INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Disequilibrium Fine-Mapping of a Rare Allele via Coalescent Models of Gene Ancestry

by

Jinko Graham

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

1998

Approved by Elizabeth Thompson
(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree Biostatistics

Date Aug. 17, 1998

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to UMI Dissertation Services 300 North Zeeb Road, Ann Arbor, MI 48106 or to the author.

Signature [signature]

Date August 15 1998

University of Washington

Abstract

Disequilibrium Fine-Mapping of a Rare Allele via Coalescent Models of Gene Ancestry

by Jinko Graham

Chairperson of Supervisory Committee

Professor Elizabeth A. Thompson

Departments of Statistics and Biostatistics

Genetic linkage studies based on pedigree data have limited resolution, due to the relatively small number of segregations. Disequilibrium mapping, which uses population associations to infer the location of a disease mutation, provides one possible strategy for narrowing the candidate region. We develop a coalescent model for the ancestry of a random sample of disease alleles, and use it to investigate population association as a tool for fine-mapping a rare disease. Recombination events may be placed on the ancestral coalescent, and define the recombinant classes, the sets of sampled disease alleles descending from the meiosis at which a given recombination occurred. All disease haplotypes within a recombinant class are identical by descent at the marker. This identity by descent underlies linkage disequilibrium, the allelic association that is due to genetic linkage. We first investigate factors influencing marker identity by descent in sampled disease haplotypes, and the power to detect allelic associations. We then combine Monte Carlo generation of recombinant classes with an analytic method for computation of the probability of observed disease haplotypes conditional on latent recombinant classes, to obtain a linkage likelihood for fine-scale mapping. This like-

lihood can take into account known features of population history, such as changing patterns of population growth. Single-marker disequilibrium mapping is compared with interval disequilibrium mapping, and an extension to multipoint mapping is discussed. The method and its properties are illustrated with simulated data examples, constructed to be typical of fine-scale mapping of rare diseases in the Finnish and Japanese populations. Possible departures from assumptions in applications to real diseases are discussed, along with their effect on estimated recombination fractions.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I owe special thanks to my supervisor, Elizabeth Thompson, for introducing me to the topic, and for her guidance, support, and enthusiasm for this work. For their helpful comments and interest, I would like to recognize my committee members: Steve Self, Ellen Wijsman, Joe Felsenstein, and Norm Breslow. I am particularly grateful to Joe Felsenstein and Norm Breslow for careful reading of the dissertation and many helpful suggestions, and to Ellen Wijsman for useful discussions about disequilibrium mapping and the Werner's syndrome example. Additionally, I wish to thank Bruce Bare for filling in as my Graduate School Representative on very short notice.

I am indebted to my research assistantship advisors, Ake Lernmark and Norm Breslow for their guidance throughout my studies at the University of Washington. I would also like to acknowledge Ake Lernmark for serving as my biology project advisor, and Ingrid Kockum for invaluable help with research connected to the assistantship.

Thanks are also due to participants in the Mathematical and Statistical Genetics groups at the University of Washington, who provided valuable feedback on several practice talks connected to this dissertation, and to Katrina Goddard for discussions about the Werner's syndrome example. Finally, I would like to thank my husband, Brad McNeney, for his encouragement and sense of humour.

This work was supported in part by NIH grant DK-42654.

DEDICATION

To Romeo

Chapter 1

INTRODUCTION

1.1 Motivation

This research investigates the use of allelic associations as a tool for fine-mapping a rare disease. Unlike standard linkage studies, which use data from families, affected relatives, or even haplotypes within inbred affected individuals, association studies use population genetic data. The basic premise is that the associated marker allele may either be the disease variant itself, or in close proximity. When the marker locus and disease locus are close together, the chromosomal segment or haplotype they define tends to be passed intact from parent to offspring, without fragmentation by recombination events. The closer the two loci are, the less fragmentation is expected over many such transmissions or segregations. All linkage studies use cosegregation of trait and marker loci to infer disease location. However, the relatively small number of segregations within families limits the mapping resolution of traditional linkage studies (Boehnke 1994). Further localization is therefore required to reduce the cost of subsequent cloning and sequencing efforts. *Disequilibrium mapping* methods aim to narrow the candidate regions. Allelic associations between the disease and markers in the candidate region are assumed to be a consequence of genetic linkage. Such association or linkage disequilibrium is expected when a fair proportion of marker alleles on disease haplotypes descend from the same ancestral marker allele, or are *identical by descent.*

One drawback of disequilibrium mapping is that association as a consequence of

identity by descent cannot be differentiated from association due to hidden population stratification or admixture. Focussing on markers known to be linked to the disease, and studying associations in apparently homogeneous populations reduces but may not completely eliminate the problem. Family-based association methods, such as the transmission-distortion test (Spielman et al. 1993) and related procedures (Schaid and Sommer 1993, Self et al. 1991), may be used for *linkage detection*, and are robust to associations due to admixture. In essence, these methods study the segregations of disease haplotypes from parents to affected offspring, and test for distortion of marker transmission from the Mendelian expectation under no linkage. To detect linkage, the methods require that most parental disease alleles descend from the same disease mutation, and that the corresponding disease haplotypes carry the same ancestral marker allele. This shared identity by descent at disease and marker loci is reflected by allelic association in the population.

In the remainder of this chapter, we discuss how study of the underlying process of marker and disease co-identity by descent can lead to insights into disequilibrium-based methods of linkage mapping and linkage detection. Coalescent models of gene ancestry (Kingman 1982a), or of lines of descent, provide one possible starting point. Such models are the foundation of the research in this dissertation. We also review current disequilibrium-mapping methods, and describe fine-scale mapping examples to be used throughout. In chapter 2, we extend standard coalescent theory for the ancestry of a population random sample to a random sample of disease alleles. In chapter 3, this *ancestral coalescent* is used to investigate the patterns of marker identity by descent in a random sample of disease haplotypes. Chapter 4 superposes the assignment of marker allelic types onto these patterns of identity by descent, and investigates the resulting population associations. In chapter 5, we develop a method for obtaining fine-scale linkage likelihoods, based on the underlying identity by descent at a single marker and on data from sampled disease haplotypes, assuming knowledge of the population marker allele frequencies. This single-marker approach is extended to

interval mapping in chapter 6; extensions to multipoint mapping are also discussed. Possible departures from assumptions in applications to real diseases are discussed in chapter 7, along with their effect on estimated recombination fractions.

1.2 Association as a consequence of identity by descent

Allelic associations arise when certain combinations of alleles at different loci occur together more frequently on a haplotype than would be expected under random association. This can occur for a number of reasons, including population stratification, recent admixture of populations, natural selection acting on certain alleles and therefore on accompanying haplotypes, recent population bottlenecks, genetic drift, new mutations on rare haplotypic backgrounds, arising either spontaneously or through migration, or a combination of these forces (Wijsman 1997, Weeks and Lathrop 1995).

After a single disease mutation is introduced into a population, the initial linkage disequilibrium or population association with marker alleles on the background haplotype decays over generations, as recombination events break up the haplotype. Every segregation or meiosis provides an opportunity for recombination between the marker and the mutation. At each segregation or meiosis, the probability of recombination is given by the recombination fraction, and so the rate of decay of disequilibrium will be slower for closer markers. The amount of decay depends on the number of segregations in the ancestry of the disease alleles (Thompson 1978, Arnason et al. 1977). The more segregations that relate the disease alleles, the more fragmentation of the ancestral disease haplotype, and the finer the scale of mapping (Thompson 1997). For example, because there are far more segregations relating affected members of a genetic isolate than there are relating a nuclear family, the scale of mapping is greatly decreased in studies of the genetic isolate. In essence, disequilibrium mapping uses historical recombination events to infer disease location (Edwards 1981). Markers closer to the disease are expected to retain their ancestral allele over more

segregations than farther markers, or have fewer historical recombinations. Thus, if all markers in a candidate region are equally informative (i.e., have the same number and frequency of alleles), the marker closest to the disease mutation is expected to have the strongest allelic association.

Implicit in this reasoning is the notion of a single ancestral disease mutation, and a corresponding ancestry relating the disease alleles at present. Segregations on the ancestry are analogous to segregations on a pedigree. In both instances, the disease is localized by considering recombination events occurring during these segregations, and the patterns of marker identity by descent that occur as a result. The concept of gene identity by descent unifies all methods of linkage analysis, as noted by Thompson (1997). In disequilibrium mapping, the identity by descent among a random sample of disease alleles is described by their ancestry. An ancestral perspective is therefore natural. The ancestry is, in turn, influenced by demographic history. Analyses incorporating historical and other cultural information would therefore be expected to make the most efficient use of data, as previously noted (Jorde 1995, Thompson et al. 1992, Cavalli-Sforza et al. 1994).

Even with rare diseases, there may be several disease pathways in the population, including different mutations or predisposing loci. Thus, the concept of disease specificity, the extent to which affected persons share the same disease mutation (Thompson 1997, Bishop and Williamson 1990), is important. Regardless of the study design, diseases with low specificity will be more difficult to map and detect. Disease specificity may be increased by using data from genetic isolates, such as Finland. Finland is an ideal population for disequilibrium fine-mapping of those diseases present in the population because the disease specificity tends to be high due to isolation and a relatively small number of founders. On the other hand, as with any genetic isolate, many diseases will be absent from the population. However, for those diseases that are present, the date of population founding and subsequent rate of population growth would seem to imply a sufficient number of ancestral segregations for fine localization

(e.g., Mitchison et al. 1995, Aaltonen et al. 1994, Kestilä et al. 1994, Sulisalo et al. 1994, Lehesjoki et al. 1993). By contrast, younger populations with high specificity, such as the community discussed in Houwen et al. (1994), permit disequilibrium mapping at coarser scales. In these very young populations, linkage detection may also be possible. Houwen et al. (1994) discuss linkage detection methods based on genome screens of a few affected individuals for shared segments representing possible regions identical by descent. Van der Meulen and te Meerman (1997) have simulated haplotypes under a genetic drift model to investigate properties of these screening methods.

1.3 Disequilibrium mapping methods

Since the initial success of disequilibrium mapping of cystic fibrosis (Cox et al. 1989), and Huntington's disease (Snell et al. 1989, Theilmann et al. 1989), disequilibrium mapping has attracted much attention from practitioners, and a number of authors have formulated inference approaches to the problem.

Hästbacka et al. (1992) calculated a moment-based estimator of location for diastrophic dysplasia, a rare recessive disease in Finns. They derived approximate bounds on the accuracy of the estimate by appealing to theory developed for a Luria-Delbrück model of bacterial reproduction, in which individuals reproduce by doubling at synchronized intervals. A single copy of the disease allele was assumed at population founding, t_f generations before present. Under this assumption, the probability that a disease lineage has no recombination events between the disease and marker locus is $(1-r)^{t_f}$, where r is the recombination fraction. This probability is also the expected proportion of nonrecombinant disease haplotypes. Equating this probability to the proportion of nonrecombinants in the disease sample yields a moment estimate of r. As noted by Kaplan et al. (1995), this location estimate is based on a single observation, since the proportion of nonrecombinants in the disease sample is the outcome

of one disease history only. The authors take the most frequent allele in the disease sample to be the marker allele on the ancestral disease haplotype. They also assume that the frequency of this allele is the proportion of nonrecombinant haplotypes in the sample. However, even when the most frequent marker allele is ancestral, the largest allelic class may contain recombinants, particularly when the ancestral allele is common. Later papers (Lehesjoki et al. 1993) extend the method to account for such recombinants, but it is still necessary to assume that the most frequent marker allele is ancestral. Another drawback is that the estimator does not use the information available in the demographic model about relationships among sampled disease alleles. This relatedness affects the variance of the estimated recombination fraction, and is reflected in the higher moments of the number of nonrecombinants, and hence in the higher moments of marker allele counts in the disease sample. In general, for a disease of fixed age in the population, the more related the disease sample, the more variable the number of nonrecombinants, and hence the more variable the estimated recombination fraction. Finally, the Luria-Delbrück bounds on the accuracy of the estimate are not statistically justified.

To overcome these difficulties, Kaplan et al. (1995) took a likelihood-based approach to the problem, modelling the evolution of the disease population as a Poisson branching process. Like Hästbacka et al. (1992), they assumed that the most frequent marker allele in the disease sample was ancestral when realizing the disease genealogy. The present disease copy number was taken into account by conditioning on a plausible range of values. A likelihood for recombination rates was evaluated by Monte Carlo methods, summing over realizations of the branching process consistent with this range. In this method, the branching process is used to model disease genealogy and the resulting marker allele frequencies within the disease population. These marker allele frequencies are random, and arise as a consequence of disease evolution; they are determined by the rate of recombination, the number of meioses (segregations) relating the current disease population, and the population marker al-

lele frequencies. Given the marker allele frequencies in the disease population, the marker allele counts in sampled disease haplotypes are multinomial. Xiong and Guo (1997) worked with the same multinomial likelihood formulation, re-expressed as a second order Taylor series about the expected frequencies in the disease population. They too assumed that the most frequent marker allele in the sample was ancestral. Their approximate likelihood requires knowledge of the mean and variance of the allele frequencies under an evolutionary model. A Wright-Fisher model of allele reproduction (Crow and Kimura 1970) was used to calculate these moments, thus avoiding Monte Carlo evaluation. However, unlike Kaplan et al., Xiong and Guo did not calculate moments conditional on present disease copy number. Both Xiong and Guo and Kaplan et al. formulate likelihoods in terms of marker allele frequencies in the disease population, rather than marker identity by descent in the disease sample. Thus, both methods can be classified as identity by state methods, since marker allele frequencies reflect marker identity by state.

Rannala and Slatkin (1998) adopted a similar approach which is also based on marker identity by state. However, they modelled the ancestry of a disease sample, rather than the ancestry of the entire disease population. These authors did not assume that the most frequent marker allele in the disease sample was ancestral. They used the ancestry of sampled disease alleles to obtain Monte Carlo realizations of past marker allele frequencies in ancestors. Their model of the ancestry looked backwards in time from the present rather than forwards from founding. This retrospective view allowed conditioning on present disease copy number, and so rejection sampling was unnecessary.

All of these approaches are maximum likelihood methods, although Xiong and Guo (1997) use an approximate rather than an exact likelihood formulation. All build likelihoods based on allele frequencies, whether in the disease population (Kaplan et al. 1995, Xiong and Guo 1997), or in ancestors of a disease sample (Rannala and Slatkin 1998). As such, they are based on marker identity by state. By contrast, the

method proposed in this dissertation bases the likelihood explicitly on marker identity by descent. Identity by descent unifies all linkage analysis, and such a perspective highlights how disequilibrium mapping is a natural population-level counterpart to traditional linkage methods for family data.

1.4 Fine-Scale mapping examples

Throughout the dissertation, we illustrate ideas using simulated data constructed to be typical of fine-scale mapping of rare diseases in isolated populations. Three diseases in three different populations provide the motivation. The first disease is infantile-onset spinacerebellar ataxia (IOSCA) in the Finnish population (Nikali et al. 1995). The second disease is Werner's syndrome in Japanese, which has recently been mapped and positionally cloned (Yu et al. 1996). The third disease is benign recurrent intrahepatic cholestasis (BRIC) in an isolated Netherlands fishing community. The simulated data for the BRIC-like disease is used to illustrate the coarser scale of mapping in very young populations. Houwen et al. (1994) used disequilibrium linkage detection methods based on haplotype sharing in a remote Netherlands fishing community to localize the BRIC gene to chromosome 18.

1.4.1 A Finnish variant

Infantile-onset spinocerebellar ataxia (IOSCA) is a rare, progressive, autosomal-recessive neurological disorder which has been reported in only 20 Finnish patients (Nikali et al. 1995). Mutations for several rare recessive diseases such as IOSCA are enriched in the Finnish population, suggesting the presence of disease-predisposing mutations in single members of a small number of founding individuals. This founder effect allowed Nikali et al. (1995) to apply disequilibrium fine-mapping to IOSCA, after localizing the disease mutation to a 4 cM genomic region with pedigree data. (1 cM is approximately equivalent to a recombination fraction of 0.01.)

The Finnish population is currently of size 5×10^6 people, or 10^7 haplotypes (Hästbacka et al. 1992). The ancestors of this population are thought to have immigrated to the southwest of the country some 80 generation before present (gbp) or 2000 years ago. We have assumed 1000 founding individuals, or 2000 haplotypes, in our simulations. Due to geographical and linguistic barriers, the population subsequently remained genetically isolated, with relatively little immigration (Nevanlinna 1972).

1.4.2 A Japanese variant

We introduce an example typical of fine-scale mapping of a disease allele in the Japanese population: this example is motivated by the recent mapping and positional cloning of the Werner's syndrome gene in Japanese (Yu et al. 1996), previously localized by pedigree studies to a 7.3 cM genomic region (Nakura et al. 1993). Werner's syndrome is a rare autosomal recessive disease characterized by premature onset of a number of age-related traits. Although it is now known that there are at least eight distinct mutations at the Werner's syndrome locus in present-day Japanese, we consider a single nonrecurrent mutation such as WRN4 in our simulations. WRN4 is the most frequent of the eight and represents about 51% of the mutants (Matsumoto et al. 1997). In the Japanese, the estimated allele frequency of all Werner syndrome mutations combined is between 0.002 and 0.004 (Goddard et al. 1996). The higher estimated frequency of 0.004 and a current Japanese population size of at least 120 million people or 240,000,000 alleles at a locus (JIN 1998, ISEI 1998) gives a WRN4 copy number of about 500,000.

The Japanese population has a well documented recent history, and data on population sizes are available. (Koyama 1978, ISEI 1998). The Japanese archipelago assumed its present shape around 400 gbp (10000 ybp). Soon afterwards, the era known as the Jomon period began, continuing for about 306 generations until the introduction of the Yayoi wet rice culture from the Eurasian continent around 94 gbp

(2350 ybp). The hunting and gathering people of the Jomon period lived in stationary communities and attained a high level of culture, particularly in the deciduous forests of eastern Japan. The coniferous-forested area of western Japan, by contrast, was thinly populated. Detailed archaeological analysis of Jomon settlements has led Koyama (1978) to conclude that the population of Jomon Japan reached up to 262,500 individuals at its height about 180 gbp. This population subsequently declined, as a result of climatic cooling to about 161,000 persons by the time of the Yayoi immigration (Koyama 1978). About 94 gbp, small numbers of rice-growing Yayoi immigrants began arriving in the southwestern island of Kyushu from the Eurasian continent. The warmer and wetter climate in this less populated western region was well suited to rice farming, and the technologically advanced Yayoi culture soon spread east and north. There is much debate about the extent to which the two populations mixed (e.g., Rose 1996, Hanihara 1991). A replacement theory hypothesizes that the Yayoi newcomers displaced the Jomon people to the extreme north and south of Japan and that there was little admixture, whereas a transformation theory hypothesizes that the modern Japanese evolved gradually from the Jomon, with the newcomers having little genetic impact. Finally, hybridization theories hypothesize that the newcomers mixed (to varying degrees) with the indigenous people. Under the replacement hypothesis, a reasonable first approximation is that the modern Japanese were founded about 94 gbp by a small number, say 1000, of Yayoi immigrants. Under the transformation hypothesis, however, the genetic foundations of the modern Japanese would have been laid much earlier, about 400 gbp, by a small number, say 1000, of Jomon or pre-Jomon individuals.

Whatever the origins of the modern Japanese, much later in their history, about 16 gbp, an era known as the Tokugawa or Edo period began and continued for roughly 11 generations until the Meiji reform in 1867. The Tokugawa shoguns controlled the nation through a strict feudal system. As a result, population size remained roughly constant (Benedict 1989) at about 30 million persons (ISEI 1998), in spite of

eleven generations of peace. Throughout most of this period, the Tokugawa shogunate pursued a policy of almost total seclusion from the outside world. Following the Meiji reform, however, the feudal system was abolished and Japan underwent a period of rapid transformation and population growth which continues today (Benedict 1989, ISEI 1998).

1.4.3 A community variant

Houwen et al. (1994) detected linkage between chromosome 18 and benign recurrent intrahepatic cholestasis (BRIC), a rare autosomal recessive disease, by searching for shared genomic segments among three affected individuals from an isolated Netherlands fishing community. Until recently, this community was an endogamous population, experiencing dramatic growth during the 19th and early 20th centuries. It was not until after this growth period that exogamous marriages became more common. Thus, most individuals descend from a common ancestral pool dating back to the 17th century. In spite of this shared ancestry, no single ancestor could be identified as the possible source of the six disease alleles in the study.

In our simulations, we assume a 17th century founding population of approximately 250 individuals or 500 alleles at a locus, with a single copy of the disease mutation. The present population size is assumed to be approximately 10000 persons or 20000 alleles at a locus, and the present disease copy number is 200. Under random mating, the present number of disease alleles implies that there are about 0-3 affected individuals, in accordance with the number observed in the community.

Chapter 2

THE DISEASE COALESCENT

When a disease mutation arises, it does so on a background haplotype, and is in complete association with the marker alleles of this haplotype. Subsequent recombination events occurring throughout the history of the mutation result in decay of this association. The rate of decay is determined by the recombination fraction, and the amount by the number of meioses (opportunities for recombination) on the ancestry of the current sample of disease alleles (Arnason et al. 1977, Thompson 1978). In the absence of "spurious" association due to population stratification or natural selection, association between the disease mutation and alleles of linked markers is a consequence of shared identity by descent (Thompson and Neel 1997), which may be traced through the ancestry of the sample at the disease locus. On this ancestry, recombination events between the disease locus and linked markers occur at the level of individual meioses (segregations). The ancestry itself is a realization of a random process. This ancestral process is the population-genetic process of disease allele reproduction viewed retrospectively. Kingman (1982a) developed a related model for the ancestry of a random sample of alleles from a random-mating population. In this chapter, we extend this coalescent theory to a random sample of disease alleles, descended from the same ancestral mutation. This disease coalescent conditions on past disease copy numbers, and is not restricted to a rare disease. However, disease copy numbers are realized under a rare disease assumption.

The remainder of this chapter is organized as follows. In section 2.1, we review Kingman's coalescent in a population of constant size, reproducing according to a continuous-time Moran model (Moran 1962). Section 2.2 describes a standard exten-

sion to populations growing deterministically at constant exponential rate (Kingman 1982b, Slatkin and Hudson 1991). In section 2.3, we develop the coalescent for a random sample of descendants of a disease mutation, conditional on past disease copy numbers. Just as coalescent rates for a population random sample depend on past copy numbers in the population, coalescent rates for a disease sample depend on past disease copy numbers. Section 2.4 describes in detail the stochastic process underlying disease copy numbers. For a rare disease, present as a single copy in the population a known time ago, copy numbers follow very closely a birth-and-death process. Section 2.5 investigates basic properties of this birth-and-death process, including its conditional probability generating function given the current disease copy number. Given a single copy of a rare disease mutation a known time ago, and the current disease allele count, the mean, variance, and higher moments of past numbers of disease copies may thus be computed. These moments allow approximation, to arbitrary precision, of the conditional distribution of past copy numbers by a distribution matched on an appropriate number of moments. Past disease copy numbers can then be realized via this approximating distribution. It is these copy numbers which determine coalescent rates for the disease sample. Putting these steps together, in section 2.6, we describe a way to realize coalescent times for a random sample of disease alleles. We then compare the distributions of these times under stochastic and determistic growth of past disease copy numbers. Next, in section 2.7, we derive an alternate version of the disease coalescent by viewing the disease ancestry as a subtree of the ancestry of the total population. We arrive at this version by conditioning on past numbers of ancestors of the disease and the total population, rather than on past disease copy numbers. Finally, in section 2.8, we examine the assumption of a single disease copy at population founding, given current disease copy number and population demographic information. This is accomplished by studying the age distribution of a selectively-neutral mutation, using subtree coalescent methods proposed by Griffiths and Tavaré (1998). Two diseases are considered: infantile-onset

spinacerebellar ataxia in Finns and Werner's syndrome in Japanese.

2.1 A Moran model

Consider first the history of a random sample of neutral alleles taken from the population. Random mating implies random mixing of alleles in diploid individuals. Thus, under random mating, lineages may be traced back through alleles rather than through individuals, without loss of generality. Stochastic descriptions of the underlying ancestry are based on retrospective analysis (Felsenstein 1971) of population-genetic models of allele reproduction (Crow and Kimura 1970), extended to account for populations of changing size. In this research, we model allele reproduction in continuous-time, where time is measured in "generations", which we take to be 25 years in simulations. Alleles reproduce when a single lineage splits according to a continuous-time Moran model (Moran 1962). When this reproductive process is viewed retrospectively, the topology of the ancestry of a random sample of K alleles from the current population must therefore be a binary tree, with K tips and $K-1$ vertices corresponding to times when ancestral lines coalesce (reviewed in Felsenstein 1996, Griffiths and Tavaré 1994, Hudson 1990). The ancestral coalescent thus describes both the randomness in ancestral relationships and the randomness in times at which coalescent events occur.

2.1.1 Moran reproduction

Moran (1962) introduced a model of allele reproduction for populations of constant size in which, at each successive time point, one copy is randomly chosen from the population to give birth, and one is randomly chosen to die. In our Moran model, the reproducing copy is also eligible to die. By definition, birth and death events occur at the same (Moran event) rate, and random sampling of copies for birth and for death events is the same. Measure age in terms of Moran events. Then, given that a Moran

event occurs, the age of the copy which is randomly selected to give birth at this time has the same distribution as the age of the population. To get the age distribution of a population of size N, we argue as follows. The age of a randomly selected copy is the length of time from the present back to when the copy was born. This waiting time is the number of Moran events, looking backwards in time, until the birth, and hence is geometric with mean N. Looking back in time, each extant copy has equal probability $1/N$ of being the copy that was born at the preceding Moran event. The same reasoning, applied forwards in time with death events, implies that the lifetime distribution of individual copies is geometric with mean N. An interesting consequence of this model is that the lifetime of a random parent is stochastically larger than the lifetime of a random copy. (Copies are randomly sampled to become parents, and so the more general implication is that the lifetime of a randomly sampled copy is stochastically larger than the lifetime of a random copy.) As noted above, the age distribution of a reproducing copy is geometric with mean N. After reproduction, the independent additional lifetime of this parent is also geometric with mean N.

To gain insight into the rates at which lineages coalesce, or come together in a common ancestor, we may compare inbreeding in this Moran population to inbreeding in a standard Wright-Fisher population (Crow and Kimura 1970), in which copies from the offspring generation are obtained by sampling with replacement from the parent generation. Regardless of the reproductive model, in a closed population of finite size, the level of inbreeding relative to the base population at time 0 increases with time. Let h_t be the chance that, at t time units after population founding, a random pair of copies descends from two distinct founders. Then h_t decreases by the ratio h_{t+1}/h_t per time unit. If this ratio is the same per generation in two populations, their "inbreeding effective sizes" (Crow and Kimura 1970) or rates of inbreeding per generation are said to be the same. Inbreeding effective size N_e is traditionally taken to be the size of the equivalent Wright-Fisher population, in terms of the rate of

inbreeding. For the Moran population,

$$h_{t+1} = (1 - 1/N)\,\frac{\binom{N}{2} - 1}{\binom{N}{2}}\,h_t + 1/N\,h_t = h_t\,(1 - 2/N^2).$$

To obtain h_{t+1} in the Moran population, we first condition on the event that a different allele is selected to die than to give birth at the last Moran event: this happens with probability $(1 - 1/N)$. Given that different alleles are selected to die and give birth, one of the $\binom{N}{2}$ pairs in the current population must be duplicated, and this duplicated pair cannot be the sampled pair since then the sampled pair would be identical by descent. Given that the sampled pair has 2 distinct parents, the chance that these parents are themselves not identical by descent is h_t. On the other hand, if the same copy is chosen to die and give birth at the last Moran event, the sampled pair must have two distinct parents and these must not be identical by descent, yielding the second term.

In order to compare per-generation inbreeding rates, we need to choose the number of Moran events that will represent a generation. One natural candidate is the expected age of a random parent when it gives birth. In a Moran population, this expected age is N Moran events. Equating N Moran events to a generation leads to a one-generation inbreeding ratio of

$$(1 - 2/N^2)^N \approx e^{-2/N} \approx (1 - 2/N),$$

which is the same as the ratio for a Wright-Fisher population of half the size, provided N is large. Thus, provided N is large, the inbreeding effective size of such a Moran population is approximately $N_e = N/2$.

The effective population size tells us that when generations are scaled as N Moran events the rate of inbreeding is faster than in a Wright-Fisher population of the same size N. This is a consequence of the larger variance in the number of progeny for an allelic copy in the Moran model. In a Wright-Fisher model, the progeny number is binomial with N trials and probability $1/N$, and so has variance $1 - 1/N$. By

contrast, allelic copies in the Moran model have extra-binomial variance in progeny number which is twice that of copies in the Wright-Fisher model. To see this, note that the lifetime of an allelic copy, measured in Moran events, has a geometric distribution with mean N and variance is $N(N-1)$. Given this lifetime L, the progeny number is binomial with L trials and probability $1/N$, since copies may give birth before dying. The conditional mean and variance of the progeny given L are thus respectively L/N and $L(N-1)/N^2$. As a result, the unconditional variance in progeny is $2(1-1/N)$. Note that in most natural populations, extra-binomial variance in progeny is observed, owing to a few copies with relatively large numbers of offspring (Crow and Kimura 1970).

In the above development, generations are scaled as N Moran events. This scaling leads to a faster per-generation rate of inbreeding in the Moran population than in a Wright-Fisher population of the same size. The majority of recently-developed methods for fine-scale mapping are based on the Wright-Fisher model or approximations thereof. In order to eliminate any differences between our method and these others due to different rates of inbreeding in different reproductive models, we have chosen to scale generations as $N/2$ Moran events. This alternate scaling leads to a one-generation inbreeding ratio of

$$(1-2/N^2)^{N/2} \approx e^{-1/N} \approx (1-1/N),$$

the same as for a Wright-Fisher population of the same size N. Under the $N/2$ scaling, the expected age of a copy at the time of birth is N Moran events or 2 generations. The rate of meioses per copy per generation is therefore one-half. However, throughout this dissertation, we have set meioses to occur at rate one per generation per copy in order to match the rate in a Wright-Fisher population of the same size.

In order to obtain a more convenient continuous-time analog to the discrete-time Moran model, we allow Moran events to occur randomly with rate $N/2$ per generation. We wish to characterize the ancestry of a random sample of K allele copies

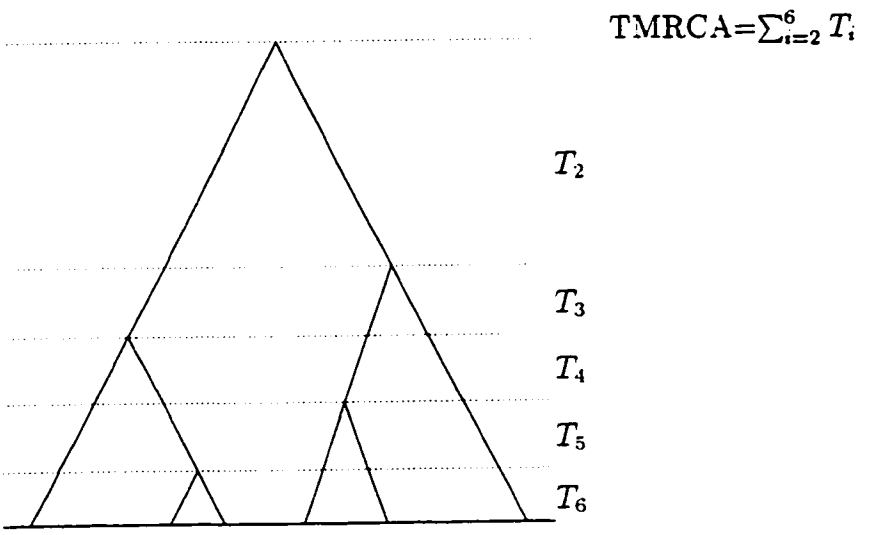


Figure 2.1: Example coalescent and notation for coalescent times. The tips of the tree represent the $K = 6$ sampled individuals at present.

from this population. To do so requires a retrospective view of the reproductive process. Ancestral lineages come together or *coalesce* in common ancestors. Coalescent events among $k = 2 \ldots K$ random lineages are random, and involve only a pair of lineages, since the instantaneous probability of more than one coalescence is negligible in this continuous-time model. Coalescences occur with rate $k(k-1)/N^2$ per Moran unit (Felsenstein 1971), or equivalently with rate $k(k-1)/(2N)$ per generation under the current scaling of a generation as $N/2$ Moran events. The resulting times T_k, $k = 2 \ldots K$, in generations, during which a sample of size K from the current population has k ancestors are then independent and exponentially distributed with rate $k(k-1)/(2N)$ (Kingman 1982a). This rate is proportional to the number of lineage pairs, and coincides with the coalescent for a Wright-Fisher population of the same size (Kingman 1982a). Figure 2.1 illustrates the coalescent and notation for $K = 6$ sampled alleles.

