THE UNIVERSITY OF CHICAGO

INFERRING HUMAN DEMOGRAPHY FROM PATTERNS OF GENETIC VARIATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS

BY
ALISON M. ADAMS

CHICAGO, ILLINOIS
AUGUST 2006

ACKNOWLEDGEMENTS

Completion of the work presented in this dissertation would not have been possible without the support and invaluable advice of my committee members: Nancy Cox, Anna Di Rienzo, Dick Hudson, Carole Ober, and Jonathan Pritchard. Special thanks are due to my advisor, Dick Hudson, for his time and patience as well as always-entertaining conversations that often had nothing whatsoever to do with demographic inference.

I am eternally grateful to the friends I've found here in Chicago, and I often wonder what I've done to deserve having such unique yet utterly compatible people in my life. It's friends like Megan and Alex, Darlene and Luke, Latishya, Jaqui, Jane, Timo, and Ben who not only have kept me sane but also have managed to extract me from the comfort of my condo on a regular basis, both of which are powerful and praiseworthy feats. Additionally, I have found extraordinary devotion in my cat, Athena, who has spent more time with me over the last five years than anyone else. I consider her adoption to be the best investment I've ever made, and her constant companionship has brought me countless hours of comfort and affection.

Of course I could not be where I am today without the unconditional love and

support of my family. Whether he's fixing my car, talking with me about politics, or lamenting a White Sox blown save, my dad is always there for me and ready to solve any problem I might encounter, even from 600 miles away. Being away from my mom and her superior back-tickling has been difficult, but it's reassuring to know she is only a phone call away and always willing to sew on a missing button or send me my favorite potato chips that can only be found in central PA. Last, but certainly not least, my brother, Seth, has always been my very best friend. I've always thought we could make the world's best Taboo® team, as our shared experiences and inside references coupled with our perpetual compatibility could make us unstoppable. I can't say enough about how critical the encouragement of all of my family members has been to my success as a graduate student, and, for that and much more, I am eternally grateful.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The recent development of analytical tools to extract information regarding human demographic history from genetic data has generated interest from a range of scientific disciplines. From an anthropological perspective, one goal is to reconcile the palaeontological record with the results of population genetic studies. However, investigation of demographic history is relevant beyond the general interest in elucidating our history as a species.

Historic changes in population size play a pivotal role in shaping neutral patterns of genetic variation. Having an understanding of what constitutes "neutral" variation is critical for determining whether selection is acting on a locus of interest. Additionally, demographic history influences genome-wide levels of linkage disequilibrium (LD), which has implications for the construction and evaluation of disease association studies. Demographic effects on LD also pose a challenge for methods of estimating the population recombination rate, as a standard model of constant population size is typically assumed for such methods.

An ideal method of demographic inference will utilize as much of the genetic data as possible, while maintaining computational feasibility. In this dissertation, several demographic inference methods are presented, each summarizing and analyzing the data in a unique way. Application of these methods to human data reveals population-specific demographic histories which often exclude the standard model of constant population size.

CHAPTER 1 INTRODUCTION TO DEMOGRAPHIC INFERENCE

1.1 Significance of human demographic history

Human demographic history, which includes historic changes in population size, is of interest to researchers across a range of disciplines. From a population genetics perspective, there is much interest in determining what constitutes a neutral pattern of genetic variation. Identifying deviation from neutral patterns of variation is one way to assess whether natural selection may have been acting on a locus of interest (Sabeti *et al.* 2002; Akey *et al.* 2004). Therefore, accurate assessment of the demographic history of a population may contribute to more accurate detection of loci under selection.

Demographic history also influences the extent of linkage disequilibrium (LD) across the genome. Disease association mapping studies rely on LD between a genetic marker and a variant associated with disease; therefore, demographic history has implications for the density of markers that would be required for association mapping efforts (Reich *et al.* 2001). Specifically, it may be desirable to utilize a population that has extended tracts of LD (perhaps due to a population bottleneck) for genome-wide association studies while choosing a population with neutral patterns of LD for fine-scale association mapping once a region of interest has been identified.

An additional consideration is the effect of demographic history on estimation of the population recombination rate. Current methods assume a constant population size when estimating ρ ($= 4Nr$, where N is the effective population size, and r is the recombination rate per generation). It has been shown that deviations from neutrality (e.g. population size changes, population structure, or migration) influence the accuracy of ρ estimates when standard approximate likelihood methods are used (Smith and Fearnhead 2005).

There is also much interest in gaining a greater understanding of human history by reconciling demographic information gleaned from population genetic data with the palaeontological record. It has been hypothesized that a bottleneck in non-African populations marked the exodus of anatomically modern humans out of Africa (Harpending and Rogers 2000; Reich *et al.* 2001). The fossil record places the most ancient modern human remains in Sub-Saharan Africa at 160 kya (White *et al.* 2003; Clark *et al.* 2003), with dispersal to Eurasia occurring approximately 60 kya (Mellars 2006).

Some fundamental questions regarding human demographic history include:

- How do historic population size-change events influence contemporary patterns of genetic variation?

- How can we use these contemporary patterns of variation to learn something about a population's demographic history?

- Can observed patterns of variability be explained by simple population size-change models?

The subsequent information and research presented in this dissertation aim to address these questions by describing the effect of demographic history on various measures of genetic variability, introducing unique demographic inference methods, and applying these methods to a number of population-specific genetic data sets.

1.2 Demographic information in genetic data

Observed patterns of genetic variation contain information concerning the demographic history of a population. This variation can be usefully summarized in a number of ways including the pairwise differences between sequences and the number of segregating sites. WATTERSON (1975) showed that the expected distributions of these quantities under a model of a single, randomly mating, diploid population of constant effective population size can easily be calculated. In the absence of recombination, the number of pairwise differences between two randomly chosen haplotypes has approximately a geometric distribution, while the expected number of segregating sites (S) in a sample can be calculated by

$$E(S) = \sum_{i=1}^{n-1} E(S_i) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{1.1}$$

where $\theta = 4N_e\mu$, N_e is the effective number of diploid individuals in the population, μ is the mutation rate per generation, and S_i is the number of segregating sites at frequency i in the sample.

TAJIMA (1989a) used models of instantaneous population size change to illustrate that both the mean number of pairwise differences and the expected number of segregating sites are influenced by such changes, with models of growth increasing both quantities and scenarios including a population bottleneck decreasing both quantities.

Slatkin and Hudson (1991) examine a model of exponential growth at a constant rate to show that distributions of pairwise differences under this type of model resemble a Poisson distribution, differing from the geometric-like distribution that would be expected under the constant size model examined by Watterson (1975) above. Rogers and Harpending (1992) also examine the effect of growth and bottleneck events on the mismatch distribution, finding that both scenarios produce waves in the distribution that travel at a predictable rate depending upon the time of the event. The differing expected values or distributions of both pairwise differences and segregating sites under disparate models of population history indicate that these data summaries are potentially useful for detecting departures from the model of constant population size.

The frequency spectrum of segregating sites is also influenced by historic changes in population size. The frequency spectrum under a constant population size model can be determined by equation 1.1. In general, models of population growth or bottlenecks tend to skew the frequency spectrum toward low- or intermediate-frequency variants, respectively. Tajima's D (Tajima 1989b) is a summary statistic based on the frequency spectrum that is the difference between θ estimated by average heterozygosity (θ_π) and θ estimated based on the number of segregating sites by using equation 1.1 (θ_w). When a population is at equilibrium, the expected value of Tajima's D is 0. However, under growth scenarios, the rapid influx of new alleles causes θ_w to grow faster than θ_π, leading to a negative value of Tajima's D. Likewise, under bottleneck scenarios, the loss of rare alleles contributes to a positive value of Tajima's D. Such trends in the value of Tajima's D allow for detection of deviation from the standard neutral model and also for parameter estimation, as described later.

An additional feature of genetic variation affected by demographic history is the extent of linkage disequilibrium (LD) across the genome. From a coalescent perspective, observed levels of polymorphism are due to an accumulation of mutations along the branches of the gene trees of the sampled sequences. Since tightly linked loci will have correlated genealogies, the allelic states of these linked loci will be in LD, while those of distant loci will be effectively independent (Nordborg and Tavare 2002). Therefore, demographic events will have an impact on levels of LD by influencing the branch lengths on which mutations have accumulated. Generally, simple scenarios of population growth tend to decrease levels of LD while bottlenecks or population structure lead to an increase in observed LD (Reich *et al.* 2001; Pritchard and Przeworski 2001).

In humans, studies of many data sets have revealed deviation from the standard constant population size model. Early work focused on the distribution of pairwise differences in mitochondrial and Y-chromosome data (Rogers and Harpending 1992; Harpending *et al.* 1998). Evidence for population expansion is also observed in studies of microsatellite loci where the observed allele-size variance and homozygosity are inconsistent with constant population size (Di Rienzo *et al.* 1998; Kimmel *et al.* 1998). More recently, studies of single-nucleotide polymorphisms (SNPs) in autosomal loci have also revealed incompatibility with a constant population size model based on either summary statistics such as Tajima's D (Pluzhinikov *et al.* 2002; Wall and Przeworski 2000), levels of LD (Reich *et al.* 2001), or the full frequency spectrum (Polanski and Kimmel 2003; Marth *et al.* 2004). Such results have prompted researchers to investigate not only whether a model of constant population size can be rejected, but also what particular model and combination of parameter values for a given model provide the best fit to the data.

1.3 Approaches to demographic inference

Some parameters of interest when addressing questions of demographic history include the times at which putative population size changes occurred, the growth rate, the ancestral effective population size (N_A), the current effective population size (N_e), and any intermediate population sizes. A number of methods have been employed in the attempt to estimate these and other demographic parameters. Basic estimates of N_e can be made from summary statistics such as the observed number of segregating sites using equation 1.1 above, but in most cases more sophisticated analyses are required.

One important class of estimators are those that result from likelihood-based inference. This type of analysis assumes that the observed data is a single random realization of a process that can be explained by a model with unknown parameters. The estimated parameter values are then taken to be those that maximize the likelihood of the observed data. Full-likelihood methods utilize all of the information contained in a data set, often at significant computational expense. Such methods have been applied to data sets consisting of loci that are not subject to recombination (Beerli and Felsenstein 2001; Nielsen 1999; Kuhner *et al.* 1998).

Related to the maximum likelihood methods are those that make use of a Bayesian framework, which requires that a prior distribution of model parameters be specified. The posterior distribution is then the likelihood of the observed data at particular parameter values weighted by the prior distribution. This type of analysis allows one to incorporate information that may have been gleaned from previous studies or anthropological sources (Stephens 2001). Examples of such analyses include estimation of coalescence times (S Tavare and Donnelly 1997), recombination rates (Nielsen 1999), mutation rates (Wilson and Balding 1998), and population size-change parameters (Thornton and Andolfatto 2006).

To avoid the computational burden of full-likelihood methods, DNA sequence data may be summarized prior to analysis, sacrificing information content for computational feasibility. Coalescent simulations may be used to determine whether various demographic scenarios are compatible with observed values of summary statistics. Such statistics could include summaries of the frequency spectrum such as Tajima's D (Pluzhinikov *et al.* 2002; Wall and Przeworski 2000), summaries of linkage disequilibrium such as D' (Reich *et al.* 2001), or combinations of summary statistics as described in Chapter 3. Additionally, Chapters 2 and 4 describe maximum likelihood methods that are applied to the full frequency spectrum of either unlinked or linked sites.

Despite the range of methods available for demographic inference, parameter estimation is not without complications. Star-like gene genealogies and a corresponding excess of low frequency variants characteristic of population growth can also be produced by selective sweeps; likewise, population contraction can mimic balancing selection by maintaining several alleles at intermediate frequencies in a given locus (Harpending *et al.* 1998). Therefore, less confounded estimates of demographic parameters may be made by analyzing data from noncoding regions which are less likely to be under natural selection (Pluzhinikov *et al.* 2002). An additional problem arises when a model or test statistic requires knowledge of genetic parameter values that are not known or are heterogeneous between loci. One approach to this issue in simulation-based studies is to treat the uncertain quantities as random variables which are then drawn from probability distributions with means corresponding to genome-wide averages (Pluzhinikov *et al.* 2002).

1.4 Summary of chapter contents

The parameter estimation methods described in the following chapters seek to incorporate as much of the observed genetic data as possible into demographic analyses while preserving computational feasibility. Application of these methods to both simulated and empirical data sets highlights the effects of population history on observed patterns of genetic variability and contributes to the construction of more accurate models of human demographic history, as the following analyses indicate that the constant population size model is often incompatible with observed patterns of variation.

Chapter 2 details a maximum likelihood approach to infer demographic history from the frequency spectrum of unlinked polymorphic sites. A method is also presented by which this approach may be adapted to accommodate data comprised of linked SNPs. Simulations reveal that large amounts of data (either large sample size or a large number of segregating sites) are required to make accurate inferences regarding demographic history. Additionally, accuracy is highly depended on the demographic scenario, with estimates improving with more ancient time of growth onset and smaller magnitude of growth. Application of this method to an African data set suggests compatibility with both constant population size as well as a range of recent and ancient growth scenarios, while a European data set suggests a bottlenecked history, with an 85% reduction in population size occurring ∼30,000 years ago.

Chapter 3 presents a demographic inference method that incorporates multiple aspects of genetic data, including the average number of segregating sites per locus ($\bar{S}$), the average value of Tajima's D ($\bar{D}$) (Tajima 1989b), and the population recombination rate as estimated by $\hat{\rho}$ (Hudson 2001), into a single combined summary statistic. Simulations illustrate that this combined summary statistic is more powerful than an individual summary statistic ($\bar{D}$, $\bar{S}$, or $\hat{\rho}$) in rejecting the null hypothesis of constant population size. Individual summary statistics of both an Italian and a Chinese data set are incompatible with the standard constant population size model. However, the combined summary statistics of both populations suggest compatibility with a range of bottleneck scenarios, ranging from a severe, shorter-lived reduction in population size beginning 20,000 years ago to a mild, longer-lasting bottleneck beginning 120,000 years ago.

Chapter 4 introduces a maximum-likelihood method of demographic inference that directly incorporates linkage between sites into the analysis. Unlike the method described in Chapter 2, which treats all segregating sites as independent and then adjusts confidence intervals to account for linkage *post hoc*, this method utilizes coalescent simulations with recombination to determine the probability of observed joint frequency spectra under a variety of demographic scenarios. Application of this method to previously analyzed data reveals generally concordant results, with this method producing smaller confidence intervals around the maximum likelihood estimates.

Finally, Chapter 5 suggests a method by which demographic information may be incorporated into estimation of the population recombination rate. Simulations illustrate that values of $\hat{\rho}$ ($= 4N_A r$) are overestimated under growth scenarios and underestimated under bottleneck scenarios when estimates are made assuming the standard constant population size model. The bias in $\hat{\rho}$ can be eliminated by incorporating the correct demographic model. Additionally, $\frac{\hat{\theta}_w}{\hat{\rho}}$ does not provide an unbiased estimate of the ratio of the mutation to recombination rate ($\frac{\mu}{r}$) under changing population size scenarios.

Chapters 2-4 of this dissertation are previously published or submitted manuscripts. Chapter 2 is from Adams and Hudson (2004), and Chapter 3 is from Voight *et al.* (2005). In Chapter 3, programs were adapted jointly by B. Voight and myself, and analyses were equally split between B. Voight and myself. Additionally, the data sets analyzed in Chapter 3 were generated by L. Frisse and Y. Qian. Chapter 4 has been submitted for publication in Genetics.

CHAPTER 2

INFERENCE USING THE FREQUENCY SPECTRUM

2.1 Introduction

Patterns of genetic variation in contemporary populations can be used to make inferences about past population size changes. Ideally, likelihood methods using the full data would be applied to make such inferences. For the case of DNA sequence polymorphism and where no recombination occurs between the variable sites, methods are available for carrying out such inferences (Beerli and Felsenstein 2001; Kuhner *et al.* 1998; Nielsen 1999). With incomplete linkage between sites, such approaches are frequently computationally infeasible. An exception is the case in which only two chromosomes are sampled at each locus, where Marth *et al.* (2003) have shown that maximum likelihood methods are feasible. These computational difficulties have led to the use of summary statistics such as Tajima's D (Tajima 1989b) for making inferences about past demography. For example, Wall and Przeworski (2000) and Pluzhinikov *et al.* (2002) tested compatibility between observed values of Tajima's D and values observed in simulations under constant-size and alternative demographic scenarios. Weiss and von Haeseler (1998) also focused on summaries of the data by implementing a likelihood approach based on mean pairwise differences and segregating sites for a model of complete linkage.

With free recombination between sites, the problem is greatly simplified. In this case, sites can be considered to be statistically independent of each other and the data are completely characterized by the number of polymorphic sites and the frequency spectrum. That is, we can represent the full data by $\mathbf{m} = (m_0, m_1, m_2, m_{n-1})$, where m_0 is the number of sites monomorphic in the sample, and, for $i > 0$, m_i is the number of polymorphic sites in which the derived allele is present i times in the sample of n chromosomes. We assume all polymorphic sites are biallelic. Then, $\sum_{i=0}^{n-1} m_i$ is L, the number of sites surveyed, and $\sum_{i=1}^{n-1} m_i$ is the total number of segregating sites in the sample, S. Also note that $m_0 = L - S$. In this case of free recombination between sites, full likelihood approaches are computationally undemanding. This case has been examined by Wooding and Rogers (2002), Polanski and Kimmel (2003), and Marth *et al.* (2004) and is also the focus of our study. We examine the statistical properties of demographic inferences based on $\mathbf{m}$ using maximum likelihood and assuming sites are independent. By utilizing the entire frequency spectrum, $\mathbf{m}$, rather than a summary statistic such as Tajima's D, this approach captures all available information in data sets consisting of unlinked polymorphic sites.

With linkage between sites, there is a statistical non-independence between polymorphic sites, and thus $\mathbf{m}$ for a set of linked sites would contain less information than for a set of unlinked sites. It follows that our results for unlinked sites gives an idea of the best one can do using $\mathbf{m}$ or summary statistics, such as Tajima's D, which can be calculated from it. It is important to note that for linked sites $\mathbf{m}$ does not completely characterize the data, and that full-likelihood which incorporates information about linkage disequilibrium, for instance, might result in better inferences.

The models examined here consist of either exponential growth or an instantaneous decrease followed by exponential growth, which require simultaneous estimation of either two or three parameters, respectively. We illustrate that, particularly for recent growth scenarios, data sets consisting of large numbers of segregating sites are required to produce good estimates based solely on frequency spectrum data. Our results provide a theoretical perspective on the feasibility of frequency spectrum-based parameter estimation with a modest amount of data, and we present methods to determine the approximate variance and covariance associated with such estimators under any demographic scenario of interest. The maximum likelihood method is also applied to three human data sets. The first is an African data set consisting of the original data set of FRISSE *et al.* (2001) as well as 40 additional locus pairs (Di Rienzo, unpublished data). The Seattle SNPs data, consisting of both African American and European data sets (http://pga.gs.washington.edu), is also examined. Each of these data sets consists of linked segregating sites within effectively unlinked loci, and a procedure is outlined by which one can extend this method to such data and construct the appropriate confidence regions associated with the estimators.

2.2 Model and methods

2.2.1 Demographic model

The demographic model considered is that of a population of constant effective size N_0 until time T when there was an instantaneous decrease to an intermediate population (N_{int}) size followed by exponential growth to the current population size (N_{rec}). As illustrated in Figure 2.1, this model involves four demographic parameters: N_0, N_{int}, N_{rec}, and T, where T is the time at which the instantaneous size change

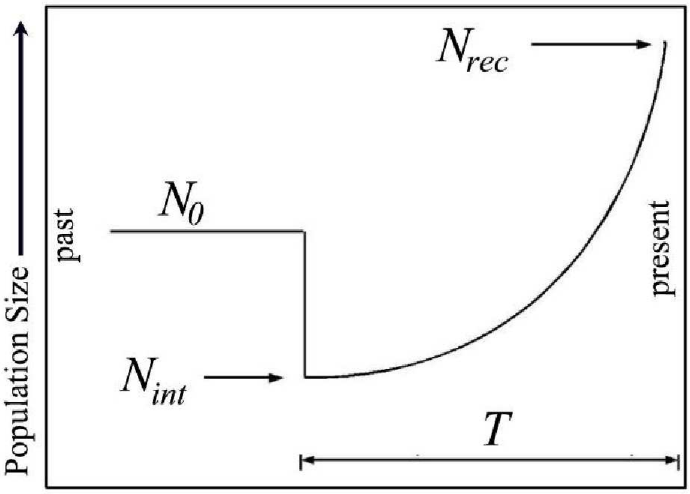


Figure 2.1: Demographic model. $f_{int}(= N_{int}/N_0)$, $f_{rec}(= N_{rec}/N_0)$, and T are the estimated parameters.

occurred. T is measured in units of $4N_0$ generations before the present. We assume the mutation rate per site, u, is small, so that the occurrence of more than one mutation occurring in the history of the sample at a single site can be ignored. We find it convenient to introduce the parameters, $f_{int} = N_{int}/N_0$ and $f_{rec} = N_{rec}/N_0$ and specify the model by $4N_0u$, f_{int}, f_{rec}, and T . This demographic model is flexible and can be generalized to the case of exponential growth with no bottleneck by setting f_{int} equal to one, or to the case of a population reduction with no recovery by setting f_{rec} equal to f_{int}. We also assume that the population is unstructured (panmictic) and that the polymorphic sites are unlinked.

2.2.2 Maximum likelihood method

The maximum likelihood approach followed here is that of Wooding and Rogers (2002) and Polanski and Kimmel (2003). Our analyses require a population survey of variation at a set of L unlinked sites. For L unlinked sites, $\mathbf{m}$ is multinomially distributed,

$$\mathrm{Prob}(\mathbf{m}) = \begin{pmatrix} & & L & \\ m_0 & m_1 & \ldots & m_{n-1} \end{pmatrix} \prod_{i=0}^{n-1} P_i^{m_i} \tag{2.1}$$

where P_0 is the probability that a site is monomorphic in the sample, and, for $i > 0$, P_i is the probability that a site is polymorphic with i copies of the derived allele. The P_i's are functions of the four parameters of the demographic model ($\theta_0, f_{int}, f_{rec}$, and T) and the sample size n .

To obtain the maximum likelihood estimates of the parameters one maximizes the right hand side of (1). We note, however, that we can write the probability of the data as

$$\begin{aligned} \mathrm{Prob}(\mathbf{m}) &= \begin{pmatrix} & & L & \\ m_0 & m_1 & \ldots & m_{n-1} \end{pmatrix} P_0^{(L-S)}(1-P_0)^S \prod_{i=1}^{n-1} \frac{P_i^{m_i}}{(1-P_0)^{m_i}} \quad (2.2) \\ &= \begin{pmatrix} & & L & \\ m_0 & m_1 & \ldots & m_{n-1} \end{pmatrix} P_0^{(L-S)}(1-P_0)^S \prod_{i=1}^{n-1} p_i^{m_i} \quad (2.3) \end{aligned}$$

where $p_i(= P_i/(1-P_0))$is the probability that a site is polymorphic with i copies of the derived allele, conditional on the site being polymorphic in the sample. P_0 and the p_i's can be written in terms of θ_0 and mean properties of sample gene trees. For example, for θ_0 small, P_0 can be expressed in terms of the mutation parameter and the mean total length of the gene genealogy:

$$P_0 \approx 1 - \theta_0 \tau(n), \tag{2.4}$$

where $\tau(n)$ is the mean total length of the gene tree of a sample of n chromosomes measured in units of $4N_0$ generations (HUDSON 1990). We define an i-branch to be a branch of the gene tree such that a mutation that occurs on the branch results in i copies of the mutation in the sample. The mean total length of i-branches in units of $4N_0$ generations, we denote by $\tau_i(n)$. Then, P_i is approximately $\theta_0\tau_i(n)$, and

$$p_i \approx \frac{\theta_0\tau_i(n)}{\theta_0\tau(n)} = \frac{\tau_i(n)}{\tau(n)} \ , \quad i > 0 \tag{2.5}$$

When time is measured in units of $4N_0$ generations, $\tau_i(n)$ and $\tau(n)$ are functions of f_{int}, f_{rec}, and T, but do not depend on N_0 or θ_0. Thus, to find the maximum likelihood estimates of the four parameters, we can first find the maximum likelihood estimates, $\hat{f}_{int}$, $\hat{f}_{rec}$,and $\hat{T}$, by maximizing $\prod_{i=1}^{n-1} p_i^{m_i}$. The maximum likelihood estimate of θ_0 , if desired, can then be obtained as $\hat{\theta}_0 = (S/L)/\hat{\tau}(n)$, where $\hat{\tau}(n)$ is the mean gene tree length with f_{int}, f_{rec}, and T set equal to the maximum likelihood estimates. In this paper, we focus on estimation of the parameters f_{int}, f_{rec}, and T, which requires maximizing $\prod_{i=1}^{n-1} p_i^{m_i}$ and does not require specifying L or m_0. Equivalently, we can consider estimation of the parameters based on the probability of $\mathbf{m}$ conditional on S, which is

$$\begin{aligned}\text{Prob}(\mathbf{m}|S) &= \frac{\text{Prob}(\mathbf{m})}{\text{Prob}(S)} \\ &= \frac{\text{Prob}(\mathbf{m})}{\begin{pmatrix} L \\ m_0 \end{pmatrix} P_0^{(L-S)}(1-P_0)^S}\end{aligned}$$

$$= \begin{pmatrix} & S & & \\ m_1 & m_2 & \dots & m_{n-1} \end{pmatrix} \prod_{i=1}^{n-1} p_i^{m_i}, \tag{2.6}$$

which does not depend on L or m_0.

To find the maximum likelihood estimates of f_{int}, f_{rec}, and T, we estimate Prob($\mathbf{m}|S$) at a set of points on a rectangular grid of values in the three-dimensional space of f_{int}, f_{rec}, and T values. For each point in the grid, we estimate the $\tau_i(n)$'s by generating 100,000 replicate gene trees with simple one site coalescent simulations. The $\tau_i(n)$'s can also be obtained as described elsewhere (Griffiths and Tavare 1998; Polanski and Kimmel 2003; Wooding and Rogers 2002), and this method can be generalized to any demographic model for which the relevant $\tau_i(n)$'s can be calculated or estimated. From the estimated τ_i's, the p_i's are calculated, which in turn are used to calculate the product, $\prod_{i=1}^{n-1} p_i^{m_i}$. Since we have ignored the problem of estimating θ_0, we do not require L, and the results are all given conditional on specified numbers of polymorphic sites.

2.2.3 Required data

Our analyses require data in the form of unlinked polymorphic sites. Ascertainment bias is not considered in this paper, so we assume that sites are randomly chosen with no prior knowledge of polymorphism and sequenced in each sampled chromosome. Frisse *et al.* (2001) sequenced ∼25kb and found 120 segregating sites in an African Hausa sample of 30 chromosomes. If we consider genome-wide polymorphism levels to be similar to that data, then approximately 104,000 sites would have to be sequenced in 30 chromosomes to assemble a data set consisting of 500 segregating sites. These 104,000 sites could be sequenced in small unlinked segments throughout the genome in order to obtain frequency spectrum data from unlinked polymorphic sites.

