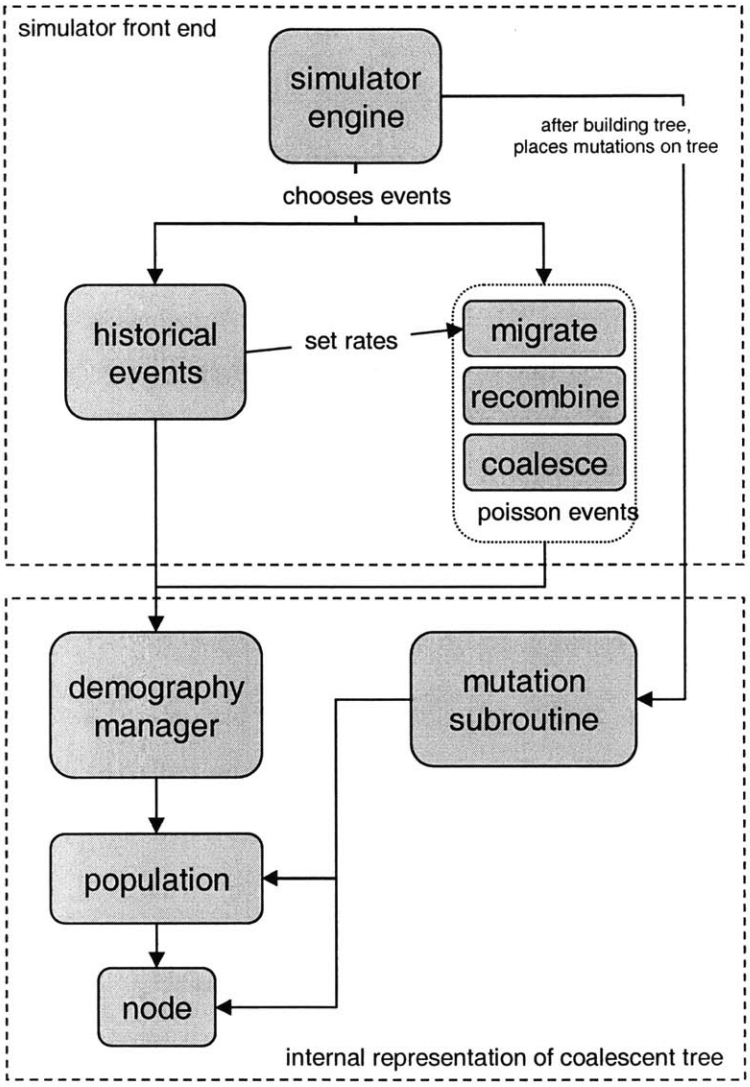

Figure 5-1: Structure of CoSi

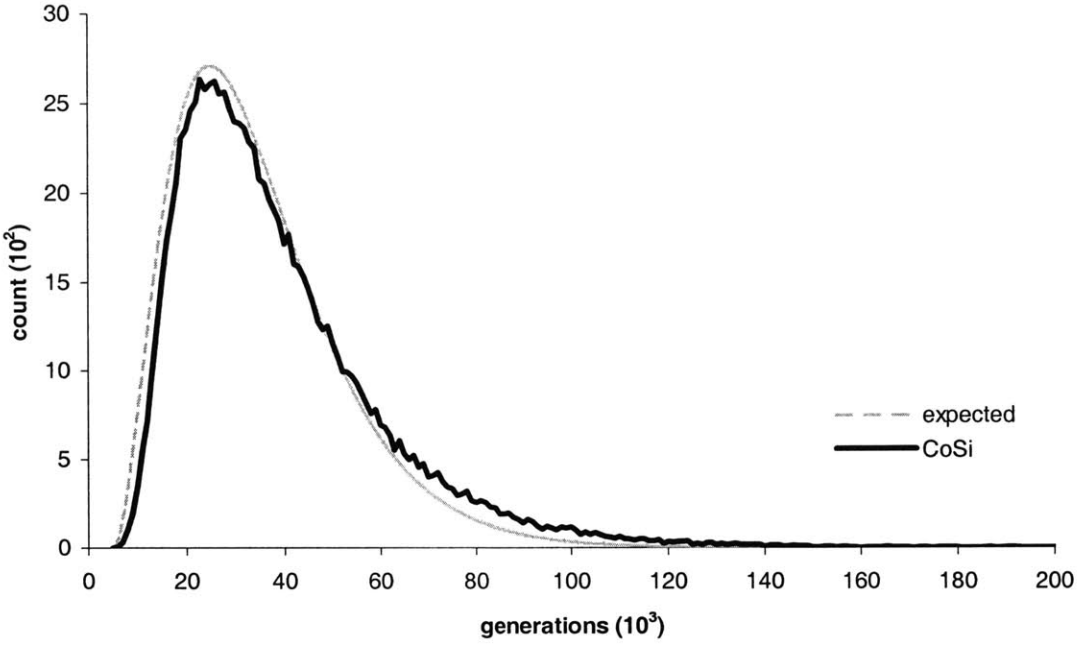

Figure 5-2: Distribution of coalescent times over 10^5 independent trials.

$\frac{4N}{1-\frac{1}{k}}$. The expected distribution is given by

$$\begin{aligned} P(t|k,\theta) &= \frac{P(t,k|\theta)}{P(k|\theta)} \\ &= \frac{(\theta\tau)^2 e^{-\tau(1+\theta)}/2}{\frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^2}, \quad \tau = \frac{t}{2N_e}, \quad \theta = 4N_e\mu \end{aligned}$$

where μ is the mutation rate and k is the number of samples.

Figure 5-2 shows the distribution of these values for 10^5 independent simulations (null model, no recombination) against the expected distribution. This simulation, and all described in this thesis, are for 100 randomly sampled chromosomes per population.

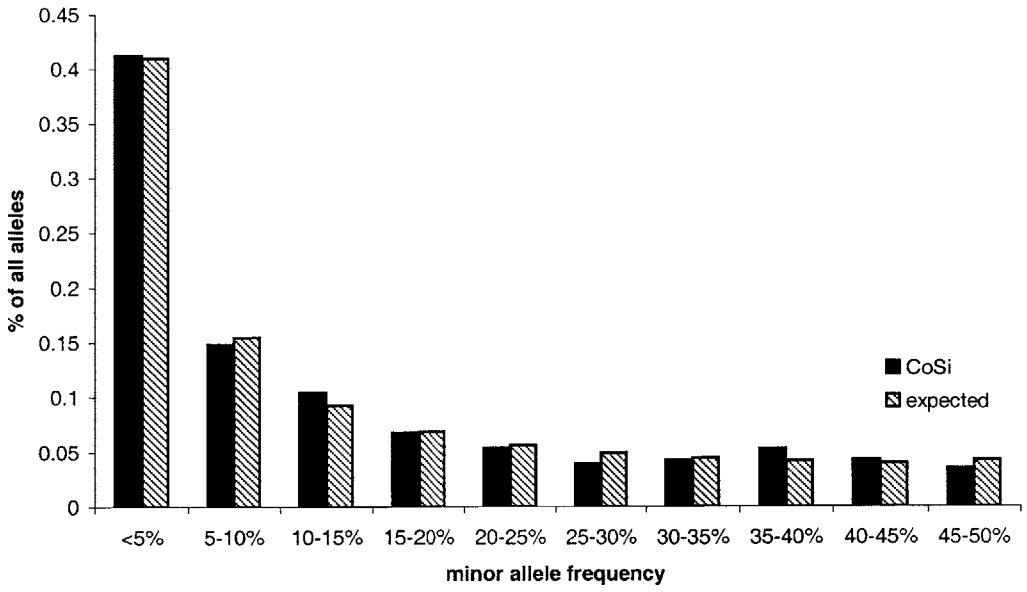

Figure 5-3: Minor allele frequency distribution for 10^2 independent trials.