2.2 *Deterministic population growth*

We next consider the coalescent for a random sample of copies drawn from a population growing deterministically at exponential rate (Kingman 1982b). Models which accommodate growth are realistic given the recent expansion of most extant human populations since the advent of agriculture (Thompson and Neel 1997, Thompson and Neel 1996, Cavalli-Sforza et al. 1994). Coalescent times in a growing population may be obtained by generating under a constant-sized population, and then rescaling to account for size fluctuations, as described by Griffiths and Tavaré (1994), or by arguing directly, as described below, or in Slatkin and Hudson (1991).

For example, suppose that K copies have been randomly selected from the present population, which has been growing exponentially at constant rate $m = e^{\lambda}$ per generation. We seek the distribution of the time to the most recent coalescent event T_K, or, more generally, the time during which there are K lineages in the ancestry of the sample. We have that the number of copies in the population at time t generations before present (gbp) is $N_P(t) = N_P(0)/m^t$. Let $Q(t) = \mathrm{P}\{T_K > t\}$. We stress here that time is being measured backwards from the present, which has been set to $t = 0$. Over a small time interval $(t, t+h]$,

$$\begin{aligned} Q(t+h) &= Q(t)\mathrm{P}\{T_K > t+h \mid T_K > t\} \\ &= Q(t)[1 - \mathrm{P}\{T_K \in (t, t+h] \mid T_K > t\}] \\ &= Q(t)\left[1 - \frac{K(K-1)}{2N_P(t)}h\right]. \end{aligned}$$

and

$$\frac{dQ}{dt} = \lim_{h\to 0}\frac{Q(t+h) - Q(t)}{h} = -Q(t)\frac{K(K-1)}{2N_P(t)}.$$

Dividing through by $Q(t)$ and integrating over $(0, s)$ gives

$$\begin{aligned} \log_e Q(t)\,|_0^s &= -\frac{K(K-1)}{2}\int_0^s \frac{dt}{N_P(t)} \\ &= -\frac{K(K-1)}{2N_P(0)}\int_0^s m^t dt \end{aligned}$$

That is,

$$\log_e Q(s) = -\frac{K(K-1)}{2N_P(0)}\frac{1}{\lambda}(m^s - 1).$$

or

$$Q(s) = \exp\left\{-\frac{K(K-1)}{2N_P(0)}\frac{1}{\lambda}(m^s - 1)\right\}. \tag{2.1}$$

Figure 2.2 plots the distribution $1 - Q$ of T_K for growth rates $m \in [1, 2]$ when $N_P(0) = 10000$ and $K = 10$. An upper bound for T_K is provided by $\log_e N_P(0)/\lambda$, the number of generations required for a single copy to grow exponentially to $N_P(0)$. Even for growth rates as small as $m = 1.04$, the distributions are markedly different from the distribution of a constant-sized population, with lineages tending to coalesce much faster, as indicated by the leftwards shift relative to the curve for $m = 1$.

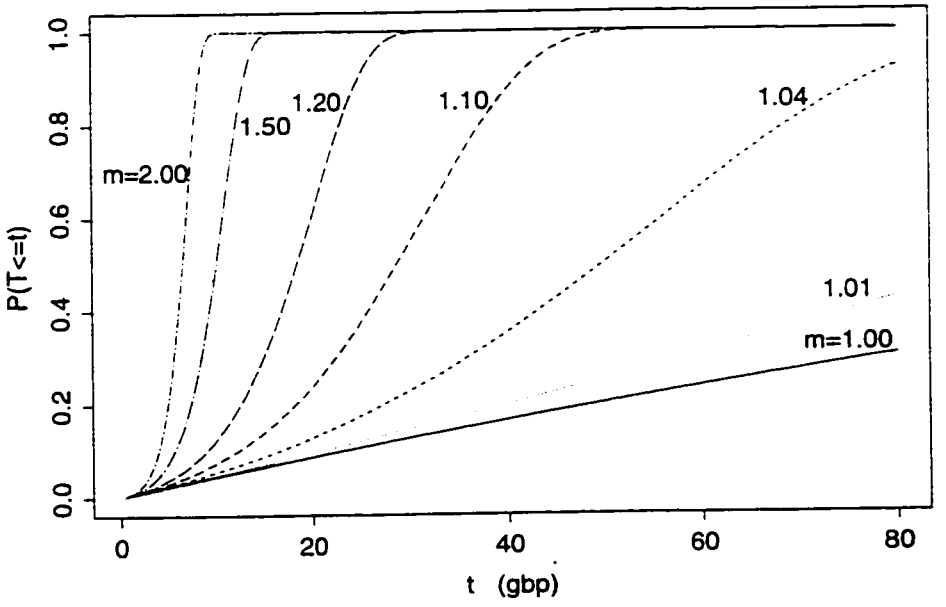


Figure 2.2: Distributions $\mathrm{P}(T_K \leq t) = 1 - Q(t)$ of the most recent coalescent time T_K, under deterministic rates $1 \leq m \leq 2$ of population growth. There are $K = 10$ randomly sampled alleles from a population of present size $N_P(0) = 10000$.

Moments for coalescence times provide further insight into the effect of population growth on the ancestry. These involve the exponential integral (Press et al. 1992), and must be evaluated numerically or through simulation. As expected, the mean and variance of the time to the most recent common ancestor TMRCA= $\sum_2^K T_i$ of the sample decreases with increasing growth rates: Figure 2.3 displays results from our simulations for $N_P(0) = 10000, K = 10$. Means and variances of the individual coalescence times T_i, $i = K \ldots 2$, also decrease (results not shown). Figure 2.4 shows that the coefficient of variation (sd/mean) for TMRCA is ~ 1 for growth rates near $m = 1$, but drops quickly and levels off as m increases. Complete coalescence is forced by $t = \log_e(N_P(0))/log_e(m)$ gbp, when only a single allelic copy remains. As the exponential growth rate m increases away from 1, this time limit becomes more constraining. Thus, the mean TMRCA decreases, and its variability decreases even faster.

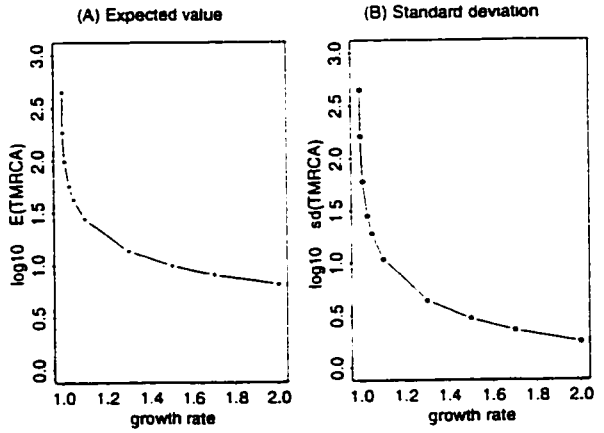


Figure 2.3: (A) Expected value, and (B) standard deviation of the time TMRCA to the most recent common ancestor of the sample, as a function of the rate m of population growth. Vertical axis is in the $\log_{10}$ scale. There are $K = 10$ randomly sampled alleles from a population of present size $N_P(0) = 10000$. Results are based on 1000 coalescent replicates.

The shape of the ancestral tree is also affected by increasing m. In particular, the time T_2 during which there are 2 sample ancestors decreases as proportion of TMRCA.

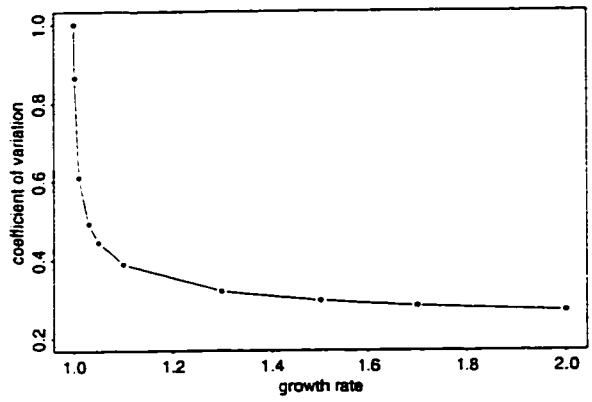


Figure 2.4: Coefficient of variation (sd/mean) of the time TMRCA to most recent common ancestor of the sample, as a function of the rate m of population growth. There are $K = 10$ randomly sampled copies from a population of present size $N_P(0) = 10000$. Results are based on 1000 coalescent replicates.

In constant-size populations, ET_2 is about half ETMRCA, but this proportion is reduced as growth rate increases, leading to more "star-shaped" trees. Figure 2.5 plots expected branch lengths, ET_i, $i = 2 \ldots K$, relative to ETMRCA for different growth rates. Most variation with growth occurs in T_K and T_2, nearest the tips and root of the ancestral tree, respectively. Although ancestral trees in growing populations are more star-shaped, they are still a long way from a complete "star", in which there is no overlap among lines of descent. It is thus useful to account for dependence among copies due to common ancestry, even in growing populations.

One final difference between constant-sized and growing populations is the dependence structure among coalescent times. Let $S_{j+1} = \sum_{j+1}^{K} T_i$ be the time (in gbp) when the ancestral process drops from $j+1$ to j ancestors, or, alternatively, the time it takes for the K sampled lineages to coalesce to $1 \leq j \leq K$ lineages. Then, given $S_{k+1} = t$, the time T_k during which there are k ancestral lineages, depends on the population size $N_P(0)/m^t$ at t, and hence on t, except when $m = 1$. The conditional distribution of T_k given $S_{k+1} = t$ for $2 \leq k \leq K$ is in fact obtained from equation (2.1)

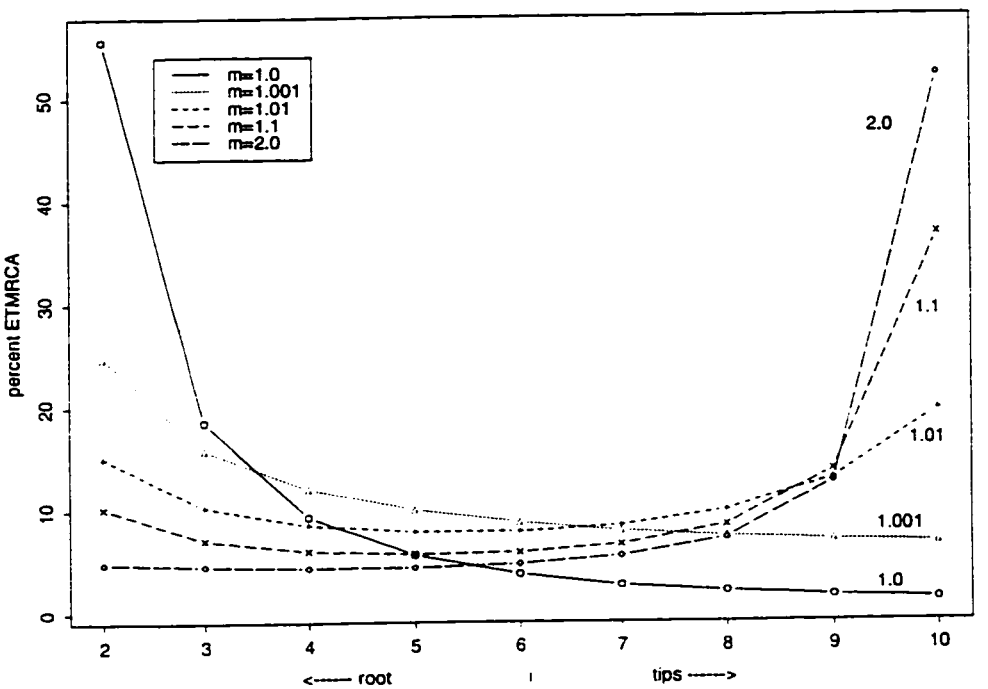


Figure 2.5: Expected coalescent times E T_i as a percentage of the expected time ETMRCA=$\sum_2^K$ ET_i to the most recent common ancestor of the sample under different population growth rates m. There are $K = 10$ randomly sampled alleles from a population of present size $N_P(0) = 10000$.

by substituting $N_P(t)$ for $N_P(0)$ and k for K:

$$\mathrm{P}\{T_k > s \mid S_{k+1} = t\} = \exp\left\{-\frac{k(k-1)}{2N_P(t)}\frac{1}{\lambda}(m^s - 1)\right\}. \tag{2.2}$$

Coalescent times in a growing population, unlike those in a constant-size population, are therefore dependent, with each T_k depending on more recent $T_{k+1}, T_{k+2}, \ldots, T_K$ through the sum S_{k+1}. As a consequence, each T_k is equally negatively correlated with $T_{k+1}, \ldots T_K$. The strength of this negative correlation increases with the rate of growth, as constraints imposed by past population size become more restrictive. As the size of the population decreases, so does the variability in coalescent times.

Historically, the coalescent has been used to model the ancestry of a random sample of alleles from the population. For disequilibrium mapping, however, interest focuses on the ancestry of a sample from a *disease subpopulation*, the current descendants of the disease mutation. As a naive first approximation, we may assume that this subpopulation grows deterministically at the appropriate exponential rate. For example, in the Finns, a population founded about $t_f = 80$ gbp, the appropriate m under exponential growth of the IOSCA disease mutation would imply $10000 = N_D(0) = N_D(t_f) \times m^{t_f}$, where $N_D(t)$ denotes the disease copy number at t gbp. If the disease mutation is further assumed to be present on a single allele at founding so that $N_D(t_f) = 1$, we have $m \approx 1.12$. However, even when the total population has grown exponentially, this approximation should be applied with caution. First, as shown by Thompson and Neel (1997), given survival, initial growth of the disease variant is expected to be faster than exponential. Second, disease growth is not smooth, especially in initial stages, due to its stochastic nature. In the next section, we avoid these problems by developing a version of the coalescent specifically for a random sample from a disease subpopulation.

2.3 Coalescent rates

In this section, we develop the coalescent for a random sample from the descendants of a disease mutation. These present descendants define the disease variant subpopulation, which is embedded within the total population. We therefore extend the standard coalescent theory from a population to a disease random sample. The overall historical pattern of population growth is assumed known, including the number of copies at population founding. We model the total population by a continuous-time Moran process (Moran 1962) with additional birth events. (For a different approach, see Slatkin and Rannala (1997), which uses a birth-and-death process as a model of population reproduction.) Population growth is modelled by increasing the birth rate rather than reducing the death rate. This choice is based on the supposition that family sizes increased with the advent of agriculture, and the accompanying lifestyle changes from nomadic hunter-gatherers to agriculturalists. Additional births are parameterized by $\lambda(t)$, the instantaneous rate of population increase at t gbp. Populations of shrinking size may be similarly accommodated by introducing additional death events, although these are not discussed here. The rate of additional births (or deaths) does not have to be constant over time. Increasing the rate of additional births decreases parental age and hence the expected length of time per meiosis within the population. As the overall birth rate increases, so does the proportion of young in the population, resulting in decreased parental age. Birth (but not death) rates therefore determine the age distribution. By constrast, deterministic population growth maintains a constant age distribution, regardless of the growth rate. Conceptually, a copy introduced by deterministic growth has an age which is drawn from the age distribution of the overall population. Hence, the expected length of time per meiosis is unchanged. Kingman (1982b) made the simplifying assumption of deterministic growth when deriving the rates of coalescence for a population random sample.

We first investigate properties of the Moran + birth reproductive model, without assuming deterministic growth. The coalescence rate for a population random sample and the inbreeding effective size are derived. Following this, we develop the coalescence rates for a disease sample, conditional on the size of the disease subpopulation.

To obtain the coalescence rate for a population random sample of size k we condition on the population size $N_P(t)$ at t gbp, $0 \leq t \leq t_f$, and then take expectations. Conditional on $N_P(t)$, the instantaneous rate of coalescence due to Moran events is

$$\frac{N_P(t)-1}{N_P(t)} \times \frac{\binom{k}{2}}{\binom{N_P(t)}{2}} \times \frac{N_P(t)}{2}.$$

The first factor is the chance that a different allele is chosen to die than to give birth, and the second is the chance that the pair of lineages for this reproducing allele and its offspring is among the lineages of the sample. Together, they comprise the probability that a Moran event at t leads to a coalescence. The third factor is the Moran event rate at t gbp. After simplifying, this yields an instantaneous conditional rate of

$$\frac{k(k-1)}{2N_P(t)}$$

per generation. On the other hand, the conditional coalescence rate due to pure birth events is

$$\frac{\binom{k}{2}}{\binom{N_P(t)+1}{2}} \times N_P(t)\lambda(t).$$

The first term is the chance that the lineages represented by the reproducing allele and its offspring are among those of the sample. Given that an allele reproduces at t gbp, the number of copies in the population immediately following is $N_P(t)+1$. The second term describes the instantaneous rate of pure birth events. After simplifying, this yields a conditional coalescence rate of

$$\frac{k(k-1)}{N_P(t)+1} \times \lambda(t)$$

attributable to pure births. Adding these rates yields the overall conditional rate of coalescence. The rates can be added because pure birth and Moran events occur

independently given $N_P(t)$. Taking expectations, the unconditional coalescence rate is therefore

$$\begin{aligned}\rho_k(t) &= k(k-1)\mathrm{E}\left[\frac{1}{2N_P(t)}+\frac{\lambda(t)}{N_P(t)+1}\right] \\ &\approx k(k-1)\mathrm{E}\left(\frac{1}{N_P(t)}\right)\left[\frac{1}{2}+\lambda(t)\right] \\ &\approx k(k-1)\frac{1}{\mathrm{E}\,N_P(t)}\left[\frac{1}{2}+\lambda(t)\right]. \qquad (2.3)\end{aligned}$$

provided $N_P(t)$, $0 \le t \le t_f$, is reasonably large. In equation (2.3), we use the approximation $\mathrm{E}\,1/N_P(t) \approx 1/\mathrm{E}\,N_P(t)$. This approximation is appropriate for growing populations, founded by a reasonable number of copies. For example, in a growing population founded by at least 100 copies, it is highly improbable that $N_P(t)$ will drop below 10 over the population history. For $N_P(t) > 10$, $1/N_P(t)$ has little curvature, and so $\mathrm{E}\,1/N_P(t) \approx 1/\mathrm{E}\,N_P(t)$. For the Moran+birth process with instantaneous pure birth rate $\lambda(t)$, $0 \le t \le t_f$,

$$\mathrm{E}\,N_P(t) = N_P(t_f) \times \exp(\int_t^{t_f} \lambda(s)\,ds)$$

(Kendall 1948). Thus, when growth is constant and $\lambda(t) \equiv \lambda$,

$$\mathrm{E}\,N_P(t) = N_P(t_f)\exp\left(\lambda \times (t_f - t)\right).$$

For instance, at $t_f = 80$ gbp, we have taken the number of founding copies in Finns to be $N_P(t_f) = 2000$; the current copy number is 10^7. Assuming the population grows at constant exponential rate, $\lambda = \log_e(10^7/2000)/80 = 0.11$. Hence, $\mathrm{E}\,N_P(t) = 2000 \times \exp\left(0.11 \times (80 - t)\right)$.

Inbreeding effective size provides an established means to compare rates of inbreeding among reproductive models. For completeness of exposition, we give formulae for effective sizes of the Moran + birth population. These formulae are derived from probabilities of identity by descent, following arguments in Crow and Kimura (1970) and Felsenstein (1995). Let h_s be the chance that, at s generations after founding,

or $t_f - s$ gbp, a random pair of alleles descends from two distinct founders. The one-generation inbreeding ratio from t_f to $t_f - 1$ is $h_1/h_0 = h_1$, and so the inbreeding effective size at $t_f - 1$ is $N_e(t_f - 1) = 1/(1 - h_1)$. In general, we define the inbreeding effective size at i gbp, $i = 1, \ldots, t_f - 1$, to be

$$N_e(t_f - i) = 1/\left(1 - h_i/h_{i-1}\right).$$

This is the size of a Wright-Fisher population with approximately the same rate of inbreeding as the Moran + birth population over the one-generation interval from $t_f - i - 1$ to $t_f - i$ gbp. Arguments similar to those in section (2.2) give

$$h_s = \mathrm{P}_{t_f - s}(T_2 > s) = \exp\left(-\int_{t_f - s}^{t_f} \rho_2(t)\, dt\right).$$

where $\rho_2(t)$ is given by equation (2.3), and the subscripted probability indicates that the coalescent process starts at $t_f - s$ gbp. Thus,

$$\frac{h_i}{h_{i-1}} = \exp\left(-\int_{t_f - i}^{t_f - i + 1} \rho_2(t)\, dt\right).$$

and the one-generation effective size at i gbp simplifies to

$$N_e(t_f - i) = 1/\left[1 - \exp\left(-\int_{t_f - i}^{t_f - i + 1} \rho_2(t)\, dt\right)\right].$$

For example, the inbreeding effective size for Finns over the one-generation interval immediately after founding is

$$N_e(79) = 1/\left[1 - \exp\left(-\int_{79}^{80} \rho_2(t)\, dt\right)\right] \approx 1739,$$

somewhat smaller than the founding copy number of 2000.

Finally, we develop the rate of coalescence for a random sample of disease alleles. Let $N_D(t)$ be the size of the disease variant subpopulation at t gbp, $0 \le t \le t_f$, and suppose that the present copy number $N_D(0)$ is known. Let $k(t)$ be the number of ancestors at t gbp of $k(0) = K$ random disease alleles sampled at present. Even if $K = N_D(0)$, $k(t) \le N_D(t)$ since some disease lineages at t gbp may go extinct by the

present. At t gbp, the instantaneous rate of Moran events is $N_P(t)/2$ per generation, and the pure birth rate is $N_P(t)\lambda(t)$. In the disease subpopulation, the pure birth rate at t gbp is $N_D(t)\lambda(t)$. Lineages split via Moran or pure births, and coalescences occur when a splitting lineage and its offspring are among the lineages of the disease sample. Given $k(t)$, $N_D(t)$, and $N_P(t)$, Moran coalescences occur with instantaneous rate

$$\frac{N_P(t)-1}{N_P(t)} \times \frac{N_D(t)}{N_P(t)} \times \frac{\binom{k(t)}{2}}{\binom{N_D(t)}{2}} \times \frac{N_P(t)}{2}. \tag{2.4}$$

The first three factors comprise the probability that a Moran event at t leads to a coalescence. The first of these is the chance that a different allele is chosen to die than to give birth, the second the chance that the one giving birth is from the disease subpopulation, and the third the chance that the lineage pair for this reproducing allele and its offspring is among the sample lineages. The fourth factor is the instantaneous Moran event rate at t gbp. After simplifying, this yields an instantaneous rate of

$$\frac{k(t)(k(t)-1)}{2N_P(t)} \times \frac{N_P(t)-1}{N_D(t)-1} \approx \frac{k(t)(k(t)-1)}{2(N_D(t)-1)} \tag{2.5}$$

per generation, provided $N_P(t)$ is of reasonable size. Coalescences due to additional births occur with instantaneous rate

$$\frac{\binom{k(t)}{2}}{\binom{N_D(t)+1}{2}} \times N_D(t)\lambda(t). \tag{2.6}$$

The first term is the chance that the lineages represented by the reproducing allele and its offspring are among those of the sample. The second term is the instantaneous rate of additional births in the disease population. After simplifying, this yields an instantaneous coalescence rate of

$$\frac{k(t)(k(t)-1)}{N_D(t)+1} \times \lambda(t)$$

attributable to additional birth events. The total coalescence rate is therefore

$$k(t)(k(t)-1)\left[\frac{\lambda(t)}{N_D(t)+1} + \frac{1}{2(N_D(t)-1)}\right] \tag{2.7}$$

per generation, provided $N_P(t)$ is not too small. This rate is awkward because it involves both $\lambda(t)$ and $N_D(t)$. A standard simplifying device is to assume that the total population grows deterministically by a factor of $m(i) = \int_i^{i-1} \exp(\lambda(t_f - s))\, ds$ over the one-generation interval from i to $i-1$ gbp, $i = 0, \ldots, t_f$. Then coalescences due to additional births do not occur randomly, and so equation (2.6) does not apply. Under deterministic growth, additonal (non-diseased) copies may be viewed to enter the population from outside at the appropriate rate. Thus, coalescences are due to Moran events, so that only the coalescent rate in equation (2.5) applies. The second factor

$$\frac{N_P(t) - 1}{N_D(t) - 1}$$

in the left-hand side of equation (2.5) speeds up the rate of coalescence relative to a random sample from the total population, and may be viewed as an ascertainment correction to account for sampling of the disease rather than the total population. Also, since $N_P(t)$ is typically large, $(N_P(t) - 1)/N_P(t) \approx 1$. Thus, except when $N_D(t)$ is very small, the rate is close to the rate $\frac{1}{2}k(t)(k(t) - 1)/N_D(t)$ that would be obtained if the disease subpopulation reproduced under a continuous-time Moran model, independently of the total population. Early in the disease history, when $N_D(t)$ is small, coalescences occur at a faster rate. Comparing equations (2.5) and (2.7), we see that the same rates obtain when $\lambda(t) \equiv 0$, as expected. When $\lambda(t) > 0$, the rate of coalescence under random growth of the total population is faster than under deterministic growth. This is not surprising given the additional source of coalescence events.

The ancestral coalescent may be realized conditional on past sizes $N_D(t)$. We use the coalescent rate in equation (2.5) rather than the rate in equation (2.7) to simplify calculations. Past sizes $N_D(t)$ are realized via a birth-and-death approximation which assumes a single copy of the disease mutation a known time ago. In the next section, we develop the distribution of these past sizes $N_D(t)$, in order to generate realizations for coalescence rates.

2.4 *Disease copy numbers*

This section describes the stochastic process underlying past sizes $N_D(t)$ of the disease population. Given past sizes $N_D(t)$, the coalescent may be realized using the rate in equation (2.5). We consider evolution forwards in time from the point t_f gbp at which a single mutant allele is assumed. The jump times for $N_D(t)$ are generated successively, from $t = t_f$ to $t = 0$, conditional on current size up to that point. For example, labelling jump times t_i, $i = 1, 2 \ldots$, in the order they occur forwards in time, and starting at t_f with $N_D(t_f) = 1$, t_1 is realized. Then, given t_1 and $N_D(t_1) > 0$, t_2 is realized, etc. At each t_i, the disease population size either increases or decreases by one. This process conditions on $N_D(t_f) = 1$, but not on present size $N_D(0)$, or disease survival to present. As before, time is expressed in generations before present (gbp). Moran and additional birth events impacting $N_D(t)$ are treated separately; combining them yields the overall process.

We consider first the instantaneous rate of Moran events impacting $N_D(t)$. Given a Moran event at time t, and $N_D(t)$ and $N_P(t)$, the chance that the disease population changes size is

$$2\frac{N_D(t)}{N_P(t)}\left(1 - \frac{N_D(t)}{N_P(t)}\right). \tag{2.8}$$

with size being equally probable to increase or decrease by one. The instantaneous rate of such Moran events is thus

$$r_M(t) = N_D(t) \times \left(1 - \frac{N_D(t)}{N_P(t)}\right) \tag{2.9}$$

per generation. Calculation of $r_M(t)$ is simplified by replacing $N_P(t)$ with its expected value, eliminating the need for explicit modelling of the total population. (An adequate approximation to the size of the total population is its expected value, provided the population size at disease founding is not too small.)

The instantaneous rate of birth events impacting the size of the disease population is

$$r_B(t) = \lambda(t)N_D(t).$$

These rates specify the distribution of times at which the disease population changes size. Let T_i denote the random time of the i^{th} (Moran or pure birth) event impacting disease population size, and define $T_0 \equiv t_f$. Then given $T_i = t_i$ and $N_D(t_i)$, the time T_{i+1} for the next event is

$$t_i - \min(M_{i+1}, B_{i+1}),$$

where M_{i+1} and B_{i+1} are, respectively, the next Moran and pure birth events impacting size. In the interim, for $t \in (T_{i+1}, t_i)$, size is constant at $N_D(t_i)$. Given t_i and $N_D(t_i)$, realization of B_{i+1} using $r_B(t_i)$ is thus straightforward. Generating M_{i+1} is more complicated, because $N_P(t)$ and hence $r_M(t)$ in equation (2.9) changes over this interval.

We generate M_{i+1} by transforming an exponential variate M^* with rate $r_M(t_i)$. The transformation rescales to account for the changing rates $r_M(t)$ in (T_{i+1}, t_i); see Appendix A. This yields

$$M^* = \int_{t_i - M_{i+1}}^{t_i} \frac{r_M(s)}{r_M(t_i)}\, ds.$$

Since $N_P(s) = N_P(t_i)\exp(\lambda(t_i - s))$ for $s < t_i$,

$$M^* = \frac{N_D(t_i)}{r_M(t_i)} \int_{t_i - M_{i+1}}^{t_i} \left[1 - \frac{N_D(t_i)}{N_P(t_i)e^{\lambda(t_i - s)}}\right] ds,$$

which evaluates to

$$M^* = \frac{N_D(t_i)}{r_M(t_i)} \left\{ M_{i+1} - \frac{N_D(t_i)}{N_P(t_i)} \frac{1}{\lambda}(1 - \epsilon^{-\lambda M_{i+1}}) \right\}.$$

implying

$$[N_P(t_i) - N_D(t_i)]\, M^* + \frac{N_D(t_i)}{\lambda} = \frac{N_D(t_i)}{\lambda} \exp(-\lambda M_{i+1}) + M_{i+1} N_P(t_i). \qquad (2.10)$$

Solving for M_{i+1} yields the next Moran event time impacting the size of the disease population.

Equation (2.10) may be solved numerically by taking M^* as the initial guess. This is a good guess provided the rate of change in $N_P(t)$ at $t = t_i$ is not too fast. In fact,

M_{i+1} can be no larger than M^*. From equation (2.9), we know that $r_M(t)$ increases with $N_P(t)$ for $t \in (T_{i+1}, t_i)$. Thus for more recent $t \leq t_i$, when $N_P(t)$ has either increased or stayed the same, $r_M(t)$ is greater than or equal to the rate $r_M(t_i)$ for M^*. Therefore, M^* is an upper bound. In addition, $N_D(t) = N_D(t_i)$ for $t \in (T_{i+1}, t_i)$, so an exponential with faster rate $N_D(t) = N_D(t_i) > r_M(t)$ provides a lower bound. When $\lambda = 0$, the total population is of constant size and the appropriate rate of Moran events given t_i and $N_D(t_i)$ is $r(t_i)$. In this case, the solution M_{i+1} should be equal to the initial guess M^*, as verified by the limit as $\lambda \to 0$ in equation (2.10). Conversely, when $\lambda \to \infty$, explosive growth of the total population cancels the usual Moran dependence among copies induced by the limit on population size. Moran dependence among disease copies is also removed when $N_D(t) << N_P(t)$, greatly simplifying the equations leading to (2.10). Thus, the solution M_{i+1} has an exponential distribution with rate $N_D(t_i)$ uncorrected for this dependence (see below).