We consider only biallelic polymorphic sites and assume that the ancestral/derived status of each allele is known. However, not having knowledge of the ancestral state makes only a minimal difference to our results (data not shown). We also assume that all sites are surveyed in a sample of n chromosomes (or $\frac{n}{2}$ diploid individuals), but more general sampling is easily accommodated. For example, one can separate a data set into a series of frequency spectra, each with a different n. A global likelihood may then be obtained for the entire data set by multiplying the result of equation [1.6] for sets of sites with different sample sizes.

2.2.4 Sample size comparison

We compare the effect of sample size and number of unlinked polymorphic sites on both the power to reject the null hypothesis of constant population size and the quality of estimates of specific demographic parameters. The number of segregating sites, when indicated, is scaled based on the average total branch length of a random gene genealogy (τ), which will vary according to sample size and demographic scenario. For example, suppose we wish to compare sample sizes of 50 and 100 chromosomes under a growth scenario where 40-fold expansion occurred beginning 10,000 years ago. In this case $\tau(50)$ (in units of $4N_0$ generations) for a sample size of 50 chromosomes is 4.93, $\tau(100)$ for a sample size of 100 chromosomes is 6.06, and $\tau(100)/\tau(50)$ is 1.23. In words, the average total branch length of a random gene genealogy is 1.23 times greater for a sample size of 100 than for a sample size of 50. This indicates that for every 500 segregating sites discovered in a sample size of 50, approximately 615 segregating sites would be found in a sample size of 100 if the same number of sites were sequenced. Thus, when we compare sample size 50 to sample size 100, we compare n=50, S=500 to n=100, S=615. Normalizing the number of sites in this way serves to facilitate comparisons between analyses of different sample size because it takes into account the expected number of polymorphic sites in the samples of each size.

2.2.5 Asymptotic properties

To investigate whether asymptotic approximations of confidence regions and variances of estimators are applicable to data sets of modest size, we first determined whether 95% confidence regions obtained from the log likelihood ratio have the expected coverage properties. Confidence regions include those points on the grid with a log likelihood ratio less than 3 or 3.9 for simultaneous estimation of 2 or 3 parameters respectively. We also compared the observed variances and covariances of the estimators with the approximate variances and covariances calculated by estimating the inverse of the information matrix

$$I_{ij} = -E(\frac{\partial^2}{\partial\xi_i\partial\xi_j}\text{log}L) \ . \tag{2.7}$$

We estimate the expected log likelihood with

$$E(\log L) = \sum_{i=1}^{n-1} P_i(f_{rec_0},\ f_{int_0},\ T_0)\text{log}(P_i(f_{rec}, f_{int}, T)) \tag{2.8}$$

with the $P_i(f_{rec_0},\ f_{int_0},\ T_0)$ estimated by coalescent simulation for a set of points on a narrow grid of f_{rec}, f_{int}, and T values around the true parameter values of f_{rec_0}, f_{int_0}, and T_0. The second partial derivatives relevant to the information matrix are then approximated from best-fit second-degree polynomial curves. The variance and covariance observed from simulation can then be compared to the appropriate terms of the inverse of I_{ij} to determine whether the asymptotic approximations apply to data sets of modest size.

2.3 Results

2.3.1 Accuracy and precision of estimated τ_i's

As described in Methods, we estimate the relevant $\tau_i(n)$'s from 100,000 replicate gene trees generated by one site coalescent simulations. We find that our $\tau_i(n)$'s, estimated from simulation, are in very close agreement to $\tau_i(n)$'s calculated numerically by the method of Polanski and Kimmel (2003). For the case of $f_{rec} = 2.0$, $f_{int} = 0.15$, and $T = 0.0375$ for a sample size of 46, the maximum likelihood parameters for the Seattle SNPs European data set, we find that our simulated $\tau_i(46)$'s differ, at most, by 0.05% from the calculated $\tau_i(46)$'s. Additionally, we calculate the log likelihood of the Seattle SNPs European data set using 10 independent $\tau_i(46)$ estimates, each resulting from 100,000 replicate gene trees, and find that the likelihoods calculated from our simulated $\tau_i(46)$'s differ little between trials, ranging from -11987.278 to -11987.302. Since the log likelihood ratio critical values relevant for our construction of confidence regions range from 3.86 to 9.1, such a negligible fluctuation would not affect our inferred acceptance regions.

2.3.2 Power curves

Power analyses were conducted using a chi-squared test with *n-2* degrees of freedom for a sample size of n chromosomes to determine the degree of growth that would be required to reject the null hypothesis of constant population size using only frequency spectrum information. For smaller numbers of segregating sites, the degrees of freedom may vary slightly, as frequency categories with expected site counts of less than 5 are collapsed. The expected frequency spectrum under the null hypothesis is calculated from

$$p_i = \frac{\frac{1}{i}}{\sum_{j=1}^{n-1} \frac{1}{j}}, \quad 1 \leq i \leq n-1 \tag{2.9}$$

(Ewens 1979) by multiplying each p_i by the number of segregating sites. The observed frequency spectrum is obtained by estimating the p_i's from 100,000 replicates for each combination of f_{int}, f_{rec}, and T values and then multinomially sampling from these simulated p_i's. For each sample size, the number of segregating sites at each f_{rec} value is scaled based on 500 polymorphic sites in a sample size of 20, as described in Model and Methods.

Recent growth beginning 10,000 years ago

Figure 2.2a shows power curves for the scenario of recent growth beginning at $T = 0.0125$ (which, for humans, would correspond to 10,000 years ago based on a generation time of 20 years and N_0 of 10,000 , roughly corresponding to the advent of agricultural society) . We consider sample sizes of 10, 20, 50, 100, and 250 chromosomes with 500 polymorphic sizes (scaled as described above) which, for a sample size range of 10-250, corresponds to a range of 398-859 sites at constant population size ($f_{rec} = 1$) to 390-1137 sites at the most extreme growth scenario considered ($f_{rec} = 250$). With sample sizes of 20 or less, the power to reject the null hypothesis of constant population size never exceeds 0.15, even with 500-fold growth. With a sample size of 50, power reaches ∼0.5 with 50-fold growth, but barely rises above 0.6 at the largest magnitude of growth considered. As sample size reaches 100, one can reliably detect 20-fold growth, and a sample size of 250 allows for a power near 1 to reject the null hypothesis with only 5-fold growth (Figure 2.2a). It is clear that recent rapid growth can be reliably detected with frequency spectrum data only with fairly large samples (>100 chromosomes), and the most modest growth scenarios may

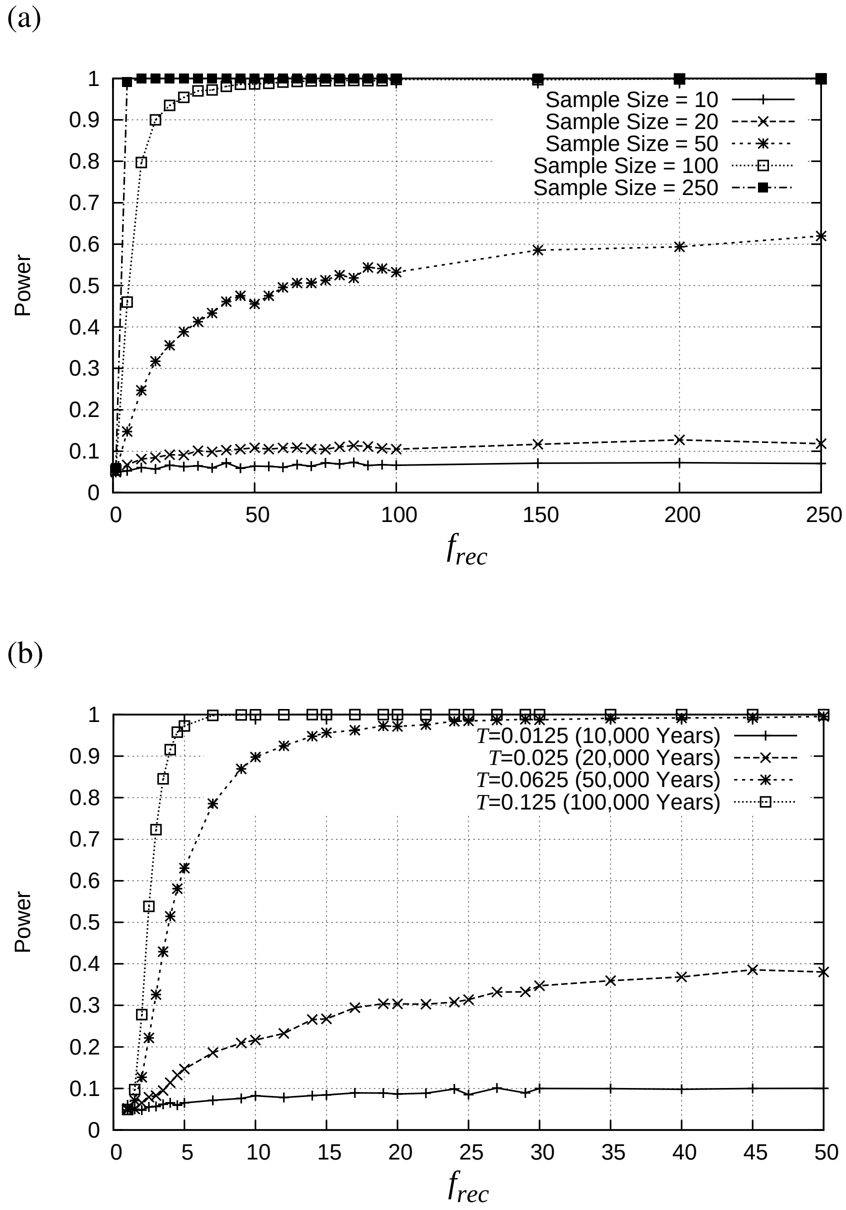


Figure 2.2: Power to detect growth with ∼500 unlinked sites. The number of sites used for each point in a curve is scaled based on 500 sites in a sample size of 20. (a) Effect of sample size on power to detect recent growth beginning 10,000 years ago ($T = 0.0125$). (b) Effect of the onset time of growth on power to detect growth with a sample size of 20.

only be detected with samples consisting of at least 250 chromosomes when data sets consist of only 500 (scaled) polymorphic sites.

More ancient growth onset

Power to reject the constant size hypothesis is also dependent upon the time that exponential growth begins, as illustrated in Figure 2.2b. While power is minimal for small sample sizes with growth beginning 10,000 years ago, power increases dramatically with more ancient growth. For example, while a sample size of 20 with 500 polymorphic sites yields virtually no power to detect any magnitude of growth beginning 10,000 years ago, if growth instead began 50,000 years ago, a sample size of 20 with the same number of sites would be sufficient to reliably detect 10-fold growth.

2.3.3 Asymptotic properties

We evaluate the distribution of our maximum likelihood estimates to determine whether asymptotic theory provides an adequate approximation of the 95% confidence regions and variance associated with our parameter estimates. Table 2.1 illustrates the proportion of maximum likelihood estimates for which the true value of the parameters lies outside the asymptotic 95% confidence region. Our simulations indicate that for large amounts of data, asymptotic theory does provide a good approximation of the 95% confidence region for the demographic scenario examined. For smaller amounts of data, the asymptotic approximation appears to be conservative, with the true parameter values lying within the 95% confidence region in ∼97-98% of the runs. We also examine a specific case corresponding to the Hausa data set, which consists of 597 sites in a sample size of 30. For a data set of this size (simulated from the Hausa MLE), we find that asymptotic approximation is especially conservative, rejecting the true parameter values in only 2.04% of the 5,000 simulated data sets.

Evaluation of asymptotic confidence region

	Sample Size = 50	Sample Size = 100
1,000 Sites	0.02332	0.01911
5,000 Sites	0.03774	0.06186
10,000 Sites	0.03750	0.05155
20,000 Sites	0.06122	0.05096

Table 2.1: Proportion of simulations where the log likelihood ratio lies outside the two-dimensional asymptotic confidence region (log likelihood ratio > 3). Each value is based on 5,000 repetitions with parameter values of $f_{rec} = 5$, $f_{int} = 0.5$, and $T = 1$.

We also determine the variance and covariance of our maximum likelihood estimators in two dimensions by both asymptotic theory and simulation (Table 2.2). In this analysis, we assume that N_{rec} is known and is 5-fold greater than $N_0(f_{rec} = 5)$, while f_{int} and T are jointly estimated. For this demographic scenario, asymptotic theory provides a good approximation for the simulated variance and covariance only when the data set consists of a large number of segregating sites.

2.3.4 Quality of estimators

We evaluate the quality of our maximum likelihood estimators by examining the distribution of the estimates under both two-dimensional and three-dimensional models.

Two-dimensional estimators

A recent growth scenario was examined in which the population size was constant until exponential growth occurred beginning 10,000 years ago. Data sets were simulated with f_{rec} ranging from 10- to 320-fold growth, f_{int} fixed at 1, and the T equal to 0.0125 (10,000 years ago based on a generation time of 20 years and N_0 of 10,000).

Asymptotic and simulated variance of two-dimensional estimators

	Variance			
	$\hat{f}_{int}$	$\hat{T}$	Covariance	Correlation
Asymptotic				
1000 Sites	0.03111	0.08806	-0.01804	-0.34
5000 Sites	0.006223	0.017612	-0.003608	-0.34
10000 Sites	0.003112	0.008806	-0.001804	-0.34
20000 Sites	0.001556	0.004403	-0.000902	-0.34
Simulated				
1000 Sites	0.016781	0.056648	-0.00745	-0.24
5000 Sites	0.005966	0.014237	-0.002720	-0.3
10000 Sites	0.002367	0.007119	-0.001810	-0.44
20000 Sites	0.001580	0.004182	-0.001040	-0.4

Table 2.2: Asymptotic variance obtained from the estimated information as described in the text (equations [2.7] and [2.8]). Results based on a demographic scenario of f_{rec}= 5.0, f_{int}= 0.5, and $T = 1.0$ and a sample size of 50 chromosomes. We assume f_{rec} is known and sites are unlinked.

For this scenario of recent growth, a sample size of at least 250 chromosomes with ∼16,000 segregating sites is required for 90% of the distribution of $\hat{f}_{rec}$ to lie within a factor of four of the true f_{rec} value for all magnitudes of growth examined. This is illustrated in Figure 2.3, which compares the $\hat{f}_{rec}$ distribution under this recent growth scenario for a sample size of 50 and 250 for 20-fold growth. As the magnitude of growth increases, $\hat{f}_{rec}$ becomes biased more severely upward. This result is similar to that obtained for growth beginning at T equal to 0.0625 (Table 2.3).

The estimates of T, however, are not subject to the upward bias seen in $\hat{f}_{rec}$. Instead, estimates of T are improved as the magnitude of growth increases (Table 2.4). Sample size also has a dramatic effect on the distribution of $\hat{T}$, as illustrated in Figure 2.4, which compares the $\hat{T}$ distribution for a sample size of 50 and 250 respectively. With 5,000 sites in a sample size of 50 chromosomes, 95% of $\hat{T}$ estimates lie within a factor of three of the true T value for ten-fold growth, with 95% of the

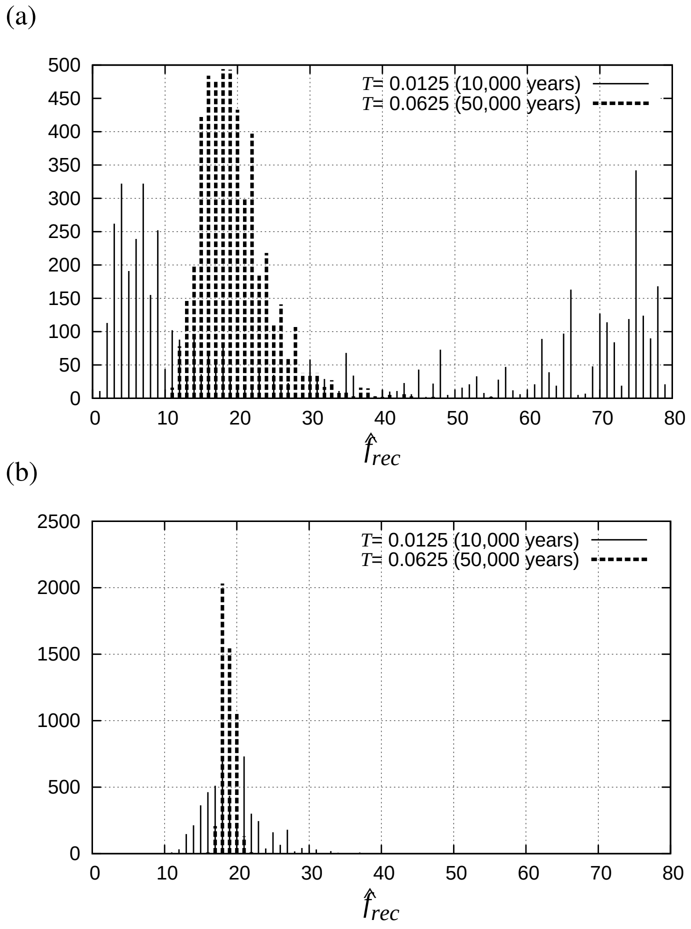


Figure 2.3: Distribution of $\hat{f}_{rec}$. Histograms are based on 5,000 simulated data sets where $f_{rec} = 20$ and f_{int} is fixed at 1. (a) 50 chromosomes; 10,000 sites (b) 250 chromosomes; 16,244 sites ($T = 0.0125$) and 20,322 sites ($T = 0.0625$).

distribution lying within a factor of 1.5 for 320-fold growth, the most severe growth scenario examined.

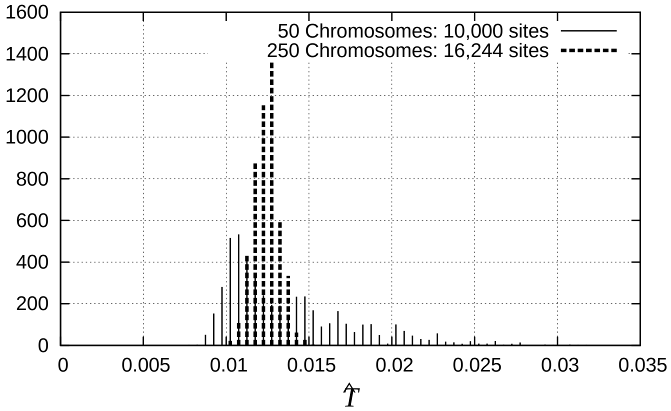


Figure 2.4: Distribution of $\hat{T}$. Histograms are based on 5,000 simulated data sets with parameters f_{rec}= 20, f_{int}= 1 (fixed), T = 0.0125 , each consisting of 10,000 polymorphic sites in 50 chromosomes or 16,244 polymorphic sites in 250 chromosomes.

We explored another growth scenario in which the onset of growth was more ancient, beginning 50,000 years ago. In this case, the quality of the estimators improved, and 90% of the $\hat{f}_{rec}$ distribution was within a factor of four of the true f_{rec} value for a sample size of 50 with 5,000 sites (as opposed to a sample size of 250 and ~16,000 sites under the more recent growth scenario). Figure 2.3 reveals the improvement in the $\hat{f}_{rec}$ estimator with the more ancient time of growth onset. As in the recent growth scenario, increasing the degree of growth both increased the bias and widened the quantiles of the $\hat{f}_{rec}$ distribution (Table 2.3). The T estimates under the more ancient growth scenario were also improved over the analogous recent growth estimates, with 95% of the $\hat{T}$ distribution within a factor of two of the true T value for all magnitudes of growth examined with data sets as small as a sample size of 50 with 1,000 sites.

Distribution of $\frac{\hat{f}_{rec}}{f_{rec}}$

f_{rec}	Mean	Std. Dev.	0.05	0.5	0.95
10	1.0674	0.2926	0.7	1.0	1.6
20	1.0925	0.3862	0.7	1.0	1.7
40	1.1629	0.5722	0.6	1.0	2.2
80	1.2980	0.8200	0.5	1.0	3.3
160	1.3906	1.0092	0.4	1.0	3.9
320	1.5072	1.1824	0.3	1.0	3.9

Table 2.3: Time of expansion is ∼50,000 years ($T = 0.0625$), and f_{int} is fixed at 1. Simulated data sets consist of 5,000 unlinked sites in a sample size of 50. The $\hat{f}_{rec}$ grid includes 40 grid points from $\hat{f}_{rec} = 0.1(f_{rec})$ to $\hat{f}_{rec} = 4(f_{rec})$.

Distribution of $\hat{T}$

f_{rec}	Mean	Std. Dev.	0.025	0.5	0.975
10	0.015254	0.007262	0.0077	0.0125	0.0365
20	0.014790	0.005698	0.0089	0.0125	0.0269
40	0.014647	0.004850	0.0089	0.0125	0.0281
80	0.014118	0.003349	0.0101	0.0125	0.0221
160	0.013706	0.002479	0.0101	0.0137	0.0197
320	0.013331	0.001947	0.0101	0.0125	0.0173

Table 2.4: Time of expansion is ∼10,000 years ($T = 0.0125$), and f_{int} is fixed at 1. Simulated data sets consist of 5,000 unlinked sites in a sample size of 50. The $\hat{T}$ grid includes 40 grid points from $\hat{T} = 0.1(T)$ to $\hat{T} = 4(T)$.

Three-dimensional estimators

We consider a three-dimensional model of a constant sized population that experienced an instantaneous decrease to 0.05 times its initial size 100,000 years in the past, followed by exponential growth until the present to a final size of five times the initial population size ($f_{rec} = 5$; $f_{int} = 0.05$; $T = 0.125$). All three parameters were estimated for 5,000 simulated data sets. Under this model, 90% of the distribution of each of the three estimators falls within a factor of four of the respective true values with data sets as small as 500 sites in a sample size of 30 (Table 2.5). If a large data set consisting of 10,000 polymorphic sites in 50 chromosomes were available, 95% of the estimates of all three parameters would lie within a factor of 1.5 of the true parameter values. As would be expected, estimates of any of the three parameters are improved by fixing one of the parameters at its true value (data not shown), indicating that incorporation of prior knowledge of one of the parameters would be beneficial.

2.3.5 Applications

We apply the maximum likelihood method to data obtained from an African Hausa population (Di Rienzo, unpublished data) as well as to the African American and Eu-

Distribution of three-dimensional MLEs

	Mean	Std. Dev.	0.05	0.5	0.95
$\hat{f}_{rec}$	6.017479	4.233278	1.5	5	15
$\hat{f}_{int}$	0.056695	0.042797	0.01	0.045	0.14
$\hat{T}$	0.112958	0.038210	0.055	0.115	0.175

Table 2.5: Simulated data sets consist of 500 sites in a sample size of 30 chromosomes, where $f_{rec} = 5.0$, $f_{int} = 0.05$, and $T = 0.125$. The three-dimensional grid includes $\hat{f}_{rec}$ values from 0.5 to 14.5 (at 0.5 intervals), $\hat{f}_{int}$ values from 0.01 to 0.15 (at 0.005 intervals) and $\hat{T}$ values from 0.025 to 0.225 (at 0.01 intervals).

Analysis of Hausa and Seattle SNPs data sets

	$\hat{f}_{rec}$	$\hat{f}_{int}$	$\hat{T}$	Likelihood[†]	p-value[‡]
Hausa Data Set					
MLE	3.1	1	6.1	-1411.14	0.304
Constant Population Size	1	1	-	-1413.34	0.204
SSNPs Afr. Am. MLE	1.9	1	0.27	-1411.69	0.113
SSNPs Afr. Am. Data Set					
MLE	1.9	1	0.27	-15448.17	$2 \text{ x } 10^{-4}$
Constant Population Size	1	1	-	-15544.63	$<< 1 \text{ x } 10^{-4}$
SSNPs Eur. Data Set					
MLE	2.0	0.15	0.0375	-11987.28	0.2015
Constant Population Size	1	1	-	-12022.41	$< 1 \text{ x } 10^{-4}$

Table 2.6: Results obtained as described in text.

[†] Note that this is not true likelihood since the SNPs are not entirely unlinked
[‡] *p*-values calculated from χ^2 goodness-of-fit test where the distribution of the χ^2 test statistic is simulated for each data set, accounting for linkage as described in the text.

ropean (CEPH) samples of the Seattle SNPs data set (http://pga.gs.washington.edu).

Hausa data

The Hausa data set consisted of the data of Frisse *et al.* (2001) in conjunction with additional unlinked locus pairs (Di Rienzo, unpublished data), which resulted in a data set consisting of 30 chromosomes and 597 polymorphic sites in an African sample, the Hausa of Cameroon. The sites in this data set include linked polymorphic sites within 50 effectively unlinked loci, but in the maximum likelihood analysis we treat each site as though it provides independent information. As seen in Table 2.6, $\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$ for this data set.

We perform a χ^2 goodness-of-fit test on the Hausa data set to determine whether the maximum likelihood parameters can be accepted as an explanation of the Hausa data. However, this test assumes that each site is independent, which is not the case for this data set. Because the linkage between sites will affect the 95% critical value of the χ^2 test statistic, we determine the critical value of the test statistic distribution for this data set by coalescent simulation with recombination (Hudson 1983, 2002). We simulate 5,000 data sets, each consisting of 30 chromosomes and 50 unlinked loci. The input parameters for the simulation included Watterson's estimate of θ (estimated to be 0.0012 per bp), the recombination rate (estimated to be 5.99 x 10^{-4} per bp), the average locus length (10,286 bp), and a gene-conversion to crossing-over ratio of 2. Polymorphic sites within the middle 8,000 bp were ignored to mimic the locus pair data (Frisse *et al.* 2001). Because the ancestral/derived status of each allele was not considered, each simulated frequency spectrum was folded at frequency 0.5 prior to performing the χ^2 goodness-of-fit test. Based on these simulations, we find the 95% critical value of the χ^2 test statistic to be 39.39, as opposed to a critical value of 23.68 (14 degrees of freedom) if all sites were independent. The χ^2 goodness-of-fit test statistic for the Hausa data set under its maximum likelihood estimate of $\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$ is 26.30 ($p = 0.304$), indicating that this scenario can not be rejected at the 0.05 significance level. Note, however, that this demographic scenario would have been rejected without properly accounting for the linkage within the data set. We also consider the equilibrium model of constant population size for this data set and obtain a χ^2 goodness-of-fit test statistic of 29.24 ($p = 0.204$). Based on an analysis of 10 locus pairs (a subset of the 50 locus pairs examined here), Frisse *et al.* (2001) also concluded that the Hausa data set is consistent with the equilibrium model.