5.2 Uniform recombination

Uniform recombination completes the basic coalescent simulator. Recombinations, like coalescences, are Poisson events with parameter λ,

$$\lambda = rl \sum_{i \in P} n_i$$

where r is the recombination rate, l is the length of the chromosome, P is the set of all populations, and n_i is the number of nodes in population i.

Mutations are placed on the chromosome in an exponentially distributed manner. The number of mutations for each segment unbroken by recombination is proportional to the time in all the branches of the genetic tree for that segment.

The expected number of SNPs with frequency f is proportional to

$$\frac{1}{fk}$$

where k is the number of samples [25, p.26]. Figure 5-3 shows the graph of allele frequency over 100 runs (null model, $r = 1 * 10^{-8}$).

5.3 Non-uniform recombination

In lieu of limiting the user to predefined recombination models, we allow complete freedom by defining the recombination rate with a piecewise model specified by the user in a text file. The total recombination rate is determined by integrating the rate over the region, and the recombination events are distributed accordingly.

A more extensive look at non-uniform recombination rate will be given in Section 6.1.

5.4 Population events: single populations

The events in the next two sections are described briefly, along with implementation details when relevant.

5.4.1 Population size changes

Two types of population size changes were implemented. Populations can either undergo an instantaneous size change, or can change their size exponentially over a specified number of generations.

5.4.2 Bottlenecks

A bottleneck is a specific instance of population size change. Bottlenecks are events where the population size is severely reduced for a short period. Many coalescences occur during the bottleneck and as a result, many variations become fixed or are lost. As described in Section 2.4.2, the out of Africa event was a major bottleneck event.

The inbreeding coefficient F defines the severity of the bottleneck. F is the probability that two randomly chosen chromosomes share a common ancestor during the bottleneck. F can be calculated from the length of the bottleneck t and the size of the population during the bottleneck N,

$$F = 1 - e^{-\frac{t}{2N}}.$$

The value of F completely determines the effect of the bottleneck. In other words, the effect of a bottleneck can be considered without the specific knowledge of t and N, and can thus be considered an instantaneous event. To simulate this event, we set the length t to one generation, and calculate the appropriate population size. Using this size (which will be improbably small), we then calculate how many coalescences occur in that generation.

5.5 Population events: multiple populations

Each population is a list of chromosomes we are currently simulating. Coalescences can only happen within a population, so to represent a migration we move a chromosome from one population to another (Figure 5-4).

A migration rate can be specified for every pair of populations $\{i, j\}$, which is the probability per chromosome that a chromosome migrates from i to j in each generation. This is implemented as a Poisson event, along with coalescence and recombination.

Larger scale moves between populations are implemented as user-specified events. A split occurs when a population breaks into two separate populations; an admixture occurs when two populations combine. Because the coalescent works in the reverse direction, a split is represented by combining two populations. An admixture is represented by creating a new population, with a certain probability that each chromosome in the original population will join the new population.

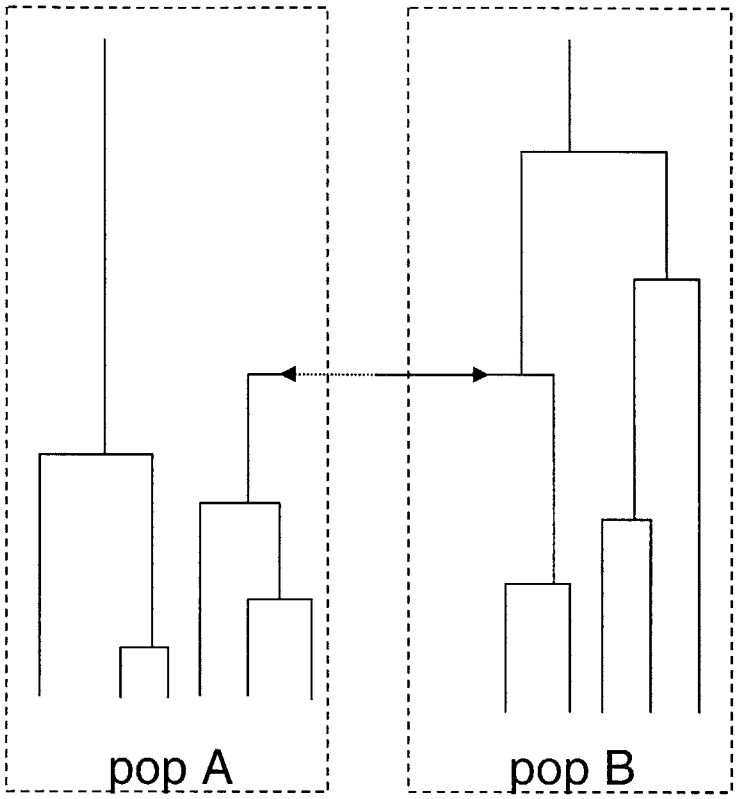

Figure 5-4: **Multiple populations and migration.** During a migration event (in the forward direction), one chromosome moves from population B to population A (dotted arrow). The simulator represents this event by moving a chromosome from population A to population B (solid arrow) and allowing it to coalesce (and recombine, not shown here) with any of the chromosomes in population B.

Chapter 6

Using CoSi

This section describes experiments using CoSi on complex models. First we examine models of recombination, then we simulate a comprehensive population model.

All simulated data shown in this section is ascertained in a TSC-like method described in Section 3.5.

6.1 Developing a recombination model

We use CoSi to build a model of non-uniform recombination. We start with uniform recombination, which we will refer to as H_0 for recombination, and build a model using nested hypotheses. Because we are using a constant population, we compare our results with empirical African-American data, which appears to fit the constant population model. All simulations are 100kb in length, with 100 independent trials.

For the constant recombination rate in H_0 we use the standard value $r = 1 * 10^{-8}$. We also use the mean recombination rate for hypothesis H_1 described below, $1.6 * 10^{-8}$, denoted as model $H_0^{1.6}$. LD patterns for H_0 are shown in Figure 6-1 and block statistics are shown in Table 6.1 (along are the block statistics for the next two models).

The first variation that we explore, H_1, is regional variation across the genome. We draw our rates from a distribution based on the Gabriel et al. dataset [6]. Each of these segments has a uniform recombination rate. As seen in Figure 6-2, regional variation creates only a slight increase in LD.