Disease population sizes may be realized prospectively, conditional on a single mutant allele at t_f, as described. This is computationally intensive because of the potentially large number of events between t_f and the present. More importantly, the scheme does not permit conditioning on the current number of copies of the disease allele, $N_D(0)$, except by rejection sampling of realizations. The next section describes a birth-and-death process that approximates disease reproduction, and also allows past sizes to be generated retrospectively, conditional on $N_D(0)$. A retrospective scheme saves computational time when realizing $N_D(t)$ conditional on $N_D(0)$ and $N_D(t_f) = 1$.

2.5 A birth-and-death approximation

Provided the disease is rare, $N_D(t)$ may be approximated by a birth-and-death process with instantaneous death rate $\mu^* = 1/2$ and birth rate $\lambda^*(t) = \lambda(t) + 1/2$. This can be seen by comparing the Moran event rate $r_M(t)$ in equation (2.9) to the total event rate

$N_D(t)$ in a birth-and-death model with birth and death rates of 1/2 per generation. The only difference is the correction factor $1 - N_D(t)/N_P(t)$, which is due to Moran dependence among disease alleles imposed by the constraints of total population size. This dependence reduces the rate of events compared to a birth-and-death model, in which alleles reproduce independently. However, for a rare disease, $N_D(t) << N_P(t)$ throughout history, and disease alleles reproduce essentially independently. The birth-and-death process also approximates disease evolution in the Moran + birth model. With a rare disease, Moran events virtually never involve disease alleles both dying and reproducing. Jump times for changes in disease size are thus essentially the same as birth or death event times in the disease population.

We first review basic attributes of the birth-and-death approximation, such as the distribution of the number of progeny and the number of descendants, and the probability generating function. These basic attributes are used to develop formulae for the conditional moments of the past size $N_D(t_1)$ of the disease population at t_1 gbp, given the size at more recent t_2. Following this, we discuss realization of conditional past sizes throughout history, given the present size $N_D(0)$.

2.5.1 Basic attributes

Let X_t be the number of births experienced by a copy of a rare disease allele over a time interval of length t generations. Then X_t is Poisson with mean

$$\int_0^t [\lambda(s) + 1/2]\ ds$$

(Cox and Miller 1977). The progeny number for a rare disease allele is X_L, where lifetime L is exponential with mean 2. Thus, progeny distribution is given by E P($X_L = i \mid L$), where i is the number of progeny observed. When the population grows at constant rate e^λ per generation, the progeny distribution for a rare disease allele is an exponential mixture of Poissons, and is therefore geometric. In this case,

$$\mathrm{P}\,(X_L = i) = (1-p)^i \times p,$$

where $p = 1/((1+\lambda))$ is the probability that an event is a death. For example, if we assume that the Finnish population grew at constant exponential rate from 2000 allelic copies at founding 80 generations ago to the present number 10^7, we obtain $\lambda(t) \equiv \lambda = 0.11$. Thus the progeny number for a rare disease allele has a mean of about 1.2, and a variance of about 2.7. When population growth varies, no general statement can be made about the progeny distribution.

Consider a copy of a rare disease allele at time t_i gbp. and define $t_0 \equiv t_f$. Then the number $Y_i(s)$ of descendants s generations later at $t_i - s$ gbp has a zero-modifed geometric distribution (Kendall 1948), with

$$\mathrm{P}\,(Y_i(s) = 0) = \xi_i(s) \tag{2.11}$$

$$\mathrm{P}\,(Y_i(s) = n) = [1-\xi_i(s)]\,[1-\eta_i(s)]\,\eta_i(s)^{n-1} \quad n \geq 1, \tag{2.12}$$

where

$$\begin{aligned}
\xi_i(s) &= 1 - \frac{\exp(-\rho_i(s))}{W_i(s)}, \\
\rho_i(s) &= \int_{t_i-s}^{t_i} [\mu^*(t) - \lambda^*(t)]\; dt, \\
W_i(s) &= \exp(-\rho_i(s)) \left[1 + \int_{t_i-s}^{t_i} \exp(\rho_i(t))\; \mu^*(t)\; dt\right] \quad \text{and,} \\
\eta_i(s) &= 1 - \frac{1}{W_i(s)}.
\end{aligned}$$

Since $Y_0(s) = N_D(t_f - s)$, this also gives the unconditional distribution of disease population size at $t_f - s$ gbp.

Under constant population growth (i.e., $\lambda^*(t) \equiv \lambda^* = \lambda + 1/2$), the descendant distribution reduces to

$$\mathrm{P}\,(Y_i(s) = 0) = \mu^* B(s), \qquad \mathrm{P}\,(Y_i(s) = n) = [1-\lambda^* B(s)]\,[1-\mu^* B(s)]\,[\lambda^* B(s)]^{n-1}$$

(Feller 1968), where

$$B(s) = \frac{1 - e^{(\lambda^*-\mu^*)s}}{\mu^* - \lambda^* e^{(\lambda^*-\mu^*)s}}.$$

The unconditional expected value of disease copy number at $t_f - s$ gbp is $\exp(\lambda(t_f - s))$. A disease mutation introduced into the Finnish population at founding 80 generations ago has probability $\mathrm{P}(Y_0(80) = 0) = \mathrm{P}(N_D(0) = 0) = 0.82$ of extinction by the present. This is in excellent agreement with simulations under the Moran and pure birth model. Simulations were run as described in section 2.5. Of the 10000 replicates, only 1793 (18%) survived to the present. The conditional distribution of current allelic copy number given survival to the present is plotted in Figure 2.6, along with the approximating geometric distribution from the birth-and-death process. The geometric approximation is actually slightly heavier-tailed because there are no constraints on the number of descendants imposed by the size of the total population, but the difference is imperceptible. The simulations show that conditional on present survival, the chance of achieving a current copy number b exceeding the approximately 10^4 extant IOSCA alleles in Finns is 0.87: unconditionally, the chance is only 0.16. Given survival to present, the chance that current copy number is within a plausible range (say, 8000 to 10000) of the extant IOSCA copies is about 0.10. Without conditioning on survival, the chance is only 0.02.

A key characterization of the birth-and-death approximation is the marginal probability generating function (pgf). For the process starting at $t_0 = t_f$ gbp and running s generations to $t_f - s$ gbp, this is

$$g_0(x, s) = \frac{\xi_0(s) + [1 - \xi_0(s) - \eta_0(s)]\, x}{1 - \eta_0(s)x} \tag{2.13}$$

(Kendall 1948). In general, the marginal pgf for the process starting at t_i gbp and running s generations to $t_i - s$ gbp is

$$g_i(x, s) = \frac{\xi_i(s) + [1 - \xi_i(s) - \eta_i(s)]\, x}{1 - \eta_i(s)x}. \tag{2.14}$$

When growth is constant, $g_0(x, s)$ simplifies to

$$\frac{\mu^*(1-x) - (\mu^* - \lambda^* x)e^{-(\lambda^* - \mu^*)s}}{\lambda^*(1-x) - (\mu^* - \lambda^* x)e^{-(\lambda^* - \mu^*)s}} = [xA_2(s) + A_3(s)] \,/\, [xA_4(s) + A_5(s)]$$

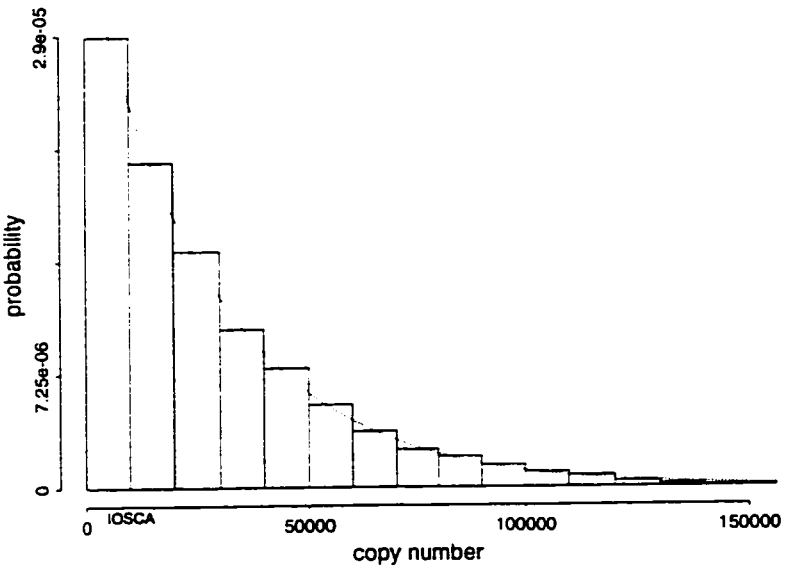


Figure 2.6: Histogram of the present copy number of an IOSCA-like mutation in Finns, given one allelic copy at founding $t_f = 80$ gbp. and survival of the mutation to the present. Dotted line is the corresponding density under a birth-and-death process approximation.

(Cox and Miller 1977), where $A_1(s) = \exp(-(\lambda^* - \mu^*)s)$, $A_2(s) = \lambda^* A_1(s) - \mu^*$, $A_3(s) = \mu^* - \mu^* A_1(s)$, $A_4(s) = \lambda^* A_1(s) - \lambda^*$, and $A_5(s) = \lambda^* - \mu^* A_1(s)$.

2.5.2 Conditional moments

We use the birth-and-death approximation to derive the moment generating function and hence the moments for past numbers of a rare disease allele at t_1, conditional on the number $N_D(t_2)$ at more recent t_2, and on $N_D(t_f) = 1$. The derivation follows Thompson et al. (1992), but is for a birth-and-death process in continuous-time rather than for a branching process in discrete-time.

The joint pgf for $N_D(t_1)$ and $N_D(t_2)$, $t_1 > t_2$, is

$$g(z, w, t_1, t_2) = \mathrm{E}\,(z^{N_D(t_1)} w^{N_D(t_2)}) = \sum_r \sum_j z^r w^s \,\mathrm{P}\,(N_D(t_1) = r, N_D(t_2) = j).$$

This may be written

$$g(z, w, t_1, t_2) = g_0\left(g_1(w, \delta)z, t_1\right). \tag{2.15}$$

in terms of the time difference $\delta = t_1 - t_2$, and the marginal pgf's g_0 and g_1 in equations (2.13) and (2.14), respectively. For a general random variable X, $\mathrm{E}\prod_{i=0}^{n-1}(X-i)$ is the n^{th} *factorial moment.* We obtain the n^{th} factorial moment for $N_D(t_1)$ given $N_D(t_2) = j$ by dividing the coefficient C_j of w^j in the series expansion of

$$\frac{\partial^n}{\partial z^n}\, g(z, w, t_1, t_2)\,\Big|_{z=1}$$

by $\mathrm{P}\left(N_D(t_2) = j\right)$, or $\mathrm{P}\left(Y_2(\delta) = j\right)$ in equation (2.12). For example, under constant growth, we get

$$\mathbf{E}\left[N_D(t_1) \mid N_D(t_2) = j, N_D(t_f) = 1\right] = c_1(t_1, t_2) \times j + c_2(t_1, t_2),$$

where

$$\begin{aligned} c_1(t_1, t_2) &= \Big\{A_5(t_f - t_2)\left[A_2(\delta)\left(A_5^2(\delta) - A_4(\delta)A_4(t_f - t_2)\right) + A_3(\delta)A_4(\delta)A_5(\delta)\right] - \\ &\qquad A_3(\delta)A_5^2(\delta)A_4(t_f - t_2)\Big\} \times \left\{A_1(\delta)A_5(\delta)A_5(t_f - t_2)(\lambda^* - \mu^*)^2\right\}^{-1}, \\ c_2(t_1, t_2) &= \left[A_2(\delta)A_4(\delta)A_5^2(t_f - t_2) - A_3(\delta)A_5(\delta)A_4^2(t_f - t_2)\right] / \left[A_1(\delta)(\lambda^* - \mu^*)^2\right]. \end{aligned}$$

The coefficient $c_1(\cdot, t_2)$ increases monotonically from 0 at $t_1 = t_f$, to 1 at $t_1 = t_2$; the coefficient $c_2(\cdot, t_2)$ is quadratic with a maximum at $t_1 = t_2/2$ and with boundary values $c_2(t_f, t_2) = 1$ and $c_2(t_2, t_2) = 0$. Although not reported, even more complicated expressions may be derived for higher moments, or for moments under changing growth. When growth rates change over time, these moments indicate that for $N_D(t_2)$ moderately large (e.g., > 500), and t_1 and t_2 close (e.g., $\delta = 1$ generation apart), local rates apply, and event rates elsewhere in history have little influence.

Figures 2.7, 2.8, and 2.9 plot expected past sizes of a disease population given a present disease allele count of $N_D(0) = b$ in the Finns, assuming a single mutant copy at founding $t_f = 80$ gbp and constant growth of the total population. Results are given for $b = 10000$ (IOSCA), 5000, and 500, respectively, and are in the $\log_{10}$

scale. In all three plots, initial expected growth of the disease population is faster than exponential, since surviving disease mutations are those which have, on average and by chance, high initial growth. For a disease with low present copy number such as $b = 500$, slower-than-exponential growth is expected to follow the initial burst. Conditional on survival, the expected copy number of a rare Finnish allele, present as a single copy at founding, is 27778. Following arguments analogous to Maynard Smith (1971), this is just the unconditional expected number $\exp(\lambda t_f) = 5000$ of descendants of a single copy after t_f generations, divided by the probability $1 - \mathrm{P}(N_D(0) = 0) = 1 - 0.82$ of non-extinction, from equation (2.11). The conditional standard deviation given survival is 28468, and of the same magnitude. Thus, there is insufficient evidence to reject the hypothesis that a disease allele with current copy number 10000 was present as a single copy at $t_f = 80$ gbp. The distribution of the age of a selectively-neutral allele given its current copy number is explored in section (2.8).

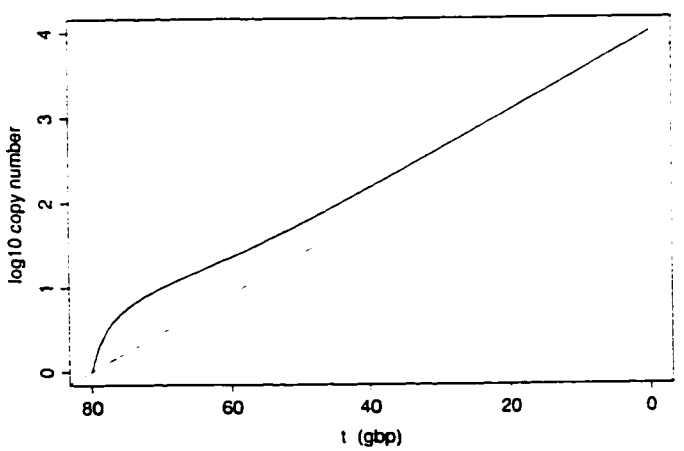


Figure 2.7: Log_{10} expected copy numbers $N_D(t)$ in Finns (solid line), given $N_D(0) = b$ and $N_D(t_f) = 1$; b=10000 (IOSCA). The diagonal dotted line is the $\log_{10}$ unconditional expected copy number.

Figure 2.10 illustrates the derivative of the curve in Figure 2.7, up to a multiplica-

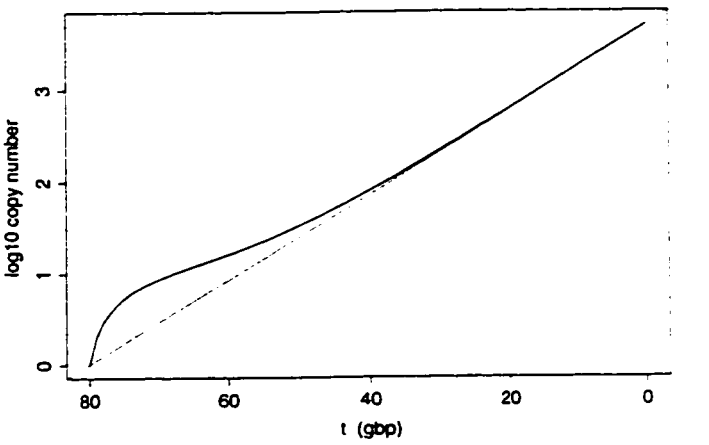


Figure 2.8: Log_{10} expected copy numbers $N_D(t)$ in Finns (solid line) given $N_D(0) = b$ and $N_D(t_f) = 1$; b=5000. The diagonal dotted line is the $\log_{10}$ unconditional expected copy number.

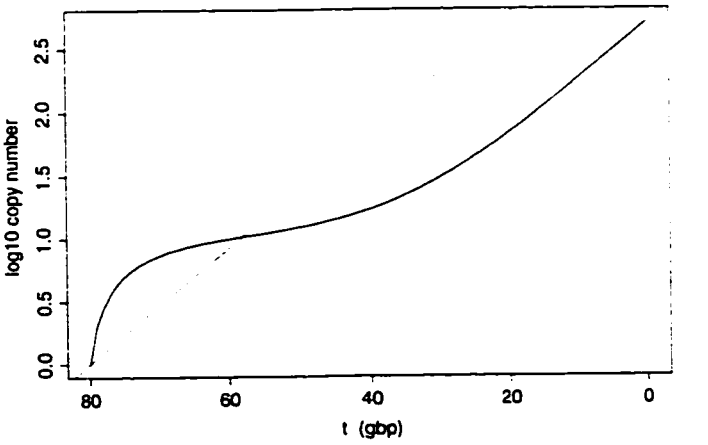


Figure 2.9: Log_{10} expected copy numbers $N_D(t)$ in Finns (solid line) given $N_D(0) = b$ and $N_D(t_f) = 1$; b=500. The diagonal dotted line is the $\log_{10}$ unconditional expected copy number.

tive constant $\log_e(10) = 2.3$. The curve is the derivative, with respect to time, of the natural logarithm of conditional copy number given $N_D(0) = 10000$, and $N_D(t_f) = 1$. This derivative is analogous to $\lambda^* = \lambda + 0.5 = \log_e(10^7/2000) + 0.5 = 0.61$, the instantaneous overall rate of growth in the approximating birth-and-death process. This "conditional growth rate" starts near the unconditional instantaneous rate of total events $\lambda^* + \mu^* = 0.61 + 0.5 = 1.11$ at 80 gbp, but drops quickly to a level below the unconditional rate around 70 gbp, and then climbs slowly to match the unconditional rate by 40 gbp. The sag in the conditional rate 70-40 gbp indicates slower-than-exponential growth after the initial burst due to the present IOSCA copy number being less than the roughly 25000 copies expected given survival to the present. Note that a generalized pure birth process with instantaneous rate equal to the conditional growth rate would be guaranteed to survive to the present. Such a process would have the same first moment as the conditional process, but it is unclear whether higher moments would be comparable. In fact, we shall see later that higher moments are *not* comparable.

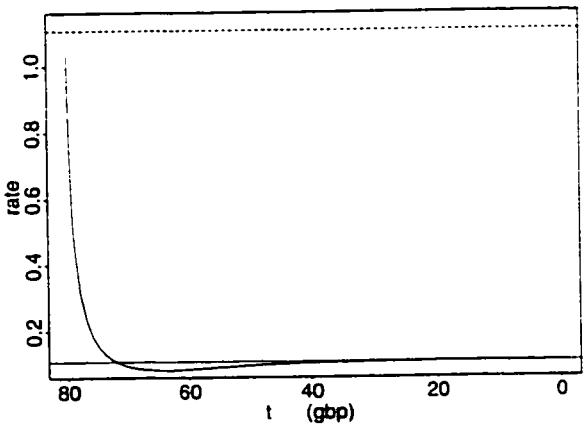


Figure 2.10: Derivative of the natural logarithm of the conditional expected copy number in Finns given $N_D(0) = 10000$ and $N_D(t_f)$=1. Solid line is this instantaneous conditional growth rate. Dotted line, unconditional instantaneous growth rate; dashed line, unconditional instantaneous total-event rate.

Figure 2.11 plots conditional variances $\mathrm{V}\{N_D(t) \mid N_D(0) = b, N_D(t_f) = 1\}$ for $b = 10000$, 5000, and 500. Conditional variance is relatively small at first due to small copy number, but increases over time with the number of copies, before decreasing to zero at present, as required by the conditioning. In general, conditional variability is small compared to the mean (e.g. the maximum s.d. for IOSCA is about 160 when expected copy number is about 3600), and increases with the present copy number b because there are more allelic copies in the system.

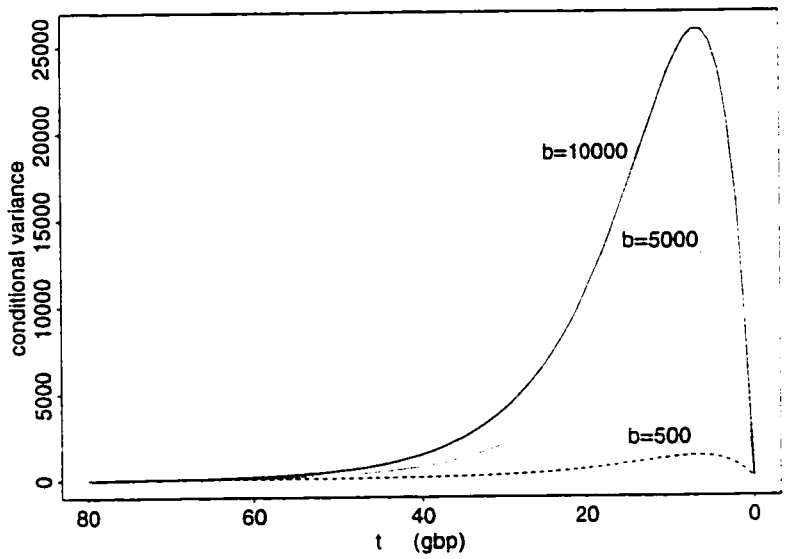


Figure 2.11: Conditional variance of past numbers $N_D(t)$ in Finns given $N_D(0) = b$ and $N_D(t_f) = 1$, for $b = 10000$ (IOSCA), 5000, and 500.

Figure 2.12 illustrates skewness coefficients for IOSCA in Finns at one generation intervals, setting $N_D(t) = \mathrm{E}\{N_D(t) \mid N_D(0) = 10000, N_D(t_f) = 1\}$ when obtaining conditional moments for $N_D(t+1)$ given $N_D(t)$ via equation (2.15). The skewness of a gamma distribution with matching first two moments is also shown, but not that of an approximating normal distribution, for which the value is always 0. The conditional distribution is expected to be positively skewed early in history when copy

number is low, since the disease mutation must survive to the present. For more recent times, individual contributions from the large number of independently reproducing disease copies average, leading to a more symmetric distribution. Positive skewness thus decreases with disease population size or with generations since founding, but in general more closely matches that of the gamma than the normal approximating distribution.

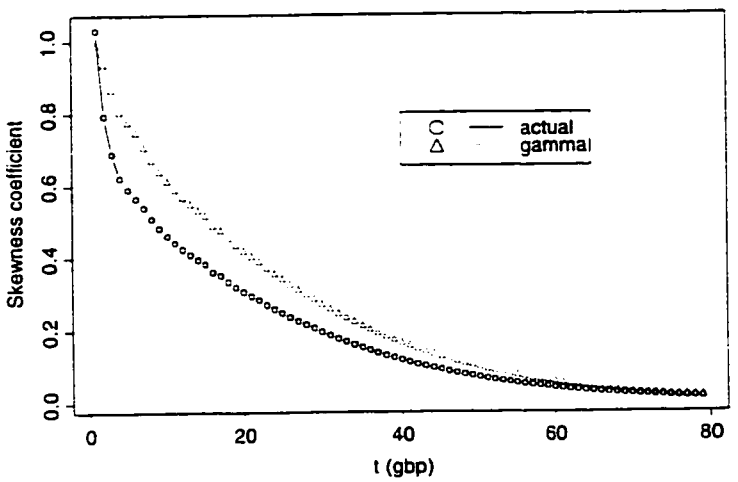


Figure 2.12: Skewness of $N_D(t+1)$ in Finns given $N_D(t) = \mathrm{E}\,(N_D(t) \mid N_D(0) = 10000, N_D(t_f) = 1)$. o: skewness under the conditioned birth-and-death process; Δ: skewness of a gamma distribution with matching mean and variance. Skewness of a normal distribution with matching mean and variance is zero, and thus not shown.

2.5.3 Realization of $N_D(t)$

Past sizes of the disease population conditional on the present number could be realized by simulating forwards in time from founding and rejecting realizations not reaching $N_D(0)$. Almost all realizations would then be rejected. Alternatively, realizations within a reasonable range of $N_D(0)$ could be used. This would still be inefficient. For example, when modelling the growth of IOSCA in Finland, and taking

realizations within 20% of $N_D(0) = 10000$, about 98% of realizations are discarded. Suppose another birth-and-death process $\{\widetilde{N}_D(t),\ 0 \le t \le t_f\}$ could be found with conditional distribution $\{\widetilde{N}_D(t) \mid \widetilde{N}_D(0) = b, \widetilde{N}_D(t_f) = 1;\ \ 0 \le t \le t_f\}$ matching $\{N_D(t) \mid N_D(0) = b, N_D(t_f) = 1;\ \ 0 \le t \le t_f\}$. If this process discarded fewer realizations when conditioning via rejection sampling, efficiency could be improved. Close matching of the conditional and unconditional total event and growth rates would be required. The first two moments of a process specify the instantaneous birth and death rates; thus, $\mathrm{E}\{\widetilde{N}_D(t)\}$ would have to be close to $\mathrm{E}\{N_D(t) \mid N_D(0) = b, N_D(t_f) = 1\}$ and $\mathrm{V}\{\widetilde{N}_D(t)\}$ would have to be close to $\mathrm{V}\{N_D(t) \mid N_D(0) = b, N_D(t_f) = 1\}$. Equating expectations and variances thus gives an indication of implied instantaneous total event and growth rates in the hypothetical process. Figure 2.13 plots these implied rates for IOSCA in Finns. The total event rate at present in the hypothetical process would have to be *negative*. Furthermore, total event rates at other times would be up to three times higher than $\lambda^* + \mu^* = 1.12$, the rate for the actual unconditional process. Any efficient process would be expected to have initial death rate (at 80 gbp) near 0, since copies must survive to present. Comparison of instantaneous total event and growth rates shows that although hypothetical death rates start at 0, they rise rapidly and very soon become greater than the death rate $\mu^* = 1/2$ for the actual unconditional process. It is thus doubtful that a more efficient birth-and-death process exists, so it is not pursued further.

Alternatively, past sizes given $N_D(0)$ could be realized backwards in time at one generation intervals, providing sufficient resolution for the coalescent rates of equation (2.5). For example, in the Finns, one could look in at times of $t = 1, \cdots, 79, 80$ gbp, and realize backwards successively from $\{N_D(t+1) \mid N_D(t), N_D(t_f) = 1\}$. Thus, given $N_D(0)$ and $N_D(t_f) = 1$, $N_D(1)$ may be realized. Similarly, given $N_D(1)$ and $N_D(t_f) = 1$, $N_D(2)$ may be realized. Continuing this process back in time from 0 to t_f gives a sample path for $\{N_D(t) \mid N_D(0), N_D(t_f) = 1\}$ at one-generation resolution. The joint pgf in equation (2.15) allows the conditional distribution of past size at $t+1$

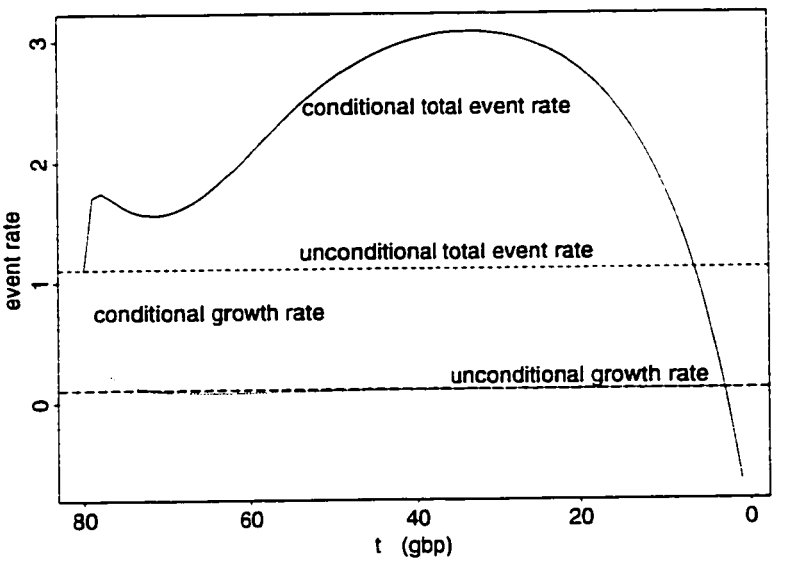


Figure 2.13: Instantaneous event rates for past numbers $N_D(t)$ in Finns. Conditional rates are the derivative of the natural logarithm of conditional expected copy number given $N_D(0) = 10000$ and $N_D(t_f) = 1$. Solid line, conditional total event rate; dotted line, conditional growth rate; short dashed line, unconditional total event rate; long dashed line, unconditional growth rate.

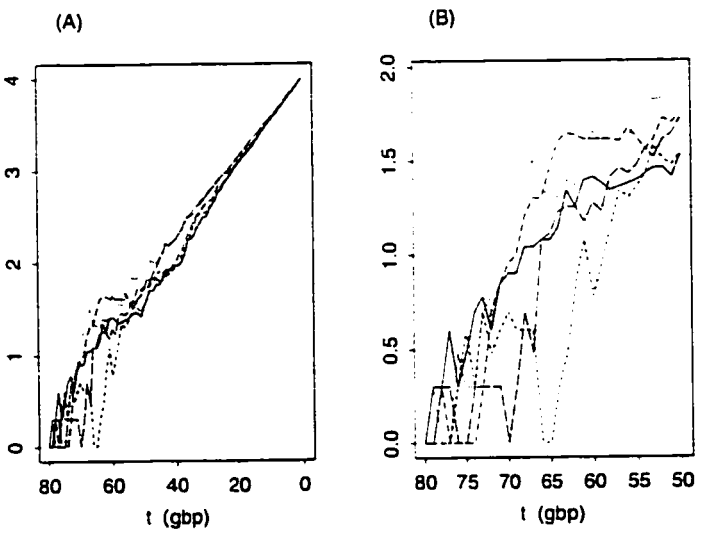


Figure 2.14: Realizations of past disease copy numbers in the Finnish example. Vertical axis, $\log_{10}$ scale. Past sizes are conditional on present size $N_D(0) = 10000$, and founding size $N_D(t_f) = 1$. (A) Five sample paths over all $t_f = 80$ generations since founding; (B) the paths in the first thirty generations only, 80-50 gbp.

gbp given the size at t gbp to be approximated to arbitrary precision by a distribution matched on an appropriate number of moments. Figure 2.14 shows five sample paths for an IOSCA-like allele. Copy numbers are generated from a normal-gamma mixture having the same first three moments as the actual distribution. The first two moments of the normal and gamma components of the mixture have been chosen to match those of the actual distribution; the mixing proportion has been chosen to obtain the right third moment. Numbers are plotted in the $\log_{10}$ scale, and cannot go below 1 because the disease allele survives to the present. However, some sample paths hit the single-copy boundary several times. The most recent of these provides a tighter upper bound on TMRCA than t_f. If the mutation is in fact younger than t_f, more such hits would be expected in order to compensate for present copy number being lower than expected given survival to the present, assuming a single mutant copy at t_f. Hence, when present copy number is more consistent with a younger mutation than t_f, stochastic modelling of $N_D(t)$ adjusts for this.

Figure 2.15 shows realized past sizes of a WRN4-like allele in the Japanese. In both populations, but especially in the Japanese, variability over realizations in more recent generations is dominated by temporal changes in the numbers themselves, giving the appearance of deterministic growth (2.15A). However, Figure 2.15B shows that there is indeed variability in more recent generations.