Seattle SNPs

We also examine both the African American and the European samples of the Seattle SNPs data set (http://pga.gs.washington.edu). These data sets consist of

12,587 and 7,712 total SNPs across 138 loci in the African American and European samples, respectively. We considered only those SNPs that were sequenced in the entire panel of 48 (African American) or 46 (European) chromosomes to facilitate evaluation of confidence regions and goodness-of-fit by simulation. Additional analyses incorporating more of the SNPs are described in the Discussion. The frequency spectrum of non-synonymous SNPs has been shown to differ from that of synonymous SNPs (Cargill *et al.* 1999; Fay *et al.* 2001; Wooding and Rogers 2002), so we also removed all SNPs that result in an amino-acid coding change to minimize the inclusion of those SNPs subject to non-neutral evolutionary processes. This resulted in a final data set of 5,892 SNPs for the African American data set and 4,211 SNPs for the European data set. We applied our maximum likelihood method to these data sets, treating all SNPs as unlinked, and found that the three-dimensional maximum likelihood estimates of f_{rec}, f_{int}, and T are $\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$ for the African American data set and $\hat{f}_{rec} = 2$, $\hat{f}_{int} = 0.15$, and $\hat{T} = 0.0375$ for the European data set (Table 1.6). These estimates suggest a scenario of very slow growth over a long period of time with no bottleneck for the African Americans and a fairly recent population bottleneck with $\sim$13-fold recovery for the Europeans.

To determine whether the demographic model we consider is compatible with the Seattle SNPs data, we simulate the distribution of the goodness-of-fit test statistic for this data set as described for the Hausa data set. For these simulations, each data set consisted of 48 chromosomes and 138 loci. The input parameters were that of the Hausa data set, except substituting the average length of a Seattle SNPs locus. In this case, each locus was simulated fixing the number of segregating sites to be the average number of segregating sites per locus in the African American or European Seattle SNPs data set, so each simulated data set contained the same total number of segregating sites as our observed Seattle SNPs African American or European data set. Because the Seattle SNPs data sets do not specify the ancestral/derived status of each allele, each simulated frequency spectrum is again folded. The 95% critical values of the distribution were found to be 48.68 (African) and 137.36 (European) as opposed to 35.17 (23 degrees of freedom) and 33.92 (22 degrees of freedom) if all sites were unlinked.

Using these simulated critical values, a χ^2 goodness-of-fit test indicates that the maximum likelihood parameters produce an expected frequency spectrum that is not significantly different from the observed Seattle SNPs European data (χ^2= 122.98*;* p = 0.2015). Therefore, we can accept our simple bottleneck model as a reasonable explanation for this data set. The same test indicates that a constant population size model is not compatible with the European data (χ^2= 207.286*;* $p < 1$ x 10^{-4}).

However, the χ^2 goodness-of-fit test on the African American data set reveals that the frequency spectrum predicted by the maximum likelihood estimates of f_{rec}, f_{int}, and T is significantly different from the empirical Seattle SNPs African American frequency spectrum (χ^2= 86.64*;* p = 2x10^{-4}); therefore, our simple demographic model can not be accepted as a complete explanation of the African American data set, although the fit is better than that predicted by the constant population size model (χ^2= 268.66*;* $p << $ 1x10^{-4}). Figure 2.5 provides a visual comparison of the observed Seattle SNPs frequency spectrum to the frequency spectra predicted by both the maximum likelihood parameters and constant population size parameters, indicating that the lack of fit of the maximum likelihood parameters does not seem to be confined to any particular non-singleton frequency class. However, our demographic model with the maximum likelihood parameters appears to provide a better fit to the data than the equilibrium model, particularly in the singleton class. In addition, we note that a χ^2 goodness-of-fit test shows that the Hausa data are compatible with f_{rec} = 1.9, f_{int} = 1, and T = 0.27 , the estimates obtained from the Seattle SNPs African

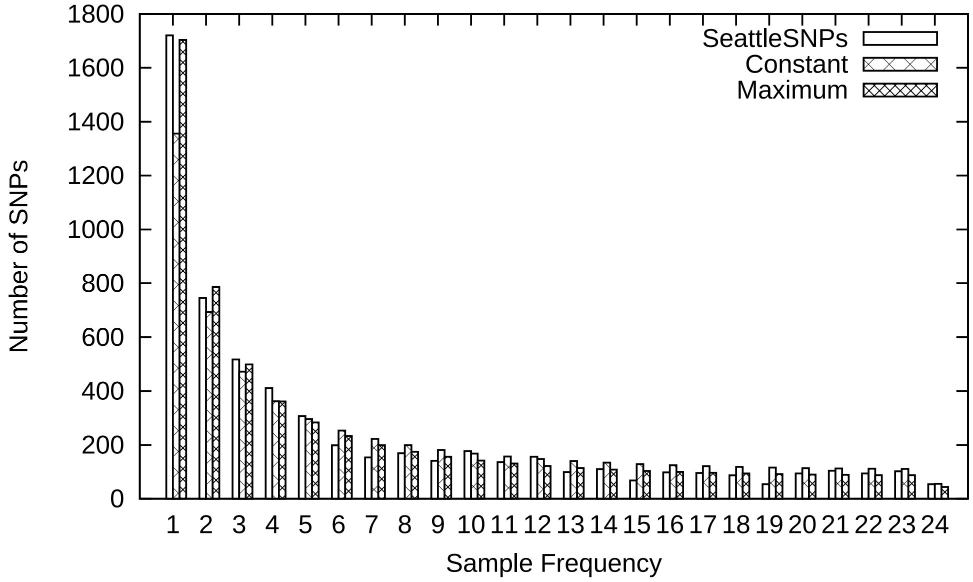


Figure 2.5: African American Seattle SNPs folded frequency spectra comparison. Empirical Seattle SNPs frequency spectrum and the expected frequency spectrum for demographic parameters corresponding to the Seattle SNPs maximum likelihood estimate (f_{rec}=1.9, f_{int}=1, and $T = 0.27$) and constant population size. The number of SNPs at a sample frequency of i is equal to the total number of SNPs (5,892) times p_i (folded). For constant population size, p_i's are obtained from Equation [2.9], and, for the maximum likelihood parameters, p_i's are obtained from simulation as described in the text.

American data set ($\chi^2 = 33.46$; $p = 0.113$).

Because the sites in the Hausa and Seattle SNPs data sets are not entirely unlinked, asymptotic approximation of confidence intervals is not appropriate. We simulate 10,000 data sets as described above for both the Hausa and Seattle SNPs data sets, using their respective maximum likelihood estimates for input parameters, and apply the maximum likelihood method to the folded frequency spectrum of each simulated data set. For the Hausa and Seattle SNPs African American data sets, we estimate both f_{rec} and T, fixing f_{int} at 1, which was the maximum likelihood estimate for both data sets. All three parameters were estimated for the data sets simulated from the Seattle SNPs European parameters. The ratio of the log likelihood at the maximum likelihood parameters to the log likelihood at the parameters from which the data set was simulated could then be calculated. From the log likelihood ratio distribution, we determine the 95% critical value to be 3.86 for the Hausa data set and 4.85 for the Seattle SNPs African American data set as compared to the asymptotic critical value of 3.0 for two-dimensional maximum likelihood estimates. The 95% critical value of 9.1 was found for the European data set as compared to the asymptotic critical value of 3.9 for three-dimension estimation. Using the critical values from simulation, we can easily reject the constant size population model for the Seattle SNPs African American and European data sets since the log likelihood ratios are 96 and 35, respectively (Table 1.6).

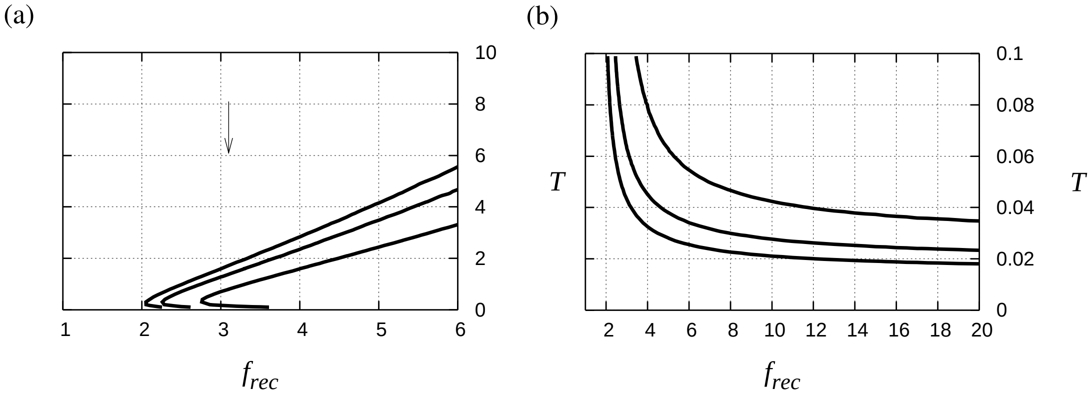


Figure 2.6: Hausa confidence region. The third dimension, f_{int} is fixed at 1. (a) Maximum likelihood estimate (MLE) is indicated by the arrow ($\hat{f}_{rec}$=3.1, $\hat{T}$ = 6.1). (b) Focus on recent growth times with expanded f_{rec} range. The leftmost, middle, and rightmost contours represent the 95%, 99%, and 99.9% confidence intervals (3.86, 6.38, and 10.08 log likelihood units, respectively).

Figure 2.6 provides a visual representation of the 95%, 99%, and 99.9% confidence regions of the Hausa data set obtained by including all parameter values for which the log likelihood ratio is $\leq$ 3.86, 6.38, and 10.08, respectively. Likewise, Figure 2.7 illustrates the analogous confidence regions for the Seattle SNPs African American (2.7a) and European (2.7b) data sets.

2.4 Discussion

Our power analyses on models with exponential growth beginning 10,000 years ago illustrate that the frequency spectrum does not provide sufficient information to reject the null hypothesis of constant population size when either small sample sizes (< 50 chromosomes) or small numbers of unlinked sites (< 1,000) are available. This result should serve as a cautionary note to researchers interested in demographic models involving expansion as recent as 10,000 years. Prior knowledge of the model of

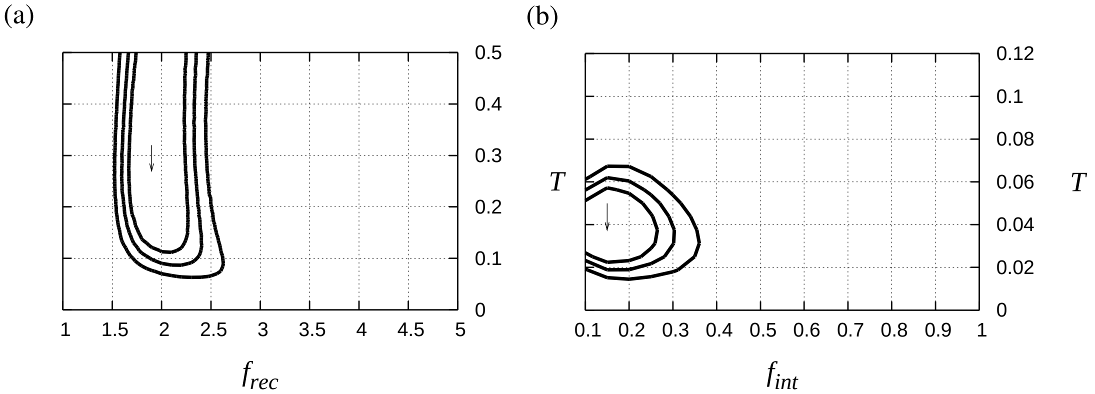


Figure 2.7: Seattle SNPs Confidence Regions. (a) African American data set, with MLE indicated by the arrow ($\hat{f}_{rec} = 1.9$, $\hat{T} = 0.27$). The third dimension, f_{int} is fixed at the MLE of $\hat{f}_{int} = 1$. (b) European data set, with MLE indicated by the arrow ($\hat{f}_{int} = 0.15$, $\hat{T} = 0.0375$). The third dimension (f_{rec}) is fixed at the MLE of $\hat{f}_{rec} = 2.0$. The innermost, middle, and outermost contours surrounding the MLE represent the 95%, 99%, and 99.9% confidence regions, respectively.

interest should also be considered when determining whether the frequency spectrum retains the requisite information for demographic inference, as the power to detect departures from constant size increases with both the extent of growth (Figure 2.2a,b) and the time since the onset of growth (Figure 2.2b).

Application of the maximum likelihood method on recent growth scenarios reveals that data sets consisting of at least 250 chromosomes with at least 10,000 scaled segregating sites (15,838-17,140 segregating sites depending upon the true f_{rec} value) are required for the $\hat{f}_{rec}$ distribution of to have 95% critical values that fall within a factor of four of the true f_{rec} value if growth began as recently as 10,000 years ago ($T = 0.0125$). Unless large sample sizes and many unlinked sites are surveyed, the frequency spectrum alone provides little information about the magnitude of growth that has occurred relatively recently. As f_{rec} increases, the frequency spectrum becomes more distinct from that which would be expected under a constant size scenario. However, with increasingly extreme recent growth, the frequency spectrum becomes less distinguishable from that of other severe growth scenarios, and it becomes more difficult to estimate the f_{rec} parameter with frequency spectrum information alone.

While it is difficult to accurately estimate f_{rec} for scenarios of recent growth, T can be estimated with more modest amounts of data. The distribution of $\hat{T}$ has 95% critical values that fall within a factor of four of the true T value for sample sizes as small as 50 chromosomes and 5,000 sites, as compared to a sample size of 250 and 15,838-17,140 sites required to estimate f_{rec} to the same accuracy. Additionally, $\hat{T}$ is not subject to the upward bias seen in $\hat{f}_{rec}$, and estimates of T actually improve with increasing f_{rec}. Estimates of both f_{rec} and T improve as the onset of growth becomes more ancient. This observation is consistent with our observation that power to reject the null hypothesis of constant population size with frequency spectrum data increases with scenarios of more ancient growth (Figure 2.2b).

Simultaneous estimation of all three parameters results in estimator distributions where 90% of the estimates lie within a factor of four of the true parameter values with data sets as small as 500 segregating sites in a sample size of 30 for a model where the population decrease and subsequent expansion began 100,000 years ago and the present population is only five times the initial population size. These estimates benefit from both a more ancient time of growth onset and a modest magnitude of growth that is not subject to the upward bias seen in more severe growth scenarios. The ability of the frequency spectrum alone to elucidate the time and magnitude of population size change events is, therefore, greatly dependent upon the underlying demographic model. While ancient demographic events may be inferred relatively accurately from contemporary frequency spectrum patterns, more recent and severe episodes of growth are problematic for this method and require exceedingly large amounts of unlinked data. For these recent growth scenarios, it is possible that more informative estimates could be obtained by using a method that uses linked polymorphic sites and considers additional aspects of the data such as levels of linkage disequilibrium.

Evaluation of the asymptotic properties of our maximum likelihood estimators indicates that asymptotic theory provides a reasonable approximation of the confidence intervals associated with the estimators. As we illustrate with the Hausa and Seattle SNPs data sets, it is also possible to construct these confidence intervals around a maximum likelihood estimate through simulation. By simulating data sets that closely match the properties of the observed data set, one can estimate the critical value of this log likelihood ratio distribution and construct corresponding confidence regions. This procedure is particularly relevant when asymptotic approximation is not appropriate, such as when the segregating sites in a data set are not unlinked.

We apply the maximum likelihood method to both the African Hausa data set and the African American and European samples of the Seattle SNPs data set. In both the Hausa and the Seattle SNPs data sets, the segregating sites are not entirely unlinked, but the maximum likelihood analysis treats them as though each site provides independent information. However, we illustrate how one may use simulation to construct confidence regions and use goodness-of-fit tests that take into account the linkage between sites.

In the Seattle SNPs African American data set, the simulated 95% confidence interval clearly allows for rejection of the constant population size model, since the log likelihood of observing the data is almost 100 units less with the constant size parameters than with the estimated parameters. The maximum likelihood estimates of $\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$ based on the Seattle SNPs African American data correspond to a slow, ancient growth scenario where growth began over 200,000 years ago to a present size of $\sim$2 times the initial population size.

The simulated 95% confidence region around the Seattle SNPs African American maximum likelihood estimates, shown in Figure 2.7a, includes only a very narrow range of f_{rec} values within 1.6 to 2.5. However, the confidence region includes a wide range of T values ranging from as recent as 80,000 years ($T = 0.1$) to the most ancient time examined, 800,000 years ($T = 1$), assuming a generation time of 20 years and an N_0 of 10,000. Even with the most recent compatible T value, it is not surprising that this data set allows for rejection of the constant size hypothesis with only an estimate of 2-fold growth. Our power analyses show that a data set consisting of 50 chromosomes has a power of 0.9 to reject the constant size hypothesis with only 1,000 unlinked sites for 2-fold growth beginning 100,000 years ago. While the Seattle SNPs data set does not consist of entirely unlinked sites, our analysis included more than 5,000 polymorphic sites across 138 loci, which should allow for comparable power.

Despite the compact confidence region (Figure 2.7a), visually reasonable fit of the frequency spectrum under the maximum likelihood parameters to the observed data (Figure 2.5), and compatibility with the Hausa data set, a χ^2 goodness-of-fit test indicates that our simple three-dimensional demographic model with the maximum likelihood estimates obtained from the Seattle SNPs African American data set is incompatible with the Seattle SNPs data ($p = 2 \times 10^{-4}$), even when linkage is taken into account. The visual comparison between the observed and maximum likelihood frequency spectra in Figure 1.5 seems to indicate that the number of singletons expected under the maximum likelihood parameters is very close to the observed value, and therefore the incompatibility must be due to some combination of the other frequency categories. The loci in the Seattle SNPs data set were chosen because of their role in inflammatory pathways, and may reflect the action of evolutionary forces other than population size changes. Although we removed those SNPs that result in amino acid coding changes, which are more apt to be subject to natural selection, it is still probable that this data set includes SNPs that are mildly deleterious and may influence the frequency spectrum toward greater numbers of low-frequency variants and mimic evidence of growth.

The African American population sampled for the Seattle SNPs data set may also be subject to population structure and admixture, which could affect the frequency spectrum and confound our inference about demographic history (Ptak and Przeworski 2002). To determine the effect of European admixture on maximum likelihood estimates obtained from an African data set, we randomly combine 6 Italian chromosomes (Di Rienzo, unpublished data) with the 30 Hausa chromosomes at each of the 50 locus pairs of the Hausa data set, which resulted in a total data set of 657 polymorphic sites in 36 chromosomes. This represents approximately 17 percent European admixture, which is consistent with admixture estimates obtained from African American populations (Parra *et al.* 1998). Admixture of this proportion had virtually no effect on the maximum likelihood estimates or confidence intervals based on the original Hausa data set (data not shown). Regardless, that does not eliminate the possibility that either the true population structure could involve admixture in different proportions or that admixture in a larger data set such as the Seattle SNPs would produce a more prominent effect. The frequency spectrum of the Seattle SNPs data set is certainly not consistent with an equilibrium model of constant population size, although the degree of growth predicted is less than some previous reports based on African populations (Pritchard *et al.* 1999; Aris-Brosou and Excoffier 1996). However, our estimate of two-fold growth beginning as recently as 80,000 years ago is consistent with a recent study based on the frequency spectrum in an African American population (Marth *et al.* 2004).

The maximum likelihood parameters estimated from this data set are consistent with the Hausa data set, which contains noncoding loci that are less likely to be subject to confounding factors such as selection. However, this analysis does not preclude population structure within Africa as a potential influence on the maximum likelihood estimates of the Hausa data set. The maximum likelihood estimates from the Hausa data set ($\hat{f}_{rec} = 3.1$, $\hat{f}_{int} = 1$, and $\hat{T} = 6.1$) correspond to a scenario of slow, ancient three-fold growth, beginning several million years ago. However, the confidence region associated with this data set (Figure 1.6a) is consistent with a wide range of growth scenarios, including both the demographic history estimated from the Seattle SNPs data set ($\hat{f}_{rec} = 1.9$, $\hat{f}_{int} = 1$, and $\hat{T} = 0.27$) and also $f_{rec} = 1$, which corresponds to constant population size. Additionally, Figure 2.6b provides a close-up view of the acceptance region of Figure 2.6a, considering only more recent values of T where the onset of growth occurs no more than 80,000 years ago. If we focus on these T values, it is clear that our confidence regions on this data set do not exclude scenarios of 20-fold or more growth, provided that the time of onset is correspondingly more recent. For example, if we believe that the Hausa population has undergone growth greater than 5-fold, then our analysis indicates that the growth must have begun no earlier than $T = 0.045$ (36,000 years ago if N_0 is 10,000 assuming a generation time of 20 years). Growth of that magnitude or larger is rejected at the 1% level (Figure 2.6b) for values of T greater than 0.045 and less than 3 (approximately 36,000 to 2.4 million years ago).

Our analysis of the Seattle SNPs European data set reveals an estimated demographic history of $\hat{f}_{rec} = 2.0$, $\hat{f}_{int} = 0.15$, and $\hat{T} = 0.0375$, which corresponds to an 85% reduction in population size at $T = 0.0375$ (30,000 years ago assuming N_0 = 10,000 and a 20 year generation time) and then approximately 13-fold exponential growth to a current population size of twice the ancestral size. In constructing our data set, we exclude all SNPs that are not successfully typed in every chromosome to facilitate construction of appropriate confidence regions and estimation of χ^2 critical values through simulation. However, we note that if all SNPs that were typed in at least half of the sampled chromosomes (7,410 SNPs) were included in this analysis, we get only a slightly different estimate ($\hat{f}_{rec} = 1.25$, $\hat{f}_{int} = 0.2$, and $\hat{T} = 0.05$) that differs by less than 2 log likelihood units from our maximum likelihood estimate based on the filtered data (-20668.96 versus -20670.93 when all SNPs are included). Since the 95% confidence region includes all parameter values within 9.1 log likelihood units from the maximum, it is not likely that this filtering of the data would result in a significant shift in our acceptance region.

A χ^2 goodness-of-fit test indicates that frequency spectrum produced by the estimated parameters ($\hat{f}_{rec} = 2.0$, $\hat{f}_{int} = 0.15$, and $\hat{T} = 0.0375$) is a reasonable match to the observed Seattle SNPs European frequency spectrum with a p-value of 0.2015. The constant size model is both rejected by the goodness-of-fit test and excluded by the simulated likelihood ratio confidence region for this data set (Table 2.6, Figure 2.7b). These results implicate a bottlenecked history for this European data set, which is consistent with previous studies (Marth *et al.* 2003, 2004) and the 'Out of Africa' model for human population history (Harpending *et al.* 1998). Since the Seattle SNPs European data set is composed of the same coding loci as the Seattle SNPs African American data set, it seems reasonable that the lack of agreement between the frequency spectrum predicted by the maximum likelihood parameters and the observed African American frequency spectrum is more likely to be due to population structure than the presence of slightly deleterious variants in the data set. Of course, the good fit of the European maximum likelihood parameters does not preclude the possibility of population structure or selection within the European data set as well.

As we have indicated earlier, there are a number of confounding factors to consider when attempting to infer demographic history based on frequency spectrum information, including population structure (past or present) and selection. An additional complication that is not considered by these analyses is ascertainment bias and genotyping error. It has been shown that ascertainment bias can lead to large errors in maximum likelihood-based demographic inference (Kuhner *et al.* 1998; Wakeley *et al.* 2001). Polanski and Kimmel (2003) have also shown that exclusion or misclassification of low-frequency SNPs can result in estimated growth rates that are significantly lower than the true value. Note, however, that the sites represented in the Di Rienzo and Seattle SNPs data sets were chosen without prior indication of polymorphism status. Therefore, the analyses on these data sets would not be influenced by ascertainment bias due to using a discovery sample for SNP identification. However, we cannot exclude the possibility that genotyping errors have biased our inferences.

2.5 Conclusions

Analysis of this maximum likelihood method indicates that demographic inferences can be drawn from frequency spectrum data when sufficient amounts of data are available. Asymptotic theory or simulation can be used to determine the variance and covariance associated with these estimators to determine whether the maximum likelihood estimates would be meaningful for a particular demographic model and amount of data that may be available. However, our results show that very large amounts of data may be required to obtain practical confidence regions, particularly in models involving recent growth. For growth beginning as recent as 10,000 years ago, the power to reject the hypothesis of constant population size is very low with sample sizes of less than 20 chromosomes. In order to make accurate inferences under this type of recent growth model using the frequency spectrum alone, both large samples ($>$ 100 chromosomes) and a large number of unlinked sites ($>$ 5,000 sites) are required, although estimators improve as the time of onset of growth becomes more ancient. In scenarios of extreme growth, there is also a severe bias in $\hat{f}_{rec}$, even with large amounts of data. However, T can be estimated with more modest amounts of data, and $\hat{T}$ is not subject to the bias seen in $\hat{f}_{rec}$, indicating that one may obtain reasonable estimates of the time of population size change events, even if the magnitude is biased. This maximum likelihood method incorporates all available information contained in unlinked polymorphic sites, and parameter estimation methods based on summaries of the frequency spectrum require even larger amounts of data to be equally as informative. Therefore, for scenarios where the entire frequency spectrum of modest data sets does not provide an adequate amount of information, it may be necessary to incorporate additional aspects of linked data in order to improve estimates of demographic parameters.

Application of the maximum likelihood method to three human data sets implicates differing demographic histories for African versus European data sets. The African Hausa data set is compatible with a wide range of growth scenarios, ranging from slow, ancient growth, to some scenarios of very recent, rapid growth. However, we can reject episodes of greater than five-fold growth beginning more than 36,000 and less than 2.4 million years ago on the basis of this data set. The Seattle SNPs African American data set also supports a model of growth, although a goodness-of-fit test indicates that the best-fit model of ancient, slow growth is not sufficient to explain observed frequency spectrum. Maximum likelihood analysis of the Seattle SNPs European data set reveals that the best-fit model is one of a population bottleneck occurring approximately 30,000 years ago, reducing the population to 15 percent of the ancestral size, followed by 13-fold growth to a current population size that is twice the ancestral size.

CHAPTER 3

INFERENCE USING MULTIPLE SUMMARY STATISTICS

3.1 Introduction

Elucidating how and when populations change in size is an important element in reconstructing evolutionary history, as these changes often reflect crucial events in the history of a species, such as range expansions, environmental changes, admixture between groups (Lahr and Foley 1998). Also, making inferences based on population variation data typically requires the specification of a demographic model. Such applications include detecting the signature of natural selection or estimating recombination rates from patterns of linkage disequilibrium (LD) (Akey *et al.* 2004; Williamson *et al.* 2005; Smith and Fearnhead 2005; Stajich and Hahn 2005). Finally, better knowledge of demographic histories in human populations is particularly important for whole genome LD-based association studies (Goldstein and Chikhi 2002; Reich *et al.* 2002).