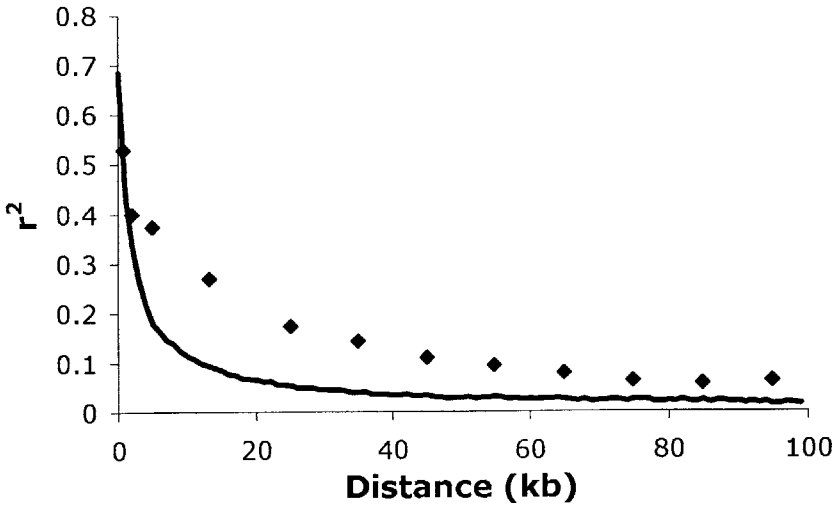

Figure 6-1: LD versus physical distance for uniform recombination, H_0

	Average block size (% Genome coverage)					
	all blocks		> 20 kb		> 10 kb	
H_0	5.9	(84.8)	23.9	(5.37)	15.2	(23.5)
$H_0^{1.6}$	1.9	(29.6)	-		12.2	(1.6)
H_1	2.2	(31.3)	22.4	(0.7)	13.2	(5.0)
H_2	10.5	(70.8)	33.6	(34.4)	21.9	(57.4)

Table 6.1: Block statistics for H_0,H_1, and H_2 recombination models. Gabriel et al. [6] observed block sizes of 9kb in African populations.

Next, we vary the local recombination rate by adding hotspots (H_2), in addition to using the distribution of H_1. We add hotspots with an exponential distribution with mean spacing 9kb, and distribute the recombination probability as a 10% background rate and a 90% hotspot rate. LD patterns and block statistics are shown in Figure 6-3.

6.2 Comprehensive model

The model shown in Figure 6-4 is the comprehensive model developed by S.F. Schaffner and used in this section. This model is used with both the H_0 ($r = 1x10^{-8}$) and H_2 recombination model described in Section 5.1. The empirical data referred to in the next section are from Gabriel et al. [6], which surveyed 51 autosomal regions

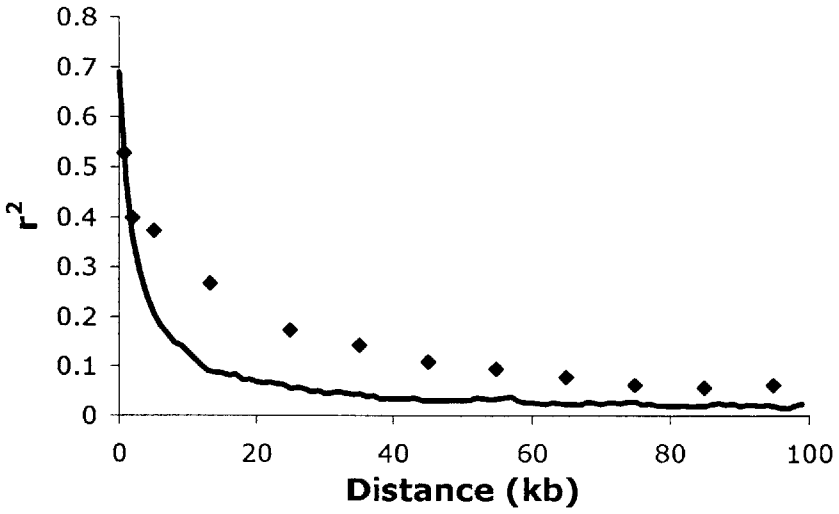

Figure 6-2: LD patterns for regional uniform recombination rates chosen from a distribution, H_1

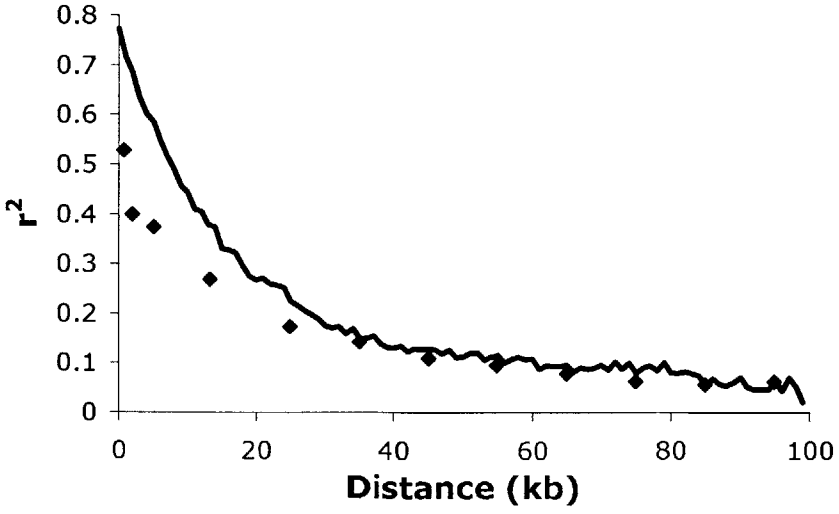

Figure 6-3: LD patterns for hotspot recombination model H_2

spanning 13 Mb.

6.2.1 Allele frequency

Figures 6-5, 6-6, and 6-7 show the allele frequency distribution for the European, Asian, and African-American populations, against the empirical values. The allele frequencies are for all markers found across all populations. The African-American population has low monomorphic alleles as expected, since the bottlenecks in non-African populations tend to fixate alleles, resulting in more monomorphisms.

6.2.2 LD versus physical distance

The LD graphs referred to in this section use only alleles with minor allele frequency $> 20\%$, for both empirical and simulated data. Figures 6-8 and 6-9 show the LD patterns for the model with constant recombination. The extent of LD is higher than the null model (Figure 3-4) but does not show as much high LD as the empirical data. Figures 6-10 and 6-11 show the patterns for the complete model. LD levels at shorter distances are elevated, possibly due to the lack of gene conversion, but the LD values at distances greater than 40kb appear to approximate the empirical data well.

6.2.3 Fraction of ancestral chromosomes

For empirical data, the probability that an allele of a given frequency is ancestral, $P(ancestral)$, is found by comparing human sequences to chimpanzee sequences. For a constant size population, the probability that an allele represents the ancestral state is equal to its frequency. The African population history is the closest to a constant population, and $P(ancestral)$ appears to have the expected linear relationship. For European and Asian populations, however, $P(ancestral)$ deviates from the linear value (Figure 6-12). This deviation indicates a bottleneck in the history of these populations.