Hereafter, stochastic growth of the disease population will be incorporated into coalescent rates for the disease sample by conditioning on realizations of $\{N_D(t) \mid N_D(0) = b, N_D(t_f) = 1;\ t = 1, 2, \ldots, t_f - 1, t_f\}$ at one-generation intervals. One-step realizations of past disease copy numbers will be generated successively backwards in time, from 0 to t_f, using an approximating normal-gamma mixture, with first three moments matching those of the actual distribution under a birth-and-death model of disease population size. A birth-and-death model of disease population size is appropriate for a rare disease allele. In the next section, we discuss how coalescent times are generated given past disease copy numbers. We also compare coalescent

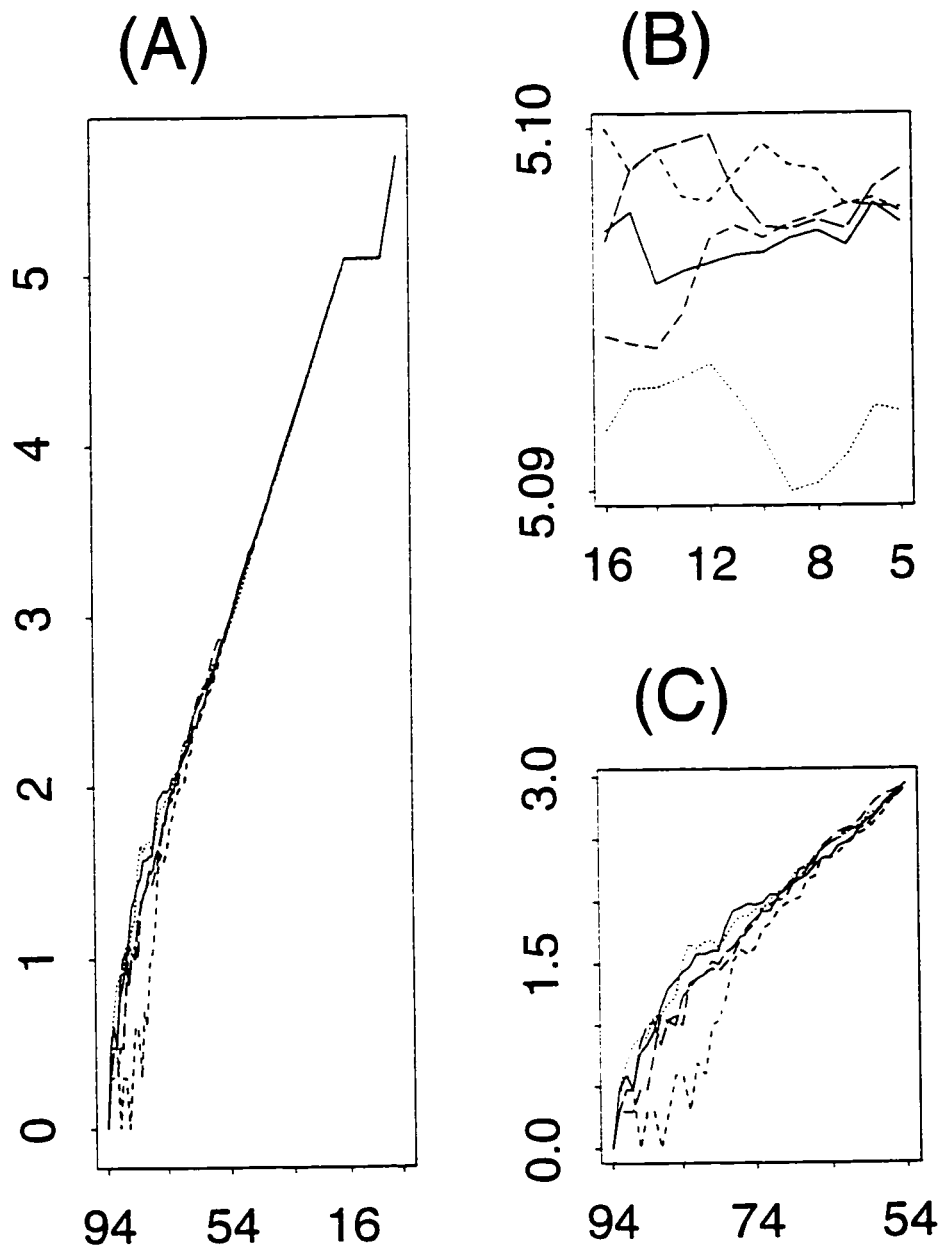


Figure 2.15: Realizations of past disease copy numbers in the Japanese example. Vertical axis, $\log_{10}$ scale; horizontal axis, generations before present (gbp). Past sizes are conditional on present size $N_D(0) = 500000$, and founding size $N_D(t_f) = 1$. (A) Five sample paths over all $t_f = 94$ generations since founding; (B) the paths during the Edo period, 5-16 gbp; (C) the paths in the first 40 generations after founding, 94-54 gbp

times under stochastic and deterministic growth of $N_D(t)$.

2.6 *Realization of coalescent times*

To realize the coalescent for a random sample of disease alleles, we use the instantaneous rates in equation (2.5) and insert realized past copy numbers $N_D(t)$, conditional on present size $N_D(0)$, and on $N_D(t_f) = 1$. Since past copy numbers are given in one generation steps only, they are interpolated between generations. For moderately large sizes of the disease population (i.e., $N_D(t) > 30$), intragenerational growth is assumed to be exponential, and $N_D(t) - 1 \approx N_D(t)$. For smaller sizes early in disease history, where super-exponential growth occurs, numbers are linearly interpolated.

Rescaling methods outlined in Appendix A are used. Coalescent times t^* are first generated under constant rate $r(0)$ (between coalescent events) and then rescaled to account for size fluctations in the disease population. From equation (2.5), the appropriate rate $r(s)$ in equation (A.1) is

$$\frac{k(s)(k(s)-1)}{2(N_D(s)-1)} \approx \frac{k(s)(k(s)-1)}{2N_D(s)},$$

for $N_D(s) > 30$, with $N_D(s)$ growing exponentially, and

$$\frac{k(s)(k(s)-1)}{2(N_D(s)-1)},$$

for $N_D(s) \leq 30$, with $N_D(s)$ growing linearly. To get the actual coalescent time, we solve for t in Equation (A.1). The integral is written as a sum over one-generation intervals, running from backwards in time from 0 to t_f gbp. Each summand is evaluated, until the running sum at the interval $(i, i+1)$ exceeds t^*. Thus, $t \in (i-1, i)$ gbp. We then solve for t exactly. Specifically, since we have already computed the integral

$$\int_0^{i-1} \frac{r(0)}{r(s)}\, ds,$$

we can solve

$$t^* - \int_0^{i-1} \frac{r(0)}{r(s)}\, ds = \int_{i-1}^{t} \frac{r(0)}{r(s)}\, ds$$

for t.

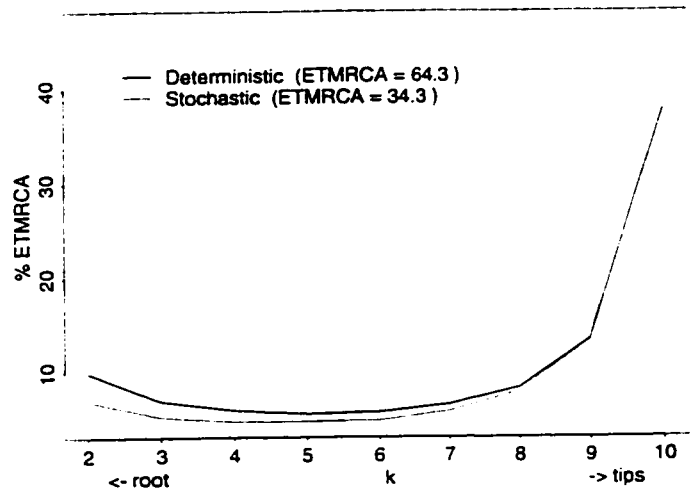


Figure 2.16: Expected coalescence times ET_k as a percentage of ETMRCA= $\sum_2^K ET_k$ for an IOSCA-like mutation in Finns: Deterministic (solid line) versus stochastic (dotted line) growth of the disease population. ETMRCA under deterministic growth, 64.3 gbp: under stochastic growth, 34.3 gbp.

Figure 2.16 compares the expected ancestral tree shapes of two samples of 10 copies of an IOSCA-like allele in Finns: one sample taken from a disease population growing deterministically by a factor of $m = 1.12$ per generation, and the other taken from a disease population growing stochastically within the total population. Recall that the deterministic growth factor of $m = 1.12 = N_D(0)^{1/t_f}$ has been calculated to match current IOSCA copy number, assuming one IOSCA mutant at founding $t_f = 80$ gbp (see section 2.2). Hereafter, such growth will be described as *deterministic growth* of the disease population. On the other hand, random growth of the disease population conditional on $N_D(0)$ and $N_D(t_f) = 1$ will be described as *stochastic growth* of the disease population. This type of growth is modelled in section 2.5.3. Figure 2.16 displays E $T_k / \sum_{i=1}^K$ E T_i, the proportion of expected time to complete coalescence (ETMRCA) taken by E T_k, expressed as a percentage. Expected values are based on 10000 coalescent replicates. The ancestral tree under stochastic growth

of the disease is slightly more star-like than the tree under deterministic growth, with coalescences near the tips taking proportionally more of the time to complete coalescence, and those at the roots taking proportionally less time. A more striking difference is the much shorter ancestral tree for the stochastically-growing disease population, with ETMRCA=34.3 generations compared to 64.3 generations under deterministic growth. The reasons for this are as follows. Consider a population with a size bottleneck at some point in the past. The bottleneck results in more inbreeding than would be expected under a smooth but comparable overall rate of population growth. Under stochastic growth, disease copy number fluctuates about its expected value. This stochastic fluctuation is, in effect, a series of population bottlenecks. Deterministic exponential growth smooths over these bottlenecks. The result is less inbreeding, and hence increased coalescent times under deterministic growth. Figure 2.17 illustrates, on the absolute scale, the longer but similarly-shaped tree.

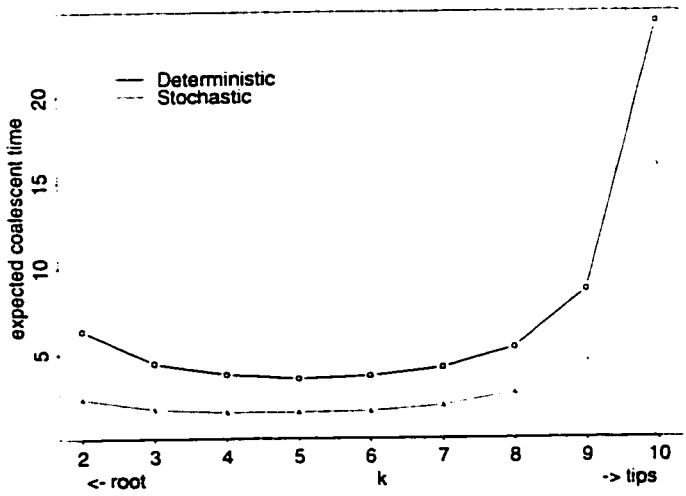


Figure 2.17: Expected coalescence times ET_k for an IOSCA-like mutation in Finns: Deterministic (solid line with circles) versus stochastic (dotted line with triangles) growth of the disease variant.

2.7 *Another version of the coalescent*

Another way to approach the disease coalescent is by viewing the ancestry of the disease population as a subtree of the ancestry of the total population. The coalescent of the disease sample is thus embedded within this subtree. Let $n(t)$ and $b(t)$ denote respectively the number of ancestors of the total population and the disease subpopulation at t gbp: $n(0) = N_P(0)$ and $b(0) = N_D(0)$. Given $n(t)$, the coalescence rate for the entire population at t gbp is

$$\frac{n(t)(n(t)-1)}{2N_P(t)}.$$

The coalescence rate of a subtree of size $b(t) \geq 2$ is just the product

$$\frac{(n(t)-1)b(t)}{2N_P(t)} \tag{2.16}$$

of the rate for the entire population and the probability $b(t)/n(t)$ that a coalescing lineage is on the subtree. Conceptually, coalescence events within the entire population are allocated to the part of the population identified with the disease subtree. Our interest is in the disease subtree and the coalescence rate for a random sample of size $k(0) = K$ from the $b(0)$ disease alleles at present. Suppose for the moment that no assumption is made regarding a single disease mutant at known time t_f gbp. Then the chance that a coalescing pair of disease lineages at t gbp is in the sample ancestry of $k(t)$ lineages is

$$\binom{k(t)}{2} \Big/ \binom{b(t)}{2}. \tag{2.17}$$

Hence, the appropriate coalescence rate for the sample given $N_P(t)$, $b(t)$, and $n(t)$ is

$$\frac{(n(t)-1)b(t)}{2N_P(t)} \times \binom{k(t)}{2} \Big/ \binom{b(t)}{2} = \frac{k(t)(k(t)-1)}{2N_P(t)} \times \frac{n(t)-1}{b(t)-1}. \tag{2.18}$$

The second factor $[n(t)-1]/[b(t)-1]$ on the right-hand side of the equation may be viewed as an ascertainment correction to Kingman's coalescent rate for a random sample of alleles from the total population. This ascertainment correction accounts

for the fact that the $k(t)$ sample lineages lie on a disease subtree with $b(t)$ lineages, so that coalescence should be faster than in a random sample from the total population.

In order to assume a single disease mutant at t_f gbp, we must condition on complete coalescence of the subtree by t_f. One way to achieve this would be to first generate conditional coalescence times $\{T_i \mid \sum_j T_j \leq t_f\}$, $i = b(0), \ldots 2$, for the subtree given $\{n(t) : t \geq 0\}$ and $\{N_P(t) : t \geq 0\}$, and then ask whether the coalescent event at each T_i involved a lineage pair from the sample. Given a coalescence on the subtree at $t = \sum_{j=i}^{K} T_j$ gbp, the chance that it involves a sample lineage is $\binom{k(t)}{2} / \binom{i}{2}$, from equation (2.17). Conditional coalescent times $\{T_i \mid \sum_j T_j \leq t_f\}$ could be generated by rejection sampling of the unconditional T_i, $i = b(0) \ldots 2$.

Although this approach is valid, it is difficult to implement because it requires the numbers of ancestors $n(t)$ for $t \geq 0$, and also the conditional coalescent times for the subtree $\{T_i \mid \sum_j T_j \leq t_f\}$, $i = b(0) \ldots 2$. Realizing the ancestral numbers $n(t)$ and $b(t)$ is difficult because of their large value at present. To generate the corresponding coalescents, time is, in effect, transformed so that the same coalescent rate applies between coalescences (although not across them). In the ancestry of a large number of copies, the majority of coalescences take place very quickly, especially in the transformed scale. Recent branches on the ancestral tree are so short in the transformed scale that problems with computational underflow and machine precision occur when simulating them.

2.8 Age of mutations

In this section, coalescent methods described by Griffiths and Tavaré (1998) are used to examine the age distribution of two rare recessive disease mutations, IOSCA and Werner's syndrome, that are assumed to be selectively neutral. In the absence of a selective advantage for disease heterozygotes, selective neutrality is reasonable for rare recessive diseases, since most disease alleles reside in heterozygotes. Heterozygote ad-

vantage has been been proposed as a possible explanation for the large copy number observed in some recessive diseases such as cystic fibrosis and phenylketonuria (e.g. Quinton 1994, Scriver 1994). However, Thompson and Neel (1997) point out that these arguments do not condition on disease survival to the present. These authors show that such conditioning is sufficient to explain the relatively high copy numbers for many recessive diseases, and that positive selection is therefore unnecessary. Inferences of the age of a selectively-neutral mutation can thus provide insight into the existence of a single disease allele at population founding, a useful simplifying assumption in disequilibrium mapping.

Following Griffiths and Tavaré (1998), in a population currently of size n copies in which there are b copies of the disease mutation, let J_0 denote the number of ancestors of the total population when the disease mutation arose. Consider the coalescent for the total population, and let T_j be the time during which there were j of its ancestors. Then Griffiths and Tavaré (1998) show that the distribution of J_0 given b and n is

$$\mathrm{P}\{J_0 = j\} = \frac{j p_{n,j}(b)\mathrm{E}T_j}{\sum_{k=2}^{n-b+1} k p_{n,k}(b)\mathrm{E}T_k}, \qquad 2 \le j \le n-b+1, \tag{2.19}$$

where $p_{n,j}(b)$ denotes the probability of b current descendants of a random allele chosen from among j ancestors. The intuition behind this equation is that, given survival of the disease variant, the probability that the mutation occurred when there were j lineages on the ancestry of the total population is proportional to the length $j \times \mathrm{E}T_j$ of that part of the ancestral tree. Conditioning on coalescent times for the total population and taking expected values yields that the probability of the mutation occuring when there were j ancestral lineages is proportional to $j \times \mathrm{E}T_j$. The restricted sum in the denominator arises because at one extreme, when all nonmutant lineages coalesce by the time at which the mutation arose, there are $j = 2$ ancestors; at the other extreme, when no lineages coalesce, there are $j = n - b + 1$ ancestors. To get $p_{n,j}(b)$, consider the number of ways to assign n unlabelled descendants to j labelled ancestors. Every ancestor must have a descendant. Hence, the problem is one of

placing $j-1$ partitions corresponding to the j labelled ancestors in $n-1$ spaces between the n unlabelled descendants. There are $\binom{n-1}{j-1}$ ways of doing this (Feller 1968, page 38). Fixing b descendants of the disease mutation at present leaves $n-b$ nonmutant descendants to assign to $j-1$ nonmutant ancestors. The quantity $p_{n,j}(b)$ is therefore given by

$$p_{n,j}(b) = \frac{\binom{n-b-1}{j-2}}{\binom{n-1}{j-1}}.$$

To obtain information about the age of the mutation from these equations, we introduce $n(t)$, the number of ancestors of the n current alleles in the total population at time t generations before present. The maximum number of ancestors at time t is the total population size $N_P(t)$ at that time. When J_0, the number of ancestors at the time the mutation arose, is greater than $n(t)$, the disease mutation must be younger than t. However, when the mutation is younger than t, J_0 must be greater than or equal to $n(t)$. In other words, the probability that the mutation is younger than t is at least $\mathrm{P}\{J_0 > n(t)\}$, and at most $\mathrm{P}\{J_0 \geq n(t)\}$. Solving these equations for t yields bounds for the α^{th} quantiles t_α of the age distribution. These bounds are typically quite narrow, provided $n(t_\alpha)$ is not too small. For example, solving $\mathrm{P}\{J_0 \geq n(t)\} = 0.5$ and $\mathrm{P}\{J_0 > n(t)\} = 0.5$ gives a lower and upper bound, respectively, on the median age of the mutation. These probabilities can be expanded and evaluated empirically. One possible expansion is

$$\mathrm{P}\{J_0 > n(t)\} = \sum_{i=1}^{N_P(t)} \mathrm{P}\{J_0 > i \mid n(t) = i\}\mathrm{P}\{n(t) = i\} = \sum_{i=1}^{N_P(t)} \mathrm{P}\{J_0 > i\}\mathrm{P}\{n(t) = i\}.$$

The events $\{n(t) = i\}$ and $\{J_0 > i\}$ are independent because the number of ancestors at time t offers no information about the relative lengths of coalescent times and hence about the number of ancestors J_0 when the mutation arose.

The following two examples apply these formulae to IOSCA in Finns and Werner's syndrome in Japanese, respectively.

2.8.1 *IOSCA in Finns*

Expected coalescence times for equation (2.19) were evaluated through simulations of 1000 coalescent replicates. The size of the Finnish population was assumed to be constant at $N_P(t_f) = 2000$ copies prior to founding at $t_f = 80$ gbp. For more accuracy at small numbers j of lineages, the following relation was used:

$$\mathrm{E}\, T_j = \mathrm{E}\,(T_j \mid S_{j+1} \geq t_f)\,\mathrm{P}\,(S_{j+1} \geq t_f) + \mathrm{E}\,(T_j \mid S_{j+1} < t_f)\,\mathrm{P}\,(S_{j+1} < t_f)\,.$$

where $n = N_P(0)$ is the total number of Finns at present, $S_j = \sum_{i=j}^{n} T_i$, and $\mathrm{E}\,(T_j \mid S_{j+1} \geq t_f) = 2N_P(t_f)/(j(j-1))$. The probability that IOSCA is younger than the $t_f = 80$ generations is at least 0.86, and the median age of the mutation is at most 68 generations. A younger median age than founding is consistent with the disease being unobserved outside Finland. These calculations show that a single selectively neutral mutation at founding would in fact be expected to lead to a copy number greater than the 10000 currently observed for IOSCA. There is thus no need to invoke selection in order to explain the observed current copy number. Figure 2.18 shows the empirical distribution of the number of ancestors $n(t_f)$ at founding obtained by evaluating

$$\mathrm{P}\,(n(t_f) = j) = \mathrm{P}\,(S_{j+1} \leq t_f) - \mathrm{P}\,(S_j \leq t_f)$$

in the simulations. At a single locus, all 10^7 current Finnish copies are estimated to descend from between 400-450 founding alleles, roughly 20-23% of the assumed 2000 originals.

2.8.2 *Werner's syndrome in Japanese*

The same methods adapted to a larger population were used to estimate the median age of WRN4, the most common mutation for Werner's syndrome in the Japanese. As described in section 1.4.2, we have assumed a present WRN4 copy number of 500,000, and a total Japanese copy number of approximately 240×10^6.

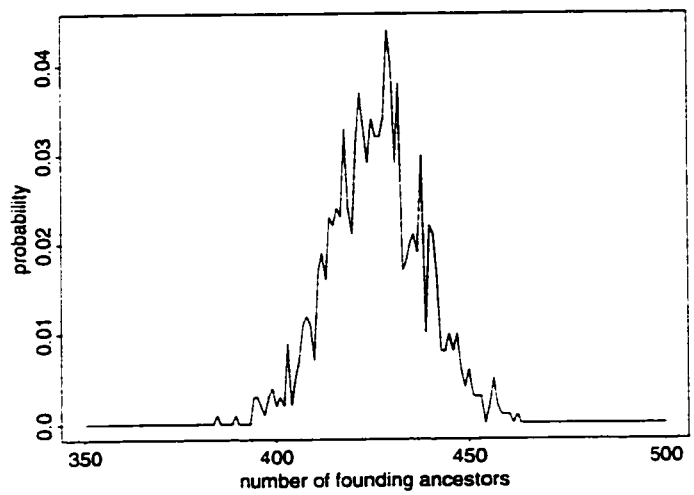


Figure 2.18: Empirical distribution of the number $n(80)$ of ancestral founding copies in Finns. The empirical distribution is based on 1000 ancestral replicates.

Simulations under the replacement hypothesis (section 1.4.2) show that the chance that WRN4 is younger than founding is at least 0.50. Thus, the median age is at most the assumed $t_f = 94$ generations. Not suprisingly, under a birth-and-death process model of WRN4 reproduction, the assumed current number of WRN4 alleles is about the same as the expected number given survival to the present. The age of WRN4 under a transformation hypothesis (section 1.4.2) was also investigated. Under this hypothesis, the modern Japanese were founded by a small number, say 1000, of Jomon individuals arriving roughly 400 generations ago. Allele copy numbers at 94 gbp are thus different under the transformation and replacement hypotheses. Simulations under the transformation hypothesis show that the chance that WRN4 is younger than 400 generations is small, and give a median age of about 411 generations.

As for the Finnish example, simulations assume a population of constant size before founding. However, with the Japanese, the probability estimates themselves, rather than the expected values in the formula (2.19), are Monte Carlo-based; i.e., observed branch lengths take the place of expected branch lengths and the resulting

conditional probability is averaged over ancestral realizations. Monte Carlo estimates of the probability were used because the large present copy number in Japanese leads to computational underflow problems when calculating expected coalescence times. The numerous tip branches on the coalescent are far too small, particularly in the most recent 5 generations of rapid growth. The probability that the mutation is younger than 5 generations is negligible since there are now 500,000 copies. Thus, in the most recent 5 generations, a reversed Wright-Fisher model has been used to scale down the copy number for each ancestral realization.

Hereafter, we assume the replacement hypothesis of Japanese origins, and fix a single copy of the WRN4 allele in a founding Yayoi population of 2000 alleles in 350 B.C. (94 generations ago).

Chapter 3

RECOMBINATION ON THE COALESCENT

As noted in section 1.2, compared to a pedigree, there is a large number of meioses on the ancestral coalescent. This greatly increases the chance of a recombination event between the disease and marker, and allows for a much finer scale of mapping. In this chapter, we discuss the process of recombination on the ancestral coalescent, and the resulting marker identity by descent in the disease sample. This identity by descent gives rise to disease associations at closely-linked markers.

3.1 Single marker recombinant classes

Consider a marker at a fixed small recombination fraction r from the disease mutation. Once the coalescent of the disease sample is realized, using methods outlined in section 2.6, recombination events between the disease and marker locus may be placed on the ancestral tree, branch-by-branch, as follows. We take a branch of the ancestry of length G generations to represent G meioses and therefore G opportunities for recombination between the disease and marker. The probability of at least one recombination event on the branch is thus $1-(1-r)^G \approx 1-e^{-Gr}$. This overall approach of realizing the coalescent of the sample of disease alleles and then placing recombination events on it is the same as that taken by Thompson and Neel (1997). Tree shape dictates where recombination events are most likely to occur on the ancestry. Define the tip branches of an ancestral tree to be the K most recent branches tracing back in time from the K sampled allelic copies. Consider the conditional probability of a recombination event on the tip branches, given that a recombination event occurs.

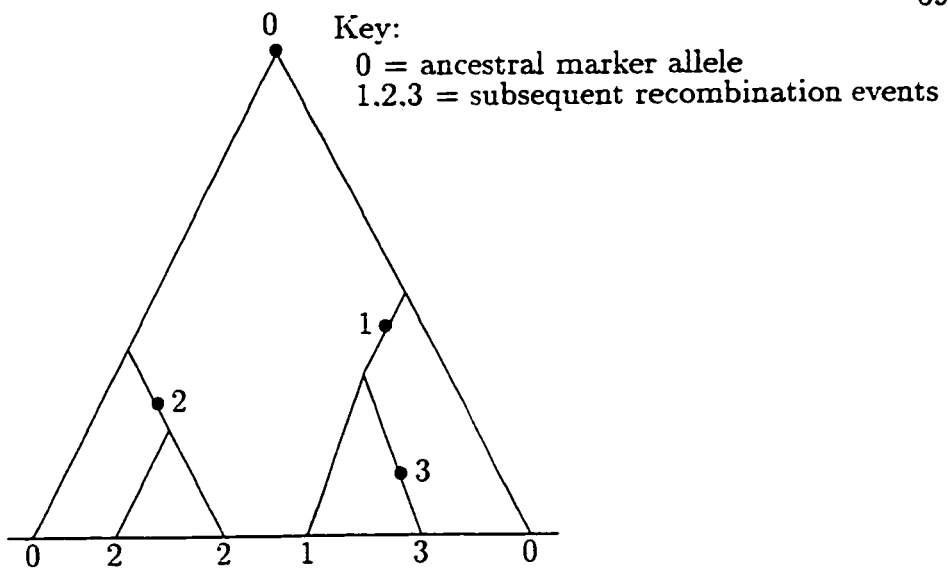


Figure 3.1: Definition of the recombinant classes of the sampled disease haplotypes, with reference to a single linked marker. There are $K = 6$ disease haplotypes, with identifiers 0,2,2,1,3,0 respectively. Thus there are four recombinant classes, two of size 1 and two of size 2.

This probability changes with the rate of growth of the disease population because the shape of the ancestral tree changes. In rapidly-growing disease populations, more of the total length of an ancestral tree is comprised of the tip branches. Thus, given that a recombination event occurs, there is more chance that it is on the tip branches in a rapidly-growing disease population than in a slowly-growing population.

At a marker locus, we define a *recombinant class* to be a subset of the current sample of disease haplotypes descended from a given recombination event. As shown in Figure 3.1, these single-marker recombinant classes partition the sample, all members within a class being identical by descent at the marker locus, excluding marker mutation. Thus all members of each recombinant class share a marker allele. The marker allelic types of different recombinant classes are drawn independently according to population marker allele frequencies. The marker allelic type of the ancestral recombinant class has the same distribution as the other recombinant classses, since

the disease mutation occurs randomly on a background haplotype. When generating single-marker recombinant classes, we need only consider the presence or absence of recombination events on a given branch of the ancestral tree. For single-marker mapping, multiple recombination events on the branch have the same effect on recombinant classes as a single recombination event. (However, multiple recombination events on a branch, occuring in separate genomic segments defined by multiple markers, do affect multi-marker recombinant classes. However, the extension of the concept of a recombinant class to multiple markers is deferred to section 6.2.) Of course, recombinant classes are latent variables: they cannot be observed, since no marker is infinitely polymorphic.

At any given marker, the size and number of recombinant classes reflects the underlying ancestral tree shape and length. For example, fast-growing disease populations with shorter ancestral trees should retain more ancestral marker alleles than slower-growing populations. Although early recombination events will be rare on these shorter ancestral trees, any that do occur will have less chance than those on longer ancestral trees of being followed by recombination events closer to the tips. Thus, nonancestral recombinant classes formed by early recombination events in rapidly-growing variant populations will, when they occur, tend to be bigger than those in disease populations of more stable size.

Figure 3.2 summarizes simulations at recombination fraction $r = 0.01$ for two disease populations of current size $N_D(0) = 10000$, growing deterministically at exponential rates $m = 1.12$ and $m = 1.01$. Simulations are based on 1000 realizations of a coalescent for a sample of $K = 20$ disease alleles. Both the increased number of ancestral marker alleles and the increased size of recombinant classes formed by early fortuitous recombinations are evident for the faster growing population. In Figure 3.3, disease growth is fixed at rate $m = 1.12$, approximating IOSCA in the Finns, as described in section 1.4.1. The figure is based on 2000 coalescent realizations. For closer markers, the number of ancestral alleles is increased and the number of recombinant

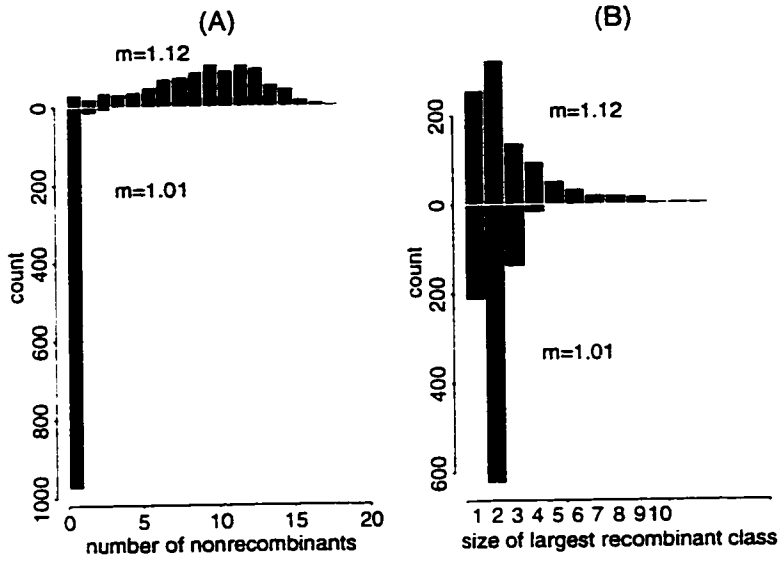


Figure 3.2: Effect of deterministic growth of the disease population on recombinant classes. There are $K = 20$ disease haplotypes sampled from a current disease population of size $N_D(0) = 10000$. The disease and marker loci are separated by a recombination fraction of $r = 0.01$. Light shaded bars correspond to a disease growth rate of $m = 1.12$; dark shaded bars correspond to $m = 1.01$. (A): Distribution of number of nonrecombinants. (B): Size of largest nonancestral recombinant class. Results are based on 1000 coalescent realizations.

classes is decreased compared to farther markers. Nonancestral recombinant classes are also less numerous but, when they occur, they tend to be larger than those for distant markers. This is due to recombinant classes formed by early recombination events which are not followed by subsequent recombinations. However, in the limit, as the recombination fraction $r \to 0$, the probability of recombination and hence nonancestral recombinant classes goes to zero. Simulations for Figure 3.3 use recombination fractions $r = 0.01$ and 0.03, and are based on 2000 coalescent realizations.