Motivated by the excess of rare variants observed in mitochondrial DNA data, attention initially focused on models of ancient population growth and on the idea that population expansions may have accompanied either the dispersal out of Africa or the emergence of new tool technology in the Upper Palaeolithic (Slatkin and Hudson 1991; Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Sherry *et al.* 1994; Rogers and Jorde 1995; Weiss and von Haeseler 1998). However, the accumulation of nuclear sequence variation surveys showed that this simple growth model was consistent with the observed frequency spectrum only for a subset of the loci (Hey 1997; Wall and Przeworski 2000; Przeworski *et al.* 2000). Likewise, LD surveys revealed marked differences in the rate of LD decay in African compared to non-African populations (Tishkoff *et al.* 1996; Reich *et al.* 2001; Frisse *et al.* 2001). These results together with the higher levels of sequence variation in African compared to non-African populations led to the proposal that population size reduction, such as bottlenecks, account for patterns of variation and LD in non-African populations (Wall and Przeworski 2000; Reich *et al.* 2001; Frisse *et al.* 2001). This bottleneck was hypothesized to correspond with the dispersal of modern humans out of Africa (Reich *et al.* 2001).

However, the investigation of formal bottleneck models has typically employed a single aspect of genetic variation data, either the allele frequency spectrum (Wall and Przeworski 2000; Marth *et al.* 2004), or patterns of LD (Reich *et al.* 2001; Marth *et al.* 2003). This raised the question of whether such models were indeed consistent with the data when multiple aspects of genetic variation were considered simultaneously (Ardlie *et al.* 2002a,b; Eswaran *et al.* 2005). Specifically, it is not known whether simple bottleneck models can generate the marked differences in LD levels seen between Africans and non-Africans with only a limited reduction in polymorphism levels outside Africa. While previous work suggested that variation in recombination rate may explain the decay in LD observed in a multiethnic sample (Reich *et al.* 2002), it is not obvious that it could also explain the differences between

Africans and non-Africans.

Ideally, making inferences about population history should be based on data from a large number of unlinked and neutrally evolving loci and on statistical methodology that makes efficient use of all or most of the information in the data. Full re-sequencing studies, in which the sequence of the surveyed segments is determined for every individual in the sample, represent one scheme for generating data sets in which multiple aspects of sequence variation are characterized. With regard to data analysis, full likelihood methods have been successfully applied to non-recombining data (Y-chromosome or mitochondrial DNA) to reconstruct population histories (Kuhner *et al.* 1998; Nielsen 1999; Beerli and Felsenstein 2001). However, for regions with recombination, the currently available methods are computationally infeasible. As a result, a variety of statistics, each summarizing different aspects of genetic variation data, may be used (Weiss and von Haeseler 1998; Wall and Przeworski 2000; Pluzhinikov *et al.* 2002), with the subsequent reduction in information content traded for computational tractability. It is still desirable to combine the results of tests based on individual statistics, as the joint distributions of multiple summaries of the data should contain more information than the marginal distributions of multiple single summaries considered separately.

We previously developed a full re-sequencing scheme in which pairs of tightly, but not completely, linked segments, referred to as "locus pairs", are surveyed (Frisse *et al.* 2001). This study design aims to maximize the information content for a given amount of sequencing effort because, by skipping the intervening segment, many more independent loci can be surveyed. Using this scheme, we previously surveyed 10 noncoding regions in three human population samples, Hausa of Cameroon, Italians and Chinese. Here, we survey an additional 40 locus pairs in the same samples. This data set allows the simultaneous characterization of polymorphism levels, allele frequency spectrum and LD in each sample; in addition, it obviates the need to correct for ascertainment bias, with its associated uncertainties and possible loss of information (Nielsen 2004; Kreitman and Di Rienzo 2004; Soldevila *et al.* 2005). In choosing only noncoding regions distant from genes, we limit the possibility that our analysis of demographic history will be confounded by the effects of natural selection.

To analyze these data, we implement an approach to determine p-values associated with several observed summaries of genetic data considered jointly over a grid of demographic parameter values. These summaries include the average Tajima's D ($\bar{D}$) and the variance of Tajima's D across loci ($\widehat{Var[D]}$) (Tajima 1989b), the average number of segregating sites across loci ($\bar{S}$), and an overall composite likelihood estimator of the population cross-over rate parameter ($\hat{\rho}$) (Hudson 2001). By combining p-values obtained from these individual statistics into a single statistical test, we greatly improve the power to reject demographic scenarios incompatible with the data. While it is well established that other demographic features apply to these populations (e.g. population subdivision and gene flow) (Rosenberg *et al.* 2002; Wakely and Lessard 2003), we chose to focus solely on population size changes to reduce modeling complexity. We explore an extensive grid of the demographic parameter space which revealed a confidence set of relatively simple bottleneck models which explain the patterns of variation in the non-African samples. Our results are the first to combine aspects of genetic variation from allele frequency spectrum, LD and polymorphism levels within noncoding autosomal regions to infer the history of human populations. Because our data set was collected without ascertainment, it may be useful for validating the results of SNP genotyping surveys.

3.2 Materials and methods

3.2.1 DNA samples and data collection

Sequence variation was surveyed in DNA samples from the same three human populations investigated in Frisse et al. (2001): 15 Hausa from Yaounde (Cameroon), 15 individuals from Central Italy, and 15 Han Chinese from Taiwan. In addition, one common chimpanzee DNA sample was also sequenced at each region. This study was approved by the Institutional Review Board of the University of Chicago.

We selected 40 unlinked genomic regions for re-sequencing using the "locus pair" approach (Frisse *et al.* 2001): for each unlinked region we sequenced two segments of approximately 1 kb separated by approximately 8 kb. The selection of genomic targets was aimed at regions that did not contain nor were tightly linked to known or strongly predicted coding regions. Most surveyed segments also did not contain and were not tightly linked to noncoding regions strongly conserved between human and mouse (as determined by inspection of the VISTA genome browser). These regions were selected as described in (Frisse *et al.* 2001) except that here we deliberately included regions with a broader range of cross-over rates and %G+C content. The local cM:Mb ratio was obtained based on the interval defined by the two closest flanking markers on the Decode genetic map (Kong *et al.* 2002). The average and variance of cM:Mb across the 50 segments (i.e., 40 new locus pairs and the 10 in (Frisse *et al.* 2001)) are: 1.31 and 0.83. The average and variance of %G+C across the 50 locus pairs are: 38.3 and 46.6. Detailed information on each surveyed segment is provided in the Supporting Text. PCR and sequencing was performed as previously described (Frisse *et al.* 2001; Wall *et al.* 2003). All sequencing reactions were run on automated capillary sequencers (ABI3100 and ABI3700). Sequence reads were scored using Polyphred (Nickerson *et al.* 1997); all putative polymorphisms and software-derived genotype calls were visually inspected and individually confirmed.

3.2.2 Testing demographic models

For each demographic model of interest, we performed a separate test for each summary statistic of genetic variation. In addition, for some of the models (equilibrium and bottleneck) we also calculated a test statistic, C, which combines the p-values of multiple summary statistics as follows:

$$C \;=\; -2\sum_{i=1}^{k} \ln(p_i) \qquad (3.1)$$

where p_i is the estimated p-value of the i^{th} summary statistic of k summary statistics.

For models defined by more than one demographic parameter (i.e. simple growth and bottleneck models), these tests were performed over a grid of parameter values. The combinations of parameter values that are compatible with the observed values of the test statistic(s) constitute the accepted portion of the parameter space for each model. For simple growth models, the test was based on Fu and Li's D^* (Fu and Li 1993), whereas for bottleneck models, the test was based on combining p-values from multiple summary statistics, as discussed below. The p-values, p_i, for each individual summary statistic were estimated from Monte Carlo simulations using a modification of the program ms (Hudson 2002), as follows. We used coalescent simulations to generate 50,000 replicates, each consisting of 50 independent locus pairs, for each combination of parameter values; mutation and recombination rates were allowed to vary across locus pairs as described below. Samples of sequences 10kb in length were generated in which the intervening 8kb were ignored to mimic the locus pair data. The probability, P, of observing a value greater than that found in the data was estimated by simulations and converted to a two-tailed p-value by applying the formula $1 - 2 \cdot |0.5 - P\,|$.

The p-values for the combined test statistic C were estimated using the empirical distribution of the statistic from simulations. For each combination of parameter values, we record the values of each summary statistic in each replicate and generate the distribution of these simulated values. For each replicate, we treat the value of each summary statistic as the "observed" value and determine its p-value relative to the empirical distribution from the remaining 49,999 replicates. For each replicate, we combine these p-values to calculate a value of C. By following this procedure for each of the 50,000 replicates (for a single demographic scenario of interest), we obtain a distribution of the combined statistic. This distribution can be used to estimate a one-tailed p-value for the observed value of C.

3.2.3 Mutation and recombination rate model

We assume an infinite sites model, where we model the variation in mutation rate across locus pairs using a Gamma(12.46, 2.11×10^{-9}) distribution. The mean and variance for this distribution matched the observed mean and variance for the mutation rates estimated based on human-chimpanzee sequence divergence in our locus pair data (assuming 6 Mya since divergence and a generation time of 25 years). The 90% central interval of this distribution is $(1.54 \times 10^{-8}, 3.96 \times 10^{-8})$ with $E\mu = 2.63 \times 10^{-8}$.

We model the variation in the crossing-over rate, c, across locus pairs using a Lognormal(-18.148, $(0.5802)^2$) distribution; cross-over rate was assumed to be homogeneous within each locus pair. The 90% central interval of this distribution is ($.51 \times 10^{-8}$, 3.41×10^{-8}). The median of this distribution matched the overall recombination rate for the Hausa data (1.31×10^{-8}) based on the composite likelihood estimator, $\hat{\rho}$, of Hudson Hudson (2001). Because we cannot accurately estimate the variance in recombination rate across surveyed segments as short as 10 kb, we matched the variance of the Lognormal distribution to the variance of cM:Mb values estimated from the Marshfield genetic map for the interval containing each locus pair (Yu *et al.* 2001). We acknowledge that this model may capture some, but not all, of the recombination rate variation estimated across the human genome (McVean *et al.* 2004).

3.2.4 Summary statistics

We summarize the locus pair data in terms of the average Tajima's D ($\bar{D}$), the variance of Tajima's D ($\widehat{Var[D]}$), the average Fu & Li's D^* ($\bar{D}^*$), the average number of segregating sites ($\bar{S}$), and the average nucleotide diversity across the 50 locus pairs ($\bar{\pi}$), as well as $\hat{\rho}$, an overall estimate of the population crossing over parameter (4Nc), as obtained by composite likelihood (Hudson 2001). Because there is not enough information in our data to estimate accurately $\hat{\rho}$ and the gene conversion parameters (Ptak *et al.* 2004), we assume a model of gene conversion with rate (f) twice that of cross-over and tract lengths exponentially distributed with mean (L) 500 bp and estimate $\hat{\rho}$. Alternative models of gene conversion (f=10, L=55 bp) based on sperm typing data Jeffreys and May (2004) yielded qualitatively similar results (data not shown).

3.3 Results

3.3.1 Summaries of variation and tests of the equilibrium model

We re-sequenced 40 unlinked locus pairs in 15 individuals from each of three population samples: Hausa, Italians, and Chinese. The results of this survey are analyzed together with data for an additional 10 unlinked locus pairs previously re-sequenced in the same population samples (Frisse *et al.* 2001), for a total of 50 unlinked locus pairs. The average surveyed length per locus pair was 2,365 bp (for a total of 118,259 bp surveyed in each individual), and the average unsurveyed intervening segment was 7,921 bp long.

The values of summary statistics used for demographic testing are shown in Table 3.1, with a synopsis of the summary statistics for the 40 new locus pairs presented in the Supporting Text (Table S1). The allele frequency spectrum was summarized by the average and variance of Tajima's D and Fu and Li's D* across loci, polymorphism levels are summarized by the average number of polymorphic sites ($\bar{S}$) across loci, and LD decay was summarized in terms of an overall composite likelihood estimator of the population cross-over rate parameter $\hat{\rho}$ (Hudson 2001). The results of this expanded data set are in qualitative agreement with those from our previous survey (Frisse *et al.* 2001) and with other similar data sets (Przeworski *et al.* 2000; Wall and Przeworski 2000; Akey *et al.* 2004; Stajich and Hahn 2005). With regard to the allele frequency spectrum, the Hausa show a skew towards rare variants and a low variance across loci while both non-African samples have an excess of intermediate frequency variants and high variance across loci. Also, polymorphism levels and LD decay are higher in the Hausa compared to both non-African samples, but this difference is greater for LD decay (1.9- to 3.2-fold) than polymorphism levels (1.6-fold).

In order to determine if the levels of LD decay and the frequency spectrum were consistent with a model of constant population size, we conducted coalescent simulations under equilibrium to determine the p-values of the observed summary statistics. We obtained the effective population size, denoted N_A, for each population using an estimator of the population mutation rate parameter ($4N_A\mu$) based on the number of polymorphic sites and sample size (Watterson 1975), and an estimate of μ based on sequence divergence between human and chimpanzee for the 50 locus pairs. Each summary statistic for the Hausa data is consistent with the equilibrium model (Table 3.1). However, for the non-African populations, the skew towards intermediate frequency variants and the elevated LD are incompatible with a simple equilibrium model; a combined statistic based on $\bar{D}$, $\widehat{Var[D]}$, and $\hat{\rho}$, obtained by using equation 3.1, is significant for both the Italian ($p \leq 0.0148$) and Chinese data ($p \leq 0.0052$).

Observed summary statistics

	$\bar{D}$	$\widehat{Var[D]}$	$\bar{D}$*	$\bar{S}$	$\bar{\pi}$ (%)	$\hat{\rho}$
Hausa	-0.20	0.55	-0.17	11.1	0.110	0.0006
Italian	0.28*	1.19**	0.18	7.1	0.085	0.0003
Chinese	0.18	1.08*	0.05	6.9	0.079	0.0002*

Table 3.1: Summaries are based on polymorphism data from 50 locus pairs. A designation of * or ** represents a p-value < 0.05 or 0.01, respectively, under an equilibrium model.

3.3.2 Estimating N_A under a growth model

Even though a model of constant population size could not be rejected for the Hausa, human populations certainly experienced rapid growth recently and, perhaps, also in more ancient times. Thus, the negative but nonsignificant values of Tajima's

D and Fu and Li's D* in the Hausa may simply reflect limited power and suggest that some expansion models are appropriate for this population. Following the approach in (Pluzhinikov *et al.* 2002), we considered a model in which an ancestral population at equilibrium size N_A grows exponentially beginning t_{onset} generations in the past at rate α, such that the present population size is $N_A e^{\alpha t_{onset}}$ Slatkin and Hudson (1991). To test this model, we fixed the ancestral population size for each combination of demographic parameter values, such that the expected number of segregating sites matched the average number observed in the Hausa sample Pluzhinikov *et al.* (2002).

Unlike in (Pluzhinikov *et al.* 2002), we estimated the best-fit growth parameters for the Hausa data, α and t_{onset}, along with the associated point estimate of N_A, via approximate maximum likelihood (ML) based on the summary statistic, Fu & Li's D^*. We focused on the average D^* across the 50 locus pairs, denoted $\bar{D}^*_{obs}$, as it was previously shown to be the most informative for discriminating between equilibrium and growth models (Pluzhinikov *et al.* 2002). For each demographic growth model, we obtained distributions of $\bar{D}^*$ by simulation, and estimated the probability that $|\bar{D}^* - \bar{D}^*_{obs}| < 0.001$, and then choose the model where this probability is highest. We call this the approximate maximum likelihood estimate (MLE) of the growth parameters, α and t_{onset}, compatible with the Hausa data based on $\bar{D}^*_{obs}$. Note that we refer to this as approximate ML on a summary statistic, because we do not use the full data, and because we approximate rather than obtain the probabilities exactly. We found that the model with the highest overall probability was at an α of 0.75×10^{-3} and t_{onset} of 1,000 generations; this corresponds to a model with ~2–fold growth starting 25,000 years ago, assuming a generation time of 25 years, from an ancestral population size of 10,659. We present confidence sets of α and t_{onset} for which $\bar{D}^*_{obs}$ are consistent with the observed Hausa data in Figure 3.1. The span of acceptable models is consistent with previous reports (Pluzhinikov *et al.* 2002), with a slight reduction in confidence set due to the inclusion of additional data.

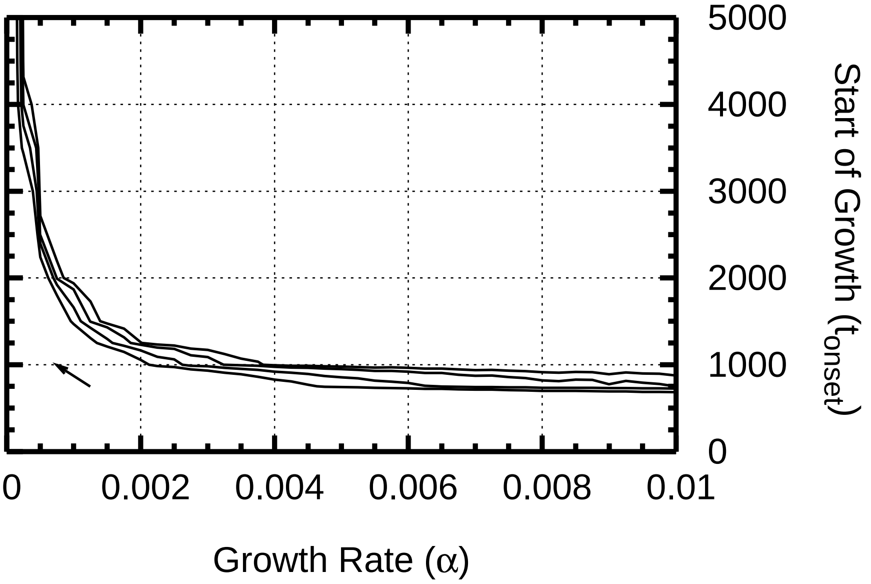


Figure 3.1: Confidence sets for pairs of parameters (t_{onset}, α) based on the average Fu & Li's D^* across 50 locus pairs for the Hausa sample. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost. The parameter set which maximizes the likelihood of $\bar{D}^*_{obs}$ is indicated by the arrow.

In order to asses the uncertainty in N_A, we obtain a range of N_A consistent with the MLE of $\hat{\alpha} = 0.75 \times 10^{-3}$ and $\hat{t}_{onset} = 1{,}000$ as follows. We performed additional coalescent simulations as described earlier, where we used the ML parameters for the demographic history and gradually lowered or raised the value of N_A until $\bar{S}$ was incompatible with the observed data at the 5% level. We found these high and low values of N_A to be 9,450 and 12,300, respectively. Later, we will utilize this information to assess the effect of our choice of N_A in testing bottleneck models.

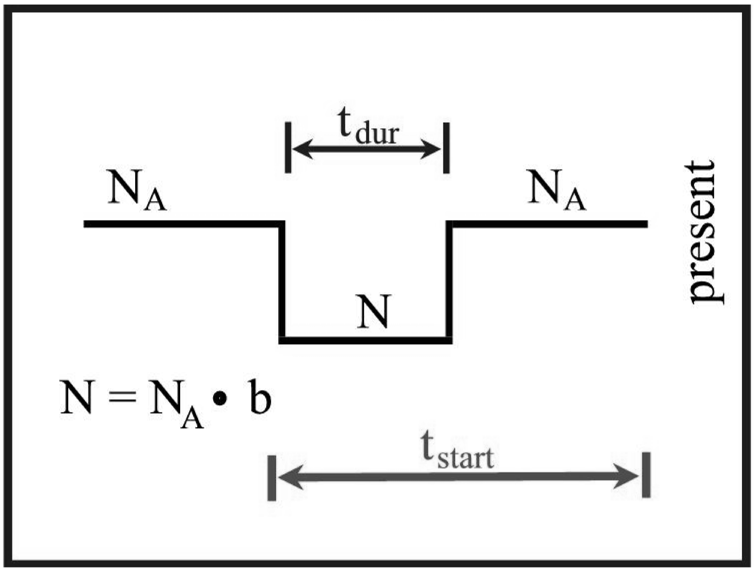


Figure 3.2: Diagram of the bottleneck model. An ancestral population at equilibrium size N_A undergoes a reduction at time t_{start} generations in the past and remains at a constant size $b{\cdot}N_A$ for t_{dur} generations; at time $t_{start} - t_{dur}$ generations ago the population size recovers to size N_A. The set of models include severities ranging from $b = 0.005, 0.05, 0.1, 0.15, 0.2, ..., 0.5$, bottleneck durations ranging from $t_{dur} = 0, 100, 200, ..., t_{start}$, and total demographic epochs of $t_{start} = 20, 40, 80$ and 120 kya.

3.3.3 Testing bottleneck models in the non-African data

The positive $\bar{D}$ values and large $\widehat{Var[D]}$ along with the low polymorphism and high LD levels observed in the non-African populations (Table 3.1) suggest that models including a reduction in population size may be compatible with the data. We considered one family of bottleneck models for these data, where a population of constant size N_A instantaneously shrinks in size to $b \cdot N_A$ at time t_{start} generations before the present. The population remains at that size for t_{dur} generations and then instantaneously recovers to its original size, as illustrated in Figure 3.2.

Under the assumption that non-African populations originated from an ancestral population in Sub-Saharan Africa, we set the ancestral population size in the bottleneck simulations to the values of N_A obtained by ML based on the Hausa data and the simple growth model (N_A = 10,659). This assumption has important implications for our subsequent inferences about compatible bottleneck scenarios. We then used coalescent simulations to estimate the p-values for each summary statistics, for each point on a grid of bottleneck severities (b), bottleneck duration (t_{dur}), and time since the beginning of the bottleneck (t_{start}). This allows defining the portion of the multidimensional parameter space that is compatible with the data.

By combining p-values of different summaries as described by equation [3.1], we can make use of multiple aspects of the data to narrow the confidence region of compatible parameter values. The value of such an approach is depicted in Figure 3.3. We found that, for all possible combinations of two or more summary statistics, the combination of $\bar{D}$-$\bar{S}$-$\hat{\rho}$ was the most powerful to discriminate between bottlenecks and a constant size model, over the parameter range depicted in Figure 3.3. Therefore, we use the combination of $\bar{D}$-$\bar{S}$-$\hat{\rho}$ in our subsequent analyses of bottleneck models.

The confidence sets for the Italian and Chinese data for a t_{start} value of 40,000 years and N_A = 10,659 are shown in Figure 3.4b,e; in all cases, the accepted portion of the parameter space tends to lie on the diagonal of the plots indicating that bottleneck severity and duration have inversely related effects on patterns of variation. The Italian data are compatible with a range of bottleneck models that include shorter and more severe bottlenecks (e.g., $b = 0.1$, $t_{dur} = 400$ generations) at one end and longer and milder bottlenecks (e.g., $b = 0.4$, $t_{dur} = 1600$ generations) at the other. If t_{start} = 80,000 years ago, this range is slightly shifted to the right, including both longer and less severe bottlenecks (Supporting Text, Figure S4). For the Chinese data, if t_{start} is 40,000 years, the compatible parameter space is similar to that of the Italian data except that it includes slightly more severe bottleneck scenarios (Figure 3.4b,e). The most severe and longest bottleneck occurs where $b = 0.005$ and $t_{dur} = 300 - 600$ generations, but fewer combinations of parameter values corresponding to mild bottlenecks are accepted. If t_{start} = 80,000 years, milder bottlenecks can not be

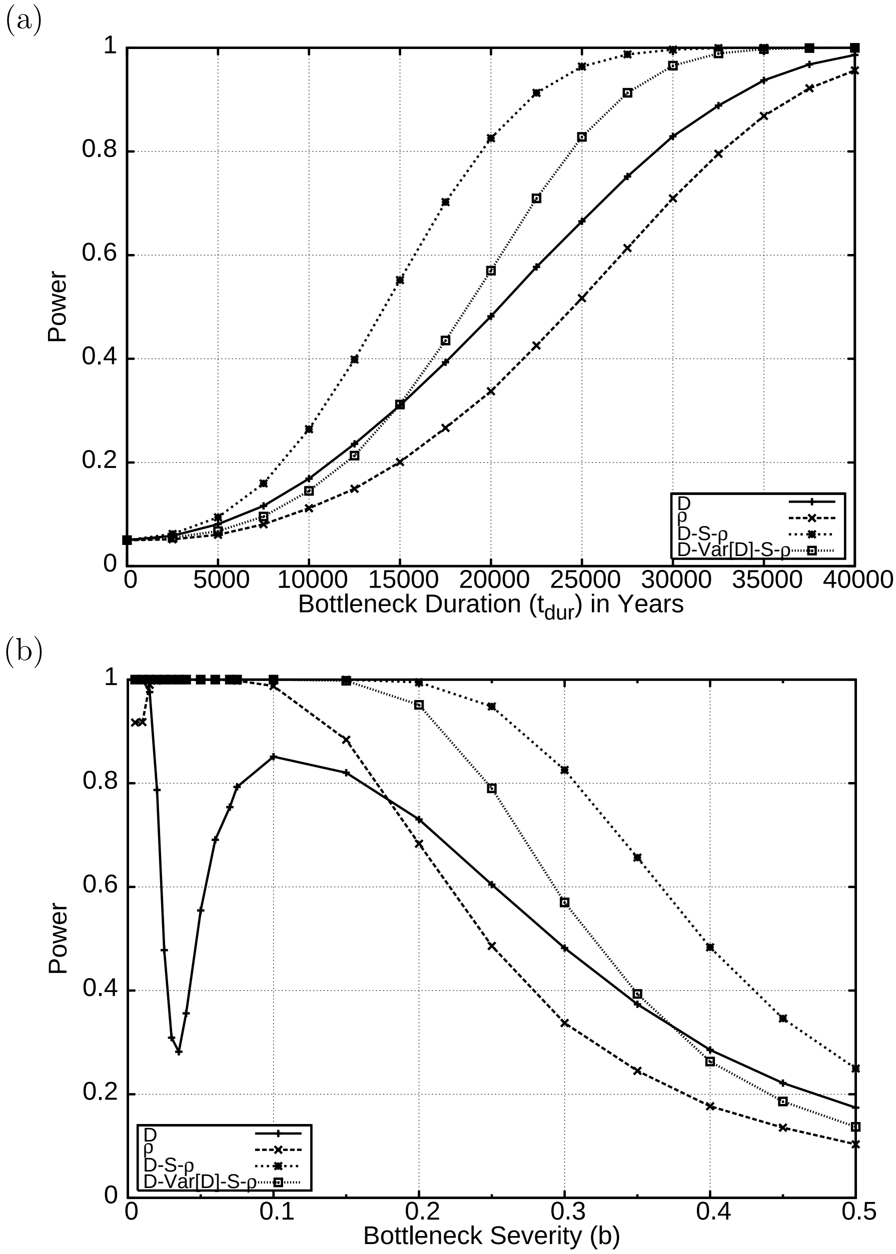


Figure 3.3: Power of combining multiple summary statistics. The power to reject a constant size demographic model for combined summaries with an N_A of 10,659 under bottleneck models with (a) a 70% reduction in N_A and a total time of 40,000 years for various t_{dur} and (b) bottleneck duration of 20,000 years for a total demographic epoch of 40,000 years for various bottleneck severities. The type-I error rate was held at 5%.

rejected, and even a long-lasting and mild bottleneck with $b = 0.4$ can not be rejected (Figure 3.5). For N_A = 10,659 and for any value of t_{start}, no bottleneck shorter than 100 generations is accepted in either population.