As seen in Figure 6-13, the CoSi model produces patterns in the African and non-African populations that are similar to the empirical observations.

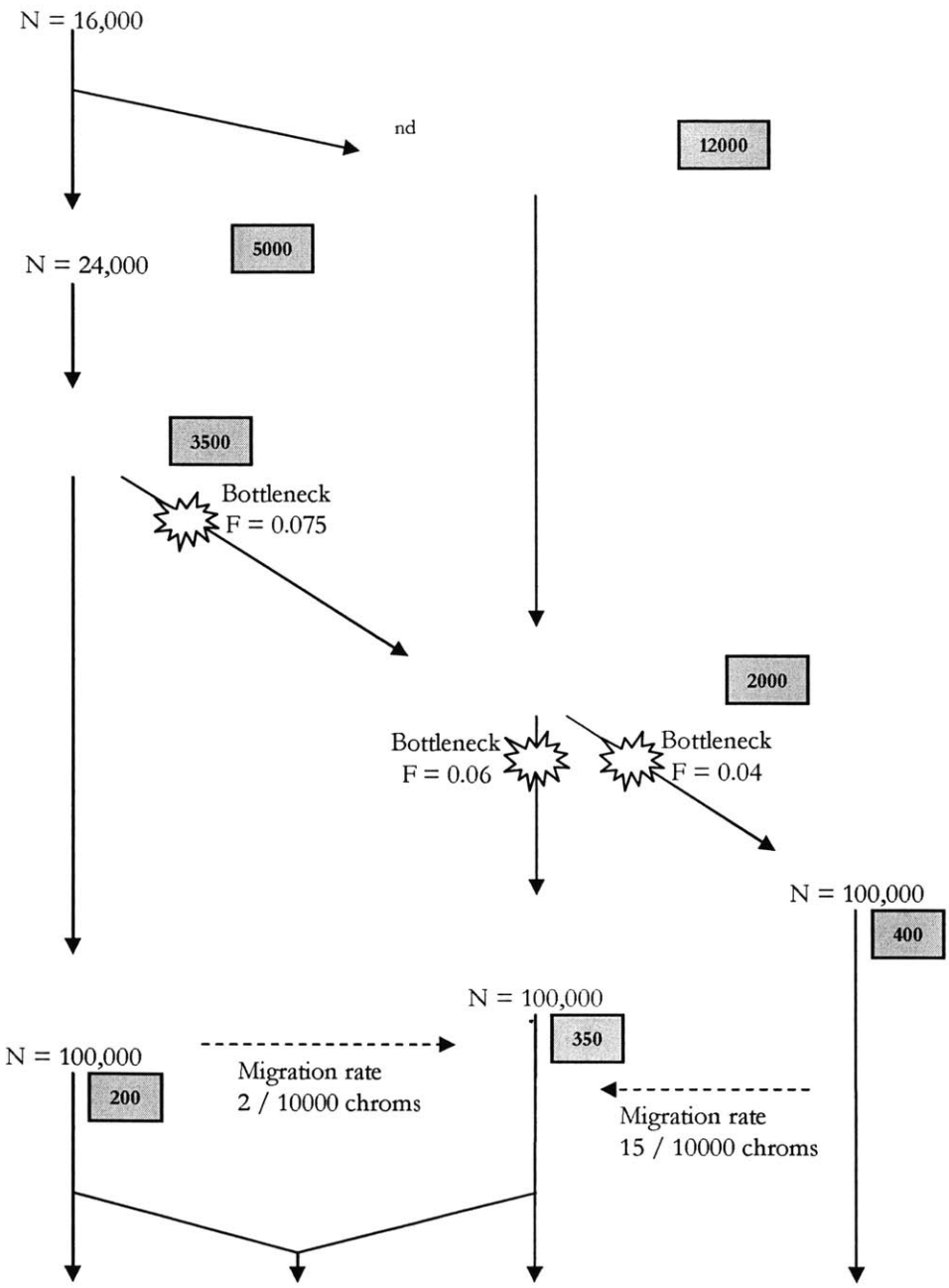

Figure 6-4: **Population model used in Section 6.2**. The gray boxes indicate the time of the event, where the number indicates how many generations ago the event occurred. The migration rates are given in the number of migrations per chromosome per generation.

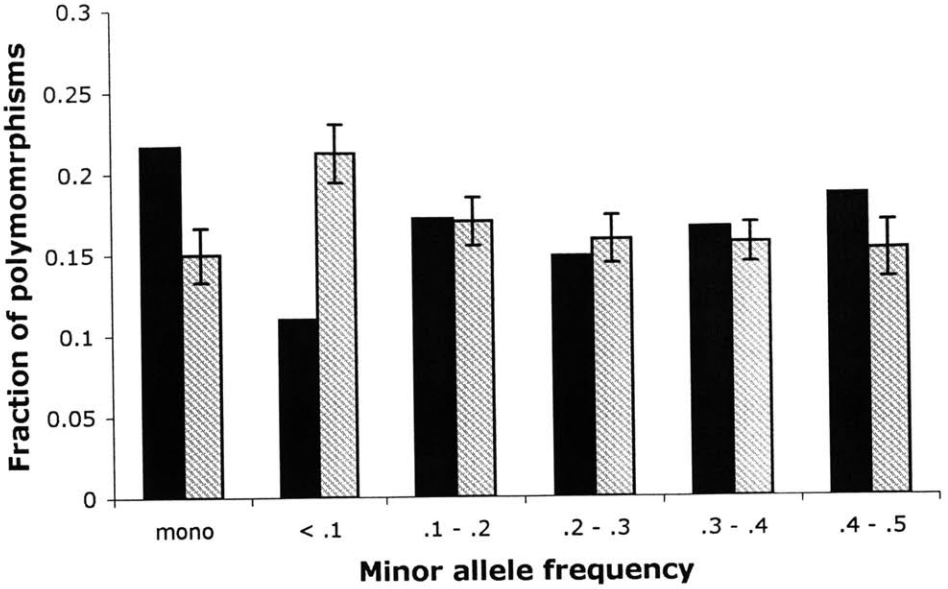

Figure 6-5: **Allele frequency of simulated European population.** The gray bars represent the allele frequencies observed in empirical data. Error bars represent standard deviation.