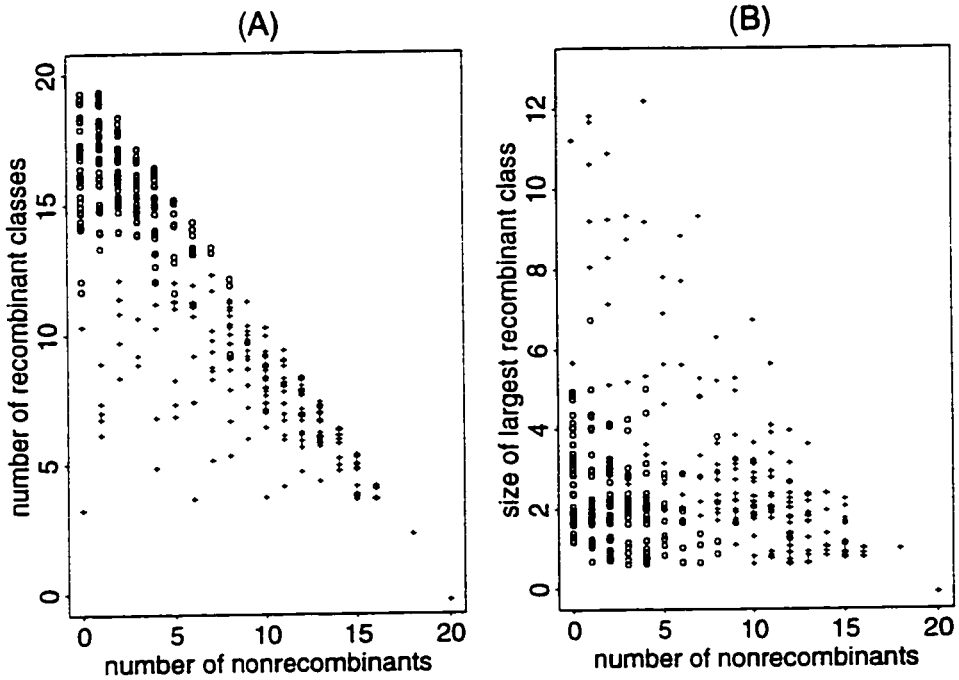


Figure 3.3: The effect of reduced recombination fraction on recombinant classes. There are $K = 20$ disease haplotypes sampled from a disease population of size $N_D(0) = 10000$. The disease population is assumed to grow deterministically at exponential rate $m = 1.12$. Recombination fractions of 0.01 (+) and 0.03 (o) are considered. (A): Joint distribution of number of nonrecombinants and recombinant classes. (B): Size of largest nonancestral recombinant class versus number of nonrecombinants.

Large recombinant classes result from early recombination events which are not

followed by recombinations closer to the tips of the ancestral tree. Ancestral trees with longer root than tip branches would therefore be expected to yield large recombinant classes for marker loci lying within a certain range of the disease locus. Markers within this range must be close enough to the disease mutation so that early recombinations on long root branches are not followed by further recombinations. However, the markers must not be so close to the disease that recombination is precluded. The longer the tree, the closer the marker needs to be in order to achieve a large nonancestral recombinant class, since too high a recombination rate leads to subsequent fragmentation of any early recombinant classes. Figure 3.4 confirms that, at small recombination fractions, the probability of a large recombinant class is increased for more stable compared to rapidly-growing disease populations. (Here "large" is defined as bigger than 20% of the sample.) The figure also illustrates that a single marker provides limited information for differentiating between multiple mutations at the disease locus and multiple large recombinant classes; this point will be discussed further in section 7.2.1. Even disease populations with relatively short ancestral trees, such as IOSCA in Finns, have a nonneglible probability of a large nonancestral recombinant class for some values of the recombination fraction.

3.2 Conserved ancestral region

By assumption, sampled disease alleles are identical by descent with respect to the original disease mutation. As meioses accumulate on the ancestry that relates the sample, recombination events fragment the ancestral haplotype, reducing the length of the core genomic region shared identical by descent. The length of this conserved ancestral region is thus another consequence of recombination on the coalescent. Assuming no interference, recombination along the chromosome may be modelled as a Poisson process with rate 1 per Morgan (Haldane 1919). Under this model, the length in cM of the conserved region, given the number n of meioses on the coalescent, is a

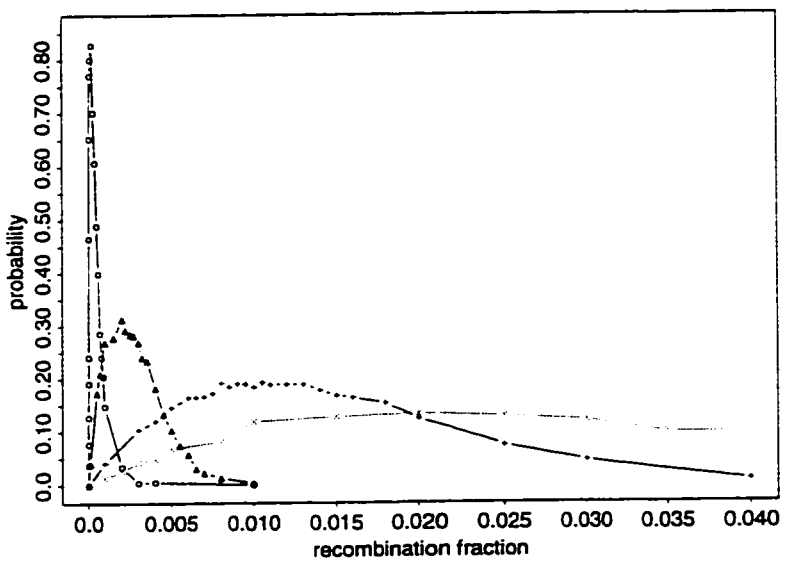


Figure 3.4: Probability of a large nonancestral recombinant class: N=10000, K=20. Solid lines are deterministic growth at rates of 1.00 (o), 1.01 (Δ) and 1.1 (+). Dotted line ($\times$) is stochastic variant growth of IOSCA in Finns. Results are based on 2000 coalescent realizations.

gamma random variable with shape parameter 2 and rate $n/100$ (Boehnke 1994). This calculation reflects the fact that the coalescent provides an implicit pedigree whose total branch length gives the number of meioses separating sampled copies. Figure 3.5 summarizes the empirical distribution of the conserved length for a sample of $K = 10$ disease alleles taken from a disease population of current size $N_D(0) = 10000$. The empirical distribution is based on 2000 coalescent replicates. The 5th and 95th percentiles, median, and mean length of the conserved region are plotted for IOSCA growing stochastically in Finns, and for disease populations undergoing deterministic exponential growth by a factor of $m = 1.05, 1.12$, and 1.5 per generation. The deterministic growth rate $m = 1.12$ has been selected to match the current IOSCA copy number, assuming the disease mutation was introduced into the population at founding 80 gbp. Under deterministic growth at this rate, the mean length is about 0.40 cM. The mean number of meioses on the ancestry is about 474. Under stochastic

growth, the mean length is increased to about 0.75 cM. and the expected number of ancestral meioses is about 271. The number of meioses is reduced because stochastic fluctuation reduces the coalescence times, as discussed in section 2.6. Increasing the growth rate decreases the total number of meioses separating sampled copies, leading to a length of conserved region of increased mean and variance. However, even under rapid growth of the disease, such as $m = 1.5$, the expected number of ancestral meioses is about 163, and the mean length is 1.2 cM. Given the large variance of this shared length (the 5th and 95th percentiles of the distribution differ by a factor of about 15), a wide range of outcomes is plausible for a given disease across replicate populations. Hence, genome screening of disease haplotypes may prove useful for some diseases in some populations, but not for others.

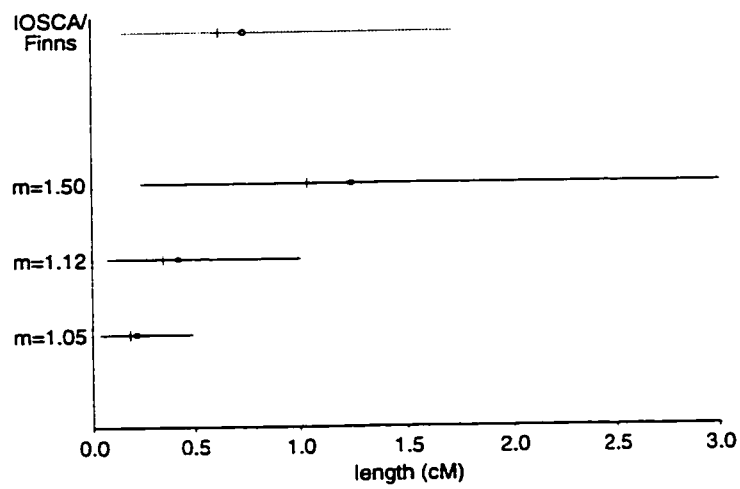


Figure 3.5: 5th and 95th percentiles (ends of horizontal lines), median (|) and mean (o) length of conserved ancestral region based on 2000 replicates; $K = 20$, $N_D(0) = 10000$. Solid lines are deterministic growth at growth rates 1.5, 1.12 and 1.05. Dotted line is for stochastic growth corresponding to IOSCA in Finns.

Houwen et al. (1994) successfully applied such a screening strategy to identify

candidate genomic regions for benign recurrent intrahepatic cholestasis (BRIC) in an remote Netherlands fishing community (see section 1.4.3). Suppose that two unrelated affected individuals are observed. If the disease variant has grown exponentially at constant rate, simulations show that the length of the conserved region shared by the $K = 4$ disease haplotypes is expected to be about 5.9 cM, and that 90% of the time this length will be somewhere in the range of 0.9-14.6 cM. If the disease variant has grown stochastically, the expected length is increased to 7.1 cM, with a 90% chance of falling in the interval 1.3-17.0 cM. Again, the approximate 15-fold difference between the 5th and 95th percentiles indicates large variability in the outcomes across replicate populations. Comparison of mean lengths in this example and the example of IOSCA in Finns indicates that the impact of stochastic growth in disease copy number in this rapidly-expanding community is less than the impact in the more slowly-growing Finnish population. In the Finnish example, the mean length under deterministic disease growth is 53% of the mean length under stochastic growth. In the example for the Netherlands fishing community, the mean length under deterministic growth is 83% of that under stochastic growth.

3.3 Sources of variability

This section investigates the variance contribution of the coalescent. The variance of a random function $W = W(\mathbf{A}, \mathbf{R})$ of the ancestral process $\mathbf{A}$ and recombination process $\mathbf{R}$ may be decomposed as

$$\mathrm{V}(W) = \mathrm{E}\,\mathrm{V}(W \mid \mathbf{A}) + \mathrm{V}\,\mathrm{E}(W \mid \mathbf{A}).$$

We interpret the component $\mathrm{V}\,\mathrm{E}(W \mid \mathbf{A})$ as the variance due the ancestral coalescent. Two random quantities W are considered: the effective number of recombination events on the ancestral coalescent of the disease sample, and the length of the conserved ancestral haplotype. The effective number of recombination events is defined to be the number of branches of the ancestral coalescent with at least one recombination

event in a specified chromosome segment. Thus if the number of disease haplotypes sampled is K, an upper bound for this effective number is $2(K-1)$, the number of branches on the tree. For this investigation, we make the simplifying assumption that the disease population grows deterministically at constant exponential rate. This deterministic rate is calculated assuming one copy of the disease allele at founding, and $N_D(0)$ copies at present. The rate therefore smooths over the expected patterns of disease growth, given survival to the present. These growth patterns are investigated in section 2.5.2. Typically, the deterministic rate is higher than the overall growth rate of the total population. Conditional on survival of the disease mutation, there is rapid initial growth in the expected number of disease copies. Once sufficient copy number is achieved, however, the expected size of the disease subpopulation increases at rates similar to the total population.

Hence, for a fixed sample of K disease haplotypes, the relevant parameters are the disease growth rates m, and the recombination fractions r. Our interest is in parameter values appropriate to fine-scale mapping in human populations; for r, this implies values less than 0.05. For disease growth rates m, a conservative lower bound is $m = 1.01$, reflecting the rapid initial growth expected for surviving disease mutations, and the expansion of human populations after the development of agriculture. Indeed, for populations founded relatively recently, such as the Finns and Japanese, we would not expect rates to be much less than $m = 1.03$.

Figure 3.6 summarizes the percentage of the total variance due to the coalescent in the effective number of recombinations. Results are for a sample of $K = 50$ haplotypes drawn from a disease population of current size $N_D(0) = 10000$, and are based on 1000 coalescent realizations. Figure 3.6(A) presents results for growth rates within the range of interest, with the recombination fraction fixed at $r = 0.02$. Figure 3.6(B) gives results for recombination fractions appropriate to fine-mapping, with the growth rate fixed at $m = 1.04$. Most of the variability is attributable to the recombination process. For example, when $r = 0.02$ the variance contribution of the coalescent is

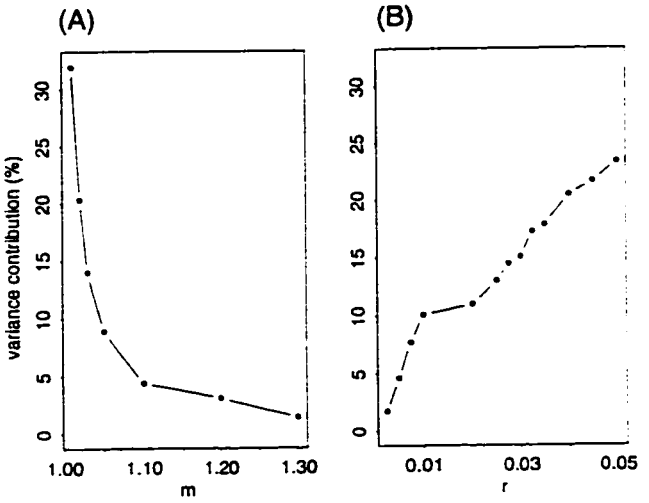


Figure 3.6: Variance contribution of the coalescent, as a percentage of total variance, for the effective number of recombinations on the ancestral tree of $K = 50$ haplotypes, drawn from a disease population of current size $N_D(0) = 10000$. Variance contribution as a function of (A) the deterministic rate m of disease growth for a fixed recombination fraction $r = 0.02$, and (B) the recombination fraction r for fixed growth rate $m = 1.04$. Curves are based on 1000 coalescent realizations.

at most 32% at $m = 1.01$, and decreases rapidly as the growth rate increases. When the growth rate is fixed at $m = 1.04$, and r varies, the variance contribution of the coalescent is no more than 24% at $r = 0.05$.

The variance contribution of the coalescent is even smaller when considering the conserved ancestral region in sampled disease haplotypes. The comparison involves only the growth rate, and for all values within the relevant range, the variance contribution is less than 1%. Taken together, these results suggest that the recombination process rather than the coalescent process dominates the variability of the random quantities that are of interest in disequilibrium fine-mapping with human populations. As a consequence, disequilibrium analyses which ignore variation in coalescent times may nevertheless lead to reasonable inferences about disease location, particularly when mapping at scales of $r \leq 0.02$ in young expanding populations. In other words, as long as expected coalescent times are about right, the exact model for these times is relatively unimportant.

Chapter 4

RECOMBINANT CLASSES TO HAPLOTYPES

Barring natural selection and population stratification, marker identity by descent underlies linkage disequilibrium. We have shown that in a random sample of disease haplotypes, identity by descent at a single marker is summarized by the recombinant classes. However, it is marker allelic classes rather than recombinant classes that are observed. In the absence of marker mutation on the coalescent, the allelic classes at a given marker are the sets of sampled disease haplotypes bearing marker alleles of the same type. Thus, because of their identity by descent, all marker alleles within a recombinant class belong to the same allelic class. In addition, a single allelic class may be comprised of several recombinant classes. To see this, consider the recombination events that define two recombinant classes. Under random mating, disease haplotypes recombine with random haplotypes from the population. If the marker alleles on these randomly-selected haplotypes are the same, the marker allelic type for both recombinant classes will be the same. The allelic type of a recombined marker allele is random, and reflects population marker allele frequencies at the time of the defining recombination. Throughout, we make the simplifying assumption that these frequencies remain constant over disease history. This chapter reviews the connection between single-marker recombinant classes and disease associations, and then investigates the power and level of tests of association. Association detection may be contrasted with estimation of disease location and accuracy of mapping, topics which are pursued in chapter 5.

4.1 *Single marker associations*

The connection between recombinant classes and disease associations is best illustrated by the following simple example. Consider a diallelic marker which has equifrequent alleles in the population. As shown in Figure 3.1, suppose that in a sample of $K = 6$ disease alleles there are two recombinant classes of size 2 and two of size 1. To obtain the haplotypes, marker alleles are randomly and independently assigned to each recombinant class, according to their population allele frequency. The ancestral recombinant class is no different from the others because the mutation arises on a random background haplotype. For example, in one possible assignment, both of the larger recombinant classes could end up with the first marker allele, and the remainder could end up with the second. This would lead to a slight association. Out of 6 sampled haplotypes, $2 + 2 = 4$ or 67% would carry the first marker allele, in contrast to the 50% expected in the general population.

4.2 *Power to detect association*

To investigate the power to detect association, recombinant classes were generated under varying recombination rates, for samples of varying size, drawn from disease populations growing at varying rates. Marker allele types were then assigned to these recombinant classes, according to population frequencies. Marker polymorphism also varied, but for simplicity all markers had equifrequent alleles. To detect association, marker allele counts in a sample of $K = 25$ disease haplotypes were compared to those in a control sample of 100 alleles via chi-squared tests with significance level $\alpha = 0.05$.

Figure 4.1 plots the power to detect an association as a function of marker polymorphism for different rates of disease growth. The results are based on 2000 coalescent realizations from a disease population of current size $N_D(0) = 10000$. The marker is at recombination fraction $r = 0.01$ from the disease locus. Markers with up to

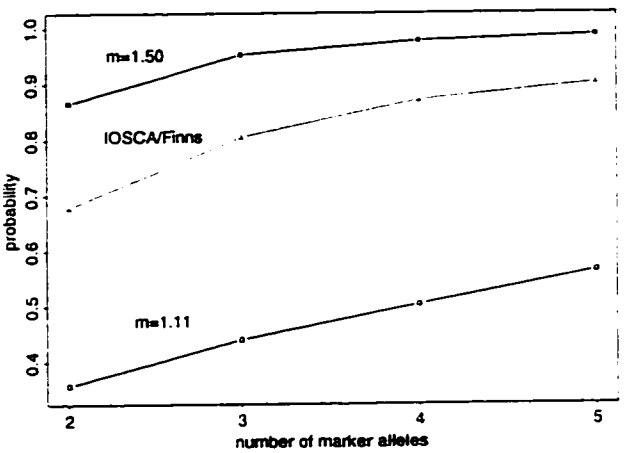


Figure 4.1: Probability of detecting association at $r = 0.01$ as a function of marker polymorphism. There are $K = 20$ disease alleles and 100 control alleles. Disease alleles are sampled from a disease population of current size $N_D(0) = 10000$. Results are based on 2000 coalescent realizations. Solid lines are deterministic disease growth at rates of $m = 1.50$ and $m = 1.12$; dotted line is stochastic growth of IOSCA in Finns. Expected number of nonrecombinants: 16.8 for $m = 1.50$, 6.9 for $m = 1.11$, 12.6 for stochastic growth.

5 equifrequent alleles are considered. The sample size of $K = 25$ ensures approximately valid chi-squared tests because the expected number of copies of each allelic type is at least 5 under no association. However, general conclusions would not be altered by using smaller sample sizes and exact permutation tests. Power curves are shown for samples taken from disease populations growing deterministically at constant exponential rates of $m = 1.5$ and 1.12, and stochastically like IOSCA in Finns (see section 1.4.1). The deterministic exponential growth rate of $m = 1.12$ has been selected to match current IOSCA copy number, assuming that the disease mutation was introduced as a single copy at founding 80 gbp. Note that the size of the disease sample ($K = 25$) is slightly high for IOSCA. The IOSCA allele frequency of about 1/1000 in the current Finnish population of 5×10^6 persons implies that between 0-10 nonconsanguineous homozygotes, or a sample of 0-20 disease alleles, would be expected with 95% probability.

When the growth of disease copy number is deterministic, at constant exponential rate $m = 1.12$, the number of nonrecombinants tends to be lower than when growth is stochastic. Under deterministic growth at rate $m = 1.12$, the expected number of nonrecombinants is 6.89, and under stochastic growth it is 12.61. There are fewer recombination events under stochastic growth because the ancestral tree is shorter. Stochastic growth of disease copy number reduces coalescence times, as discussed in section 2.6. Expected total tree length is 589 generations under stochastic growth, and 991 generations under $m = 1.12$. Similarly, the expected time to most recent common ancestor is 35 generations under stochastic growth, and 64 generations under $m = 1.12$. When disease copy number grows rapidly at the deterministic exponential rate $m = 1.5$, the present disease copy number of $N_D(0) = 10000$ implies a single disease mutation at rougly 23 gbp, rather than 80 gbp. Ancestral trees tend to be even shorter under $m = 1.5$ than under stochastic growth of IOSCA, with an expected total length of 351 generations, and an expected TMRCA of 20 generations. The expected number of nonrecombinants is correspondingly increased to 16.76.

Not suprisingly, the ability to detect an association is greatest when the number of nonrecombinants is highest, at high rates of disease growth. The power curve for stochastic growth of IOSCA is closer to the curve for $m = 1.5$ than to the curve for $m = 1.12$, reflecting the the more similar expected total lengths of the ancestral trees. Figure 4.1 also shows that power increases with marker polymorphism, particularly when the disease grows more slowly, and the number of nonrecombinant disease haplotypes is reduced.

Power was also evaluated empirically for a diallelic marker with equifrequent alleles, but this time as a function of recombination fraction. All other parameters remained the same, except that a disease sample of more realistic size $K = 10$ for IOSCA was drawn. A chi-square test of association was used. The expected number of copies of each allelic type under no association is 5, and so the asymptotic chi-square distribution is reasonable. Empirical probabilities of detecting association are

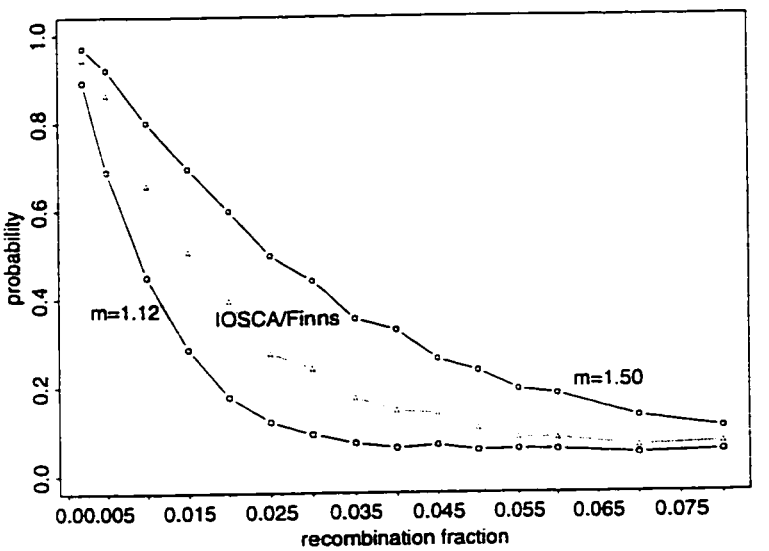


Figure 4.2: Probability of detecting association as a function of recombination fraction for a diallelic marker with equifrequent alleles. There are $K = 10$ disease alleles and 100 control alleles. Haplotypes are sampled from a disease population of current size $N_D(0) = 10000$. Results are based on 2000 coalescent realizations. Solid lines, deterministic disease growth: dotted line, stochastic growth of an IOSCA-like mutation in Finns.

summarized in Figure 4.2. Results are based on 2000 coalescent replicates. Power increases as the recombination fraction decreases, regardless of disease growth, but drops off faster for lower growth rates. Under stochastic growth of IOSCA, association may be detected at least half the time with diallelic markers up to about $r = 0.015$ from the disease locus. This upper detection limit is close to the $r = 0.01$ limit previously proposed (Kaplan et al. 1995, Boehnke 1994). When $K = 10$, the power curve under stochastic disease growth is roughly midway between the curves under deterministic growth at rates $m = 1.5$ and 1.12. When $K = 10$, the expected total tree length under stochastic growth is about 271 generations, roughly halfway between 163 and 474 generations, the expected lengths under $m = 1.5$ and 1.12, respectively.

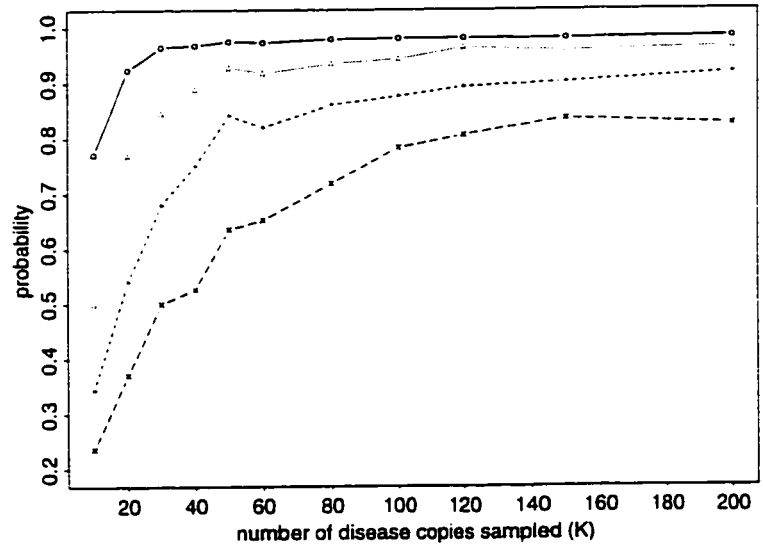


Figure 4.3: Probability of detecting association as a function of disease sample size K, for a WRN4-like mutation in Japanese. A diallelic marker with equifrequent alleles is used. There are 100 control alleles. Results are based on 1000 coalescent realizations. Curve for $r = 0.005$, solid line (o); $r = 0.01$, dotted line (Δ); $r = 0.015$, short-dashed line (+); and $r = 0.02$, long-dashed line (×).

The power as a function of K, the number of sampled disease alleles, is also of interest. Figure 4.3 displays empirical curves for different recombination fractions.

The data are simulated to resemble a sample of WRN4 copies from the Japanese, under the replacement hypothesis described in section 1.4.2. WRN4 is assumed to have grown stochastically from a single allele at founding 94 gbp. A diallelic marker with equifrequent alleles, and a control group of 100 alleles are used. Results are based on 1000 coalescent realizations. Probabilities range from $0.2 - 1.0$, and so the Monte Carlo standard error is expected to be as high as 0.016. The plot shows that within $r = 0.005$ of the disease mutation there is reasonable power (i.e., $\geq 80\%$) to detect association with as few as $K = 25$ disease copies. When sample size is increased to $K = 30$, association may be reliably detected up to $r = 0.01$ from the mutation. However, even with diallelic markers as far away as $r = 0.02$, reasonable power can still be achieved provided about $K = 150$ or more disease alleles are sampled.

4.3 False-positive rates

When analysing contingency tables in association studies, investigators often reduce the number of allelic categories, and the sparseness of the table, by pooling all alleles but the one or two that are most frequent in the disease sample. The idea is that the most frequent alleles in the disease sample represent the ancestral alleles for one or two major mutations. Applying a chi-squared test to such a collapsed contingency table would be expected to inflate type I errors (Goddard et al. 1996, Terwilliger 1995). As pointed out by J. Felsenstein (personal communication), comparing this chi-squared statistic to a reference distribution with $r + c - 3 = r - 1$ degrees of freedom, where r is the number of alleles and $c = 2$, yields a conservative test. The test is conservative because grouping rows and columns of the contingency table to maximize the chi-square statistic leads to a chi-square distribution with degrees of freedom $r + c - 3$ (Williams 1952).

Figure 4.4 illustrates the effects of selective grouping under no association. Chi-squared tests are used to compare 100 disease haplotypes to 200 control haplotypes.

Results are based on 15000 and 300000 realizations, for tests which pool all alleles but the one and two most frequent in the disease sample, respectively. Type I error can be inflated up to five times the nominal rate when the two most frequent alleles in the disease sample are retained in the contingency table, and up to three times the nominal rate when only the most frequent allele is retained. These results are reflected by the chi-squared distribution of the conservative test statistic, which has increasing degrees of freedom in the number of marker alleles.

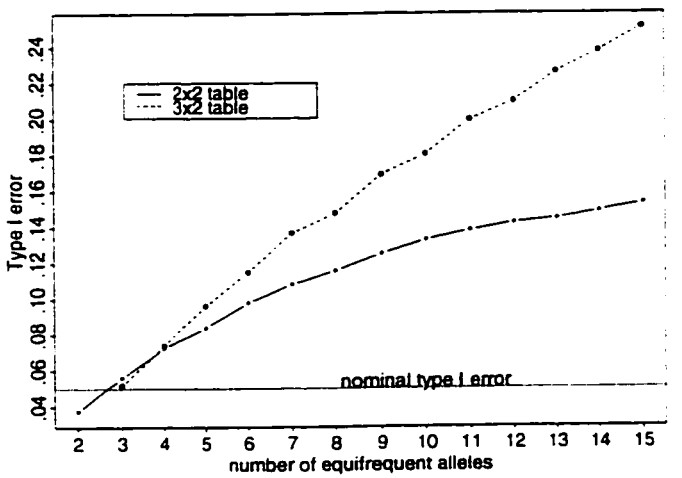


Figure 4.4: Type 1 error incurred from selective grouping on all but the one (2×2 table, solid line) or two (3×2 table, dashed line) most frequent alleles in the disease sample. There are 100 sampled disease haplotypes and 200 control haplotypes. Yates continuity correction is used for the 2×2 table. Results are based on 15000 and 300000 realizations for the 2×2 and 3×2 tables, respectively.

Chapter 5

SINGLE-MARKER MAPPING

This chapter develops a linkage likelihood for disease location based on marker *identity by descent*, as summarized by recombinant classes. Our approach thus differs from previous likelihood approaches, all of which are formulated in terms of marker *identity by state*. Most of these likelihood methods, including ours, involve simulation of disease haplotype histories or ancestries. Kaplan et al. (1995) provided the first linkage likelihood. They use forwards simulation of disease history, with rejection sampling of histories which do not lead to current disease allele counts within a specified range. Rannala and Slatkin (1998) simulate ancestries of disease alleles, conditional on the current disease allele count and the sampling of a specified number of disease haplotypes. To bypass simulation of disease histories and the resulting marker allele frequencies, Xiong and Guo (1997) developed an approximate likelihood. The approximation uses the first and second moments of these allele frequencies, calculated under a Wright-Fisher population genetic model. However, moments are not conditional on current disease allele count.

5.1 Notation

Suppose K disease-bearing haplotypes are sampled from a population whose growth is described by known demographic parameters $\Delta = (t_f, \{\lambda(t) : 0 < t \leq t_f\})$, where time t_f and population growth rates $\lambda(t)$ are as defined previously. The recombinant class information on sampled haplotypes can be summarized as $\mathbf{X}$, a vector of recombinant class counts indexed by the size of the recombinant class (Figure 3.1). The element

X_i of $\mathbf{X}$ is the number of recombinant classes of size i, $1 \leq i \leq K$. The labelling of the recombinant classes themselves is irrelevant.