We also considered values for t_{start} of 20,000 and 120,000 years (Figures 3.6 and 3.7). In all cases, the lower t_{start} values showed a confidence set that was shifted toward scenarios of longer and more severe bottlenecks. Conversely, at higher t_{start} values, more severe bottlenecks were rejected in favor of milder bottleneck scenarios.

To assess the effect of the uncertainty associated with the estimates of N_A, we repeated the above analyses by using different values of N_A, obtained from estimating the uncertainty around N_A from the Hausa ML growth models, described above. As shown in Figures 3.4-3.7, the effect of N_A on the accepted parameter space is substantial. As expected, for the larger value of N_A, the accepted portion of the parameter space is reduced such that only relatively severe and long bottlenecks are compatible with the data, whereas a larger range of less severe bottlenecks are compatible with the smaller value of N_A.

3.4 Discussion

By re-sequencing unlinked non-coding regions, we assessed patterns of polymorphism levels, frequency spectrum and LD for the same set of genomic segments and population samples. To achieve greater resolution of different demographic scenarios, we use an analytical approach that combines information from individual summary statistics of sequence variation; computer simulations showed that combinations of summaries allow for more powerful tests of each demographic scenario. Rather than focusing on a single best-fitting demographic model, we construct an acceptance region of the parameter space that is compatible with the demographic model of interest–in this case population growth or bottleneck–thus providing an inclusive picture of the uncertainty in inferences of human demography. A major new conclusion of this analysis is that the non-African population samples are compatible with simple bottleneck models even when multiple aspects of sequence variation are considered simultaneously. Consistent with our previous analysis (Pluzhinikov *et al.* 2002), the Hausa sample from Sub-Saharan Africa is compatible with both the equilibrium model and with relatively recent population growth.

Modeling human population history is central to a variety of questions in human biology, but most recently the search for signatures of natural selection has given new importance to this line of inquiry (Akey *et al.* 2004; Stajich and Hahn 2005; Jensen *et al.* 2005; Zhu and Bustamante 2005). The impact of natural selection on the human genome can be detected by contrasting patterns of neutral variation, i.e. those shaped solely by demography, to those observed at test loci that may be shaped by natural selection in addition to demography. Traditionally, this contrast utilized the theoretical predictions of the standard neutral model, in which the population was assumed to be constant in size and randomly mating. However, studies of human variation have shown genome-wide departures from this model suggesting that human demography is complex (Wall and Przeworski 2000; Reich *et al.* 2002; Marth *et al.* 2003; Adams and Hudson 2004). Thus, the development of a more realistic null model of evolutionary neutrality is necessary for improving inferences about natural selection (Akey *et al.* 2004; Stajich and Hahn 2005).

Several conditions must be satisfied to achieve these goals. One is the availability of sequence variation data for many unlinked and neutrally evolving regions. While several whole genome variation data sets are available, these consist mainly of genotyping data for ascertained polymorphisms (The International HapMap Consortium 2003; Hinds *et al.* 2005). Re-sequencing data are also available, but they tend to focus on gene regions that may have been targets of selection and, hence, are less suitable for demographic inference (Akey *et al.* 2004; Williamson *et al.* 2005; Stajich and Hahn 2005). An additional challenge derives from the complexity of human demography and the fact that realistic models are defined by multiple unknown demographic parameters. This implies that, for any given value of one parameter (e.g. bottleneck severity), there may be a range of values for the other parameters (e.g. time of onset and duration of bottleneck) that are equally consistent with the data. It is particularly important in this context to make efficient use of the information in the data. While it may be useful to generate point estimates of the demographic parameters, it is even more important to obtain the multi-dimensional confidence set if specific hypotheses about human evolution are to be tested.

The present study represents an important step toward improving our inferences about human demography. Though the present data set is not as large as other resequencing surveys (Akey *et al.* 2004; Stajich and Hahn 2005), it was specifically designed for demographic inference and will provide a useful reference for analyses of gene regions. This is because, in an attempt to select neutrally evolving regions, we focused on segments that neither contain nor are tightly linked to coding regions. Also, most of these segments neither contain nor are tightly linked to noncoding sequences conserved between human and mouse. Our scheme for data collection aimed at maximizing the information content of the data so that multiple aspects of genetic variation could be analyzed for the same set of independent loci. Owing to the use of ethnically identified samples, we could provide evidence for different demographic histories in different populations.

Our analytical approach also improves on previous studies of human demography. First, it provides a full characterization of the uncertainty around the best-fitting model by identifying the portion of the multi-dimensional parameter space that is consistent with genetic variation data in each population. The inclusion of multiple aspects of genetic variation by combining the p-values for different summary statistics provides greater power than any single summary alone, allowing us to reduce substantially the accepted space for each model. Our study is based on an extensive exploration of the demographic parameter space including onset, duration, and severity of the bottleneck. It is important to note that the reduction in bottleneck parameter space was greatly aided by our inference about N_A based on the Hausa data. Because the N_A is restricted, the range of compatible values for summary statistics that depend on N_A (i.e. $\hat{\rho}$ and $\bar{S}$) is also constrained.

An important limitation of our analysis is that we considered only models of randomly mating populations. Although this is a common assumption in modeling studies of population size change, it is unlikely to be satisfied by human populations, even if geographically defined (Rosenberg *et al.* 2002; Harding and McVean 2004). In fact, it is possible that population structure alone could account for the observed patterns of human variation (Wall and Przeworski 2000; Wakely and Lessard 2003; Akey *et al.* 2004; Stajich and Hahn 2005). Interestingly, the addition of $\widehat{Var[D]}$ into the bottleneck analysis results in a further reduction of the accepted parameter space (Figures 3.8-3.11), even though combining this statistic with $\bar{D}$, $\bar{S}$, and $\hat{\rho}$ reduces the power to reject the constant size model (Figure 3.3). This suggests that additional features, such as population structure, are required to produce $\widehat{Var[D]}$ values that are more consistent with our data. Though it is desirable and certainly more realistic to include elements of population structure in models of human demography (Wakeley *et al.* 2001), there is insufficient data to indicate the most plausible family of such models. For these reasons, testing simple growth and bottleneck models is a reasonable first step toward developing more complex and realistic models. Obviously, if changes in population size and population structure were considered jointly rather than separately, the accepted range of values for the growth and bottleneck parameters is likely to be different.

A main new conclusion of this study is that simple bottleneck models can explain the non-African data even when multiple aspects of genetic variation are considered simultaneously. Several previous studies of human sequence variation had modeled specific bottleneck scenarios on the basis of either frequency spectrum information (Wall and Przeworski 2000; Marth *et al.* 2004; Akey *et al.* 2004; Adams and Hudson 2004; Stajich and Hahn 2005), LD decay (Reich *et al.* 2001), or polymorphism levels (Marth *et al.* 2003). Wall and Przeworski (Wall and Przeworski 2000) analyzed full re-sequencing data and proposed that a bottleneck and selective sweeps at some loci could explain the frequency spectrum observed in non-Africans, but did not provide information regarding the likely parameter values. The frequency spectrum was used also by Marth et al. (Marth *et al.* 2004) to estimate a best-fit bottleneck model for Europeans and East Asians. We used our simulation scheme to estimate the probability of the Italian and Chinese data for the corresponding best-fit models of Marth et al. (Marth *et al.* 2004). In our parameterization, the best fit model for the Asian sample in Marth et al. corresponds to an N_A of 10,000, b of 0.3, t_{dur} of 400 generations, a t_{start} of 90,000 years; note that this model includes growth after the bottleneck to a size of 25,000. The best fit model for the European sample in Marth et al. corresponds to an N_A of 10,000, b of 0.2, t_{dur} of 500 generations, a t_{start} of 87,500 years, with growth after the bottleneck to a size of 20,000. Using our simulation scheme, our data turned out to be incompatible with these models ($p < 0.0001$). It should be noted, however, that Marth et al. (Marth *et al.* 2004) analyzed a data set of ascertained SNPs and attempted to correct for the resulting bias. Hence, the discrepancy between the two studies may be due to incomplete ascertainment correction and highlights the value of re-sequencing data.

Based on the frequency spectrum observed in a large re-sequencing study of genes involved in inflammation, Akey et al. (Akey *et al.* 2004) concluded that the European data was consistent with a bottleneck starting 40,000 years and a bottleneck intensity, as measured by the inbreeding coefficient (F), of 0.175. This best-fit model can be translated to a range of models in our notation by using

$$F = 1 - \left(1 - \frac{1}{2 \cdot b \cdot N_A}\right)^{t_{dur}} \tag{3.2}$$

This bottleneck model corresponds to a number of points that are well within the accepted portion of the parameter space for our non-African data (for example, $b =$ 0.2 and $t_{dur} = 820$ generations assuming our best-fit N_A of 10,659). Because only the best-fit model is reported by Akey et al. (Akey *et al.* 2004), the overall agreement between these two data sets cannot be assessed.

Similar conclusions were obtained through an analysis of pairwise LD data of ascertained SNPs in a European population sample (Reich *et al.* 2001); however, a narrow portion of the parameter space was investigated. We determined that there are points in our accepted parameter space that correspond to the estimated time of onset and F reported by Reich et al. (Reich *et al.* 2001), indicating agreement between the two methods and data sets. Finally, a recent analysis of re-sequencing data from a pool of ethnically diverse samples detected evidence for very recent population growth (Williamson *et al.* 2005). While this model is compatible with our Hausa data, it does not provide a good explanation for the Italian and Chinese data, hence, pointing to the need for population-specific demographic inferences.

3.5 Appendix: Bottleneck Figures

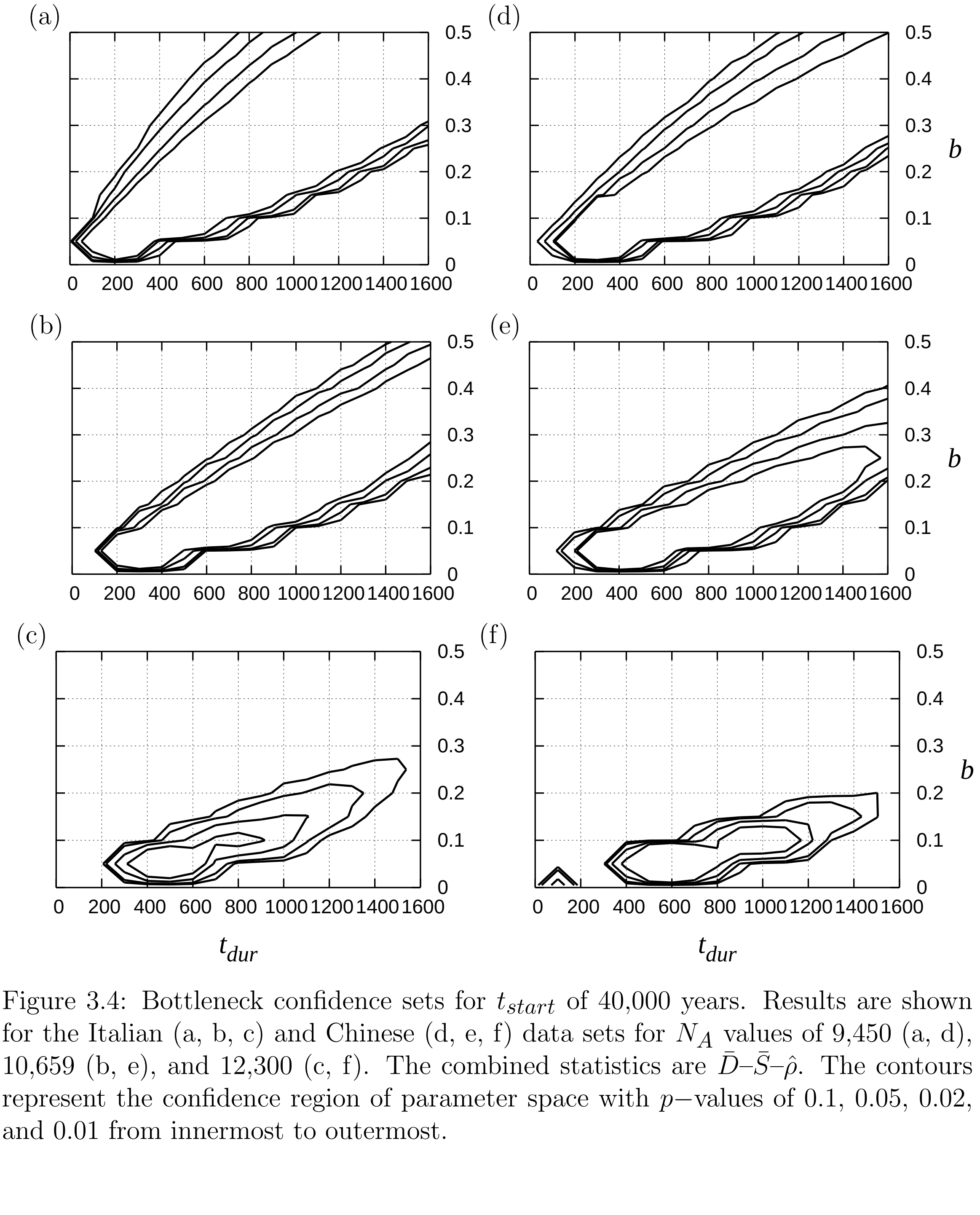


Figure 3.4: Bottleneck confidence sets for t_{start} of 40,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

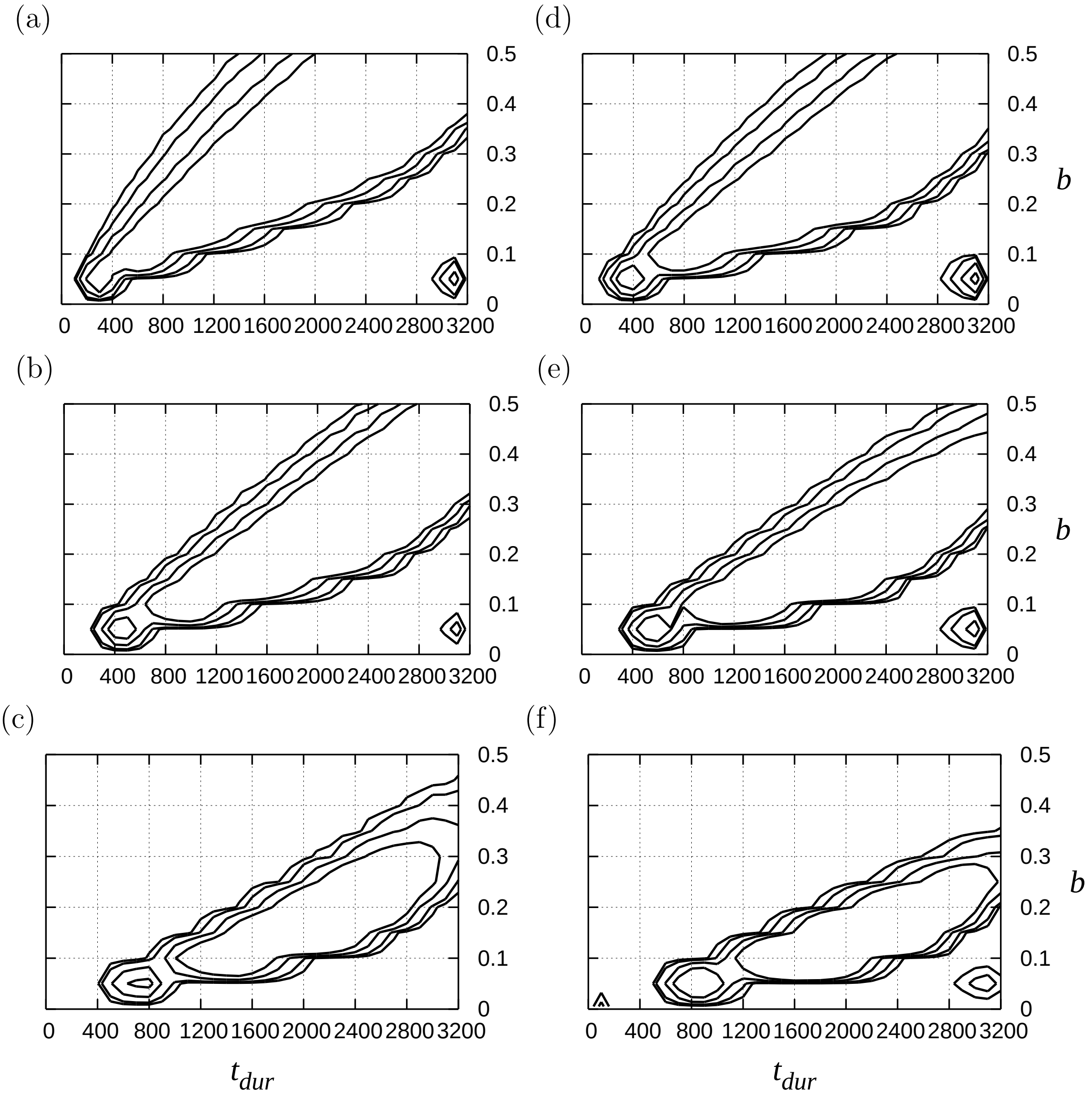


Figure 3.5: Bottleneck confidence sets for t_{start} of 80,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

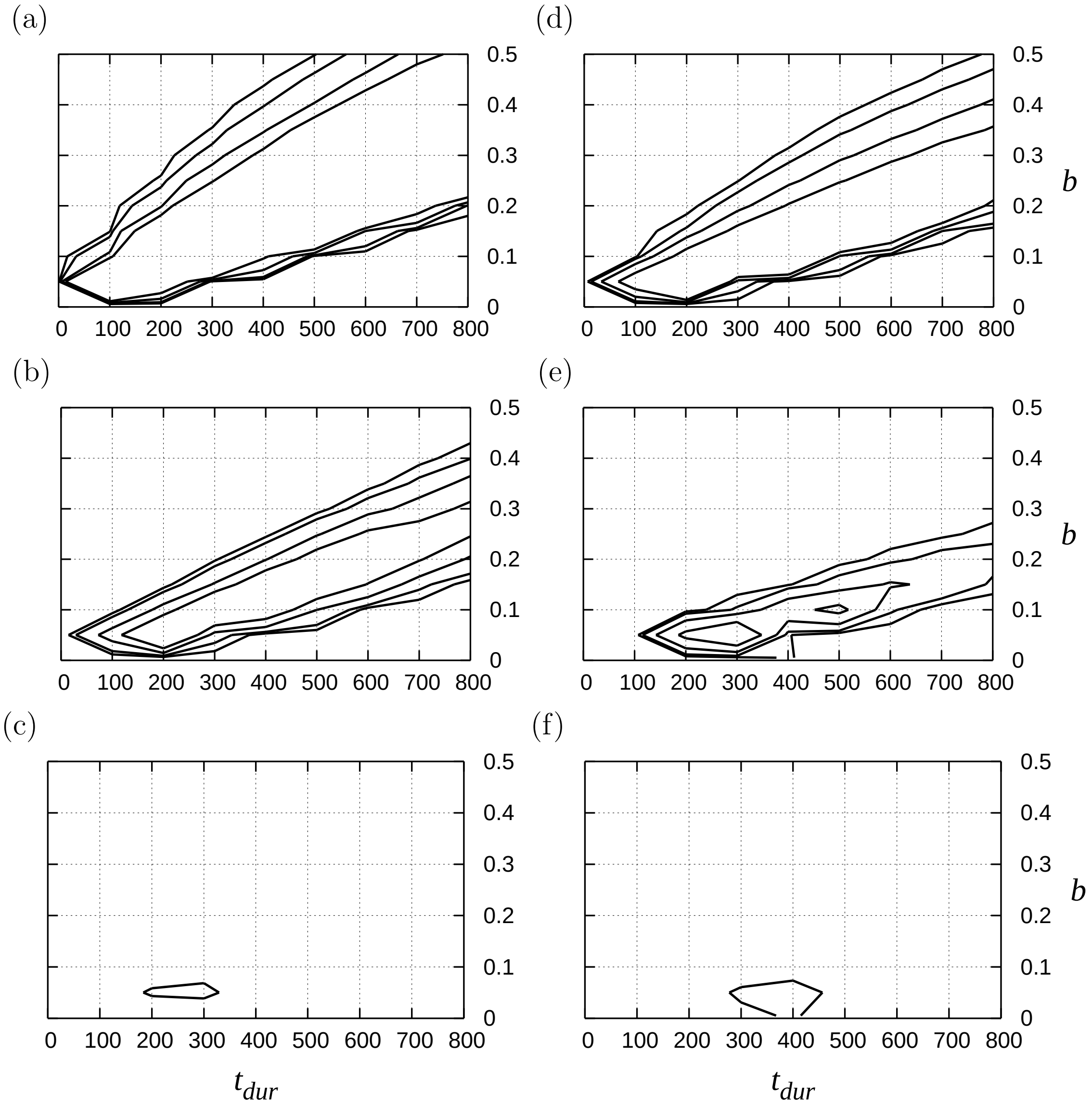


Figure 3.6: Bottleneck confidence sets for t_{start} of 20,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

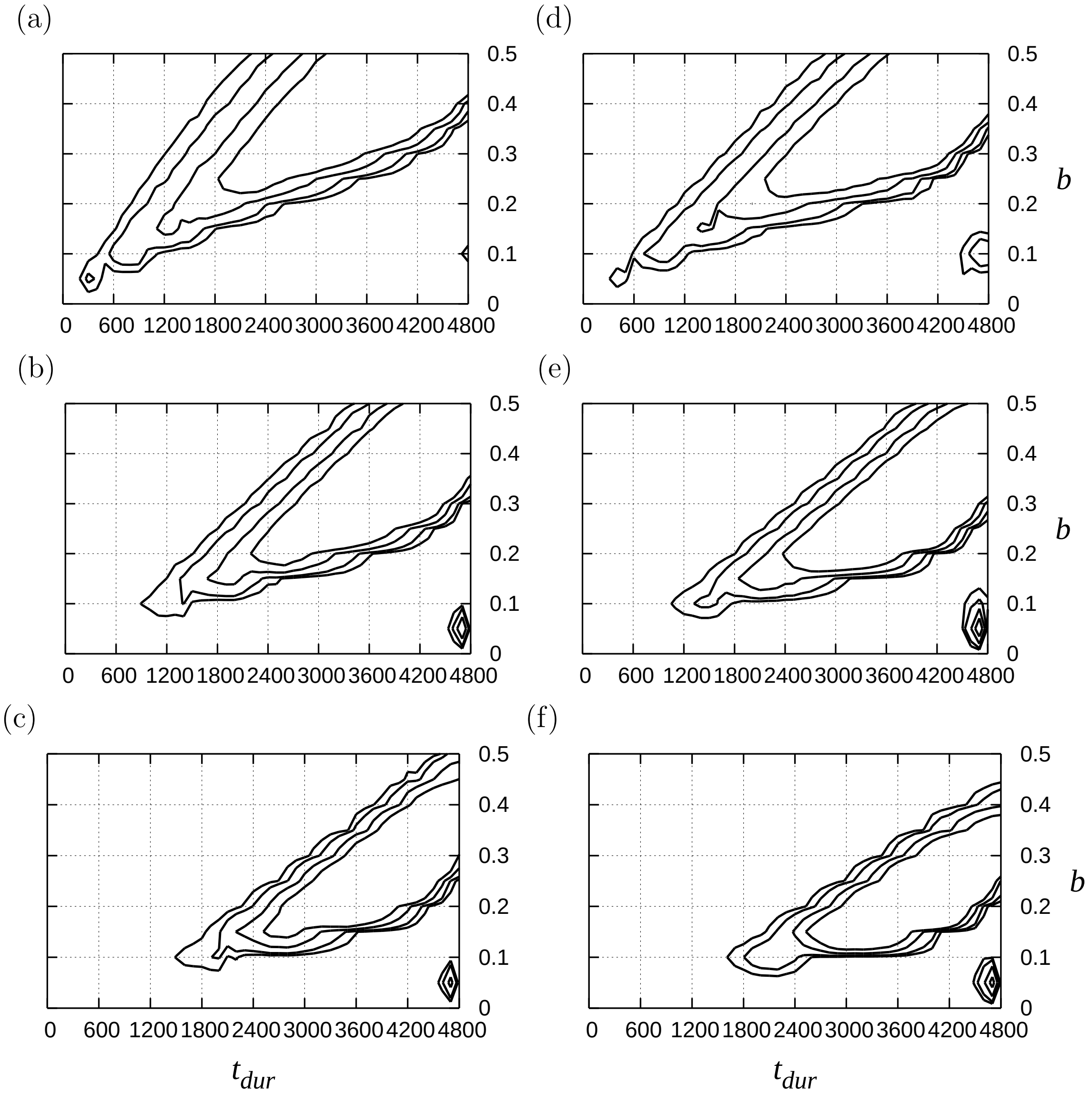


Figure 3.7: Bottleneck confidence sets for t_{start} of 120,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

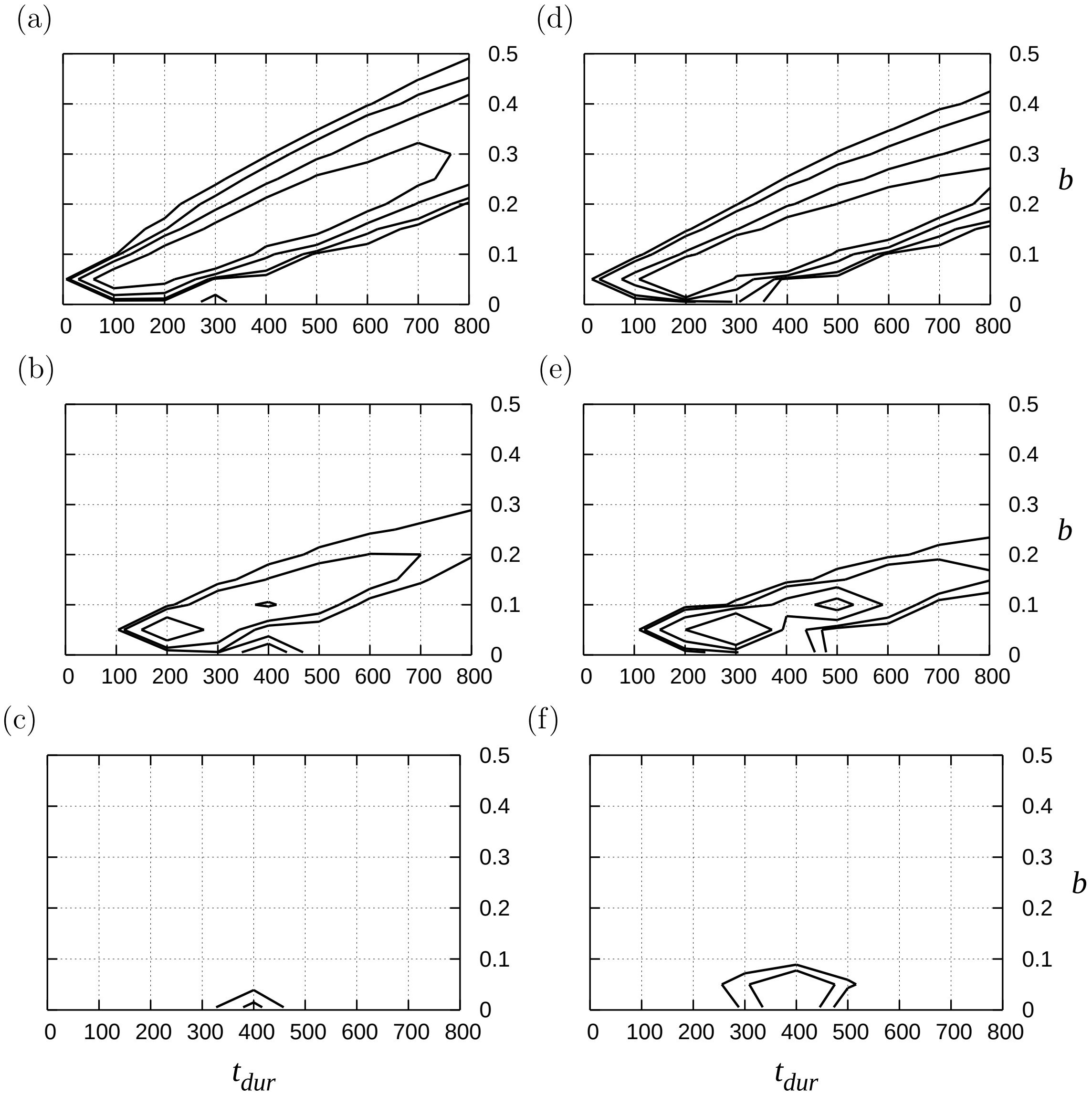


Figure 3.8: Bottleneck confidence sets (including $\widehat{Var[D]}$) for t_{start} of 20,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\widehat{Var[D]}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