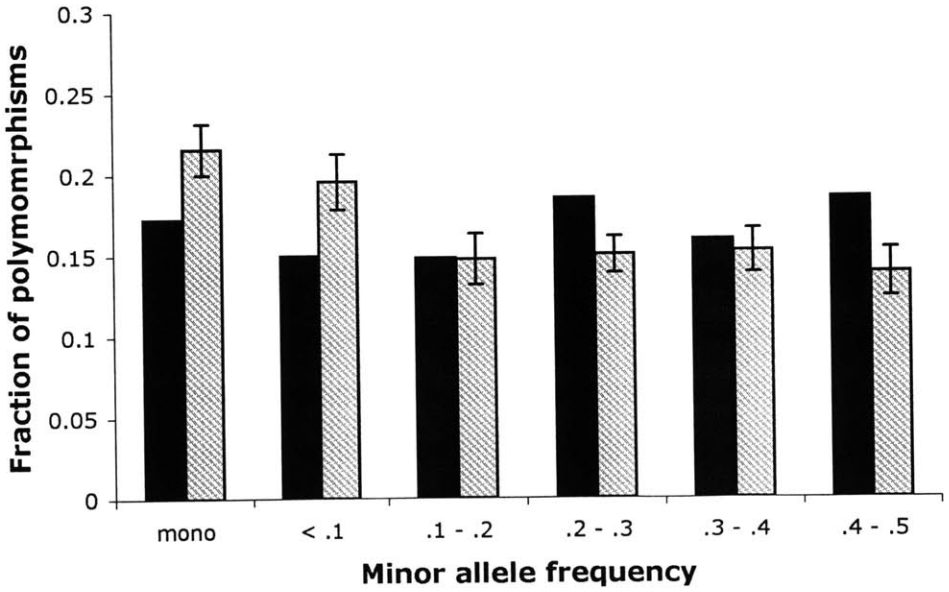

Figure 6-6: **Allele frequency of simulated Asian population.** The gray bars represent the allele frequencies observed in empirical data. Error bars represent standard deviation.

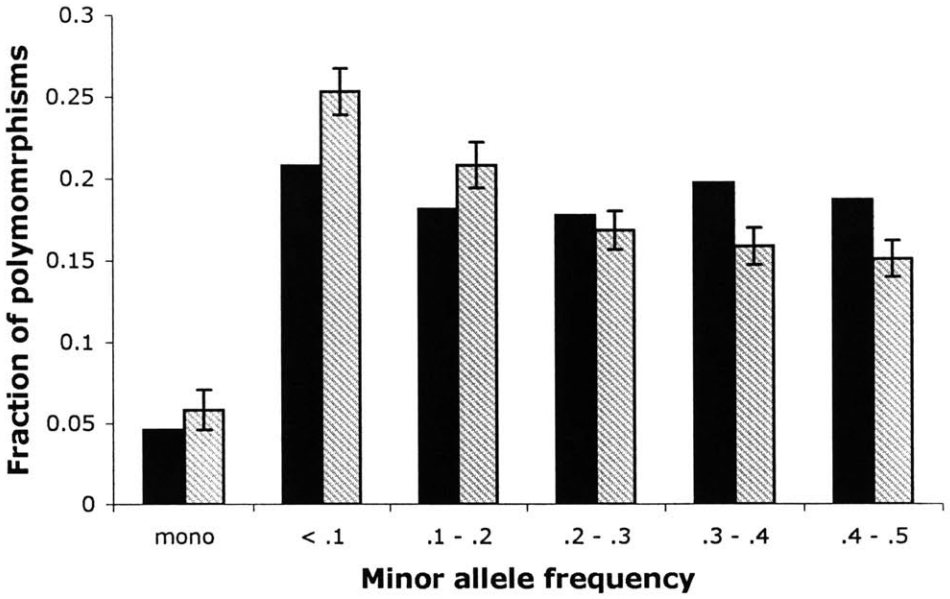

Figure 6-7: **Allele frequency of simulated African-American population**. The gray bars represent the allele frequencies observed in empirical data. Error bars represent standard deviation.

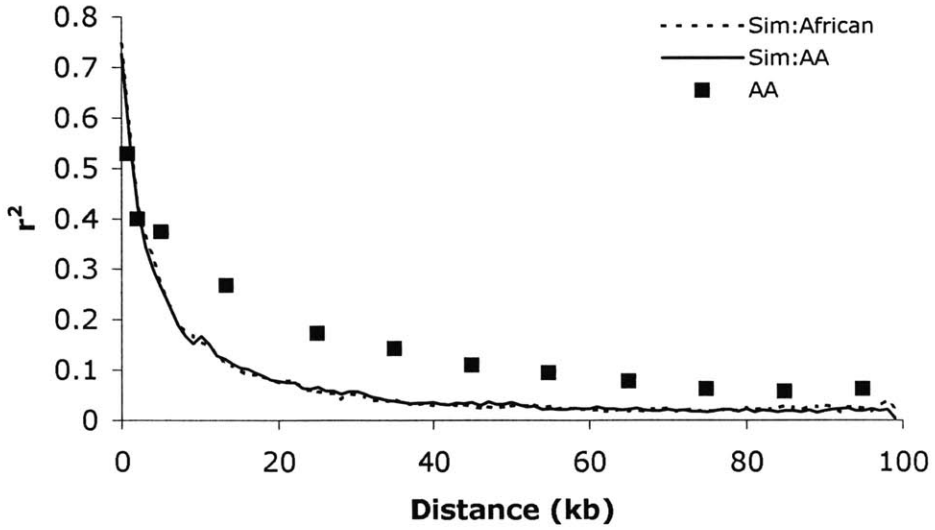

Figure 6-8: **LD for African populations with uniform recombination.** In this and following figures and tables, AA and Sim:AA refer to the empirical and simulated African-American datasets, respectively.

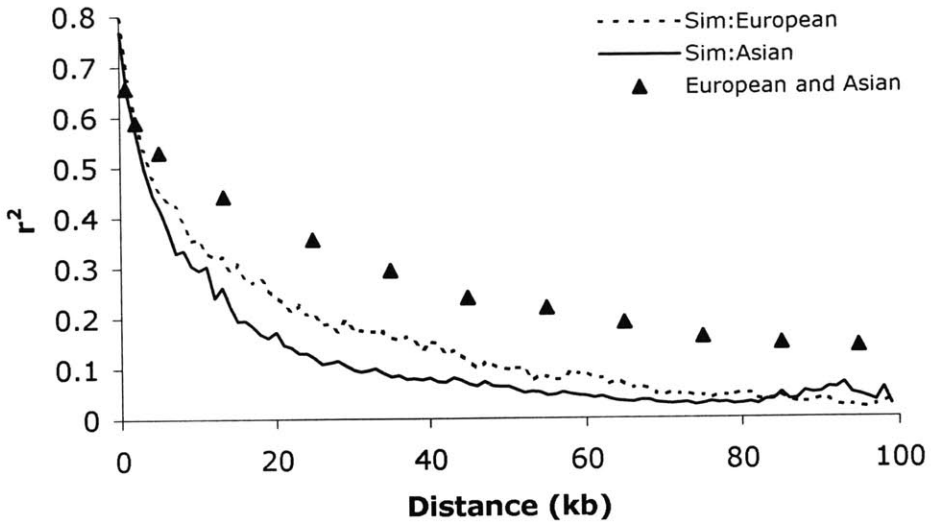

Figure 6-9: LD for European and Asian populations with uniform recombination.