Consider an m-allele marker at recombination frequency r from the disease locus. The population allele frequencies at the marker locus, $\mathbf{q} = (q_1, q_2, \ldots, q_m)$, are assumed to have remained constant over time. Let $\mathbf{Y}$ be the vector of marker allele counts for the sample; Y_j is the number of sampled disease copies carrying marker allele j, $j = 1, \ldots, m$.

5.2 *Single-marker likelihood*

The probability for the data $\mathbf{Y}$, or likelihood for the recombination frequency r, can be written

$$L(r) \;=\; \mathrm{P}_{\mathbf{q},r,\Delta}(\mathbf{Y}) = \sum_{\mathbf{X}} \mathrm{P}_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})\, \mathrm{P}_{r,\Delta}(\mathbf{X}) \tag{5.1}$$

where $\mathrm{P}_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$ is the conditional probability of the data $\mathbf{Y}$ given $\mathbf{X}$, the marker identity-by-descent information, and $\mathrm{P}_{r,\Delta}(\mathbf{X})$ is the probability of $\mathbf{X}$. Equation (5.1) uses the patterns of marker identity by descent to partition the likelihood. As noted by Thompson (1997), partitioning on gene identity by descent forms the basis of inference in genetic linkage analysis, whether the data arise from pedigree, sib-pair, or, in our case, population-genetic studies.

The identity-by-descent information $\mathbf{X}$ is a function of the coalescent ancestry, $\mathbf{A}$, and the recombination events, $\mathbf{R}$, on its branches:

$$L(r) \;=\; \mathrm{P}_{\mathbf{q},r,\Delta}(\mathbf{Y}) \;=\; \sum_{(\mathbf{A},\mathbf{R})} \mathrm{P}_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}(\mathbf{A},\mathbf{R}))\, \mathrm{P}_{r}(\mathbf{R} \mid \mathbf{A})\, \mathrm{P}_{\Delta}(\mathbf{A}). \tag{5.2}$$

As discussed in chapter 2, the ancestry is, in turn, a function of the disease population size $\mathbf{N_D} = \{N_D(t), 0 \leq t \leq t_f\}$ over disease history. Thus, we may rewrite equation 5.2 more fully as:

$$\begin{aligned} L(r) \;&=\; \mathrm{P}_{\mathbf{q},r,\Delta}(\mathbf{Y}) \\ &=\; \sum_{(\mathbf{A},\mathbf{N_D},\mathbf{R})} \mathrm{P}_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}(\mathbf{A}(\mathbf{N_D}),\mathbf{R}))\, \mathrm{P}_{r}(\mathbf{R} \mid \mathbf{A}(\mathbf{N_D}))\, \mathrm{P}(\mathbf{A} \mid \mathbf{N_D})\, \mathrm{P}_{\Delta}(\mathbf{N_D}). \end{aligned}$$

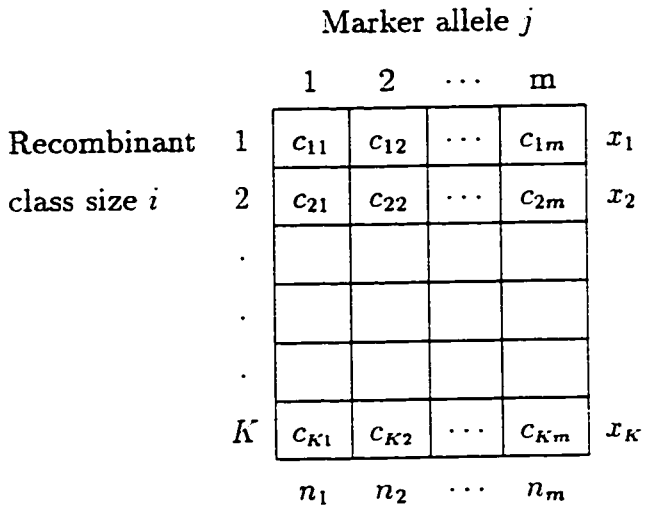


Figure 5.1: Notation for configuration $C = \{c_{ij}\}$ of recombinant classes. Note that $y_j = \sum_{i=1}^{K} i\, c_{ij}$.

Equation (5.1) suggests Monte Carlo evaluation of the likelihood for r given observed allelic classes $\mathbf{Y} = \mathbf{y}$ by sampling recombinant classes $\mathbf{x}$ from $\mathrm{P}_{r,\Delta}(\mathbf{X})$ and averaging $\mathrm{P}_{\mathbf{q}}(\mathbf{y} \mid \mathbf{x})$, over the realized values $\mathbf{x}$ of $\mathbf{X}$. Realizations of $\mathbf{X}$ are obtained by Monte Carlo simulation, as outlined in the previous section. (By contrast, an alternate disequilibrium-mapping method being developed by J. Felsenstein and colleagues (personal communication) uses an approach based on Markov Chain Monte Carlo sampling of labeled ancestries.)

To evaluate the likelihood, we require

$$\mathrm{P}_{\mathbf{q}}(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathrm{C}} \mathrm{P}\,(\mathrm{C} \mid \mathbf{x}), \tag{5.3}$$

where $C = \{c_{ij}\}$ denotes a configuration of recombinant classes consistent with $\mathbf{x}$ and $\mathbf{y}$ such that c_{ij} recombinant classes of size i are assigned to allele j. Figure 5.1 shows the notation. The row totals for C are given by $\mathbf{x}$, since $x_i = \sum_j c_{ij}$. The column totals are given by $\mathbf{n}$, where n_j is the the number of recombinant classes assigned to marker allele j. The number y_j of sampled disease haplotypes carrying marker allele j is $y_j = \sum_{i=1}^{K} i\, c_{ij}$, the inner product of the recombinant class sizes and the j^{th}

column of the matrix C. The setup for determining possible configurations C is thus analogous to that for determining possible tables in Fisher's exact test (Fisher 1970), except that conditioning is on the row totals $\mathbf{x}$ and column inner products $\mathbf{y}$, rather than on the row totals $\mathbf{x}$ and column totals $\mathbf{n}$.

The probability $\mathrm{P}(\mathrm{C} \mid \mathbf{x})$ is obtained as follows. Given $\mathbf{x}$, the i^{th} row of C is multinomial with parameters x_i and $\mathbf{q}$, and thus has probability

$$\frac{x_i!}{\prod_{j=1}^{m} c_{ij}!} \times \prod_{j=1}^{m} q_j^{c_{ij}}.$$

The product of these independent multinomial distributions over the rows (recombinant class sizes) therefore gives

$$\mathrm{P}(\mathrm{C} \mid \mathbf{x}) = \frac{\left(\prod_{i=1}^{K} x_i!\right) \times \left(\prod_{j=1}^{m} q_j^{n_j}\right)}{\prod_{i=1}^{K} \prod_{j=1}^{m} c_{ij}!}.$$

so that equation (5.3) becomes

$$\mathrm{P}_q(\mathbf{y} \mid \mathbf{x}) \;=\; \sum_{\mathrm{C}} \mathrm{P}(\mathrm{C} \mid \mathbf{x}) \;=\; \left(\prod_{i=1}^{K} x_i!\right) \times \sum_{\mathrm{C}} \frac{\prod_{j=1}^{m} q_j^{n_j}}{\prod_{i=1}^{K} \prod_{j=1}^{m} c_{ij}!}. \qquad (5.4)$$

To evaluate $\mathrm{P}(\mathbf{y} \mid \mathbf{x})$, we enumerate all tables or configurations C of recombinant classes consistent with $\mathbf{x}$ and $\mathbf{y}$. A variant of the network algorithm of Mehta and Patel (1983), in which a path through a network represents a consistent table, is used. Details are given in Section 5.3 below.

Statistical uncertainty is measured by variance of estimators or curvature of the observed or expected likelihood surface. However, here we have no explicit form for the likelihood. Instead, a bootstrap approach may be applied, but this is also non-trivial, since we have a single realization of the coalescent process of disease ancestry, and hence of the underlying recombinant classes. Thus, adopting a *nonparametric bootstrap* approach, and sampling with replacement from the current disease sample amounts to bootstrapping one observation, and does not reflect the variation in recombinant classes across replicate populations. The recombination process contributes

to the variation in recombinant classes, and so inferences about the recombination fraction must account for the variation. We therefore adopt a *parametric bootstrap* approach, and generate realizations of $\mathbf{Y}$ under the MLE, reestimating r for each such realization. Support intervals for r are defined by dropping down from the maximum of the original likelihood surface. The interval capturing the appropriate percentage of the (parametric) bootstrapped MLE's is selected as the confidence interval.

We illustrate the approach with the Japanese example for a sample of $K = 50$ disease haplotypes. Two markers, M2 and M3, separated by a recombination frequency 0.01 flank the disease, which is located at recombination frequency 0.006 from M2 and $0.01 - 0.006 = 0.004$ from M3. (For fine-scale mapping such as this, recombination frequencies are effectively additive.) Each marker has four equifrequent alleles. Estimated single-marker likelihood surfaces, calculated over a grid of recombination fractions spaced at one-tenth of a percent (0.001), are based on 10,000 Monte Carlo replicates. With one user on a Pentium 133 MHz, the single-marker map for 20 recombination frequencies ranging from 0.001-0.020 takes about 15 minutes. Bootstrap confidence intervals are based on 200 bootstrap samples.

The simulated data are $\mathbf{y}_1 = (2, 2, 44, 2)$ for the marker M2, and $\mathbf{y}_2 = (46, 2, 1, 1)$ for M3. Lod scores ($\log_{10}$ likelihood-ratios) for M2 are shown in Figure 5.2. The MLE is $\hat{r} = 0.005$, with associated 95% confidence interval 0.0016-0.0110. The confidence interval corresponds to about a one lod-unit support interval, just over what would be computed assuming a chi-squared approximation to the distribution of minus twice the $\log_e$ likelihood-ratio. The likelihood surface for the marker M3 (results not shown) has more curvature about its MLE of $\hat{r} = 0.003$ than the likelihood surface for M2. Increased curvature is expected at lower recombination fractions because ancestral recombinant classes tend to be larger, and sampled marker alleles more dependent. (Likelihood curves based on positively-correlated data have curvature that increases with dependence.) For marker M3, the 95% bootstrap confidence interval is 0.0004-0.0110, and corresponds to a lod-score difference of about 2. The confidence interval

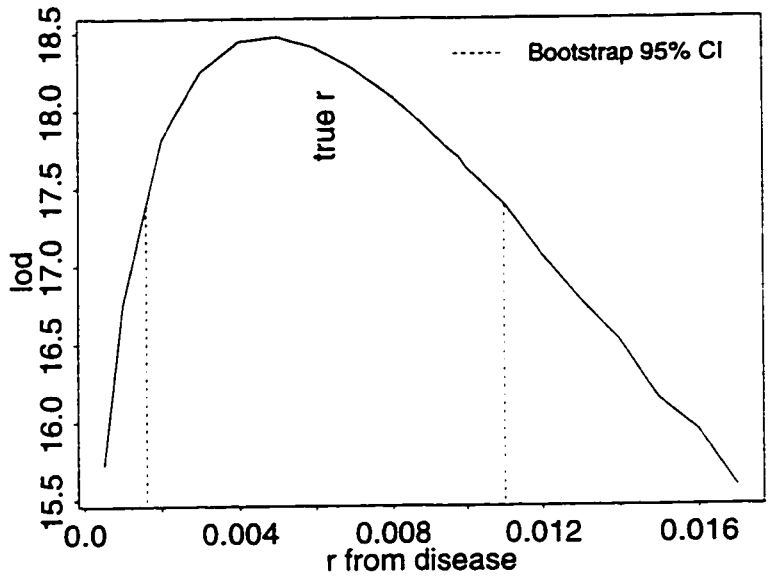


Figure 5.2: Single-marker lod-score ($\log_{10}$ likelihood-ratio) curve and parametric-bootstrap confidence interval for r. There are $K = 50$ disease haplotypes, and lod scores are estimated from 10000 Monte Carlo realizations.

for M3 thus represents a larger drop in lod than the confidence interval for M2, as would be expected with increased dependence under the smaller recombination fraction. As the recombination fraction gets smaller, confidence intervals based on the chi-squared approximation are expected to become increasingly too narrow. The chi-squared approximation assumes independence among marker alleles, but positive correlation increases as the recombination frequency decreases.

5.3 *Evaluating* $\mathrm{P}_q(\mathbf{y} \mid \mathbf{x})$

This section gives details on the network algorithm used to evaluate $\mathrm{P}_q(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x})$ in Section 5.2. Mehta and Patel (1983) develop a network algorithm to compute p-values for Fisher's exact test on contingency tables. Their application requires enumeration of all tables consistent with given marginal totals; these consistent tables

are efficiently represented as paths through the network. Our application is similar, but conditioning is on row totals x_i and column inner products (weighted column sums) $y_j = \sum_{i=1}^{K} i c_{ij}$, rather than on row totals x_i and column totals n_j (Figure 5.1). A further difference is that, in our application, it may be impossible to assign the sampled recombinant classes ($\mathbf{x}$) to alleles in a manner compatible with the observed allelic classes $\mathbf{y}$. For example, the size of the largest recombinant class could be larger than the size of the largest allelic class, in which case $P_{\mathbf{q}}(\mathbf{y} \mid \mathbf{x}) = 0$.

The number of paths through a network can be much larger than the number of nodes. For instance, one network for a diallelic marker had 11,774,790 paths but only 524 nodes. A network is thus an efficient representation of the often large number of tables consistent with marginal quantities. The nodes represent updated column margins; specifically, these are the recombinant classes remaining to be assigned to an allelic type. The edges represent the columns of cells in the table, the counts c_{ij} of recombinant classes of size i assigned to allelic type j. A convenient feature of the network representation is that tables are easily extracted by traversing paths. As a path is traversed, the probability of the table is determined by multiplying predetermined edge weights (see equation (5.4)).

Consider a simple example for a three-allele marker with allelic classes of size $y_1 = 9$, $y_2 = 7$, and $y_3 = 6$, and corresponding allele frequencies $\mathbf{q} = (q_1, q_2, q_3)$. Suppose that there are $x_1 = 2$ recombinant classes of size 1, $x_2 = 1$ recombinant class of size 2, and $x_3 = 6$ recombinant classes of size 3. Thus $\mathbf{x} = (2, 1, 6)$ and $\mathbf{y} = (9, 7, 6)$. The three possible tables associated with $\mathbf{x}$ and $\mathbf{y}$ are shown in Figure 5.3.

To obtain the possible tables, recombinant classes are assigned to small allelic classes before large allelic classes. Within any allelic class, the largest recombinant classes are assigned first. Assuming allelic classes in the table are listed in decreasing order of size from largest to smallest, we start at the bottom right corner of a table and works upwards and leftwards to build the network. Figure 5.4 gives the network for the example.

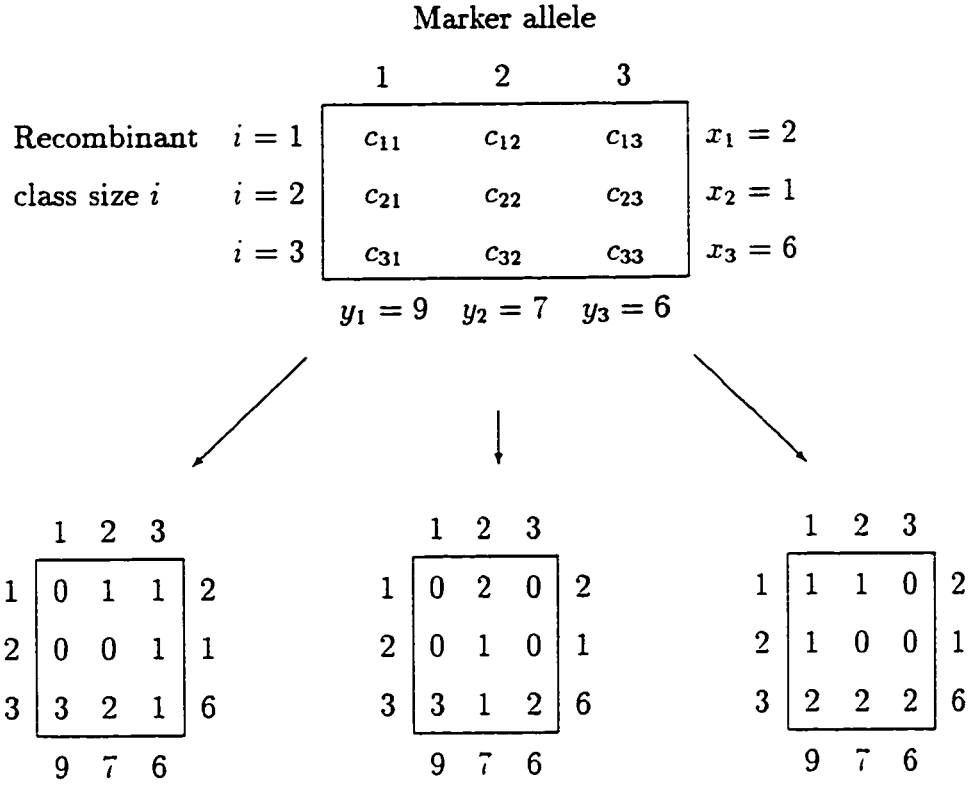


Figure 5.3: The general configuration table, and the three possible configurations for the example with $\mathbf{x} = (2, 1, 6)$ and $\mathbf{y} = (9, 7, 6)$. The network is shown in Figure 5.4.

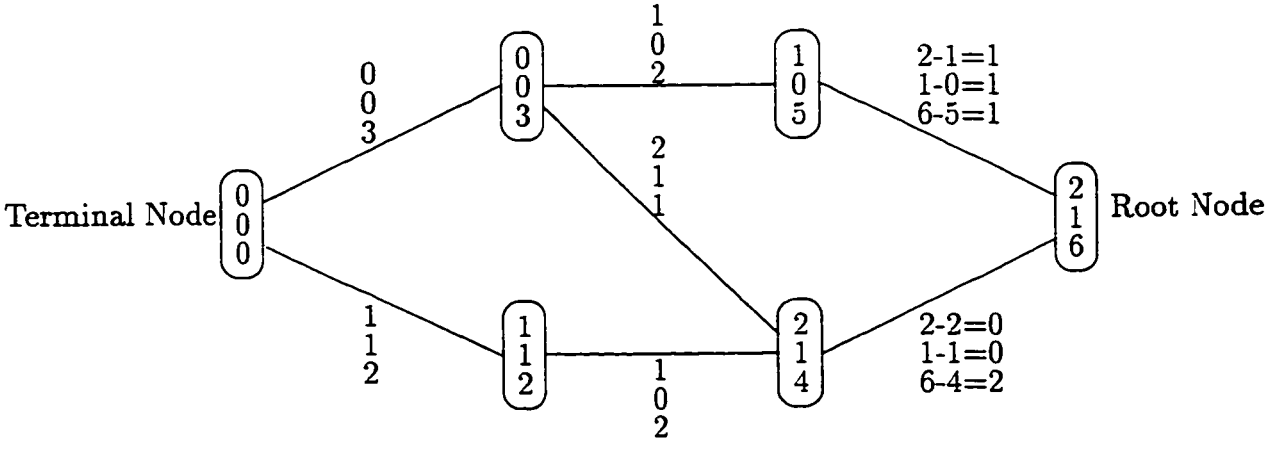


Figure 5.4: The network diagram for the example with $\mathbf{x} = (2, 1, 6)$ and $\mathbf{y} = (9, 7, 6)$. The corresponding configurations are shown in Figure 5.3

5.3.1 *Steps of the algorithm*

Let $\mathbf{x}^*$ be the current row margin and $\mathbf{y}^*$ the current column margin. We build the network from the right (or root) starting with $\mathbf{x}^* = \mathbf{x}$ and $\mathbf{y}^* = \mathbf{y}$. Place $\mathbf{x}^*$ in the network as the initial or root node. In the example, we start with $\mathbf{x}^* = (2, 1, 6)$ as the root node and $\mathbf{y}^* = (9, 7, 6)$. Suppose the largest recombinant class is of size I and the marker has m alleles. In the example, $I = 3$ and $m = 3$.
Starting with $i = I$ and $j = m$:

1. Compute a feasible range for c_{ij}. For instance, in the example, the feasible range for c_{33} is $\{1, 2\}$. Details on computing the feasible range are given in section 5.3.2.

2. Select a possibility from within this range for c_{ij}. For illustration, suppose $c_{33} = 1$ is selected.

3. Update $\mathbf{x}^*$ and $\mathbf{y}^*$: $\mathbf{x}_i^* \rightarrow \mathbf{x}_i^* - c_{ij}$, $\mathbf{y}_j^* \rightarrow \mathbf{y}_j^* - i \times c_{ij}$. For instance, after $c_{33} = 1$ is selected, $\mathbf{x}^*$ changes from $(2, 1, 6)$ to $(2, 1, 5)$ and $\mathbf{y}^*$ changes from $(9, 7, 6)$ to $(9, 7, 3)$.

4. Move up the column: $i \rightarrow i - 1$. Repeat steps 1.2.3 for the new (i, j). For instance, after selecting c_{33}, we move up the third column to select c_{23}.

5. When the top of a column of the table has been reached ($i = 1$), place $\mathbf{x}^*$ in the network as a node and connect it to the previous node that gave rise to it. If $\mathbf{x}^*$ is already in the network, do not add it again, just make the connection to the previous node. In the example, the choice $c_{33} = 1$, gives rise to unique possibilities $c_{23} = c_{13} = 1$, with an associated network node $\mathbf{x}^* = (1, 0, 5)$ after completion of the column $i = 3$. (Figure 5.4).

6. Go to the next column of the table: $i \rightarrow I$ and $j \rightarrow j-1$. For instance, after the top of the column is reached and c_{13} has been selected, we move left in the table to the bottom of the second column and select c_{32}.

7. Repeat steps 1,2,3,4,5 for the new column.

8. When the table is completed ($i = j = 1$; $\mathbf{x}^* = (0,0,0)$), or the selected path terminates prematurely, return to the last cell in the table at which unexplored values in the feasible range exist, and repeat steps from step 2.

In the example, the choice $c_{33} = 1$ determines a single feasible path and configuration. Returning to $i = j = 3$, the alternate choice, $c_{33} = 2$ also gives rise to unique possibilities $c_{23} = c_{13} = 0$ and the third-column, associated node $\mathbf{x}^* = (2,1,4)$. There are then two possible choices for c_{32}, each providing a feasible configuration (Figure 5.4).

Once a network has been constructed, the difference in elements of nodes flanking each edge defines the appropriate column of a table. The three paths in Figure 5.4 from the root to terminal node correspond to the three possible tables, whose columns can be read off the edges. The third column has two possible assignments, (1,1,1) and (0,0,2). If the third column is (1,1,1), the second column is (1,0,2). If the third column is (0,0,2), the second column is either (2,1,1) or (1,1,2). If the second column is (1,0,2) or (2,1,1), the first column is (0,0,3). If the second column is (1,0,2), the first column is (1,1,2). The probability of a table, up to the multiplicative constant $\prod_{i=1}^{K} x_i! = \prod_{i=1}^{I} x_i!$, is determined by the product of edge weights along the path, where the weight for an edge corresponding to the j^{th} column is $q_j^{n_j} / \prod_{i=1}^{K} c_{ij}! = q_j^{n_j} / \prod_{i=1}^{I} c_{ij}!$ (equation (5.4)) Summing these probabilities over all paths results in $\mathrm{P}_{\mathbf{q}}(\mathbf{y} \mid \mathbf{x})$ (equation (5.3)).

5.3.2 *A feasible range for c_{ij}*

The number c_{ij} of recombinant classes of size i assigned to the j^{th} marker allele has four restrictions:

1. The number of disease copies $i \times c_{ij}$ defined by the cell can be no more than the size y_j^* of the appropriate allelic class: i.e., $i \times c_{ij} \leq y_j^*$.

2. The number of recombinant classes c_{ij} of a given size i assigned to a marker allele j can be no more than the number of recombinant classes x_i^* of that size; i.e., $c_{ij} \leq x_i^*$.

3. The number of disease copies $i \times x_i^* - i \times c_{ij}$ remaining for the rest of the row i to the left of the cell c_{ij} (that is, for columns $1 \ldots j-1$ of row i) can be no more than the pooled size $\sum_{l=1}^{j-1} y_l^*$ of the corresponding allelic classes; i.e., $i \times x_i^* - i \times c_{ij} \leq \sum_{l=1}^{j-1} y_l^*$.

4. The number of disease copies $y_j^* - i \times c_{ij}$ remaining for the rest of the column j above the cell c_{ij} (that is, for rows $1 \ldots i-1$ of column j) can be no more than the number of copies $\sum_{l=1}^{i-1} l \times x_l^*$ defined by the corresponding recombinant class sizes: i.e., $y_j^* - i \times c_{ij} \leq \sum_{l=1}^{i-1} l \times x_l^*$.

The first two restrictions imply the upper bound of the feasible range for c_{ij} given above,

$$c_{ij} \leq \min(y_j^*/i,\ x_i^*),$$

while the second two imply the lower bound,

$$c_{ij} \geq \max\left(x_i^* - \frac{1}{i}\sum_{l=1}^{j-1} y_l^*,\ \frac{1}{i}[y_j^* - \sum_{l=1}^{i-1} l \times x_l^*]\right).$$

Thus, in the example, the feasible range for c_{33} is $\{1, 2\}$ since

$$c_{33} \leq \min\left(\frac{6}{3}, 6\right) = 2,$$

and

$$c_{33} \geq \max\left(6 - \frac{1}{3} \times (9+7), \frac{1}{3} \times [6 - (1 \times 2 + 2 \times 1)]\right) = \max\left(\frac{2}{3}, \frac{2}{3}\right) = \frac{2}{3}.$$

Provided $\mathbf{x}$ and $\mathbf{y}$ are compatible, at least one path in the network will terminate.

Assigning values within the feasible range to cells in order of largest to smallest recombinant class and smallest to largest allelic class limits, but does not eliminate, nonterminating paths in the network. A small example, with $\mathbf{x} = (2, 1, 2)$ and $\mathbf{y} = (4, 4, 2)$ is shown in Figure 5.5. It is necessary that $c_{33} = 0$, but then the feasible range for c_{23} is $\{0, 1\}$. The choice $c_{23} = 1$ leads to a path reaching the terminal node and a feasible configuration, but $c_{23} = 0$ leads to an empty feasible range at $i = j = 2$ (Figure 5.5). In contrast, all paths of a network in the Mehta and Patel (1983) method reach the terminal node because the application requires singletons rather than recombinant classes to be assigned to cells.

5.4 Monte Carlo properties

For the method presented in this dissertation, where coalescents of the disease sample, and thence recombinant classes, are realized independently, the Monte Carlo error is readily assessed, and is primarily a function of Monte Carlo sample size. Note, however, that the effective Monte Carlo sample size for a likelihood $L(r)$ varies with the value of r.

If the hypothesized location of the disease gene is much closer to a marker than the true location, many of the recombinant class realizations simulated at the hypothesized location will be inconsistent with the data $\mathbf{y}$ due to an ancestral recombinant class that is larger than the largest allelic class. Thus, for example, if 10,000 realizations of $\mathbf{x}$ are used at each location at which a likelihood is to be estimated, sample sizes close to 10,000 will be realized only for locations in the neighborhood of the true recombination fraction or locations with recombination frequencies that are larger than the truth (Figure 5.6). Figure 5.6 shows the number of compatible realized

(A) $c_{23} = 1$:

		Marker allele 1	2	3	
Recombinant	$i = 1$	$c_{11} = 1$	$c_{12} = 1$	$c_{13} = 0$	$x_1 = 2$
class size i	$i = 2$	$c_{21} = 0$	$c_{22} = 0$	$c_{23} = 1$	$x_2 = 1$
	$i = 3$	$c_{31} = 1$	$c_{32} = 1$	$c_{33} = 0$	$x_3 = 2$
		$y_1 = 4$	$y_2 = 4$	$y_3 = 2$	

(B) $c_{23} = 0$:

		Marker allele 1	2	3	
Recombinant	$i = 1$	$c_{11} = -$	$c_{12} = -$	$c_{13} = 2$	$x_1 = 2$
class size i	$i = 2$	$c_{21} = -$	$c_{22} = \emptyset$	$c_{23} = 0$	$x_2 = 1$
	$i = 3$	$c_{31} = -$	$c_{32} = 1$	$c_{33} = 0$	$x_3 = 2$
		$y_1 = 4$	$y_2 = 4$	$y_3 = 2$	

Figure 5.5: Configuration tables showing a case where there is (A) one non-terminating path ($c_{23} = 1$) and (B) one terminating path ($c_{23} = 0$). The data are $\mathbf{y} = (4, 4, 2)$, and the recombinant class vector is $\mathbf{x} = (2, 1, 2)$. $\emptyset$, empty feasible range; $-$, undefined feasible range.

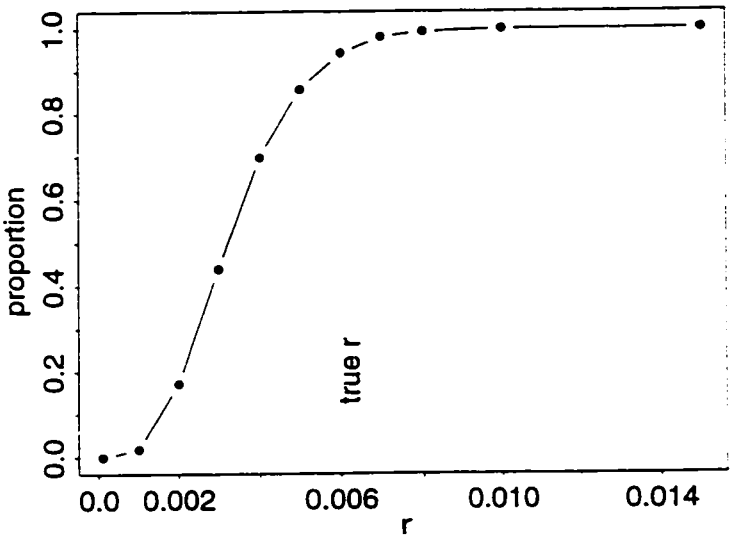


Figure 5.6: Proportion of realized $\mathbf{x}$ compatible with observed $\mathbf{y} = (2, 2, 44, 2)$, as a function of hypothesized r. There are $K = 50$ disease haplotypes, the true r-value is 0.006, and the marker has four equifrequent alleles. The curve is based on 10,000 realizations of $\mathbf{x}$ at each value of r.

($\mathbf{x}$,$\mathbf{y}$) pairs as a function of the hypothesized r in one such Monte Carlo simulation for the Japanese example. The data $\mathbf{y} = (2, 2, 44, 2)$ are those considered previously (Figure 5.2). For values of r that are substantially smaller than the true value, only a small fraction of the 10,000 Monte Carlo realizations results in $\mathbf{x}$-values compatible with data $\mathbf{y}$.