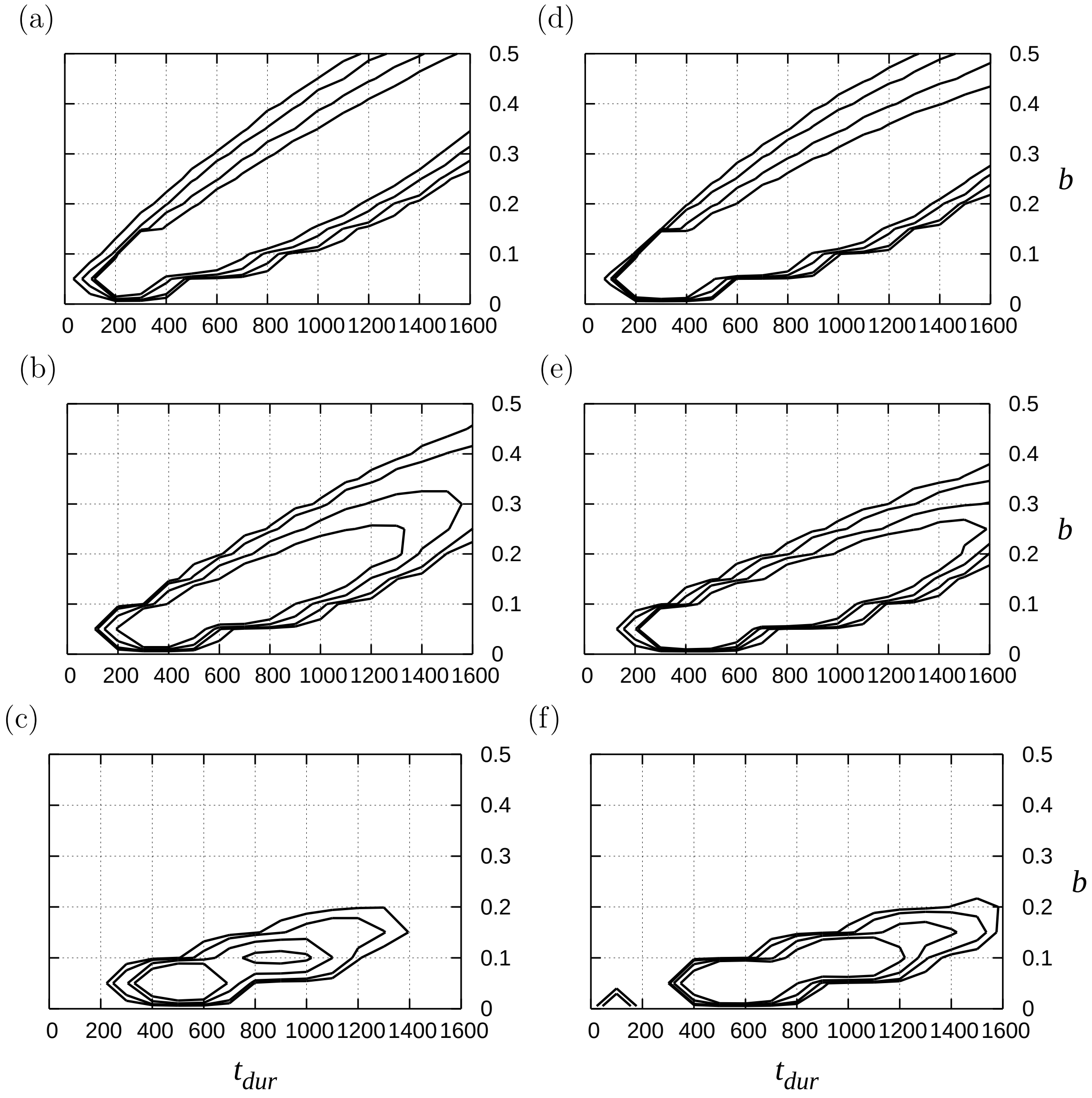


Figure 3.9: Bottleneck confidence sets (including $\widehat{Var[D]}$) for t_{start} of 40,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\widehat{Var[D]}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

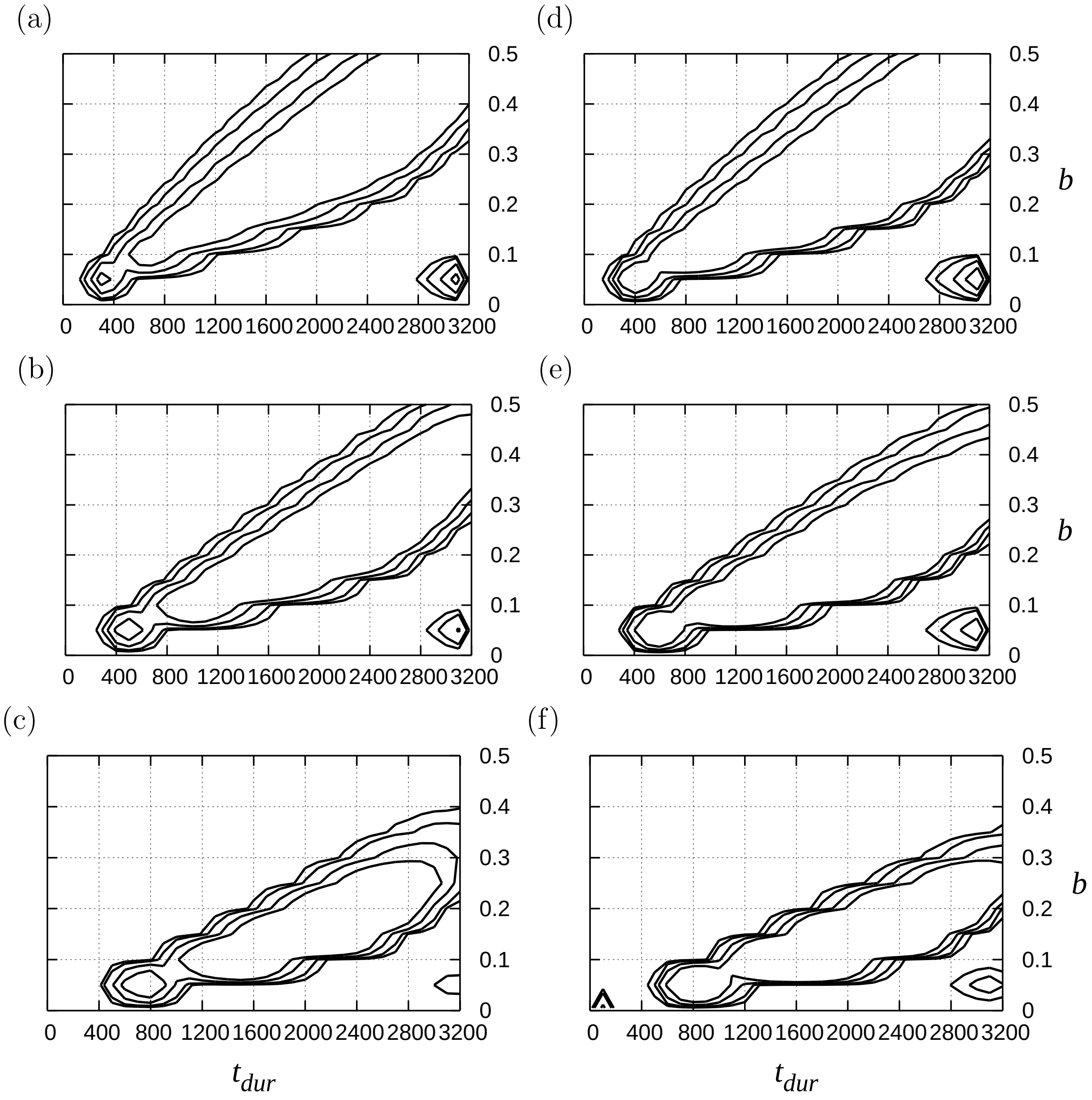


Figure 3.10: Bottleneck confidence sets (including $\widehat{Var[D]}$) for t_{start} of 80,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\widehat{Var[D]}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

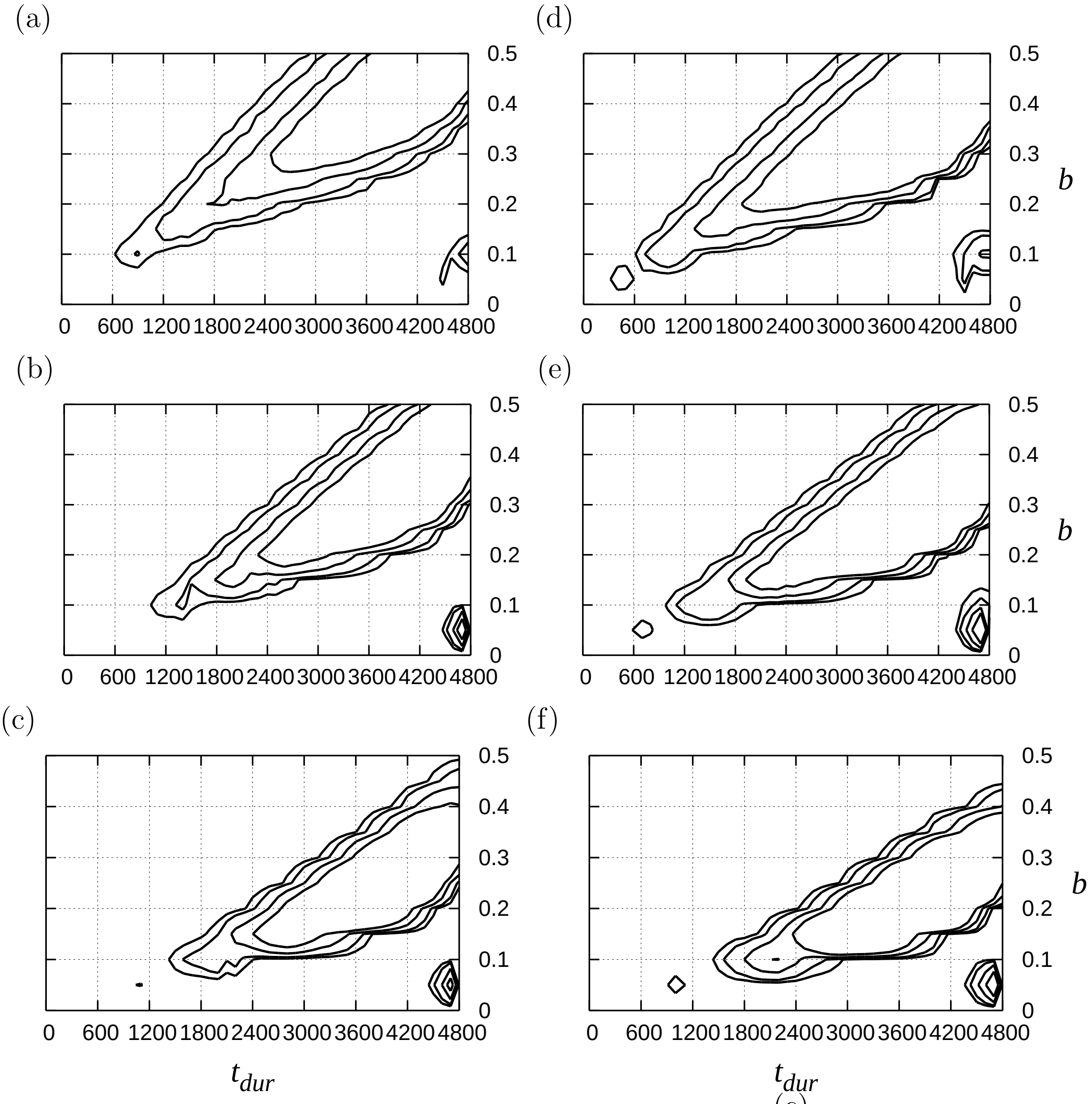


Figure 3.11: Bottleneck confidence sets (including $\widehat{Var[D]}$) for t_{start} of 120,000 years. Results are shown for the Italian (a, b, c) and Chinese (d, e, f) data sets for N_A values of 9,450 (a, d), 10,659 (b, e), and 12,300 (c, f). The combined statistics are $\bar{D}$–$\widehat{Var[D]}$–$\bar{S}$–$\hat{\rho}$. The contours represent the confidence region of parameter space with $p-$values of 0.1, 0.05, 0.02, and 0.01 from innermost to outermost.

CHAPTER 4

INFERENCE USING JOINT SNP FREQUENCIES

4.1 Introduction

Historic changes in population size influence many characteristics of extant polymorphism, including levels of variation, the frequency spectrum, and linkage disequilibrium. As a consequence, many analyses of polymorphism data, whether to infer evolutionary forces such as selection (Haddrill *et al.* 2005; Akey *et al.* 2004; Sabeti *et al.* 2002; Hudson *et al.* 1994) or to use linkage disequilibrium in association mapping (Reich *et al.* 2002), are sensitive to assumptions about demography. Therefore, there has been much interest in extracting information regarding demographic history from genetic data, especially with the availability of publicly accessible population-specific resequencing data.

When making inferences regarding demographic history, it is advantageous to include as much of the information contained in genetic data as possible. Full-likelihood methods are available for data sets not subject to recombination (Beerli and Felsenstein 2001; Kuhner *et al.* 1998; Nielsen 1999), however such methods are not feasible when polymorphic sites are not completely linked. Therefore, in the interest of computational feasibility, previous methods have focused on a single summary of genetic data such as Tajima's D (Wall and Przeworski 2000; Pluzhinikov *et al.* 2002) or patterns of LD (Reich *et al.* 2001; Marth *et al.* 2003). An alternative strategy is to utilize multiple summaries of genetic data in either a combined p-values (Voight *et al.* 2005) or approximate Bayesian (Thornton and Andolfatto 2006) framework. Other recent methods have sought to incorporate more of the available information into demographic analyses while maintaining computational efficiency by using maximum likelihood on the frequency spectrum of unlinked polymorphic sites (Wooding and Rogers 2002; Polanski and Kimmel 2003; Marth *et al.* 2004) and adjusting confidence regions when linkage between sites exists (Adams and Hudson 2004).

Here we present a method that represents a middle ground between full-likelihood on nonrecombining regions and maximum likelihood on polymorphic sites that are assumed to be unlinked. Our maximum likelihood method directly incorporates linkage between polymorphic sites into the estimation of demographic parameters. We summarize the joint frequency spectrum at a given locus with S segregating sites by $\mathbf{j} = \{j_1, j_2, \ldots, j_{n-1}\}$, where j_i is the number of polymorphic sites where the derived allele is at frequency i in a sample of n chromosomes and $\sum_{i=1}^{n-1} j_i = S$. Note that this summary of the data is identical to that described in Adams and Hudson (2004) for a single locus. However, here we have a separate $\mathbf{j}$ for each locus since we directly incorporate linkage between sites within a locus into our analysis. By using coalescent simulations with recombination (Hudson 2002), we can estimate the probability of an observed joint frequency spectrum for each locus on a grid of demographic parameter values and choose the combination of parameter values that maximizes the product of this likelihood for a set of unlinked loci.

We also present an extension of this method where the data may be summarized in terms of the frequency and position of S linked segregating sites within a locus. In this case, the data is summarized by $\mathbf{p} = \{p_1, p_2, \ldots, p_S\}$ and $\mathbf{q} = \{q_1, q_2, \ldots, q_S\}$, where p_i is the position, in base pairs, of the i^{th} SNP, and q_i is the frequency of the i^{th} SNP in the sample of n chromosomes.

Because this method utilizes coalescent simulations with recombination to determine the likelihood of the data under different demographic scenarios, including the proper levels of recombination is critical to our analysis. Recombination rates are typically estimated assuming a standard constant population size model. However, violations of this standard model can result in a biased estimates of the population crossing-over rate, ρ (=$4Nr$, where N is the ancestral population size in a changing population size scenario). Specifically, such deviations lead to overestimates of ρ with population growth and underestimates of ρ with population bottlenecks (Smith and Fearnhead 2005). Therefore, in order to ensure that each demographic scenario is accepted or rejected solely on the basis of demographic history and not recombination, we adapt the approach of Hudson (2001) to obtain demography-specific composite likelihood estimates of ρ for each data set as described in Smith and Fearnhead (2005).

We apply our method to three data sets: African Hausa and Italian data sets previously published in Voight *et al.* (2005) and the European sample from the Seattle SNPs data set (http://pga.gs.washington.edu). We find that the Hausa support a history of very slow, very ancient growth as well as a range of more recent, rapid growth scenarios. The confidence regions of the two European data sets indicate a mild to moderate population bottleneck occurring ∼40,000-80,000 years ago.

In order to assess the utility of directly incorporating linkage into our analyses, we compare our results from the Hausa and Seattle SNPs European data sets to the previous method of Adams and Hudson (2004), which summarizes the data by the frequency spectrum. We refer to the Adams and Hudson (2004) method as a composite likelihood method because it treats all sites as independent, regardless of linkage. Additionally, we compare our results for the Italian data set to the results of Voight *et al.* (2005), whose analysis utilized a combined statistic based on p-values of Tajima's D (Tajima 1989b), the number of segregating sites per locus, and an estimate of the population crossing-over parameter, ρ (Hudson 2001).

4.2 Model and methods

4.2.1 Demographic model

For the Hausa data, we consider a growth model as described previously in Adams and Hudson (2004). Under this model, a population is at a constant population size, N_0, until time T before the present. At time T, the population undergoes exponential growth to a present size of N_{rec}. We define two demographic parameters of interest: $f_{rec} = N_{rec}/N_0$ and T, the time at which the exponential growth began, in units of $4N_0$ generations before the present. Unless otherwise noted, we will assume an ancestral population size of 10,000 and a generation time of 25 years when converting our demographic time parameters to approximate year equivalents.

In analyzing the European data sets, we use a bottleneck model as described in Voight *et al.* (2005). This is a simple bottleneck model where a population of constant size, N_0, instantaneously collapses to size $b \cdot N_0$ for a period of time and then instantaneously recovers back to size N_0. This model is defined by the time of onset of the bottleneck, t_{start}, the duration of the bottleneck, t_{dur}, and the severity of the bottleneck, b.

We assume an infinite-sites mutation model, with a mutation rate of u per generation. We also assume that the recombination rate per generation, r, is constant. Estimates of θ ($=4N_0u$) and ρ ($=4N_0r$) per locus are determined as described later in this section. For all analyses, the rate of gene conversion is assumed to be twice that of crossing-over, and conversion tract lengths are exponentially distributed with a mean of 500bp.

4.2.2 Maximum likelihood method

We first consider the case where data consists of a population survey of n chromosomes at a single locus of length L (base pairs) that contains S segregating sites, where we assume all segregating sites are biallelic. We do not consider ascertainment bias, so we assume that the entire length of the locus is resequenced in each sampled chromosome. The probability of the joint frequency spectrum, $P(\mathbf{j})$, is a function of the demographic parameters (f_{rec} and T or t_{start}, t_{dur}, and b) as well as θ ($=4N_0u$). We will use $\mathbf{d}$ to denote the demographic parameters of interest, where $\mathbf{d}$ represents a combination of either f_{rec} and T or t_{start}, t_{dur}, and b. To estimate $P(\mathbf{j})$, we generate a large number of gene genealogies with recombination, and, for each genealogy, G, we estimate $P(\mathbf{j}|G)$, the conditional probability of $\mathbf{j}$ given G. Our method for estimating $P(\mathbf{j}|G)$ is described in the following paragraphs.

When the mutation rate per site, u, is small enough to disregard the possibility of more than one mutation occurring in the sample at a single site, we can express $P(\mathbf{j})$ in terms of θ ($=4N_0u$) and relative branch lengths of a gene genealogy. Each genealogy is obtained by using coalescent simulations with recombination, as described previously (Hudson 2002). We designate G to represent a single genealogical history, including recombination events, simulated using specified values of θ, ρ, and $\mathbf{d}$. Linkage between sites results in a correlation between gene genealogies of each segment of a locus, where segments result from recombination events. If recombination events break up a locus into M segments, and l_m is the length of the m^{th} segment in base pairs, then we calculate

$$\tau(G) = \sum_{m=1}^{M} (l_m/L) \cdot \psi_m \tag{4.1}$$

where ψ_m is the total branch length of the gene tree, in units of $4N_0$ generations, for segment m. The expected number of segregating sites in the sample is $\theta\tau(G)$.

A branch of this gene tree is defined to be an i-branch if a mutation that occurs on that branch results in i copies of the mutation in the sample, and the sum of the length of i-branches in the mth segment is denoted $\psi_{i,m}$. The total length of i-branches in a single simulated replicate is then denoted by $\tau_i(G)$ and obtained by summing the product of $\psi_{i,m}$ and the segment length (l_m/L) over all m segments, as follows:

$$\tau_i(G) = \sum_{m=1}^{M} (l_m/L) \cdot \psi_{i,m} \qquad . \tag{4.2}$$

The expected number of segregating sites at frequency i in the sample is then $\theta\tau_i(G)$. Then, conditional on the sample genealogy, j_i is poisson distributed with mean $\theta\tau_i(G)$, and

$$p(j_i|G) = \frac{\theta\tau_i(G)^{j_i} e^{-\theta\tau_i(G)}}{j_i!} \qquad . \tag{4.3}$$

The conditional probability of the joint frequencies, $\mathbf{j}$, is then

$$P(\mathbf{j}|G) = \prod_{i=1}^{n-1} p(j_i|G) \tag{4.4}$$

for a single replicate, and the unconditional probability of $\mathbf{j}$, $P(\mathbf{j};\, \mathbf{d})$, is then

$$P(\mathbf{j};\mathbf{d}) = \mathrm{E}[P(\mathbf{j}|G)] \tag{4.5}$$

which we estimate by

$$\frac{\sum_{y=1}^{Y} P(\mathbf{j}|G_y)}{Y} \quad , \tag{4.6}$$

where $P(\mathbf{j}|G_y)$ denotes $P(\mathbf{j}|G)$ for the yth replicate genealogy generated under $\mathbf{d}$, and Y is the total number of replicates. We find that Y must be at least 2-4 million to ensure a sufficiently smooth likelihood surface.

To obtain an approximate maximum likelihood estimate (MLE) of $\mathbf{d}$ (either f_{rec} and T or t_{start}, t_{dur}, and b), we evaluate $P(\mathbf{j};\mathbf{d})$ over a rectangular grid of parameter values and choose the combination of parameter values that maximizes the likelihood of our observed data. A global estimate of $P(\mathbf{j};\mathbf{d})$ for an entire data set may be obtained by taking the product of $P(\mathbf{j};\mathbf{d})$ obtained for individual loci.

If the ancestral or derived status of a polymorphism is not known, $\mathbf{j}$ can be folded at frequency $n/2$, and a branch on a simulated gene genealogy is then an i-branch if a mutation on that branch leads to either i or $n - i$ copies of the mutation in the sample. The product in equation [4.4] is then over i equals 1 to $n/2$.

4.2.3 Incorporating position and frequency information

In the analyses reported in this manuscript, we summarize the data in terms of $\mathbf{j}$ and use the method described above. However, one may wish to incorporate more of the available information in a data set and summarize the data in terms of $\mathbf{p}$= $\{p_1, p_2, \ldots, p_S\}$ and $\mathbf{q}$= $\{q_1, q_2, \ldots, q_S\}$, where p_i and q_i are the position and frequency of the i^{th} of S SNPs, respectively.

We use the same coalescent simulation with recombination scheme described above. The total branch length of the sample gene genealogy for a given combination of demographic parameters, $\tau(G)$, is given by equation 4.1. Again, recombination events break a locus into M segments, each of length l_m, and we now consider a_m and z_m to be the beginning and ending position of the m^{th} segment, respectively.

We define $\lambda_i(G)$ to be the total length of branches in a single gene genealogy, G, where a mutation on any of those branches results in q_i descendants at position p_i. Then, $\lambda_i(G)$ is equal to $\psi_{q_i,m}$ (the sum of q_i-branches in the m^{th} segment) when $a_m \leq p_i \leq z_m$. Therefore, the probability that a segregating site lies at position p_i and has frequency q_i, conditional on the gene genealogy, G, is proportional to

$$p(p_i, q_i|G) = \theta\lambda_i(G) \qquad . \tag{4.7}$$

Then, the probability density of a given position and frequency configuration, $P(\mathbf{p},\mathbf{q}|G)$, is the product, over all S segregating sites, of the probability that a segregating site lies at each position and frequency multiplied by the probability that there are no other segregating sites, as follows:

$$P(\mathbf{p},\mathbf{q}|G) = e^{-\theta\tau(G)} \prod_{i=1}^{S} p(p_i, q_i|G) \qquad . \tag{4.8}$$

The unconditional probability, $P(\mathbf{p},\mathbf{q};\mathbf{d})$, is then obtained by taking the average of $P(\mathbf{p},\mathbf{q}|G)$ over a large number of replicate genealogies. As with the joint frequencies method, a global $P(\mathbf{p},\mathbf{q};\mathbf{d})$ may be obtained by taking the product of equation 4.8 over individual loci. An approximate MLE may then be obtained by evaluating $P(\mathbf{p},\mathbf{q};\mathbf{d})$ over a grid of $\mathbf{d}$ and choosing the $\mathbf{d}$ that maximizes $P(\mathbf{p},\mathbf{q};\mathbf{d})$.