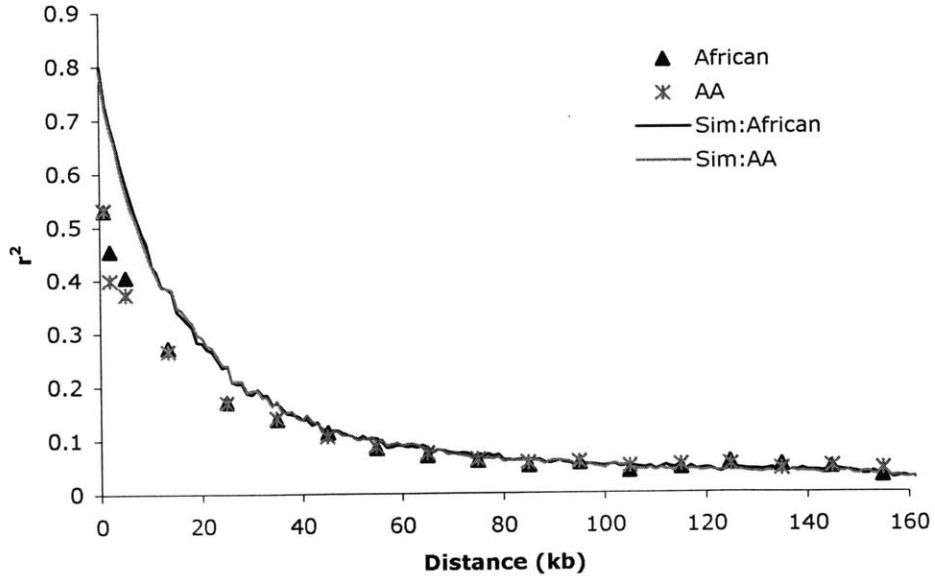

Figure 6-10: LD for African populations using the H_2 model of recombination.

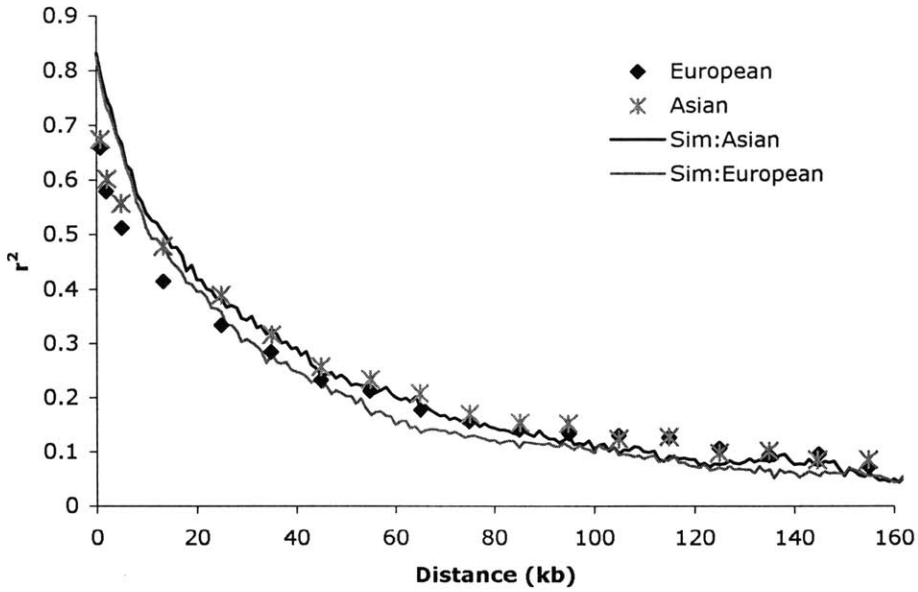

Figure 6-11: LD for European populations using the H_2 model of recombination.

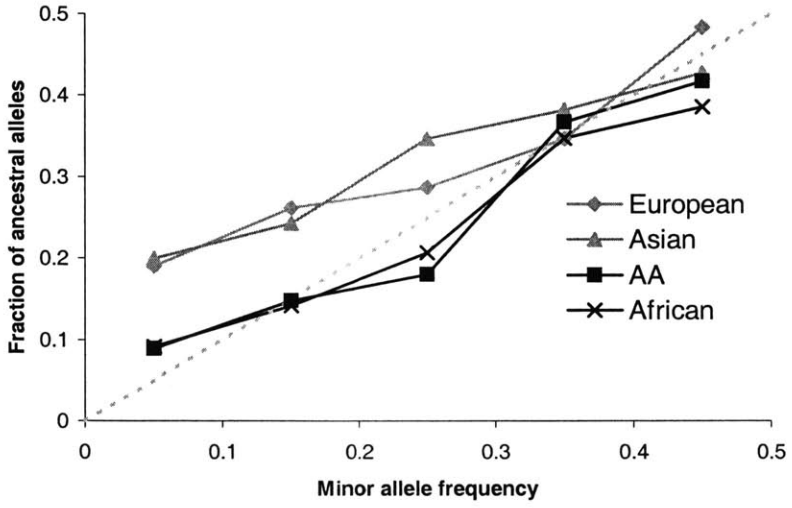

Figure 6-12: **Fraction of ancestral alleles in empirical data.** The dotted line indicates the linear relationship.

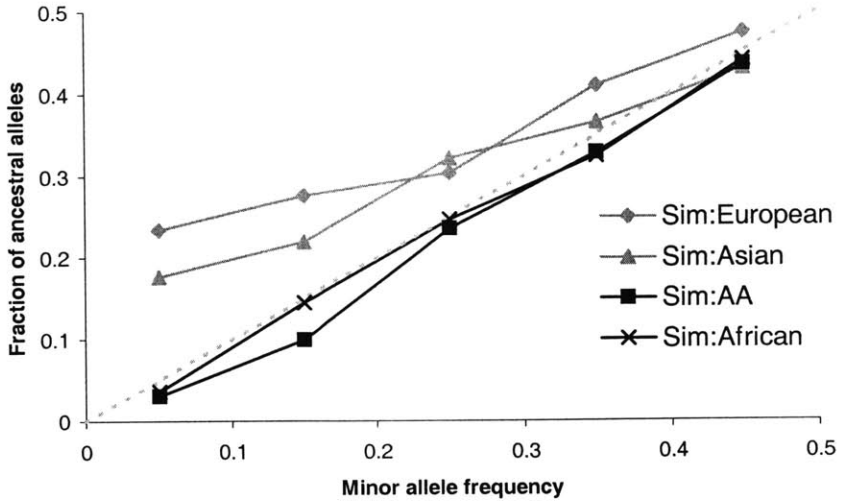

Figure 6-13: **Fraction of ancestral alleles in simulated data.** Compare this graph to Figure 6-12, which shows the corresponding empirical data.

	Sim.	Emp.
African, European	0.168	0.157
African, AA	0.010	0.016
African, Asian	0.158	0.200
European, AA	0.116	0.110
European, Asian	0.102	0.139
AA, Asian	0.109	0.162

Table 6.2: F_{ST} values for simulated and empirical data

6.2.4 Genetic distance between subpopulations

Table 6.2 shows the F_{ST} values for each pair of populations. The non-Asian comparisons represent the empirical data well, but the Asian values are low. This will later be examined by further altering the population parameters.