However, Monte Carlo error in estimated likelihoods is of most concern in the neighborhood of the maximum, where there are few incompatibilities, not for the smallest values of r where the likelihood is small. Figure 5.7 shows the Monte Carlo standard error of the estimated likelihood as a percentage of the estimated likelihood value for the above example. The estimated $\log_{10}$ likelihood is also shown. The relative Monte Carlo error is minimized in the neighborhood of the MLE at $\hat{r} = 0.005$. In absolute terms, the error in the likelihood at the maximum is about 3×10^{-5}. The maximum estimated likelihood at $\hat{r} = 0.005$ differs by 3×10^{-4} from the likelihood

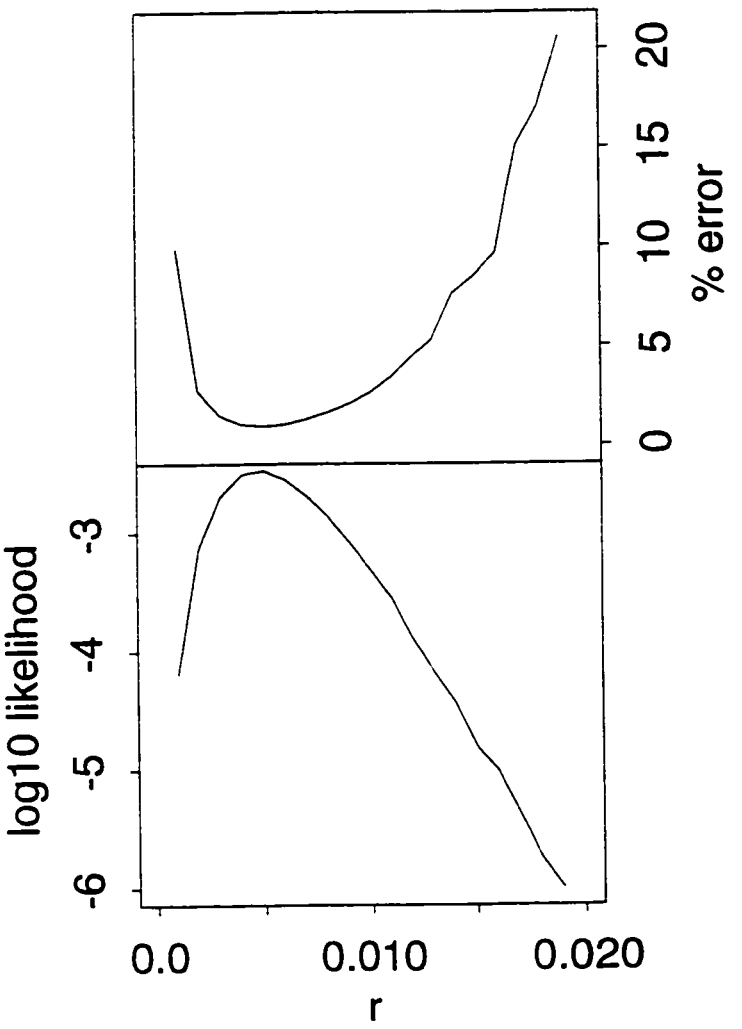


Figure 5.7: Relative Monte Carlo error as a percentage of the estimated likelihood (above), and, for comparison, the estimated $\log_{10}$ likelihood (below).

at the adjacent $r = 0.004$ on the search grid, and by 6×10^{-4} from the estimated likelihood at $r = 0.006$. The Monte Carlo error in the estimated maximum likelihood is therefore an order of magnitude smaller than the differences between estimated likelihood values in the neighborhood of the maximum. These results indicate that 10,000 Monte Carlo replicates are sufficient to locate the MLE to the resolution 0.001 used in the grid search.

Another aspect of the precision and efficiency of Monte Carlo is the choice of the Monte Carlo paradigm. We have chosen to realize recombinant classes $\mathbf{x}$ and compute analytically $\mathrm{P}(\mathbf{y} \mid \mathbf{x})$. Alternatively, allelic classes $\mathbf{Y}$ at present or in the past can be realized and scored when they are consistent with observed data (Rannala and Slatkin 1998). Generally, the more that can be computed analytically, the smaller the Monte Carlo error. To illustrate, we compare Monte Carlo sampling of recombinant classes to scoring of present allelic classes. The standard error of an estimator of a Monte Carlo likelihood based on scoring is $E = \sqrt{P(1-P)/S}$, where P is the likelihood value, and S is the Monte Carlo sample size. Given the desired accuracy, E, this may be solved for the required number of allelic class replicates, S. When the observed data are $\mathbf{y} = (2, 2, 44, 2)$, approximately 3 million allelic class replicates are required to achieve the same degree of accuracy in the neighborhood of the MLE as 10,000 recombinant class replicates. Standard errors and estimated likelihoods from Monte Carlo sampling of recombinant classes are used in this calculation. When both Monte Carlo sampling schemes are calibrated to have the same degree of accuracy, the run based on scoring takes about 150 times longer. The scoring method used for this comparison generates allelic classes $\mathbf{y}$ by randomly assigning the underlying recombinant classes to alleles. As marker polymorphism increases, Monte Carlo sampling of recombinant classes approaches rejection sampling of allelic classes. In the limit, the allelic classes are the recombinant classes.

Replicates in our Monte Carlo approach correspond to a given coalescent ancestry, $\mathbf{A}$, that is recycled over recombination frequencies, r, in the search grid. Recycling

is possible because the recombination process does not enter into the generation of ancestral coalescents (equation (5.2)). Recycling smooths the likelihood curve since recombinant class realizations at different r share the same ancestry. In addition, computational time is saved by recycling, and it may be possible to economize further by eliminating the generation of tree topologies, $\mathbf{F}$, where $\mathbf{A} = (\mathbf{F}, \mathbf{t})$. Sampling coalescent times, $\mathbf{t}$, and averaging conditional probabilities of observed allelic classes $\mathbf{Y}$ given only these times would be a Monte Carlo paradigm that would require fewer replicates than one sampling either recombinant classes $\mathbf{X}$ or allelic classes $\mathbf{Y}$. The form

$$L(r) \;=\; \mathrm{P}(\mathbf{Y}) \;=\; \int_{\mathbf{t}} \mathrm{P}(\mathbf{Y} \mid \mathbf{t})\, \mathrm{P}(\mathbf{t})\, \mathbf{dt}$$

may be contrasted both with our equation (5.2) and with the final equation on P.463 of Rannala and Slatkin (1998). Since each pair of extant ancestral lineages has the same probability of coalescing, the topology of the ancestral tree has a parameter-free distribution. The results of Harding (1971) on the probability distributions of unlabeled tree topologies $\mathbf{F}$ could provide a basis for analytic evaluation of the conditional probability of observed allelic classes given coalescent times. In principle, these probabilities of unlabeled tree topologies would be used to reweight the probability of allelic classes given a tree (topology and coalescent times) in a sum over all possible topologies:

$$\mathrm{P}(\mathbf{Y} \mid \mathbf{t}) \;=\; \sum_{\mathbf{F}} \mathrm{P}(\mathbf{Y} \mid \mathbf{A} = (\mathbf{F}, \mathbf{t}))\, \mathrm{P}(\mathbf{F}).$$

However, exact calculation of $\mathrm{P}(\mathbf{Y} \mid \mathbf{t})$ currently seems impractical, as there are many unlabeled topologies, $\mathbf{F}$ over which to sum, even with a disease sample of moderate size. Moreover, evaluation of allelic class probabilities for any given ancestry $\mathbf{A}$ is computer-intensive, since

$$\begin{aligned} \mathrm{P}(\mathbf{Y} \mid \mathbf{A} = (\mathbf{F}, \mathbf{t})) \;&=\; \sum_{\mathbf{R}} \mathrm{P}(\mathbf{Y}, \mathbf{R} \mid \mathbf{F}, \mathbf{t}) \\ &=\; \sum_{\mathbf{R}} \mathrm{P}(\mathbf{Y} \mid \mathbf{X}(\mathbf{R}, \mathbf{F}, \mathbf{t}))\, \mathrm{P}(\mathbf{R} \mid \mathbf{F}, \mathbf{t}). \end{aligned}$$

Chapter 6

MAPPING WITH MULTIPLE MARKERS

Typically, map positions of marker loci within the candidate genomic region are assumed to be known. These established maps allow the hypothesized trait gene location to be varied throughout intervals defined by the markers. The resulting likelihoods can then be plotted as a function of disease location. Use of multiple markers is desirable in pedigree analyses because of increased power to detect linkage, and improved accuracy of the location estimate (Ott 1992). When a single marker is uninformative, additional markers may provide information about recombination. In addition, markers flanking the disease provide information about recombination events on either side of the disease locus, allowing localization within the interval. Not suprisingly, these considerations apply equally to disequilibrium mapping. This chapter describes interval disequilibrium mapping, which uses allelic information on two markers assumed to flank the disease. Extensions to multipoint disequilibrium mapping are also discussed.

6.1 Interval mapping

Suppose the disease locus lies in an interval of known length $s \times 100$ cM defined by two markers. Recombination events on either side of the disease locus occur independently on the underlying coalescent ancestry, **A**. Genetic interference (Haldane 1919) is effectively irrelevant here, since the chance of recombination with markers on both sides of the disease in a single meiosis is negligible. Let r and r^* denote the recombination frequencies between the disease and first and second flanking markers

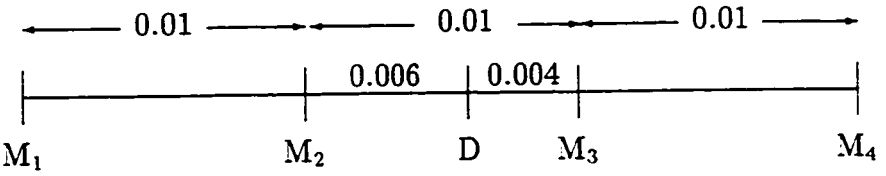


Figure 6.1: Map of four equispaced markers M1-M4, and disease, for the example of interval mapping

respectively. Under the scale typical of disequilibrium fine-mapping, recombination fractions may be considered additive, so that $r + r^* = s$. (Nonadditive recombination fractions can also be accommodated, if desired.) Further, assuming absence of allelic associations between the markers in the total population at the time of the most recent common ancestor of the sample, allelic types at the two markers are independent. The consequence is that equation (5.2) becomes

$$\begin{aligned} L(r,\ r^* = s - r) &= P_{q,r,r^*,\Delta}(\mathbf{Y}, \mathbf{Y}^*) \\ &= \sum_{\mathbf{A}} \left(\sum_{\mathbf{R}} P_q(\mathbf{Y} \mid \mathbf{X}(\mathbf{A}, \mathbf{R})) P_r(\mathbf{R} \mid \mathbf{A}) \right) \times \\ &\qquad \left(\sum_{\mathbf{R}^*} P_q(\mathbf{Y}^* \mid \mathbf{X}^*(\mathbf{A}, \mathbf{R}^*)) P_{r^*}(\mathbf{R}^* \mid \mathbf{A}) \right) P_\Delta(\mathbf{A}) \end{aligned} \tag{6.1}$$

where $\mathbf{Y}$ and $\mathbf{Y}^*$ are the data at the two markers, $\mathbf{X}$ and $\mathbf{X}^*$ the underlying recombinant classes, and $\mathbf{R}$ and $\mathbf{R}^*$ the two sets of recombinant events. Thus, simulation and computation may be done as for the single-marker case. We have assumed absence of associations between markers at the time of the most recent common ancestor of the sample. J. Felsenstein and colleagues are developing a disequilibrium-mapping method that does not require this assumption. Their method reconstructs ancestries for the alleles at linked markers (personal communication).

We illustrate interval mapping with the Japanese example. We assume four equispaced markers M1-M4, at recombination frequencies $s = 0.01$ apart, defining three intervals (Figure 6.1). The disease is located in the second interval, at recombination frequency $r = 0.006$ from M2 and $r^* = s - r = 0.004$ from M3. Each of the four

markers has four equifrequent alleles. As before, estimated likelihoods are based on 10,000 Monte Carlo replicates, and bootstrap distributions on 200 bootstrap samples. For the $K = 50$ sampled disease haplotypes, the simulated (marginal) data for M1 through M4 are, respectively, $\mathbf{y_1} = (2, 6, 3, 39)$, $\mathbf{y_2} = (2.2, 44.2)$, $\mathbf{y_3} = (46, 2, 1, 1)$, and $\mathbf{y_4} = (7.4, 33.6)$. Under the assumption of no association between markers at the time of the most recent common ancestor of the sample, the ancestral alleles defining the background haplotype at this time are independent. Thereafter, recombined marker alleles are independent since recombinations occur independently on either side of the disease locus. As a result, current marker allele status is independent, and only the marginal allelic counts for each marker are required for interval mapping.

The procedure correctly identifies the disease interval. The maximum lod score is at least 5 units above the maxima in adjacent intervals: the correct interval is 100,000 times more likely. The disease location is correctly estimated at $\hat{r} = 0.000$ from the disease. (Of course, in this case, the exact location lies on the search-grid, and happened to be correctly determined. In general, the accuracy of the estimate is limited not only by the amount of data, but also by the resolution of the grid.) The location estimate has associated 95% parametric-bootstrap confidence interval -0.0034-.0020, computed as described in section 5.2. The confidence interval is narrower than those obtained from separate consideration of flanking markers in single-marker mapping, indicating the greater power of interval mapping. The greater power is also reflected in the curvature of the likelihood surfaces at the MLE. Figure 6.2 shows the lod score surface relative to its maximum. For comparison, the single-marker log-likelihood with most curvature at its MLE – the log-likelihood for M3 – is also plotted relative to its maximum. The greater power of interval mapping compared to the single-marker lod score is not surprising given that interval mapping uses information at two markers which flank the disease. Lod scores are not plotted at the markers; at each marker, the log-likelihood for 0% recombination is $-\infty$ since more than one allelic class is observed.

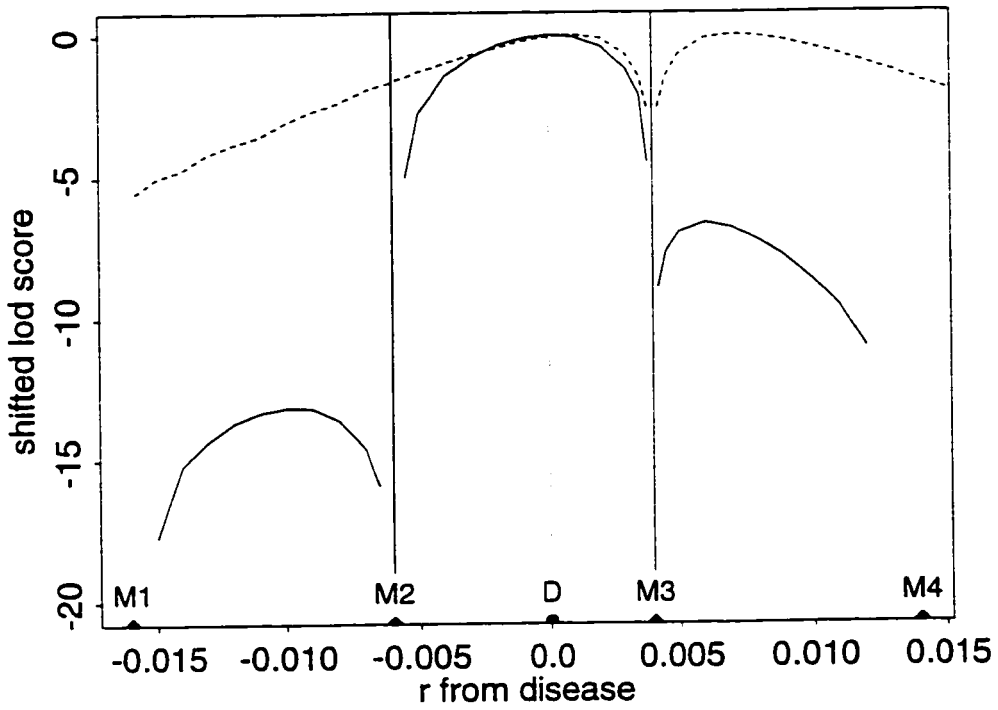


Figure 6.2: Information in interval versus single-marker mapping. Using the marker map of Figure 6.1, the interval-mapping lod-score curve is shown relative to its maximum (solid line). For comparison, the single-marker lod-score curve for marker M3 is also shown, relative to its maximum (dashed line). There are $K = 50$ disease haplotypes, and lod scores are estimated from 10000 Monte Carlo realizations.

Marker allele frequencies influence the power to detect allelic associations (Olson and Wijsman 1994, Lewontin 1988), and hence the power of disequilibrium-mapping methods. Goddard et al. (1996) discuss the limitations of disequilibrium measures when the disease locus lies in a genomic region of relatively uninformative markers. We investigate the performance of interval-mapping under these circumstances. For instance, will a disease interval defined by low-heterozygosity markers be correctly identified? Figure 6.3 illustrates a simple situation for the Japanese example, in which diallelic markers (U',U), both with population allele frequencies of 0.9, flank the disease locus (D), and separate it from an informative marker (I) with 4 equifrequent alleles in the population. Markers are spaced at recombination fraction 0.01. Adjacent markers define two contiguous intervals, U'-U and U-I. There is also one larger interval U'-I, with defining markers separated by recombination fraction 0.02. The disease is located in the center of U'-U, at $r = 0.005$ from both markers. Interval lod-curves are indicated by solid lines, and single-marker curves by symmetric dashed lines about the appropriate marker. Only single-marker lod-curves for markers U and I are plotted. Considered alone, the interval-mapping lod-curves for U'-U and U-I incorrectly suggest that the disease is located in U-I. The low heterozygosity of the uninformative markers drastically reduces power to detect linkage to the disease interval U'-U. Power is improved by including the informative marker. The maximum interval-mapping lod-score for U'-I is at least 5 units higher than the maximum for U'-U. The single-marker data for I also indicate that the disease lies outside U-I. The maximum single-marker lod-score for I is higher than the maximum interval-mapping lod-score for U-I. This example illustrates how naïve application of interval mapping may be misleading when only intervals defined by adjacent markers are used, and some intervals are defined by low-heterozygosity markers. These same cautions would apply to interval-mapping with pedigree data.

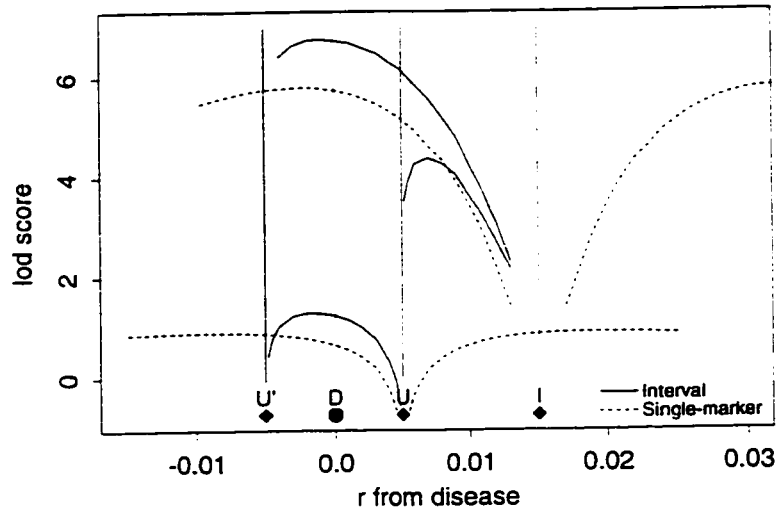


Figure 6.3: Interval (solid lines) and single-marker (short-dashed lines) lod-score curves when the disease locus is flanked by relatively uninformative markers U and U', and there is an informative marker I nearby. Population allele frequencies are $q = (0.9, 0.1)$ for markers U and U', and $q = (0.25, 0.25, 0.25, 0.25)$ for marker I. Single-marker lod scores are shown for markers U and I. The Japanese example is used, with a sample of $K = 50$ disease haplotypes. Lod scores are estimated from 10000 Monte Carlo realizations.

6.2 *Extensions to multipoint mapping*

When flanking markers are not highly polymorphic, additional information on historical recombination events between the disease and flanking markers may be gained by considering jointly several markers on each side of the disease locus. Considering markers jointly incorporates the haplotype information available across markers. In interval mapping, the more polymorphic a flanking marker, the more nearly its allelic classes determine the corresponding single-marker recombinant classes, $\mathbf{x}$. In the limit, $\mathbf{x}$ is observed, and interval mapping is fully efficient. As for interval mapping, recombination events in disjoint segments of the chromosome occur independently on the ancestral coalescent $\mathbf{A}$ (equation (6.2)). However, given the ancestry, the sets of single-marker recombinant classes are now dependent. Recombination events between the disease and a flanking marker are a subset of those between the disease and a farther marker, and thus the recombinant classes for the farther marker partition those for a closer marker.

Multipoint mapping thus requires extension of the concept of a recombinant class. The defining principle is analogous: disease haplotypes in the same multi-marker recombinant class share a common ancestor at all the markers. As before, recombinant classes are obtained by placing recombination events on the realized ancestral tree, for each disjoint chromosome segment defined by the putative disease locus and the markers.

For simplicity, we consider the case of two markers M_1 and M_2 separated by a known recombination fraction s and located to one side of the disease (Figure 6.4). However, the approach extends to any number of markers. Let M_1 be the marker that is closest to the disease. Suppose that the disease and M_1 are separated by the unknown recombination fraction r. Since M_2 is farther from the disease than M_1, recombinant classes for M_2 partition those for M_1. The marker M_2 can thus provide information about r when the underlying M_1 recombinant classes are poorly

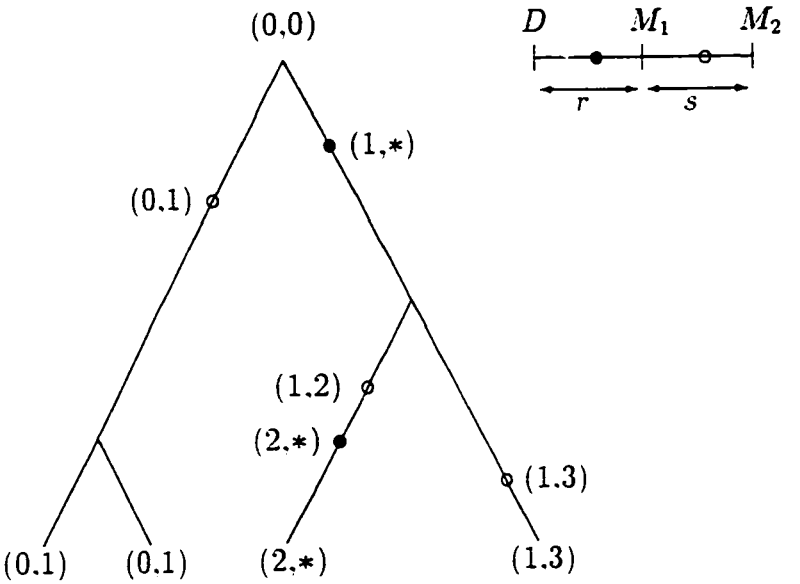


Figure 6.4: Definition of multipoint recombinant classes: the case of two markers M_1 and M_2 on one side of the disease locus D. Recombination events both between the disease and M_1 are indicated by •, as before, and those between M_1 and M_2 by ○. There are $K = 4$ sampled disease haplotypes, with haplotype identifiers (0,1), (0,1), (2,*) and (1,3), respectively.

defined. When the M_1 recombinant classes are well defined, such as when M_1 is highly polymorphic, M_2 provides little or no information about r. In the extreme case that M_1 is infinitely polymorphic, each recombinant class is of a different allelic type. The recombinant class sizes for M_1 are then observed, and no extra information about r is gained from the M_2 recombinant classes.

To obtain a realization $\mathbf{z}$ of the joint recombinant class information $\mathbf{Z}$ for M_1 and M_2, recombination events between both the disease and M_1, and between M_1 and M_2 are placed on the ancestral tree, as shown in Figure 6.4. Each type of event is indexed in the order in which it occurs as the ancestral tree is traversed forwards in time from the root to the tips. The most recent common ancestral haplotype of the K sampled copies is denoted by the haplotype vector (0,0). The first element of the vector corresponds to M_1 and the second element to M_2. Subsequent haplotypes,

formed by recombination events on the ancestral tree, are recorded in haplotype vectors and coded in the order of the recombination events defining them. When a recombination event occurs between the disease and M_1, a new (M_1, M_2) haplotype joins the disease allele. To indicate this, the M_2 element of the haplotype vector is coded with the wild-card symbol $*$. The resulting joint recombinant class information Z_i for each sampled copy i is given at the tips of the ancestral tree. In the example of Figure 6.4, there are two joint recombinant classes, $(2, *)$ and $(1, 3)$ each of size 1, and one $(0, 1)$ of size 2.

Assignment of marker alleles to the resulting recombinant classes takes into account haplotype frequencies. In particular, the recombinant class $(2, *)$ is assigned alleles based on (M_1, M_2)-haplotype frequencies. As for a single marker, we make the simplifying assumption that current marker allele and haplotype frequencies in the general population have obtained throughout the history of the disease allele. On the other hand, since recombination events between M_1 and M_2 imply independent allele status at each marker locus, allelic assignment for classes $(0, 1)$ and $(1, 3)$ is based on the marginal allele frequencies at each marker.

The two-marker likelihood for r,

$$\mathrm{P}(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{z}} \mathrm{P}(\mathbf{y} \mid \mathbf{z})\mathrm{P}(\mathbf{z}),$$

is analogous to the single-marker likelihood, where $\mathbf{y}$ is now the observed table of haplotype counts $\mathbf{Y}$, with $\mathrm{P}(\mathbf{y} \mid \mathbf{z})$ evaluated analytically. Efficient computational algorithms for calculation of $\mathrm{P}(\mathbf{y} \mid \mathbf{z})$ require future investigation. Note that the proposed method of multipoint mapping differs in some details from the method being developed by J. Felsenstein and colleagues, which reconstructs haplotypes by Markov Chain Monte Carlo sampling of marker ancestries (personal communication).

Chapter 7

ASSUMPTIONS AND DIAGNOSTICS

We have investigated population association as a tool for fine-mapping a rare disease, and developed a linkage likelihood for the problem. Our results are based on the ancestral coalescent of a random sample of rare disease alleles. This rare disease is assumed to arise from a single mutation, present as one copy at a known time in the past, which we have taken to be the time of population founding. The overall rate of population growth is assumed known throughout disease history. Selective neutrality of disease alleles is also assumed, as is random mating. Other assumptions are constant population frequencies of marker alleles and no marker mutation on the ancestral coalescent. In the case of interval mapping, it is also assumed that the recombination frequency between markers is correctly specified. This chapter discusses possible departures from these assumptions in real populations, and the effect of these departures on estimated recombination fractions. Diagnostics are also proposed.

7.1 Rare disease

Our model is for a rare disease, arising from a single mutation. Under this assumption, disease alleles reproduce essentially independently of the rest of the population, and of each other. This allows disease growth to be approximated by a birth-and-death process when realizing disease population sizes for coalescent rates. Such an approximation permits conditioning on the present disease copy number, without resorting to rejection sampling. For common diseases, however, the birth-and-death

approximation breaks down. Thus past sizes of the disease population can no longer be conveniently realized conditional on present size. However, if rejection sampling of unconditional past disease sizes (see equation 2.10) were practical, the rare disease assumption could be removed.

7.2 Disease homogeneity

Within the population, the disease is assumed to arise from a single mutation. There are three main departures from this assumption in real populations: allelic heterogeneity, locus heterogeneity, and phenocopies or sporadic cases.

7.2.1 Allelic heterogeneity

Allelic heterogeneity arises when there are multiple disease-causing mutations at the disease locus, as a result of recurrent disease mutations. Allelic heterogeneity can occur even in relatively homogeneous populations, such as the Japanese. If a single disease mutation is incorrectly assumed, sampled disease alleles will be less related than expected, biasing estimated recombination fractions upwards. Multiple mutations on differing background haplotypes in the disease sample will look like multiple recombinant classes. For example, two major mutations on different haplotypic backgrounds would appear to be an ancestral recombinant class, along with a second large recombinant class. Assuming a single disease mutation, the probability of a large nonancestral recombinant class depends on the rate of disease growth and the recombination fraction, as shown in Figure 3.4. When disease copy number increases slowly, large nonancestral recombinant classes have reasonably high probability, within a range of recombination fractions close to zero. However, rapid rather than slow growth in copy number is typical of diseases in recently-founded human populations, as discussed in section 2.5.2. Under rapid disease growth, the chance of a large nonancestral recombinant class is reduced, and there is increased power to

distinguish multiple mutations from large nonancestral recombinant classes.

One way to explore the assumption of a single disease mutation is to calculate likelihoods under this assumption for a plausible range of recombination fractions. Uniformly small likelihood values would suggest departures from the single mutation hypothesis. An ideal marker for this test would be highly polymorphic yet stable (i.e., low mutation rate), so that multiple mutations would arise on distinct background marker haplotypes. Power could be further increased by interval mapping. Then additional mutations on different background haplotypes would appear to be double recombinants, with recombinations events in both the subintervals defined by the disease and marker loci. Double-recombinant haplotypes would suggest large nonancestral recombinant classes at *both* flanking markers, a very improbable event reflected in uniformly low values of the interval-mapping likelihood.

7.2.2 Locus heterogeneity

Locus heterogeneity arises when there are multiple disease-causing loci. In some populations, a specific mutation at just one of these loci may be the major cause of the disease. In this situation, we would expect the disease sample to be "contaminated" with a few random, non-disease-predisposing alleles at the major disease locus. The sample of alleles at the disease locus would thus be less related than expected, biasing estimated recombination fractions upwards. Marker allelic types on contaminating haplotypes are drawn randomly from the total population. Suppose that the proportion of contaminating alleles at the major disease locus is known. Then sampled marker alleles are partitioned randomly, according to this proportion, with those on the sporadic haplotypes representing singleton recombinant classes. Generalization of the Monte Carlo procedure for sampling recombinant classes is then straightforward, and evaluation of the conditional probability of allelic classes given recombinant classes is the same as before. If the amount of contamination is not known, it may be possible to maximize the likelihood jointly on the recombination fraction and the

contamination proportion. The properties and practicality of such an approach, including the reduced information available about the recombination fraction, require future investigation.

7.2.3 Phenocopies

We use phenocopies to refer to alleles, at the disease locus, of individuals who develop the disease for nongenetic reasons. Conceptually, phenocopies represent the same "contamination" of the disease sample described in section 7.2.2. The effect of phenocopies on inference would thus be the same, as would the approaches to dealing with them.

7.3 Single copy at founding

Even if the disease arises from a single mutation, the assumption of a single copy at t_f, the time of population founding, may not hold. For example, the disease mutation may actually be younger than founding. Then the disease sample will tend to be more related than expected, and the estimated recombination fraction will be biased downwards. However, simulations of $\{N_D(t) \mid N_D(0), t \leq t_f\}$, the past disease population sizes conditional on present size, indicate that the method described in this dissertation is robust to a younger than assumed disease. As discussed in section 2.5.3, if $N_D(0)$ is more consistent with a younger mutation, $N_D(t)$ will tend to reach a single copy sooner than t_f. The other type of departure arises when more than one disease copy is present at founding, as might be expected with an older mutation. In this case, the disease sample will tend to be less related than expected, and the estimated recombination fraction will be biased upwards. However, Rannala and Slatkin (1998) suggest that disequilibrium likelihoods are robust to this departure. The distribution of the age of the disease mutation provides a useful diagnostic for both departures from the single-copy assumption. Current methods for selectively-neutral mutations

(Griffiths and Tavaré 1998), allow investigation of this age distribution, and are applied to the diseases and populations we study in section 2.8.

7.4 Population growth

The population growth rate $\lambda(t)$, $0 \leq t \leq t_f$ is intended to capture the effects of historical events impacting population size. Throughout the dissertation, we have only considered population growth, but population shrinkage can be similarly accommodated. Historical events include technological advances and prosperity, war, famine, emigration and immigration. Emigration and immigration are assumed to have minimal impact on population marker allele frequencies. This assumption is reasonable when the base population is relatively large. In addition, immigrant copies are assumed to be disease-free, so that all present disease copies descend from the ancestral mutation at t_f.

Realistically, it is possible that new disease variants are introduced into the population through immigration or admixture. However, as long as the majority of the disease subpopulation descends from a single ancestral mutation, the methods described in this dissertation, particularly interval mapping, are expected to yield useful information about disease location. Provided markers are equally informative, the strongest marker associations are still expected closest to the disease locus, even when some disease alleles descend from secondary mutations introduced through admixture or immigration. Of course, allelic heterogeneity of the disease may preclude the existence of population associations, in spite of known disease linkage to markers within the candidate genomic region. Disequilibrium fine-mapping is then of no help.