Due to computational requirements of both analysis of complex data sets and also constructing confidence intervals, we choose to focus on the joint frequencies method as opposed to this position/frequency extension in this manuscript. We did apply this frequency and position method to the Italian data set and found that the MLE is only a single grid point away from the MLE obtained using our joint frequencies method. However, we are not able to quantify the size of the confidence region associated with this MLE as compared to the joint frequencies method.

4.2.4 Estimation of ρ under each demographic scenario

We assumed that the recombination rate is constant across loci in a data set and used an extension of the composite likelihood estimator of HUDSON (2001) to estimate demography-specific values of $\hat{\rho}$ for each data set. The `maxdip` program described in HUDSON (2001) utilizes sample configuration probability tables that are generated by the `ehnrho` program, which uses coalescent simulations under a constant-size population model. We adapted the `ehnrho` program to accommodate the growth and bottleneck models described above, allowing sample configuration probability tables to be generated under each combination of parameter values. A separate value of $\hat{\rho}$ was then estimated from each data set for each point in the demographic parameter space by using `maxdip` with the appropriate sample configuration probability table. The coalescent simulations required to obtain the likelihood of the joint frequency spectrum at a given demographic scenario were then run with the value of $\hat{\rho}$ specific to that combination of parameter values. Both `maxdip` and `ehnrho` can be found at `http://home.uchicago.edu/∼rhudson1` .

4.2.5 Estimation of θ and grouping of loci

In order to reduce the dimensionality of parameter space to explore, we replaced θ for a given locus in [4.3] with a simple moment estimator obtained by

$$\theta = \frac{S/L}{\bar{\tau}(\mathbf{d})} \quad , \tag{4.9}$$

where S is the number of segregating sites in a locus of length L and $\bar{\tau}(\mathbf{d})$ is the average length of a gene genealogy for a sample of n chromosomes under the specified demographic parameters, $\mathbf{d}$, measured in units of $4N_0$ generations. We can obtain an estimate of $\bar{\tau}(\mathbf{d})$ by using [4.1] and determining the average over many ($\geq$ 100,000) replicates.

In the interest of computational efficiency, we grouped loci of similar recombination rate (length) and used one set of genealogies to estimate $P(\mathbf{j}; \mathbf{d})$ for all loci in a group, reducing the total number of genealogies required to analyze a data set. Because the loci in both the Hausa and Italian data sets are nearly uniform in length, we analyzed all 50 loci in a single group with a single recombination rate for each population. Since the Seattle SNPs loci are much more diverse in length, we assigned each locus to one of 16 bins, which range from 5 to 85kb. We assigned loci longer than 85kb to the 85kb bin and used only the first 85kb in our analysis.

4.2.6 Constructing confidence intervals

We can not assume that asymptotic approximation of confidence intervals is appropriate for the Hausa and Italian data sets. In order to place confidence intervals on our Hausa and Italian MLEs, we use coalescent simulations with recombination (Hudson 2002) to simulate data that mimic the locus pair data structure. For each population, we generate 1,000 loci for a sample of 30 chromosomes. Each locus is ~10kb in length, and we ignore the middle ~8kb to mimic the locus pair structure (see Locus Pair Data Sets section of Results for further description). The demographic parameters for these simulations are the MLEs for the Hausa and Italian data sets.

We choose a value of θ for each population such that the expected number of segregating sites in the ∼2kb at either end of the segment matches the average observed number of segregating sites in each population.

We then generate 5,000 bootstrap replicate data sets by randomly sampling sets of 50 loci with replacement from the 1,000 simulated loci. For each bootstrap replicate, we then apply our joint frequencies method, recording the ratio of the log-likelihood at the MLE to the log-likelihood at the parameters from which the data were simulated. Using this log-likelihood ratio distribution, we can estimate the 95% critical value and construct confidence regions by including those parameter values where the log likelihood ratio is less than the estimated 95% critical value.

4.3 Results

4.3.1 Locus pair data sets

We first consider two population samples from a data set described in Adams and Hudson (2004) and Voight *et al.* (2005) obtained from an Italian and an African Hausa population sample of 30 chromosomes each. The data set includes 50 unlinked "locus pairs". Each locus pair consists of ∼1kb resequenced at either end of a ∼10kb segment, as originally described in Frisse *et al.* (2001). In order to apply our maximum likelihood method to both the Hausa and Italian data, we adapt our simulations to ignore those segments which lie within the ∼8kb unsequenced region when calculating the total branch length, τ, and length of i-branches, τ_i, in equations [4.1], [4.2], and [4.9].

Hausa analysis

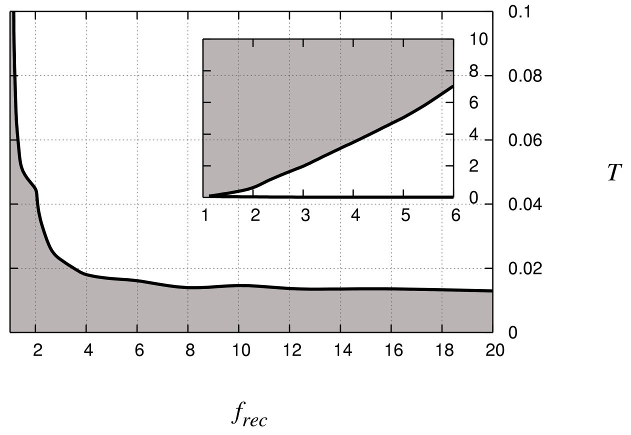
We evaluate the Hausa data over a two-dimensional grid of f_{rec} and T parameter values. Values of f_{rec} range from 1 (corresponding to constant population size) to 25 at intervals of 1, and values of T range from 0 to 10, at intervals of 0.00625 from 0 – 0.1, 0.1 from 0.1 – 1, and 1 from 1 – 10. Application of the joint frequencies maximum likelihood method to this data set results in a maximum likelihood estimate (MLE) of $\hat{f}_{rec} = 3$ and $\hat{T} = 6$. These parameters correspond to very slow, ancient three-fold growth beginning approximately 6 million years ago.

We construct a 95% confidence region using the simulation scheme described above. We estimate the 95% critical value to be 2.2 log likelihood units as compared to the asymptotic approximation of 3.0 for two-dimensional MLEs, indicating that asymptotic approximation would have been conservative for this particular data set. The 95% confidence region constructed using the critical value of 2.2 is depicted in Figure 4.1a. In order to assess the performance of our joint frequencies method, we compare the 95% confidence regions obtained using this method to that of a prior method. Figure 4.1b illustrates the previously published analysis of the Hausa data set using the composite likelihood method described in Adams and Hudson (2004).

Italian analysis

We consider a three-dimensional grid of t_{start}, t_{dur}, and b parameter values for the Italian analysis to match the parameters in Voight *et al.* (2005). We consider t_{start} values of 800, 1600, and 3200 generations, each with 9 t_{dur} grid points evenly spaced from 0 to t_{start}, inclusive. Values of b range from 0.05 to 0.5 at intervals of 0.05. We find that the MLE for this data set is $tstart = 1600$ generations, $t_{dur} = 1400$ generations, and $b = 0.2$.

(a)



(b)

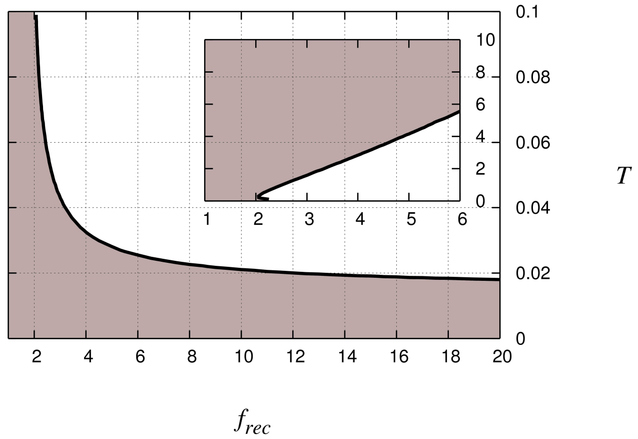


Figure 4.1: Comparing Hausa confidence regions. The shaded area in each figure represents the 95% confidence region determined from simulation (a) Confidence region using our joint frequencies method. MLE is $\hat{f}_{rec}$=3, $\hat{T} = 6$. (b) Confidence region using the method of Adams and Hudson (2004). MLE is $\hat{f}_{rec}$=3.1, $\hat{T} = 6.1$. The inset figures show a zoomed-out view of the main figures.

We note that the previous analysis of this data set by Voight *et al.* (2005) benefits from assuming an ancestral population size (N_A) and mutation rate (μ) estimated from external data. Therefore, to allow for more direct comparison between our joint frequencies method and the combining $p-$values method of Voight *et al.* (2005), we also apply our joint frequencies to the Italian data using a fixed value of θ ($=4N\mu$) based on an N_A of 10,659 and a mutation rate of 2.63×10^8, as estimated in Voight *et al.* (2005). By incorporating this additional information, our MLE is shifted slightly to $t_{dur} = 1200$ generations and $b = 0.2$.

In order to construct confidence intervals around the MLE, we estimate the 95% critical value of the likelihood ratio statistic to be 3.3 for the Italian data set when a value of θ is estimated for each locus. We also find the 95% critical value to be 2.6 when the fixed value of θ is used. We depict the confidence sets using the simulated 95% critical values for the Italian data set in Figures 4.2a,b. In comparison, Figure 4.2c depicts the 95% confidence region for the Italian data based on the analysis of (Voight *et al.* 2005).

4.3.2 Seattle SNPs

We also apply our maximum likelihood method to the European population sample of the Seattle SNPs data, which can be found at the University of Washington-Fred Hutchinson Cancer Research Center (UW-FHCRC) Variation Discovery Resource (http://pga.gs.washington.edu). At the time we accessed this resource, the data consisted of 215 loci containing a total of 13,130 SNPs. In our analysis, we included only those SNPs that were successfully resequenced in the entire panel of 46 chromosomes. Additionally, we exclude 8 loci that have been identified as genes subject to selection in this data set by Akey *et al.* (2004). We also consider only those SNPs that do not result in an amino acid coding change in order to minimize selection as a confounding

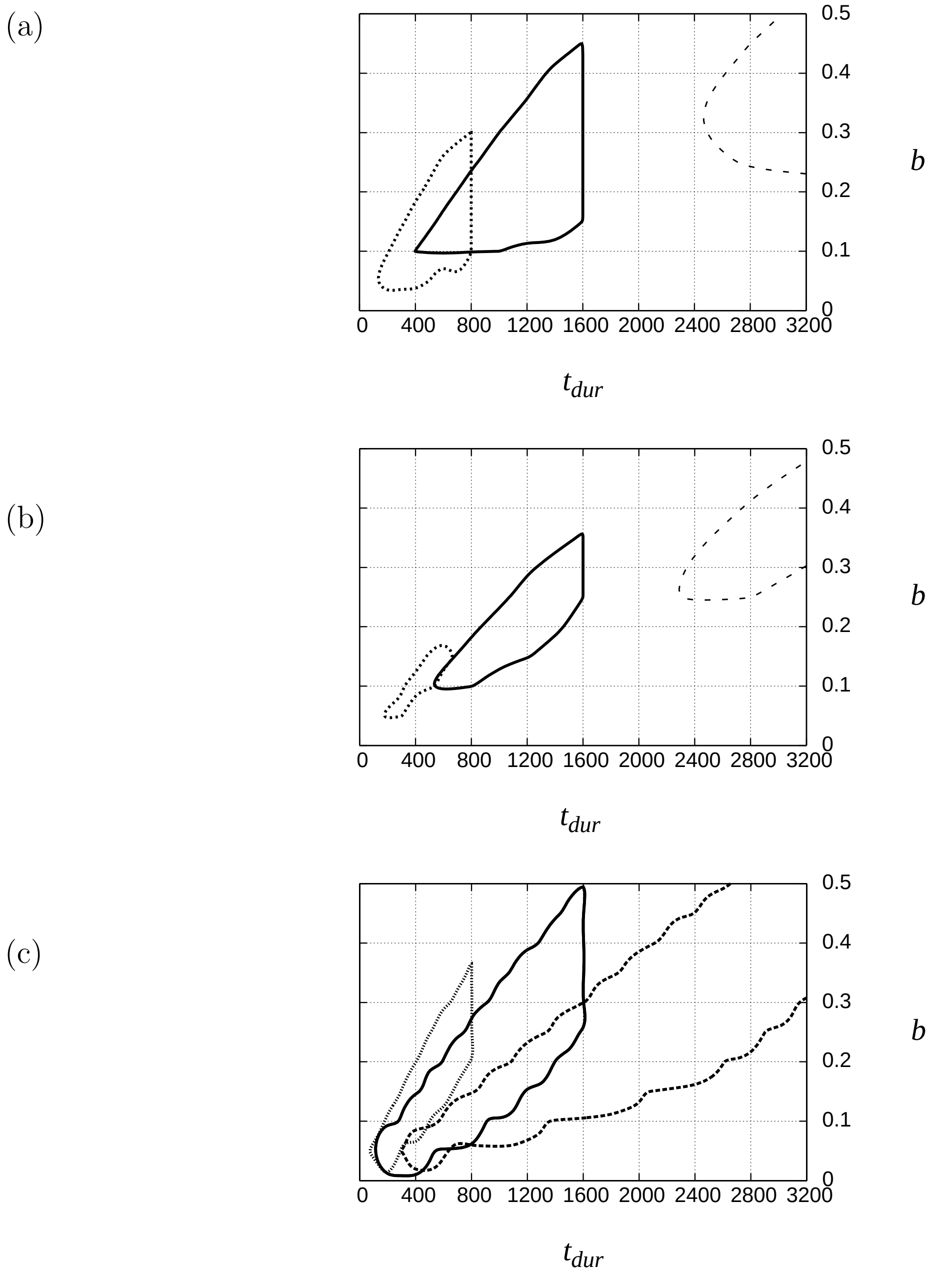


Figure 4.2: Comparing Italian confidence regions. 95% confidence regions for t_{start} values of 800, 1600, and 3200 generations are represented by dotted, solid, and dashed lines, respectively. (a) Confidence regions using our joint frequencies method. MLE is $\hat{t}_{start} = 1600$, $\hat{t}_{dur} = 1400$, $\hat{b} = 0.2$. (b) Confidence regions using our joint frequencies method with a fixed value of $\theta = 0.00112$/bp. MLE is $\hat{t}_{start} = 1600$, $\hat{t}_{dur} = 1200$, $\hat{b} = 0.2$. (c) Confidence regions using the method of VOIGHT *et al.* (2005) with an ancestral population size of 10,659.

factor in our analysis. Finally, in the interest of computational feasibility, we set the maximum locus length to be 85kb. If a locus exceeds 85kb in length, we include only the first 85kb in our analysis. The final data set that was used in the following analyses contained 6,539 SNPs across 207 loci.

Seattle SNPs results

In order to determine how much additional information regarding demographic history can be gleaned by directly incorporating linkage into the maximum likelihood analysis, we compare our method based on joint SNP frequencies to the composite likelihood method of Adams and Hudson (2004), which treats all SNPs as unlinked and then accounts for linkage *post hoc*. In order to make a direct comparison, we first re-analyze the Seattle SNPs data using the composite likelihood method of Adams and Hudson (2004), since more data has been made available since the publishing of that manuscript. Using the composite likelihood method on the expanded data set, we find that the MLE for the Seattle SNPs European data set is $\hat{t}_{start} = 3200$ generations, $\hat{t}_{dur} = 2800$ generations and $\hat{b} = 0.45$. The 95% confidence region based on the expanded data set is illustrated in Figure 4.3b, where the confidence region is determined from simulation as described in Adams and Hudson (2004).

Because of the heterogeneity of locus lengths and number of loci in the Seattle SNPs data set, our joint frequencies method is computationally expensive for this data set; therefore, we are unable to explore the full three-dimensional parameter space. Instead, we fix t_{start} at 3200 generations, which was the MLE value obtained applying the method of Adams and Hudson (2004) to this data set. We also note that the other t_{start} values considered using the composite likelihood method (800 and 1600 generations) are not contained in the 95% confidence region constructed for the Seattle SNPs European data set as described above.

With t_{start} fixed at 3200 generations, we then use our joint frequencies method to evaluate the likelihood of the Seattle SNPs data set over a two-dimensional grid of t_{dur} and b values including $t_{dur} = 0 - 3200$ at 200 generation intervals and $b = 0.05 - 0.75$ at 0.05 intervals. We find that the two-dimensional MLE using our joint SNP frequencies method is $\hat{t}_{dur} = 3200$ and $\hat{b} = 0.65$, which corresponds to a 35% reduction in effective population size that began approximately 80,000 years ago and persists to the present day.

Because analysis of even a single data set of this size and complexity is computationally demanding, even in two dimensions, we are unable to use simulations to construct confidence regions around our MLE as we did for the Hausa data set. However, Figure 4.3a provides an illustration of the likelihood surface resulting from our joint frequencies method, where each contour represents an interval of 5 log likelihood units from the maximum.

Bootstrap analysis

In order to further assess the uncertainty in our MLE for the Seattle SNPs European data set, we perform a bootstrap analysis on the 207 locus pairs. We generate 10,000 bootstrap replicates. To form a single bootstrap replicate, we randomly draw a set of 207 loci, sampling with replacement from the 207 loci in the Seattle SNPs data set. We then apply our joint frequencies method to each random "data set" and record the MLE, thus producing a set of 10,000 MLEs distributed on our grid of parameter values. Because we had recorded the $P(\mathbf{j};\mathbf{d})$ for each individual locus at each point in our demographic parameter space, this analysis requires no additional coalescent simulations. We sort the grid points in our demographic parameter space by the number of replicates that have each point as its MLE, from most to least. We find that 95% of the replicate MLEs fall on points that have at least 180 replicates

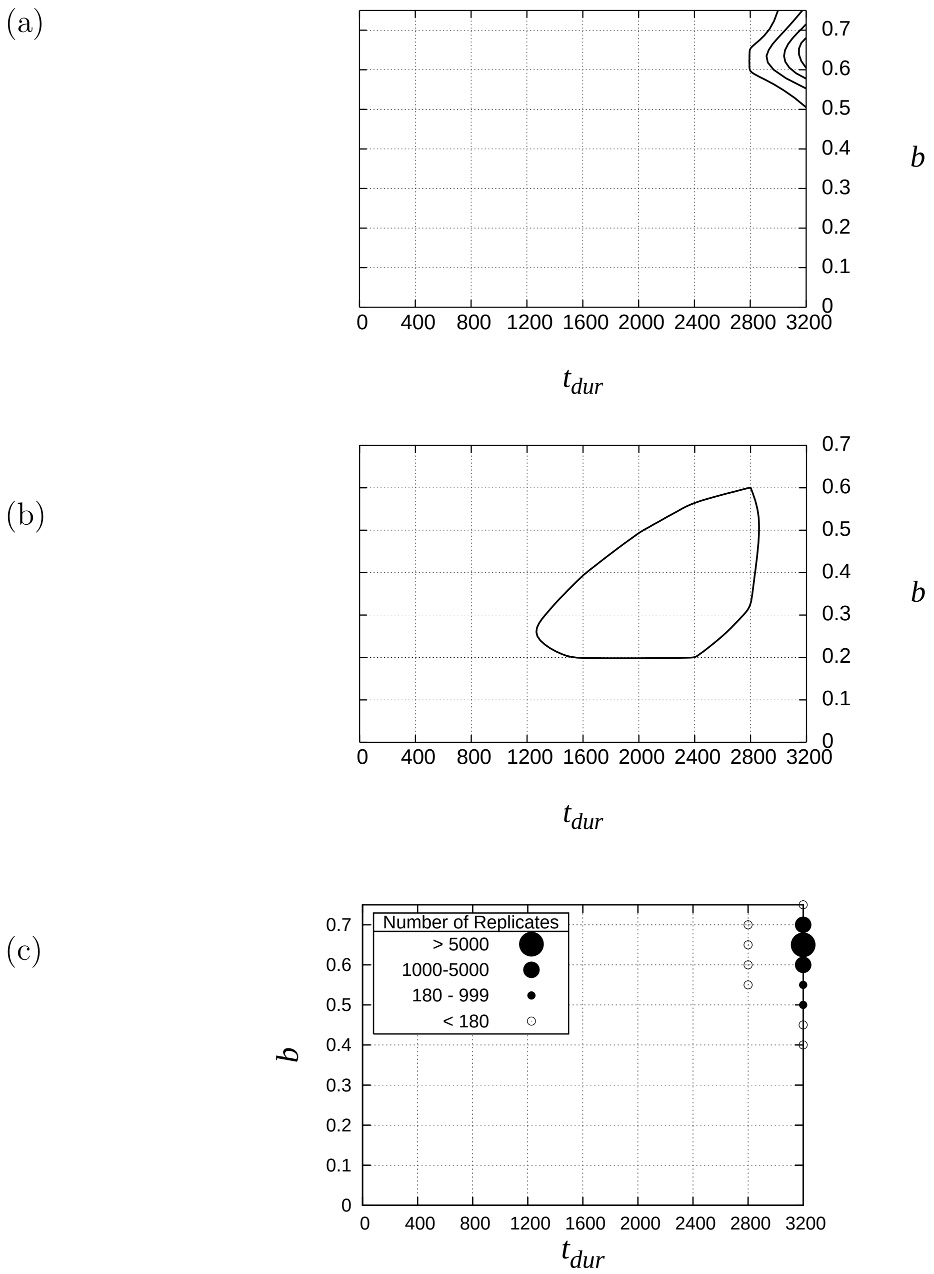


Figure 4.3: Comparing Seattle SNPs confidence regions. The third dimension, t_{start} is fixed at 3200 generations. (a) Likelihood surface using our joint frequencies method. MLE is $\hat{t}_{dur}$=3200, $\hat{b} = 0.65$. (b) Confidence region using the method of Adams and Hudson (2004). MLE is $\hat{t}_{dur}$=2800, $\hat{T} = 0.45$. (c) Bootstrap analysis. MLEs for 95% of the 10,000 replicates fall on the points represented by the solid circles. These points constitute a 95% credible region. The contours in (a) represent intervals of 5 likelihood units from the maximum, and the enclosed area in (b) represents the 95% confidence region determined from simulation.

with the same MLE, and use this cutoff to define a 95% credible region. Figure 4.3c provides a visual representation of the bootstrap results, where the size of each point indicates the number of replicates that had that combination of parameter values as the MLE and the solid circles indicate those points that constitute the 95% credible region.

4.4 Discussion

For the Hausa and Italian data sets, we find that the MLEs obtained using our joint frequencies method lie within confidence regions constructed using methods that utilize different summaries of the data. In analyzing the Hausa data, we confirm previous results that indicate this data set is compatible with a model of constant population size as well as slow, ancient growth and a range of recent growth scenarios (Adams and Hudson 2004; Voight *et al.* 2005). The MLE of $f_{rec} = 3$ and $T = 6$ is identical to that obtained using the method of Adams and Hudson (2004), although the size of the confidence region around the joint frequencies MLE is slightly reduced. In particular, more of the parameter space that includes recent growth scenarios is now rejected in favor of demographic models that more closely resemble constant population size.

In contrast, the joint frequency spectrum of the Italian data set is incompatible with the constant-size model, instead indicating a bottlenecked history. The MLE of $t_{start} = 1600$ generations, $t_{dur} = 1400$ generations, and $b = 0.2$ approximately corresponds to a 80% reduction in effective population size occurring 40,000 years ago and lasting for 35,000 years before recovering back to the ancestral size. A range of other bottleneck scenarios are compatible with the Italian data, including longer, more mild bottlenecks beginning 80,000 years ago and shorter, more severe bottlenecks beginning 20,000 years ago (Figure 4.2a). Interestingly, the 95% confidence region of the Italian data set does overlap with the 95% credible region of the Seattle SNPs for a t_{start} value of 3200 generations, perhaps providing evidence in favor of the older t_{start} value.

In analyzing the Seattle SNPs European data, the size and complexity of the data set prevented us from applying our joint frequencies method to the full three-dimensional parameter space. Therefore, we use fixed value of $t_{start} = 3200$ generations. This value was the MLE of t_{start} obtained by applying the method of ADAMS and HUDSON (2004) and also the only value of t_{start} considered that was contained within the 95% confidence region (t_{start} values of 800 and 1600 generations were rejected for this data set using the method of ADAMS and HUDSON (2004)). With t_{start} fixed at 3200 generations, we find the two-dimensional MLE using the joint frequencies method was $t_{dur} = 3200$ and $b = 0.65$, which represents a 35% reduction in effective population size occurring 80,000 years ago and persisting to the present.

We compared the results of our joint frequencies method to those of either the composite likelihood method of ADAMS and HUDSON (2004), which applies maximum likelihood to the frequency spectrum of SNPs that are assumed to be unlinked, or the combining $p-$values method of VOIGHT *et al.* (2005), which constructs confidence regions based on multiple aspects of genetic data. Each of these methods consider linkage in their respective analyses: the composite likelihood method by using simulations with recombination to adjust critical values and the combined p-values method by calculating a combined test statistic that includes an estimated value of ρ. However, each of these methods utilizes a different summary of the data, and we use direct comparison of confidence regions to evaluate the performance of the joint frequencies method.

The main plots in Figures 4.1a and 4.1b illustrate that the joint frequencies method allows for rejection of some of the recent, rapid growth scenarios that would be accepted with the composite likelihood method. However, by comparing the main plots in Figure 4.1a and 4.1b, it is clear that part of the confidence region obtained by applying the joint frequencies method to the Hausa data set is nearly identical to that of the composite likelihood method for ancient, slow growth scenarios. In fact, the MLE of $f_{rec} = 3$ and $T = 6$ is nearly identical to that reported in Adams and Hudson (2004) for this data set.

When comparing the results for the Italian data based on either the joint frequencies method (Figure 4.2a) or the combined p-values method of Voight *et al.* (2005) (Figure 4.2c), we find that the shapes of the confidence regions are similar. However, the confidence region based on our joint frequencies method is more compact than that based on combined p-values, particulary for a t_{start} value of 3200 generations. We also note that the analysis of Voight *et al.* (2005) incorporates information regarding ancestral population size (N_A) and mutation rate (μ), while our joint frequencies analysis requires no assumptions about the values of these parameters.

As a comparison, we also adjusted our method to incorporate information regarding N_A and μ by fixing a value of θ (=$4N_0\mu$) based on the estimates of Voight *et al.* (2005). We find that the size of the confidence region obtained by using the joint frequencies method is further reduced, while the MLE remains nearly the same (t_{dur} = 1200 generations and $b = 0.2$ with fixed θ versus $t_{dur} = 1400$ generations and $b = 0.2$ with variable θ as described in Methods). We note, however, that the combined p-values method of Voight *et al.* (2005) utilizes a slightly different simulation scheme where recombination and mutation parameters are drawn from distributions rather than fixed. Therefore, we can not fully evaluate the agreement between confidence regions constructed using the two methods.