6.2.5 Haplotype blocks and block coverage

See Figure 6.3 for haplotype block statistics for simulated populations. Gabriel et al. found average block lengths of 9kb in the African samples and 18kb in the European and Asian samples, with a wide range of sizes (< 1kb to 174kb) [6]. As seen in Figure 6.3, the average block size for the simulated African populations is close to the

A *Uniform recombination*

	Average block size (% Genome coverage)					
	All blocks		> 20 kb		> 10 kb	
Sim:African	2.5	(15.8)	-		11.8	(1.9)
Sim:AA	2.4	(15.8)	-		12.8	(2.2)
Sim:European	6.0	(28.9)	25.1	(5.0)	16.0	(16.0)
Sim:Asian	4.2	(24.9)	22.1	(1.1)	14.1	(8.7)

B *Recombination hotspot model*

	Average block size (% Genome coverage)					
	All blocks		> 20 kb		> 10 kb	
Sim:African	9.6	(65.5)	32.2	(32.5)	21.7	(51.9)
Sim:AA	9.5	(65.7)	31.9	(32.8)	22.2	(51.3)
Sim:European	16.0	(77.0)	41.1	(55.1)	30.8	(67.5)
Sim:Asian	14.0	(74.8)	35.8	(48.3)	26.2	(64.8)

Table 6.3: **Haplotype block statistics for complex population model**, with (A) and without (B) recombination hotspots. Gabriel et al. observed blocks of 9kb and 18kb for African and non-African populations respectively.

desired average. Non-African populations show a lower than desired block average, which will be investigated in future work. These figures also indicate that a hotspot recombination model is essential for reproducing the desired block lengths.

Chapter 7

Conclusions

This paper has described CoSi, a new coalescent simulator with the flexibility to model SNP data that resembles empirical data.

7.1 Goals met

This section reiterates the requirements set in Chapter 4, and how CoSi satisfies these requirements.

Development of CoSi A simulator, CoSi, which provides much-needed flexibility was developed as described in Chapter 5. CoSi allows for user-specified non-uniform recombination, multiple populations, and a variety of population events.

Correctness of CoSi CoSi implements the basic coalescent process correctly. Results using simple models match theoretical results, as shown in Sections 5.1-5.3.

Desired features available CoSi is capable of simulating a variety of population events for one or more populations, and can use any user-defined recombination model.

Simulated data resembles empirical data As seen in Section 6.4, CoSi can be used to generate data which closely resembles empirical data. This will allow

researchers to draw more accurate conclusions about the underlying causes of observed genomic patterns.

Software engineering CoSi is implemented in ANSI C. The program is modular, as evidenced by the design in Section 5.1. Crucial elements are abstracted from the simulator to allow for greater flexibility, such as the recombination model described in Section 6.1. The parameter file is clear and easily read (Appendix B.1).

7.2 Uses and extensions of CoSi

As seen in Section 6.4, CoSi can be used to simulate complex population models. The data have characteristics which are also exhibited by empirical data. This thesis used a model developed to match a specific dataset, but CoSi could be used to model any proposed scenario.

In Section 6.1, we use CoSi to compare various recombination models. Other models could also be evaluated, such as the "warm and cool zones" proposed by Zhang et al. [34]. The advantage of the user-defined recombination model used in CoSi is that it offers complete flexibility to the researcher.

CoSi can easily be modified to include other components such as gene conversion [1], selection [9, 10, 19], or double mutations (for example, to simulate mitochondria).

The ability to generate realistic simulation data with CoSi provides researchers with the unprecedented opportunity to use simulations in a reliable way. Studies will be able to look at the entire model at once, rather than simulating each component separately. As genomics continues to expand into larger scale studies, simulations will become indispensable for proposing new studies, and for analyzing the underlying forces that shape our genetic information.

Appendix A

Additional background information

A.1 Linkage disequilibrium

Linkage disequilibrium (LD) is the measure of allelic association between two markers, and is often measured by r^2 and D'. Both r^2 and D' are related to a quantity D which measures the difference between expected and observed haplotypes. For two alleles with values A/a and B/b, these values are defined as follows:

$$D = f_{AB} - f_A f_B$$

$$r^2 = \frac{D^2}{f_A f_a f_B f_a}$$

$$D' = \frac{|D|}{D_{max}}, \quad D_{max} = \begin{cases} min(f_A f_b, f_a f_B), D > 1 \\ min(f_A f_B, f_a f_b), D \leq 1 \end{cases}$$

where f_A is the frequency of allele A, f_a is the frequency of allele a (equal to $1 - f_A$), and so forth. Note that D' is always between 0 and 1.

A.2 Finding haplotype blocks

Gabriel et. al [6] describes a block definition that uses a confidence limit on D'. Since D' gives high values to low frequency SNPs, the limit is calculated by determining the probability that a given D' could have produced the observed data, over all values of D', and examining the upper and lower 5% bounds.

Based on their confidence limits, each pair of markers is classified as a pair in strong LD, a pair with strong evidence of recombination, or an uninformative pair. Blocks are defined as continuous stretched of segments where at least 95% of the informative pairs exhibit strong LD. The specific parameters were determined empirically.

Appendix B

CoSi inputs

Inputs to CoSi are shown in this section.

B.1 Parameter file

This is the parameter file used to generate the simulations in Section 6.2. Comparison with Figure 6-4 should be straightforward. Keep in mind that the simulator works backwards. The following section explains the recombination file.

```
# sample file
# comments have #s in front of them
# newlines don't matter.

# sets the length of the segment, in base pairs.
length 200000

# mu, here it is 2e-8
mutation_rate 0.00000002

# recomb file
recomb_file model.out

# population info

# for each population, include a line:
# pop_define pop-index pop-label

pop_define 1 european
pop_define 3 african-american
pop_define 4 asian
pop_define 5 african
pop_define 6 afr2
pop_define 7 eurafr


#init sample pops
```

```
# for each sample set, include
# pop_size pop-label pop-size
# sample_size pop-label sample-size

#european
pop_size 1 100000
sample_size 1 100

#african american
pop_size 3 100000
sample_size 3 100

#asian
pop_size 4 100000
sample_size 4 100

#african
pop_size 5 100000
sample_size 5 100


# FORMAT OF POPULATION EVENTS
# pop_event admix "label" from-pop-index to-pop-index gen to-pop-size %moved
# pop_event split "label" origin-pop-index new-pop-index gen
# pop_event bottleneck "label" pop-index gen coeff
# pop_event migration_rate "label" from-pop-index to-pop-index gen rate
# pop_event change_size "label" pop-index gen new-size
# pop_event exp_change_size "label" pop-index start-gen end_gen start_size end_size

# Comprehensive population model

# Set migration rates
pop_event migration_rate "afr->eur migration" 5 1 0 .00002
pop_event migration_rate "as->eur migration" 4 1 0 .00015

# Create African-American population
pop_event admix "african american pop" 3 7 5 100000 .2
pop_event split "european to aa" 1 7 6
pop_event split "african to aa" 5 3 7

# Agricultural expansions
pop_event change_size "agriculture - african" 5 200 24000
pop_event change_size "agriculture - european" 1 350 7500
pop_event change_size "agriculture - asian" 4 400 7500

# Mixing of African2 and European populations, followed
# immediately by Asian and European split and bottlenecks.
# Eliminate migration rate, since Asian population ceases
# to exist.
pop_event migration_rate "as->eur migration" 4 1 1997 0
pop_event bottleneck "asian bottleneck" 4 1998 .04
pop_event bottleneck "european bottleneck" 1 1999 .06
pop_event split "asian and european split" 1 4 2000
pop_event admix "african2 european admix" 1 6 2001 16000 .04

# Out of Africa event
pop_event migration_rate "afr->eur migration" 5 1 3498 0
pop_event bottleneck "OoA bottleneck" 1 3499 .075
pop_event split "out of Africa" 5 1 3500

# African expansion
pop_event change_size "african pop size" 5 5000 16000

# African2 population split
pop_event split "african split" 5 6 12000
```