We assume demographic parameters such as the time of population founding and the subsequent rate of population growth can be reasonably inferred from historical records. Expected coalescent times will then be about right, and inferences about disease location should be reliable, as discussed in section 3.3. However, the effect

of misspecifying population-growth parameters requires further investigation. It is therefore advisable to estimate disease location under a plausible range of growth parameters.

7.5 Selection

Our methods assume that the disease is selectively-neutral. For recessive diseases, selective neutrality is reasonable provided most copies reside in heterozygotes, as would be the case for a rare allele, and provided there is no heterozygote advantage such as that observed for sickle-cell anemia (Stern 1960, Neel 1951). Heterozygote advantage has in fact been proposed as a possible explanation for the large copy number observed in some recessive diseases such as cystic fibrosis and phenylketonuria (e.g. Quinton 1994, Scriver 1994). However, Thompson and Neel (1997) have shown that selection need not be invoked. Rapid expansion of human populations since the advent of agriculture, combined with survival of neutral alleles to the present, is sufficient to explain high copy number. These authors argue convincingly that selectively-neutral models are appropriate for many recessive diseases. Selective neutrality for dominant diseases of late-onset, such as Huntington's disease, is also reasonable.

One way to incorporate known selection on the disease is to generalize equations (2.4) and (2.8) by altering the probability that the allele giving birth is diseased from $N_D(t)/N_P(t)$ to $s \times N_D(t)/N_P(t)$. Then, for example, $s > 1$ implies that disease alleles have a selective advantage over normal alleles in the Moran model, while $s = 1$ implies they are selectively neutral. Under this model, equation (2.8), the chance that the disease population changes size given a Moran event at time t, becomes

$$s\frac{N_D(t)}{N_P(t)}\left(1-\frac{N_D(t)}{N_P(t)}\right)+\frac{N_D(t)}{N_P(t)}\left(1-s\frac{N_D(t)}{N_P(t)}\right).$$

The first term represents the chance that a disease allele gives birth and a normal allele dies, while the second represents the chance that a normal allele gives birth and a disease allele dies. Multiplying by the Moran event rate $N_P(t)/2$, and simplifying,

gives the analog to equation (2.9), the unconditional rate of Moran events impacting disease sizes,

$$\frac{N_D(t)}{2}\left[1+s\left(1-2\frac{N_D(t)}{N_P(t)}\right)\right].$$

For a rare disease, $N_D(t)/N_P(t) \approx 0$, and so the rate is approximately

$$\frac{N_D(t)}{2}(1+s). \tag{7.1}$$

Thus, when the overall population grows at rate λ, the growth of a rare disease is well-approximated by a birth-and-death process with combined rate $(1+s+\lambda)/2$ per copy per generation, provided s is not too far from 1. We have modelled disease selection by altering instantaneous birth rather than death probabilities. Therefore, the appropriate birth rate in the approximation is $(s+\lambda)/2$, and the death rate $1/2$. The birth-and-death approximation with selection allows realizations of past disease sizes, conditional on present size and a single disease allele at t_f. After similarly adjusting for selection in the coalescence rate of equation (2.4), we obtain the analog to equation (2.5),

$$\frac{k(t)\,(k(t)-1)}{2\,(N_D(t)-1)} \times s.$$

Inserting realized disease sizes $N_D(t)$ conditional on $N_D(0)$ and $N_D(t_f)=1$ yields the coalescence rate with selection.

As pointed out by J. Felsenstein (personal communication), the current size $N_D(0)$ of the disease population contains information about the selection coefficient s. It would therefore be possible to use $N_D(0)$ to infer s, although we do not pursue this here.

7.6 Random mating

Throughout, we have assumed an idealized random-mating population. Coalescent theory is based on this idealization, as is the random assignment of marker allelic type to recombinant classes. In real populations, however, some degree of nonrandom

mating is expected due to population structure, both past and present. For example, geographic structure may explain why the most recent coalescence event in a sample of $K = 10$ IOSCA copies was apparently 5 gbp rather than the 15-25 gbp expected under random mating (Nikali et al. 1995). It is well known that many large modern populations result from amalgamations of smaller endogamous (regional) populations. If such population substructure is known, estimates of population-genetic parameters should account for it. (For example, the recently-released program MIGRATE (Beerli 1998) accounts for substructure when estimating migration rates.) Often, however, there may be unknown population substructure, particularly in the past. To investigate the effect of past population structure on disequilibrium mapping, consider a simplified situation in which a randomly mating population is the result of the past merging of two subpopulations. Suppose that the disease variant originates in just one of these populations. Then, at more recent times, coalescences will occur with rates consistent with a random-mating population. However, at more distant times in the past, prior to population merging, the rate of coalescence should be faster due to the reduced size of the originating subpopulation. Early ancestors of the disease sample (i.e., before amalgamation) should thus be more related than expected under random mating. Consequently, sampled disease alleles will also be more related, biasing estimated recombination fractions downwards.

Consanguineous marriages are another source of nonrandom mating. Offspring from these marriages have increased rates of rare recessive diseases. When parents are closely related (e.g., first cousins), there is high probability that both disease alleles of the affected offspring derive from the recent ancestor, rather than from independent sources. Thus, affected offspring tend to carry one randomly drawn disease allele and its copy. When parents are known to be closely related, one obvious adjustment is to randomly select only one of the offspring's disease alleles for analysis. When parents are less related, or when consanguinity is uncertain, such a procedure may be overly conservative. In this case, comparing the variation of marker alleles

within and between offspring provides at least a qualitative check of consanguinity. For instance, if all offspring carried consanguineous disease alleles, marker variation between subjects would dominate variation within. On the other hand, roughly equal sources of variation would be consistent with random mating.

With a recessive disease, it may also be possible to account for consanguinity by joint estimation of the recombination fraction and the proportion of sampled individuals who carry consanguineous disease alleles. When only one disease haplotype per individual is used, the estimated recombination fraction will not be biased by consanguinity. As a first step, the recombination fraction could be fixed at this value, and the likelihood maximized over the consanguinity proportion. For example, suppose that 50 affected people or 100 disease haplotypes were sampled. Suppose further that the hypothesized consanguinity proportion was 20%, or 10 sampled individuals. Then there would be $\binom{K/2}{10} = \binom{50}{10}$ possible assignments of consanguinity among sampled individuals. For any given assignment, this would imply 80 random disease alleles in the 40 individuals selected to be nonconsanguineous, and 10 random alleles in the remaining 10 consanguineous individuals. Although a great many consanguinity assignments are possible, very few would be expected to be consistent with observed marker data in individuals, reducing computations substantially. A given consanguinity assignment is consistent with observed marker data when none of the individuals assigned to be consanguineous are marker heterozygotes. Here, we make the simplifying assumption that there are no recombination events in the meioses separating a pair of consanguineous haplotypes. This is reasonable at the scale of disequilibrium fine-mapping. (Any approach allowing recombinations in the meioses separating a pair of consanguineous haplotypes would require specification of the frequently unknown relationship between these haplotypes.) Carrying through with the example, a given consanguinity assignment would imply marker allelic classes for the 90 random disease alleles under the assignment. The probability of these implied allelic classes under the fixed recombination fraction could be evaluated as before, via

Monte Carlo methods. Summing over consistent consanguinity assignments would yield, up to a constant over consanguinity proportions and recombination fractions, a likelihood value for a consanguinity proportion of 20%, with the recombination fraction fixed. In this way, one could maximize over the consanguinity proportion, for a fixed recombination fraction. Following this, one could maximize over the recombination fraction, with consanguinity proportion fixed. Iterating between estimation of the consanguinity proportion and the recombination fraction would provide a joint maximum likelihood estimate.

7.7 Population marker allele frequencies

We have assumed constant population frequencies of marker alleles throughout disease history. This may be unrealistic for small populations, due to genetic drift. The effect of nonconstant population frequencies on estimated recombination fractions depends on the size and pattern of fluctuations in frequencies over time, as well as the shape of the ancestral tree and the recombination fraction. For example, if population marker allele frequencies are approximately constant over the period during which most recombinant classes are defined (by recombination events), results would be relatively unaffected, provided the specified allele frequencies are about right. Stable population frequencies of marker alleles are expected when the nondiseased population is large.

7.8 Marker mutation

Throughout, we have assumed that there is no marker mutation on the ancestral coalescent. This may be unrealistic for markers with a high mutation rate, such as dinucleotide repeat polymorphisms, when the disease is very old. Stable markers with low mutation rate are therefore preferred. However, it may be possible to extend the methods of this thesis to handle marker mutation. Define a mutant class

to be a subset of sampled marker alleles descending from a particular mutation at the marker, without subsequent recombination or mutation events. Mutant classes partition recombinant classes. Alleles are assigned to recombinant classes as before, and to mutant classes within recombinant classes according to a mutation model. For example, a mutation event at a dinucleotide repeat marker results in a gain or loss of repeat units to the originating dinucleotide sequence, changing the allele length. A reasonable mutation model would assign higher probability to changes that involve a single rather than several repeat units. Under such a model, the marker allelic type after a mutation depends on the dinucleotide length of the allele before the mutation. Sampling of mutant classes within recombinant classes is therefore analogous to multipoint sampling of recombinant classes for the outer marker, farther from the disease, within recombinant classes for the inner marker, closer to the disease, described in section 6.2. Once mutant classes are sampled, assignment of the mutated marker allele given the originating allele is analogous to assignment of the outer marker allele given the allele for the inner marker in multipoint mapping.

7.9 Marker maps

Map positions of marker loci within the candidate genomic region are assumed to be known. While the relative ordering of markers on established maps is generally regarded as reliable, the recombination frequencies between close markers are more prone to error, due to insufficient numbers of meioses in the pedigrees used to establish the maps. Such map misspecification could impact the inferences of interval and multi-point disequilibrium-mapping, since these likelihoods assume accurate recombination frequencies between markers (see section 6.2). Figure 7.1 investigates the effect of map misspecification in the interval-mapping example of section 6.1. The data were simulated with markers separated by a recombination fraction of 0.01. The disease locus was $r = 0.006$ from the marker at the left end of the interval, and

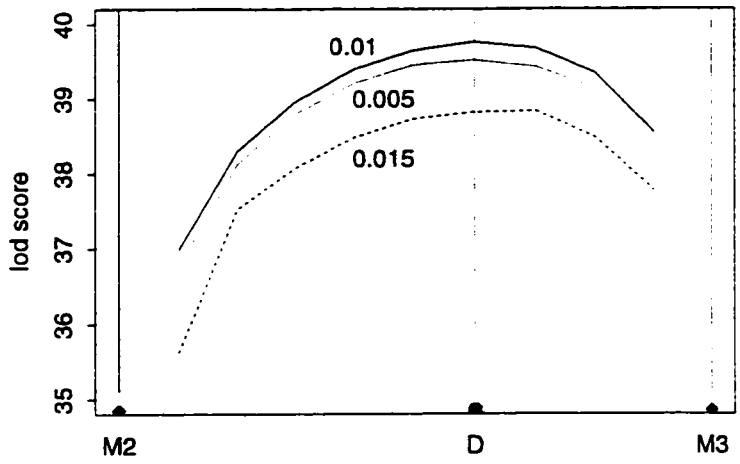


Figure 7.1: Interval-mapping lod-score curves assuming markers are separated by recombination fractions of 0.005, 0.01, or 0.015. Data were simulated with the markers separated by a recombination fraction of 0.01.

$r = 0.004$ from the marker at the right end. Figure 7.1 compares interval-mapping $\log_{10}$ likelihood-ratios or lod-scores under the assumption that the markers are separated by recombination fractions of 0.005, 0.01 (correct), or 0.015. The maximum likelihood estimate of the relative disease location is approximately the same under all three marker spacings. The relative location estimate within the interval is therefore robust to map misspecification. As expected, the correctly-specified marker spacing yields the highest lod-score curve, which is up to one unit higher than the other curves. Hence, map misspecification can reduce the power to correctly identify the genomic interval containing the disease. As an aside, once the disease gene has been identified, interval-mapping data may be used to re-estimate the recombination frequency between markers flanking the disease. The marker spacing that produces the highest lod-score curve will be the maximum-likelihood estimate of the recombination frequency separating the markers. Such estimates may be compared to established marker maps.

Chapter 8

CONCLUSIONS AND FURTHER WORK

In this dissertation we investigate linkage disequilibrium as a tool for fine-mapping a rare nonrecurrent disease mutation. We also present a linkage likelihood for estimating disease location, based on marker and disease coidentity by descent in a random sample of disease haplotypes. The approach highlights how disequilibrium mapping is a natural population-level counterpart to traditional linkage methods, which also rely on gene identity by descent. A coalescent model is developed to describe the ancestry, and hence identity by descent, of sampled *disease* alleles. This model accommodates demographic parameters, including varying population growth rates such as those for the Japanese. Generations on the ancestral coalescent of the disease sample correspond to meioses at which recombination events between the disease and marker loci may occur. These recombination events define the recombinant classes, which describe marker identity by descent in the sample. Using our methods, large Monte Carlo samples of recombinant classes for various hypothesized trait gene locations are readily obtained. We combine these Monte Carlo samples with an analytic method for computation of the probability of observed disease haplotypes, conditional on latent recombinant classes, to obtain a likelihood for fine-scale mapping. The parametric bootstrap is used throughout to obtain confidence bounds on the resulting likelihood estimates of disease location. This likelihood method extends easily from single-marker to interval mapping. Interval mapping is shown to perform substantially better than single-marker mapping in our simulated data example, correctly identifying the interval containing the disease, and providing a more precise estimate of disease location. Further extensions to multipoint mapping are also discussed.

We conclude by reviewing the assumptions made in developing the approach, along with the consequences of departures in real populations. Possible modifications and diagnostics are also proposed.

Of the three recent papers developing likelihood methods of disequilibrium mapping, Xiong and Guo (1997) does not condition on the current disease allele count, and Kaplan et al. (1995) realize this count and must reject realizations not within some acceptable range. Both Xiong and Guo (1997) and Kaplan et al. (1995) assume that the most frequent marker allele in the disease sample is ancestral. Neither we nor Rannala and Slatkin (1998) make this assumption. The present method shares with Rannala and Slatkin (1998) the approach of realizing the ancestry of a sample, conditioning on current disease allele count. The present approach to realizing coalescent ancestry is more an approximation than Rannala and Slatkin (1998), since coalescence rates use disease population sizes realized at one generation steps, based on conditional moments of disease copy number. However, these moments can be computed for any assumed patterns of population growth, and thus the demographic model can be more flexible. All these likelihood methods assume a single disease lineage at some given point t_f in time. However, the simulations in this dissertation and those of Rannala and Slatkin (1998) suggest that disequilibrium likelihoods are robust to assumptions regarding the age of the disease allele. In any case, the probability distribution of the age of the disease mutation may be investigated (see section 2.8), giving insight into the assumption.

Once coalescent ancestry is realized, the approach of this thesis differs significantly from that of Rannala and Slatkin (1998). Their method, like that of Kaplan et al. (1995) and Xiong and Guo (1997), is based on marker identity by state. They sample the marker allelic classes existing immediately after the most recent coalescent event, which is closer in spirit to sampling and scoring current allelic classes than is the present approach of realizing underlying recombinant classes, and computing analytically the probability of observed data given the recombinant class configuration. It is

the realization of recombinant classes which makes straightforward the use of multi-allelic markers. Indeed, performance and feasibility of the present method does not depend significantly on marker polymorphism, whereas Rannala and Slatkin (1998) considered only diallelic markers.

Also, Rannala and Slatkin (1998) consider only likelihoods for a single marker. Xiong and Guo (1997) accommodate multiple markers by treating the information from each marker as independent, and ignoring haplotypic information in a composite likelihood approach. Such an approach departs from the maximum likelihood paradigm. Kaplan et al. (1995) realize the marker haplotype frequencies in the total disease population, and then compute probabilities of the observed sample of disease haplotypes, given these frequencies. In principle, their approach permits likelihoods to be based on multiple polymorphic markers, although Monte Carlo efficiency will decrease rapidly as the possible diversity of haplotypes increases. The method in this dissertation extends readily to interval mapping, and by extending the definition of the latent recombinant classes will also permit multipoint mapping. However, the statistical gains and computational costs of multipoint disequilibrium mapping are uncertain, and remain to be investigated.

All these approaches, including the present approach, consider disequilibrium mapping for a rare allele. All except Xiong and Guo (1997) assume a nonrecurrent disease mutation. Common complex traits are caused by multiple loci and alleles, with effects that are possibly modified by environmental factors. Collectively, disease-predisposing alleles for a complex trait may be common in the population. Unless disease specificity can be increased by studies in genetic isolates, these approaches to disequilibrium fine-mapping are not feasible for complex, heterogeneous diseases. However, disequilibrium methods of linkage detection, such as the TDT and other family-based association methods (e.g. Spielman et al. 1993, Schaid and Sommer 1993, Self et al. 1991), may prove useful. Unlike disequilibrium mapping methods, these methods are robust to population substructure, and have been successfully ap-

plied to complex diseases such as type 1 diabetes and aplastic anemia in heterogeneous populations. Like disequilibrium mapping, family-based methods of linkage detection rely on association due to the underlying coidentity by descent at disease and markers. The more such association, the more power to detect linkage. Family-based methods such as the TDT require population association to detect linkage. In contrast, disequilibrium fine-mapping requires linkage to a candidate genomic region, and uses population associations to further localize the disease gene.

BIBLIOGRAPHY

Aaltonen J. Björses P, Sandkuijl L, Perheentupa J. Peltonen L (1994) An autosomal locus causing autoimmune disease: autoimmune polyglandular disease type assigned to chromosome 21. Nat Genet 8:83–87

Arnason A. Larsen B. Marshall WH. Edwards JH. MacIntosh P. Olaisen B. Teisberg P (1977) Very close linkage between HLA-B and Bf inferred from allelic association. Nature 268:527–528

Beerli P (1998) MIGRATE: Estimation of migration rate and effective population size. http://evolution.genetics.washington.edu/lamarc/migrate.html

Benedict R (1989) The Chrysanthemum and the Sword. Houghton Mifflin Boston

Bishop DT. Williamson JA (1990) The power of identity-by-state methods for linkage analysis. Am J Hum Genet 46:254–265

Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55:379–390

Cavalli-Sforza L, Menozzi P, Piazza A (1994) The History and Geography of Human Genes. University Press, Princeton

Cox DR, Miller HD (1977) The Theory of Stochastic Processes. Chapman and Hall, London

Cox T, Kerem B, Rommens J, Iannuzzi M, Drumm M, Collins F, Dean M, et al. (1989) Mapping of the cystic fibrosis gene using putative ancestral recombinants. Am J Hum Genet Suppl 45:A136

Crow J, Kimura M (1970) An Introduction to Population Genetics Theory. Harper and Row, New York

Edwards J (1981) Allelic association in man. In: Eriksson AW (ed) Population Structure and Genetic Disorders, Proceedings of the 7th Sigfred Juselius Foundation Symposia. New York Academic Press pp 239–256

Feller W (1968) An Introduction to Probability Theory and Its Applications: Volume I. Wiley, New York 3rd edn

Felsenstein J (1971) The rate of loss of multiple alleles in finite haploid populations. Theor Pop Biol 2:391–403

Felsenstein J (1995) Theoretical Evolutionary Genetics. ASUW Press. University of Washington, Seattle Class notes

Felsenstein J (1996) Population genetics. Unpublished class notes: University of Washington, Seattle

Fisher R (1970) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, UK 14th edn

Goddard KA, Yu CE, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al. (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. Am J Hum Genet 58:1286–1302

Griffiths R, Tavaré S (1998) The age of a mutation in a general coalescent tree. Stoch Models 14:273–295

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307–319

Haldane J (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hanihara K (1991) Dual structure model for the population history of the Japanese. Japan Review 2:1–33

Harding E (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–11

Houwen RH, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Hudson R (1990) Gene genealogies and the coalescent process. In: Dawkins R, Ridley M (eds) Oxford Surveys in Evolutionary Biology, Oxford University Press: Oxford, pp 1–44

ISEI (1998) Teachers' and textbook writers' handbook on Japan. International Society for Educational Information http://www.isei.or.jp/books/66/isei_66_contents.html

JIN (1998) Census of Japan. Japan Information Network http://www.jinjapan.org/stat/index-f.html

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kendall DG (1948) The generalized "birth and death" process. Ann Math Statist 19:1–15

Kestilä M, Männikkö M, Holmberg C, Gyapay G, Weissenbach J, Savolainen ER, Peltonen L, Tryggvason K (1994) Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19. Am J Hum Genet 54:757–764

Kingman JFC (1982a) The coalescent. Stochastic Processes 13:235–248

Kingman JFC (1982b) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in Probability and Statistics. North Holland, Amsterdam, pp 97–112

Koyama S (1978) Jomon subsistence and populations. Senri Ethnological Studies 2:1–65

Lehesjoki AE, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la A. Chapelle (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229–1234

Lewontin R (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Matsumoto T, Imamura O, Yamabe Y, Kuromitsu J, Tokutake Y, Shimamoto A, Suzuki N, et al. (1997) Mutation and haplotype analyses of the Werner's syndrome gene based on its genomic structure: genetic epidemiology in the Japanese population. Hum Genet 100:123–130

Maynard Smith J (1971) What use is sex? J Theor Biol 30:319–335

Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J Amer Stat Assoc 78:427–434

Van der Meulen M, te Meerman GJ (1997) Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In: Pawlowitzki I, Edwards J, Thompson E (eds) Genetic Mapping of Disease Genes. Academic Press, London. pp 115–136

Mitchison HM, O'Rawe AM, Taschner PE, Sandkuijl LA, Santavuori P, de N. Vos, Breuning MH, Mole SE, Gardiner RM, Jarvela IE (1995) Batten disease gene, cln3: linkage disequilibrium mapping in the Finnish population, and analysis of european haplotypes. Am J Hum Genet 56:654–662

Moran P (1962) The Statistical Processes of Evolutionary Theory. Clarendon Press, Oxford

Nakura J, Miki T, Nagano K, Kihara K, Ye L, Kamino K, Fujiwara Y, Yoshida S, Murano S, Fukuchi K, et al. (1993) Close linkage of the gene for Werner's syndrome to ANK1 and D8S87 on the short arm of chromosome 8. Gerontology 39(Suppl 1):11–15

Neel J (1951) The population genetics of two inherited blood dyscrasias in man. Cold Spring Harbor Symp Quant Biol 15:141–155

Nevanlinna HR (1972) The finnish population structure - A genetic and genealogical study. Hereditas 71:195–236

Nikali K, Suomalainen A, Terwilliger J, Koskinen T, Weissenbach J, Peltonen L (1995) Random search for shared chromosomal regions in four affected individuals: the assignment of a new hereditary ataxia locus. Am J Hum Genet 56:1088–1095

Olson J, Wijsman E (1994) Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. Am J Hum Genet 55:574–580

Ott J (1992) Analysis of Human Genetic Linkage. Johns Hopkins University Press, Baltimore

Press W, Flannery B, Teukolsky S, Vetterling W (1992) Numerical Recipes in C. Cambridge University Press, Cambridge

Quinton PM (1994) Human genetics. what is good about cystic fibrosis? Curr Biol 4:742–743

Rannala B, Slatkin M (1998) Likelihood analysis of disequilibrium mapping, and related problems. Am J Hum Genet 62:459–473

Rose M (1996) The peopling of Japan. Archaeology 48:43

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114–1126

Scriver CR (1994) Science, medicine and phenylketonuria. Acta Paediatr Suppl 407:11–18

Self S, Longton G, Kopecky K, Liang K (1991) On estimating hla/disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Slatkin M, Hudson R (1991) Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. Genetics 129:555–562

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

Snell R, Lazarou L, Youngman S, Quarrell O, Wasmuth J, Shaw D, Harper P (1989) Linkage disequilibrium in Huntington's disease: an improved localisation for the gene. J Med Genet 26:673–675

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. Am J Hum Genet 52:506–16

Stern C (1960) Human Genetics. Freeman, New York 2nd edn

Sulisalo T, Klockars J, Mäkitie O, Francomano CA, de la A. Chapelle, Kaitila I, Sistonen P (1994) High-resolution linkage-disequilibrium mapping of the cartilage-hair hypoplasia gene. Am J Hum Genet 55:937–45

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777–87

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al. (1989) Non-random association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. J Med Genet 26:676–681

Thompson E (1997) Conditional gene identity in affected individuals. In: Pawlowitzki I, Edwards J, Thompson E (eds) Genetic Mapping of Disease Genes. Academic Press, London, pp 137–146

Thompson EA (1978) The number of ancestral haplotypes contributing to a sample of B8 alleles. Nature 272:288

Thompson EA, Neel JV (1996) Private polymorphisms: how many? how old? how useful for genetic taxonomies? Mol Phylogenet Evol 5:220–31

Thompson EA, Neel JV (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet 60:197–204

Thompson EA, Neel JV, Smouse PE, Barrantes R (1992) Microevolution of the Chibcha-speaking peoples of lower Central America: rare genes in an Amerindian complex. Am J Hum Genet 51:609–26

Weeks DE, Lathrop GM (1995) Polygenic disease: methods for mapping complex disease traits. Trends Genet 11:513–519

Wijsman E (1997) Association versus linkage analysis in mental disorders. In: Blum K, Noble EP, Sparkes RS, Cull JG, Chen T (eds) Handbook of Psychiatric Genetics. CRC Press, Boca Raton, pp 11–13

Williams E (1952) Use of scores for the analysis of association in contingency tables. Biometrika 39:274–289

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531

Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, et al. (1996) Positional cloning of the Werner's syndrome gene. Science 272:258–262

Appendix A

TIME TRANSFORMATION

We describe how to transform waiting times in a time-dependent Poisson process (Cox and Miller 1977) with rate $r(t)$ to waiting times in a regular Poisson process with constant rate $r(0)$. For example, Kingman (1982b) used such a transformation to model the effects of population size on the coalescent times. The essence of the transformation is the observation that $W_1 \stackrel{d}{=} W_2 \times r_2/r_1$ for exponential waiting times W_1, W_2 with respective rates r_1, r_2.

Let t and t^* denote real and transformed time, respectively. The appropriate rescaling is given by

$$t^* = \int_0^t \frac{r(0)}{r(s)}\, ds. \tag{A.1}$$

To see this, first discretize real time into intervals, $(0,h),(h,2h),\ldots(ih,(i+1)h)\ldots$ of length h. Let $r_i = r(i+0.5h)$, and consider the resulting discretized process:

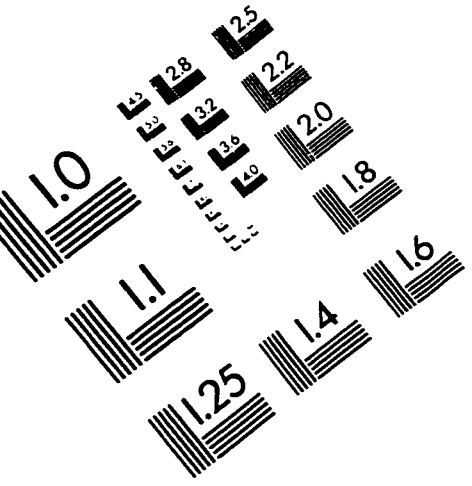
i	t in	scale	t^*
0	$(0,h)$	$\frac{r_0}{r_0}=1$	$1 \times t$
1	$(h,2h)$	$\frac{r_1}{r_0}$	$h + \frac{r_1}{r_0}(t-h)$
2	$(2h,3h)$	$\frac{r_2}{r_0}$	$h + \frac{r_1}{r_0}h + (t-2h)\frac{r_2}{r_0}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$(mh,(m+1)h)$	$\frac{r_m}{r_0}$	$h\sum_{i=1}^{m-1}\frac{r_i}{r_0} + (t-mh)\frac{r_m}{r_0}$

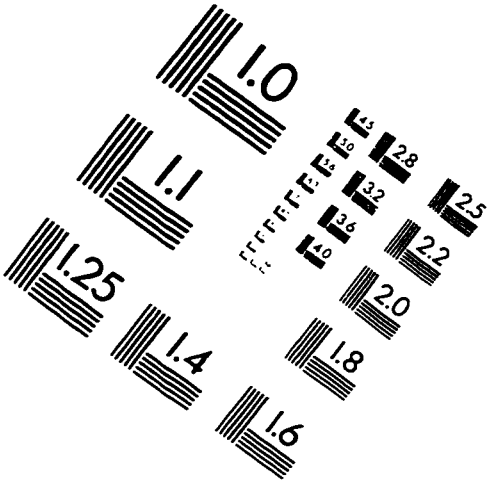
The transformation from t to t^* given in equation (A.1) is obtained by taking limits as $h \downarrow 0$.

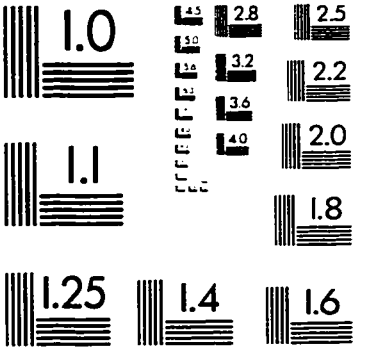
VITA

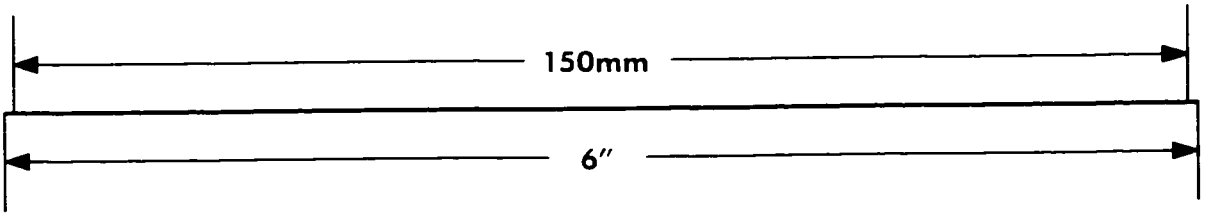
Jinko Graham received a Bachelor of Science in Mathematics in 1987 and a Masters of Science in Statistics in 1992 from the University of British Columbia. She received a Masters of Science in Biostatistics from the University of Washington in 1994. Since then she has been enrolled in the Biostatistics Ph.D. program at the University of Washington.

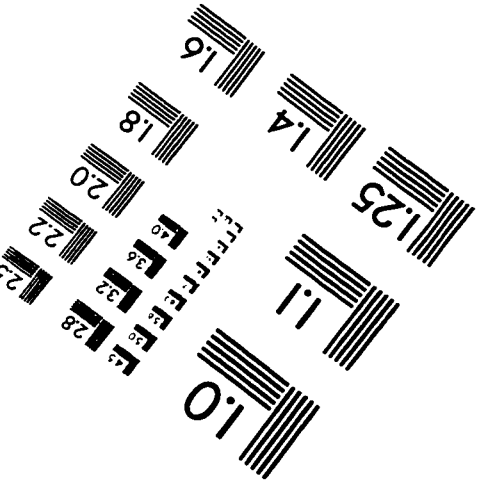
IMAGE EVALUATION
TEST TARGET (QA–3)

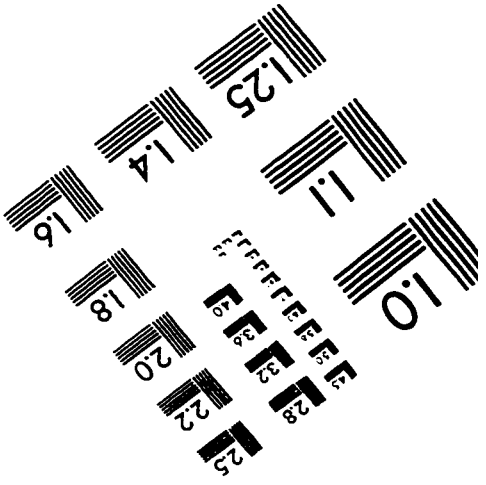












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989