B.2 Recombination input file

The format of the recombination file is a piecewise definition of the recombination model:

```
[Position (bp)]      [recombination rate (recs/bp/gen)]
```

The following is a recombination file from a model with hotspots.

```
1 1.3306e-09
79444 8.50665e-05
79445 1.3306e-09
80251 7.76718e-05
80252 1.3306e-09
81680 0.00111013
81681 1.3306e-09
112343 1.9749e-09
112344 1.3306e-09
113767 0.000157409
113768 1.3306e-09
116201 1.42287e-05
116202 1.3306e-09
130617 2.90585e-07
130618 1.3306e-09
130792 0.000124841
130793 1.3306e-09
131085 1.10916e-05
131086 1.3306e-09
131213 0.00096577
131214 1.3306e-09
133475 0.000215526
133476 1.3306e-09
134662 0.00140553
134663 1.3306e-09
137603 0.000639362
137604 1.3306e-09
140671 8.91719e-06
140672 1.3306e-09
167331 3.19626e-05
167332 1.3306e-09
169877 0.00056655
169878 1.3306e-09
173184 2.73664e-06
173185 1.3306e-09
```

Bibliography

[1] K. Ardlie, S. N. Liu-Cordero, M. A. Eberle, M. Daly, J. Barrett, E. Winchester, E. S. Lander, and L. Kruglyak. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.*, 69:582–589, 2001.

[2] R. M. Badge, J. Yardley, A. J. Jeffreys, and J. A. Armour. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Human Molecular Genetics*, 9(8):1239–1244, 2000.

[3] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, Oct. 2001.

[4] W. Enard, M. Przeworski, S. E. Fisher, C. S. L. Lai, V. Wiebe, T. Kitano, A. P. Monaco, and S. Pääbo. Molecular evoluation of FOXP2, a gene involved in speech and language. *Nature*, 418:869–872, 2002.

[5] W. J. Ewens. *Population genetics.* Methuen & Co, London, 1927.

[6] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, June 2002.

[7] R. R. Hudson. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, 109:611–631, Mar. 1965.

[8] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[9] R. R. Hudson and N. Kaplan. The coalescent process and background selection. *Phil. Trans. R. Soc. Lond. B*, 349:19–23, 1995.

[10] R. R. Hudson and N. L. Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 123:831–840, Nov. 1988.

[11] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature Genetics*, 29:217–222, Oct. 2001.

[12] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29:233–237, Oct. 2001.

[13] L. Kauppi, A. Sajantila, and A. J. Jeffreys. Recombination hotspots rather than population history dominate linkage disequilibrium on the mhc class II region. *Human Molecular Genetics*, 12(1):33–40, Jan. 2003.

[14] J. Kingman. The coalescent. *Stochastic Process and their Applications*, 13:235–248, 1982.

[15] A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, July 2002.

[16] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, 22:139–144, June 1999.

[17] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genetics*, 27:234–236, Mar. 2001.

[18] S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *Am. J. Hum. Genet.*, 71:1129–1137, 2002.

[19] C. Newhauser and S. M. Krone. The geneology of samples in models with selection. *Genetics*, 145:519–534, Feb. 1997.

[20] M. Nordborg. Structured coalescent processes on different time scales. *Genetics*, 146:1501–1514, Aug. 1997.

[21] M. Phillips, R. Lawrence, R. Sachidanandam, A. Morris, D. Balding, M. Donaldson, J. Studebaker, W. Ankener, S. Alfisi, F.-S. Kuo, A. Camisa, V. Pazorov, K. Scott, B. Carey, J. Faith, G. Katari, H. Bhatti, J. Cyr, V. Derohannessian, C. Elosua, A. Forman, N. Grecco, C. Hock, J. Kuebler, J. Lathrop, M. Mockler, E. Nachtman, S. Restine, S. Verde, M. Hozza, C. Gelfand, J.Broxolme, G. Abecasis, M. Boyce-Jacino, and L. Cardon. Chromosome-wide distribution of haplotype blocks and the role of recombination hotspots. *Nature Genetics*, 33:382–387, Mar. 2003.

[22] D. Posada and C. Wiuf. Simulating haplotype blocks in the human genome. *Bioinformatics*, 19(2):289–290, 2003.

[23] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411:199–204, 2001.

[24] J. D. Rioux, M. J. Daly, M. S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E. J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S. B. Bull, R. S. McLeod, A. M. Griffiths, A. Bitton, G. R. Greenberg, E. S. Lander, K. A. Siminovitch, and T. J. Hudson. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genetics*, 29:223–228, Oct. 2001.

[25] A. R. Rogers. Lecture notes on evolutionary genetics. http://www.anthro.utah.edu/~rogers/ant4221/Lecture/a_spectrum.pdf (PDF, 14 May 2003), Univ. of Utah, 20 Oct. 2002.

[26] S. F. Schaffner. CoalSim [two-locus coalescent simulator]. Computer program, 2003.

[27] R. A. Smith, P. J. Ho, J. B. Clegg, J. R. Kidd, and S. L. Thein. Recombination breakpoints in the human β-globin gene cluster. *Blood*, 92(1):4415–4421, Dec. 1998.

[28] N. Takahata, Y. Satta, and J. Klein. Divergence time and population size in the lineage leading to modern humans. *Theoretical population biology*, 48:198–221, 1995.

[29] N. Wang, J. M. Akey, K. Zhang, R. Chakraborty, and L. Jin. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of populations history, recombination, and mutation. *Am. J. Hum. Genet.*, 71:1227–1234, 2002.

[30] B. Weir and C. C. Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution*, 36(6):1358–1370, 1984.

[31] Z. Yang. On the estimation of ancestral population sizes of modern humans. *Genet. Res.*, 69:111–116, 1997.

[32] C. Yauk, P. Bois, and A. Jeffreys. High-resolution sperm typing of meiotic recombination in the mouse mhc e_β gene. *The EMBO Journal*, 22(6):1389–1397, 2003.

[33] S. P. Yip, J. U. Lovegrove, N. A. Rana, D. A. Hopkinson, and D. B. Whitehouse. Mapping recombination hotspots in human phosphoglucomutase. *Human molecular genetics*, 8(9):1699–1706, 1999.

[34] K. Zhang, J. M. Akey, N. Wang, M. Xiong, R. Chakraborty, and L. Jin. Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Human Genetics*, Apr. 2003.