Due to the large number and heterogeneous lengths of loci in the Seattle SNPs

European data set, simulating confidence intervals around the MLE is computationally infeasible. However, Figure 4.3a illustrates the shape of the likelihood surface resulting from the joint frequencies method. Bootstrap analysis, depicted in Figure 4.3c, provides a sense of the uncertainty in the MLE. If we consider the points where the MLEs of bootstrapped replicate data sets fall 95% of the time to be a 95% credible region, it is clear that these points represent a smaller portion of parameter space than the 95% confidence region constructed by using the composite likelihood method.

We note, however, that our Seattle SNPs MLE and the 95% credible region constructed using the joint frequencies method do not overlap with the 95% confidence region obtained using the composite likelihood method. This suggests that our simple bottleneck model is not sufficient to explain the observed patterns of variation in the Seattle SNPs data set. As a test, we applied a χ^2-like goodness-of-fit test to the frequency spectrum of the Seattle SNPs European data set as compared to the frequency spectrum expected under the MLE parameters and obtained a goodness-of-fit test statistic of 218.18. Because these sites are not independent, we can not use the χ^2 distribution to assess significance of this test statistic. Instead, we determined the 95% critical value of our test statistic by using simulations to account for linkage as described in Adams and Hudson (2004). This simulated 95% critical value is 93.32, indicating that the model defined by the MLE parameters can be rejected on the basis of the frequency spectrum ($p < 0.001$)). This model does, however, represent an improvement over the standard constant-size model for this data set ($\chi^2 = 322.97$) . A similar analysis of the Italian data set indicates that the Italian frequency spectrum is consistent with the MLE parameters obtained by applying the joint frequencies method to either the Italian data ($\chi^2 = 26.05$; $p = 0.47$) or the Seattle SNPs data ($\chi^2 = 26.05$; $p = 0.42$).

The observation that the Italian data are compatible with both the Italian MLE as well as the Seattle SNPs European MLE while the Seattle SNPs European data are not could suggest two possibilities. One is that the smaller size of the Italian data set reduces power to reject these parameters. Also possible is that the Seattle SNPs data are fundamentally different from the Italian data, indicating that different models may be appropriate. To address whether the Italian data are significantly different from the Seattle SNPs data, we first drew 5,000 random subsets of the Seattle SNPs data, each subset matching the number of chromosomes, loci, and segregating sites of the Italian data set. To create each subset, we randomly chose 15 individuals (30 chromosomes) from the 23 individuals sampled for the Seattle SNPs data set. We then randomly sampled 50 loci from the 207 Seattle SNPs loci. Finally, we randomly sampled 383 segregating sites from the subset of 50 loci in 30 chromosomes. For each subset, we calculate our goodness-of-fit test statistic as described above. From the distribution of subset test statistics, we find that test statistics as low as the observed Italian test statistic of 26.05 are rarely observed ($p = 0.0099$), indicating that the Italian data are significantly different from subsets of the Seattle SNPs European data. This could suggest that the simple bottleneck model is sufficient for the Italian data, while an adequate model for the Seattle SNPs data would require additional components. The lack of fit of the Seattle SNPs data to our bottleneck model suggests that conclusions about demographic history based on the Seattle SNPs MLE should be drawn with caution.

We recognize that our model does not account for all features that may be important in shaping observed patterns of polymorphism, such as selection and population structure. Unlike the locus pair data sets which are comprised of noncoding loci far from genes, the Seattle SNPs European data set consists of coding loci implicated in inflammatory response. While we exclude non-synonymous SNPs and putative selected loci (Akey *et al.* 2004) from our analysis, it is still possible that selection could be a confounding factor in our analysis of this data set. Additionally, we do not consider population structure in our analysis due to the additional number of parameters that would be required and the difficulty in visualizing and interpreting multi-dimensional parameter space. However, we note that our joint frequencies method may be extended to accommodate additional model features such as migration, population subdivision, or additional demographic epochs if sufficient computational resources are available.

An additional factor we do not consider is recombination rate variation within or among loci. Heterogeneous recombination rate has been shown to play an important role in shaping observed patterns of genetic variation (Fearnhead and Smith 2005; Myers *et al.* 2005; McVean *et al.* 2004). While we do not incorporate such variation in our analyses, we note that this method could easily be extended to incorporate variation in recombination rate among loci if a specific recombination model were assumed.

We also note the importance of accounting for demographic history in estimating ρ. We find that if a single value of ρ estimated under a standard constant population size model were used as the ancestral ρ value for all grid points, there would be an excess of recombination included in all analyses involving population growth and a dearth of recombination in all bottleneck analyses. Inappropriate rejection of demographic parameter space based on the level of recombination rather than an incompatible demographic scenario results in a shift of confidence regions away from those demographic histories where the ρ value estimated under a constant size model is most different from the ρ value estimated under the correct demographic model, such as very ancient growth models or long, severe bottlenecks (data not shown).

We find that application of maximum likelihood to joint SNP frequencies provides more compact confidence regions as well as generally compatible MLEs in comparison to previous methods. The general agreement between the results obtained using the joint frequencies method and previous methods further highlights the deficiency of the standard neutral model in accounting for patterns of genetic variation in many human populations. Additionally, the differences between parameters estimated for each population underscores the need for population-specific demographic inference. While simple demographic models may not completely account for observed genetic data, they often represent an improvement over the standard constant size model. However, as larger and more complex data sets become available, it is likely that additional model features such as structure, selection, or variation in recombination rate may be required to produce an adequate fit to the data.

CHAPTER 5

DEMOGRAPHY-SPECIFIC ESTIMATION OF ρ

5.1 Introduction

Levels of recombination have a direct impact on the extent of linkage disequilibrium (LD) across the genome, knowledge of which is critical for disease association mapping efforts. Population genetic data contain information regarding historic recombination events. Therefore, much effort has been focused on using genetic data to estimate the population crossing-over rate, $4N_e r$, where N_e is the effective population size and r is the per generation recombination rate. We consider a collection of demographic models that involve recent population size changes following a long-term ancestral size of N_A and consider estimation of $\rho = 4N_A r$.

Approximate likelihood methods for estimating ρ typically assume a model of constant population size (Li and Stephens 2003; Fearnhead and Donnelly 2002; Hudson 2001). Application of such methods results in biased estimates of ρ when populations deviate from the constant size model. In particular, growth scenarios tend to result in overestimated values of ρ, and bottlenecks lead to underestimated values of ρ (Smith and Fearnhead 2005). Since the constant size model is not appropriate for many populations, it is advantageous to incorporate demographic information into estimates of ρ. Here we present a method to estimate ρ under various demographic scenarios and evaluate the improvement in ρ estimates when proper demography is incorporated.

5.2 Model and method

5.2.1 Demographic model

The demographic model used here is depicted in Figure 5.1, where a population at constant size, N_A, until time T at which it may experience an instantaneous size change to an intermediate size N_{int} before immediately undergoing exponential growth to the present size N_{rec}, as described in Adams and Hudson (2004). Parameters include f_{rec}, the ratio of the present population size to the ancestral population size (N_{rec}/N_A), f_{int}, the ratio of the intermediate population size to the ancestral size (N_{int}/N_A), and T, the time of the instantaneous size change and/or onset of growth, in units of $4N_A$ generations. Note that models with an f_{int} value of 1 indicate models of growth beginning at time T with no intermediate size change. Additionally, scenarios of persistent bottleneck beginning at time T with no recovery can be modeled by setting f_{rec} equal to f_{int}. For converting times in units of $4N_A$ generations to year equivalents, we will use an ancestral population size of 10,000 and a 25 year generation time.

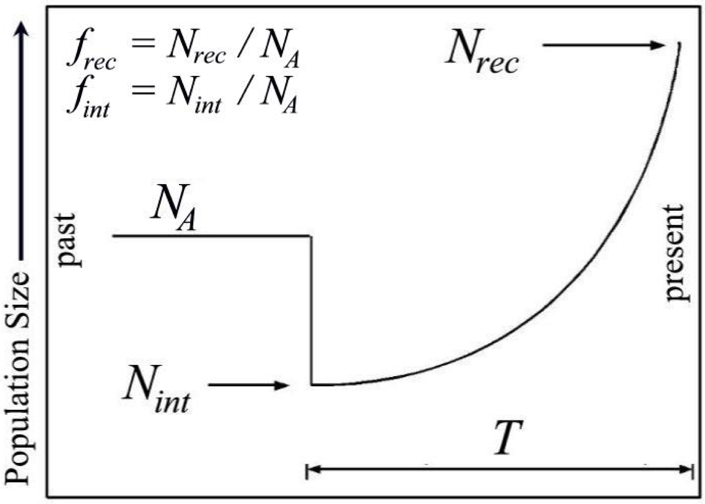


Figure 5.1: Demographic model.

5.2.2 Method

The composite likelihood method of Hudson (2001) is implemented in the `maxdip` program. This program utilizes two-locus sample configuration probability tables to estimate a constant value of $\hat{\rho}$ based on diploid genotype data. These tables are generated by the `ehnrho` program, which typically uses coalescent simulations under the standard constant-size model. However, as noted by Smith and Fearnhead (2005), `ehnrho` also allows for generation of the relevant probability tables under a variety of demographic scenarios, including models of growth or bottlenecks. Both `maxdip` and `ehnrho` can be found at `http://home.uchicago.edu/`∼`rhudson1`.

5.3 Results and discussion

Table 5.1 illustrates estimation of ρ under a range of growth models. Note that the models in Table 5.1 where T is 0.00625 or 0.0125 are within the confidence set of

Incorporating demography under growth models

			Estimates of ρ ($\hat{\rho}$)		
f_{rec}	f_{int}	T	Standard Model	Correct Demography	$\frac{\hat{\theta}_w}{\hat{\rho}} / \frac{\theta}{\rho}$
10	1	0.00625	5.42 (4.21-6.62)	5.00 (3.91-6.12)	0.97 (0.76-1.23)
10	1	0.0125	5.83 (4.52-7.20)	4.97 (3.92-6.11)	0.94 (0.73-1.19)
10	1	0.0375	7.53 (5.90-9.38)	4.98 (3.97-6.11)	0.82 (0.64-1.04)
20	1	0.00625	5.42 (4.12-6.78)	4.98 (3.77-6.21)	0.97 (0.77-1.27)
20	1	0.0125	5.97 (4.57-7.43)	5.01 (3.88-6.21)	0.92 (0.73-1.19)
20	1	0.0375	7.92 (6.22-9.81)	5.00 (3.93-6.10)	0.80 (0.64-1.01)

Table 5.1: Average and (0.025 - 0.975) central interval of the distribution of $\hat{\rho}$ and $\frac{\hat{\theta}_w}{\hat{\rho}} / \frac{\theta}{\rho}$ under growth models. Estimates of $\hat{\rho}$ are obtained using `maxdip` with sample configuration tables generated under either the standard constant size model (Standard Model) or the demographic parameters under which the data sets were simulated (Appropriate Demography). Average and quantiles are based on 1,000 data sets generated via coalescent simulation (Hudson 2002) for a sample of 50 chromosomes, each data set consisting of 50 loci that are each 10kb in length. The input value of ρ is 0.5/kb and θ is chosen for each demographic scenario such that the expected number of segregating sites per locus is equal to the number expected for $\theta = 1$/kb under the constant-size model.

demographic parameter values previously constructed for an African data set (Adams and Hudson 2004). In these cases, the average value of $\hat{\rho}$ is biased upward by approximately 10-20% when using a sample configuration table generated under the standard constant size model. In all growth models examined, ρ is overestimated when the standard model is assumed. In general, the bias becomes more severe with either a larger magnitude of growth or a more ancient time of onset. However, when the demography-specific sample configuration tables are applied, the mean of the distribution of $\hat{\rho}$ values is at or near the true value of ρ in all cases, regardless of the severity of the bias in $\hat{\rho}$ when the standard constant size model is assumed.

We also consider models of population bottlenecks, where the population experiences an instantaneous reduction in population size that persists to the present day. Those models where $f_{rec} = f_{int} = 0.25$ and $T = 0.04$ and also $f_{rec} = f_{int} = 0.5$ and

Incorporating demography under bottleneck models

			Estimates of ρ ($\hat{\rho}$)		
f_{rec}	f_{int}	T	Standard Model	Correct Demography	$\frac{\hat{\theta}_w}{\hat{\rho}}/\frac{\theta}{\rho}$
0.25	0.25	0.04	1.50 (1.10-1.98)	5.00 (3.65-6.46)	2.12 (1.54-2.84)
0.25	0.25	0.08	1.00 (0.66-1.36)	4.98 (3.48-6.50)	2.61 (1.84-3.86)
0.5	0.5	0.04	3.07 (2.44-3.66)	4.93 (3.90-5.92)	1.34 (1.09-1.67)
0.5	0.5	0.08	2.58 (2.02-3.19)	4.93 (3.87-6.09)	1.46 (1.17-1.83)
1	0.25	0.04	3.25 (2.38-4.16)	4.96 (3.64-6.38)	1.29 (0.98-1.73)
1	0.25	0.08	3.06 (2.22-3.98)	4.97 (3.64-6.43)	1.31 (0.97-1.79)
1	0.5	0.04	4.09 (3.11-5.18)	4.92 (3.72-6.25)	1.14 (0.88-1.48)
1	0.5	0.08	4.00 (3.08-5.13)	4.98 (3.82-6.37)	1.13 (0.86-1.44)
10	0.25	0.04	5.66 (4.42-6.98)	4.96 (3.90-6.09)	0.95 (0.75-1.22)
10	0.25	0.08	6.92 (5.34-8.53)	4.97 (3.90-6.09)	0.86 (0.68-1.11)
10	0.5	0.04	6.76 (5.20-8.30)	4.96 (3.85-6.02)	0.86 (0.62-0.97)
10	0.5	0.08	8.61 (6.75-10.46)	4.98 (3.99-6.01)	0.77 (0.62-0.97)

Table 5.2: Average and (0.025 - 0.975) central interval of the distribution of $\hat{\rho}$ and $\frac{\hat{\theta}_w}{\hat{\rho}}/\frac{\theta}{\rho}$ under bottleneck models. Details are as described in Table 5.1.

$T = 0.08$ have been shown to be compatible with non-African data sets (Voight *et al.* 2005). In these scenarios, the average $\hat{\rho}$ is biased downward by approximately 50-70% when the standard constant size model is assumed. For more complex models that include a population size reduction followed by exponential growth, values of ρ may be either over- or underestimated when the standard table is applied (Table 5.2). Again, when the appropriate demographic history is considered, the distributions of $\hat{\rho}$ are very similar for all demographic scenarios examined, with the mean at or near the true value of ρ.

A quantity of interest is $\frac{\theta}{\rho}$, where $\theta = 4N_e\mu$, $\rho = 4N_er$, and, therefore, $\frac{\theta}{\rho}$ equals $\frac{\mu}{r}$, the ratio of the mutation rate to the crossing-over rate. Models to account for patterns of polymorphism and divergence are often dependent on the parameters θ and ρ, both of which are composite parameters which involve the effective population size, N_e, a quantity about which we have little direct empirical information. However, we can obtain estimates of μ and r from divergence and genetic map data, respectively, and, thus, the quantity $\frac{\mu}{r}$ can be at least roughly known. It is then interesting to investigate both the accuracy of $\frac{\hat{\theta}}{\hat{\rho}}$ estimated from population genetic data under changing population size scenarios and also the agreement between population genetic estimates of $\frac{\hat{\theta}}{\hat{\rho}}$ and empirical estimates of $\frac{\mu}{r}$.

In order to determine whether the quantity $\frac{\theta}{\rho}$ can be estimated accurately without correcting for demography, we consider the ratio of $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ under the demographic scenarios considered above, where $\hat{\theta}_W$ is Watterson's estimate of θ (Watterson 1975), and $\hat{\rho}$ is the value of ρ estimated under the standard constant size model. One might hope that both $\hat{\theta}_W$ and $\hat{\rho}$ might be biased in the same way under alternative demographic scenarios, and, thus, that the ratio might be relatively unbiased.

However, we find that, under growth models, the average ratio of $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ is less than 1, indicating that population growth tends to reduce levels of linkage disequilibrium to a greater extent than it increases levels of polymorphism (Table 5.1). Like the overestimation of $\hat{\rho}$, the underestimation of $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ correlates with the severity of the growth scenario, with larger values of f_{rec} and T resulting in greater bias.

The average of $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ under bottleneck models reveals the opposite trend (Table 5.2). We find that $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ is overestimated under persistent bottleneck scenarios, confirming that the increased levels of linkage disequilibrium caused by population bottlenecks are greater than the concomitant decrease in polymorphism. Under more complex models involving both bottleneck and growth, the bias in $\frac{\hat{\theta}_W}{\hat{\rho}}/\frac{\theta}{\rho}$ depends on both the severity of the bottleneck and the magnitude of growth (Table 5.2).

In order to examine the concordance between estimates of μ and r gleaned from divergence and genetic map data, respectively, with corresponding estimates of θ and ρ from population genetic data, we consider the Italian data set described in Voight *et al.* (2005) as an illustration. For this data set, the average value of μ was found to be 2.63×10^{-8} based on human-chimpanzee divergence data, and the average value of r was 1.42×10^{-8} based on the Marshfield genetic map, indicating a $\frac{\mu}{r}$ ratio of 1.85 (Voight *et al.* 2005).

If we estimate θ for the Italian data set by $\hat{\theta}_w$ and obtain $\hat{\rho}$ using the standard constant-size model, $\hat{\theta}_w$ is 8.1×10^{-4} and $\hat{\rho}$ is 3.11×10^{-4}. Thus $\frac{\hat{\theta}_w}{\hat{\rho}}$ is 2.60, which is 1.4 times greater than the $\frac{\mu}{r}$ ratio estimated from divergence and genetic maps. This is consistent with our simulation results that indicate $\frac{\hat{\theta}_w}{\hat{\rho}}$ is an overestimate of $\frac{\mu}{r}$ under bottleneck scenarios such as those we believe are applicable to the Italian data set.

However, we can incorporate the inferred demographic history of the Italian data set into the estimate of $\frac{\mu}{r}$. The parameter set of $f_{rec} = 0.5$, $f_{int} = 0.5$, and $T = 0.08$ considered in Table 5.2 is one bottleneck scenario that has been shown to be compatible with the Italian data. We can estimate θ under this demographic model by using

$$\hat{\theta}^*_w = \frac{S/L}{\bar{\tau}} \qquad , \tag{5.1}$$

where S is the average number of segregating sites per locus, L is the average locus length, and τ is the average length of a gene genealogy simulated under the aforementioned demographic parameters. Likewise, $\hat{\rho}$ can be estimated with incorporated demographic information by using `maxdip` with the appropriate table as described above. Using these approaches, we obtain a $\hat{\theta}^*_w$ of 1.05×10^{-3} and $\hat{\rho}$ of 6.08×10^{-4}, indicating a $\frac{\mu}{r}$ ratio of 1.73, which is in much closer agreement with the empirical $\frac{\mu}{r}$ ratio than the estimate obtained without accounting for demographic history. These analyses suggest that both LD and levels of variation in the Italian data set are compatible with a simple neutral model featuring a recent bottleneck.

CHAPTER 6 DISCUSSION AND CONCLUSIONS

6.1 Factors affecting accuracy and power

Application of the demographic inference methods detailed in the previous chapters has yielded a number of important results. Highlighted in Chapter 2, the first of these is the identification of factors affecting both the accuracy of demographic parameter estimates and also the power to reject the null model of constant population size. Simulations reveal that estimates of growth parameters that are based on the frequency spectrum of unlinked sites are influenced by both the magnitude of growth and the time of growth onset. Estimates of f_{rec}, the ratio of the present population size to the ancestral population size, are biased increasingly upward with increasing values of f_{rec}. Interestingly, estimates of the time of onset of growth, T, improve with increasing values of f_{rec}. Estimates of both f_{rec} and T improve with larger values of T, which represent more ancient growth scenarios.

Additionally, power analyses indicate that the power to reject the null hypothesis of constant population size increases with both the magnitude of growth and the time of onset of growth. Sample size and number of unlinked polymorphic sites also play a significant role, as recent, rapid growth scenarios can not be consistently distinguished from constant population size with small sample sizes (< 100 chromosomes and $\sim$500 unlinked polymorphic sites), regardless of the magnitude of growth.

These results suggest that, while the frequency spectrum contains information regarding historic changes in population size, the accurate extraction of this information depends upon the magnitude and timing of such events. This highlights the need for incorporation of additional aspects of genetic data as described in Chapters 3 and 4, particularly for recent, rapid growth scenarios which show the largest bias in f_{rec} and T estimation and the lowest power to reject the null model of constant population size. Additionally, these results and others presented in Chapter 2 provide a guide to researchers who may be interested the accuracy of parameter estimates they might obtain from a data set consisting of a particular number of chromosomes and segregating sites.

6.2 Inference methods

Another notable contribution of the preceding chapters is the introduction and application of the demographic inference methods detailed in Chapters 2-4. Some major features of each of these methods are summarized in Table 6.1. Chapter 2 presents a composite likelihood method that utilizes the frequency spectrum of polymorphic sites. As indicated above, simulations indicate that the accuracy of this method depends upon the underlying demographic scenario, although accurate estimates can be made for most demographic parameters if a sufficient number of chromosomes and segregating sites is available. This method treats all polymorphic sites as independent, regardless of linkage. However, a procedure is also presented by which

Summary of methods

	CL	CPV	JFS
Computational cost	Low	Medium	High
Includes ρ and μ Variation	No	Yes	No
Major Advantage	Very efficient	Can combine any summaries	Small confidence regions
Major Disadvantage	Ignores linkage in estimation	Influenced by estimate of N_A	Computational expense

Table 6.1: The methods summarized here are the Composite Likelihood (CL) method described in Chapter 2, the Combining p-values (CPV) method described in Chapter 3, and the Joint Frequency Spectrum (JFS) method described in Chapter 4.

simulations can be used to construct confidence regions around MLEs obtained using this method. Some major advantages of this method include this ease of adjusting for linkage as well as both computational efficiency and use of the entire frequency spectrum as opposed to a single summary statistic such as Tajima's D.

A novel approach to considering multiple aspects of genetic data is taken in Chapter 3. Power analyses show that a combined summary statistic based on the average Tajima's D value, the average number of segregating sites, and an estimate of ρ over all loci of a data set provides the greatest power to distinguish bottleneck scenarios from the standard model of constant population size. This method is more computationally demanding than the aforementioned composite likelihood method because large numbers of coalescent simulations are required to assess the significance of the combined test statistic over a grid of demographic parameters. However, this method does allow for incorporation of information regarding the ancestral population size and also considers levels of variation as well as linkage disequilibrium. These features makes this method ideal for studies such as the one described in Chapter 3, which utilizes multiple population samples and employs a data collection scheme that allows for simultaneous assessment of both polymorphism and linkage.

Finally, Chapter 4 details a method that summarizes the data in terms of joint frequencies of linked SNPs. Although this method represents data in a way similar to the composite likelihood method of Chapter 2, the joint frequencies method directly incorporates linkage into the demographic analyses by using coalescent simulations with recombination as opposed to simple one-site coalescent simulations. In all cases examined, the confidence region surrounding estimates obtained using the joint frequencies method is smaller than that of either the composite likelihood or combined p-values method described above. Additionally, this method may be applied either with or without making assumptions regarding the ancestral population size and mutation rate. However, there is significant computational cost for applying this method to large data sets consisting of loci of heterogeneous length.

Although each of these methods has been used to infer specific demographic parameters that relate to population size changes, it is important to note that any of these methods may be adapted to models that include population structure or additional demographic epochs. Additionally, the composite likelihood and joint frequencies methods may easily be modified to accommodate recombination or mutation rate heterogeneity between loci, although the combined p-values method already includes these features as described in Chapter 3. While the choice of method may depend largely on computational resources or the structure of an individual data set, such flexibility allows for accommodation of new information regarding the values and distributions of nuisance parameters, such as the mutation or recombination rate, or the demographic models that may be of interest to particular researchers.

6.3 Consistency of population-specific estimates

An additional important result is the consistency of demographic parameter estimates for similar data sets across different described methods. As described in Chapters 2 and 4, an African Hausa data set was analyzed by using two of the three demographic inference methods described above as well as by using maximum likelihood on a single summary statistic, Fu and Li's D^*, in Chapter 3. The maximum likelihood estimate indicated very ancient, very slow three-fold growth beginning approximately six million years ago when either the composite likelihood or the joint frequencies method was applied. A range of recent growth scenarios as well as constant population size were also accepted, where the parameters within the joint frequencies confidence region were a subset of those within the composite likelihood confidence region. Additionally, the MLE for this data set obtained by using D^* as described in Chapter 3 lies within the confidence regions of both the composite likelihood and joint frequencies methods. The results for the composite likelihood and joint frequencies methods are summarized in Table 6.2.

The case is similar for the Italian data set, which was analyzed using the combined p-values method in Chapter 3 and the joint frequencies method in Chapter 4. A series of bottleneck models, with bottleneck onset times from 20,000 to 120,000 years ago, was included in the acceptance region constructed using the combined p-values method. Again, the confidence region constructed using the joint frequencies method included a subset of the parameter space accepted using the combined p-values method. Specifically, the MLE obtained using the joint frequencies method indicates an 80% reduction in effective population size occurring 40,000 years ago and persisting for 35,000 years before recovering to the ancestral size. This estimate lies within the combined p-values acceptance region.

Summary of MLEs

Data Set	Method	$\hat{f}_{rec}$	$\hat{f}_{int}$	$\hat{T}^{\dagger}$	$\hat{t}^{\ddagger}_{start}$	$\hat{t}^{\ddagger}_{dur}$	b
Hausa							
	CL	3.1	1	6.1			
	JFS	3	1	6			
SSNPs African American							
	CL§	1.9	1	0.27			
SSNPs European$^{\|}$							
	CL	2.0	0.15	0.0375			
SSNPs European$^{\sharp}$							
	CL§				3200	2800	0.25
	JFS§				3200	3200	0.65
Italian							
	JFS				1600	1400	0.2

Table 6.2: The results summarized here are obtained using the Composite Likelihood (CL) method described in Chapter 2 and the Joint Frequency Spectrum (JFS) method described in Chapter 4.

† Parameter is in units of $4N_A$ generations.
‡ Parameter is in units of generations.
§ Estimates were found to be incompatible with the observed frequency spectrum.
$^{\|}$ Data set is as described in Chapter 2.
$^{\sharp}$ Data set is as described in Chapter 4.

The third data set that was examined using more than one of the methods described above is the Seattle SNPs European data set, which was evaluated using the composite likelihood method of Chapter 2 and the joint frequencies method of Chapter 4. In this case, the acceptance regions are similar, but not overlapping. As described in the following section, this discrepancy could be due to the simple bottleneck model being inadequate for the Seattle SNPs data.

6.4 Consistency with palaeontological record

In addition to the consistency of estimates across methods, it is also interesting to consider the consistency of these estimates with the fossil record. For the African Hausa data set, the MLE estimates of very ancient, slow growth scenarios are not likely to be biologically relevant. However, some more recent, rapid growth models that also lie within the confidence regions constructed using each of the methods described above are consistent with an African population expansion occurring 70-80kya as posited by Mellars (2006).

The genetic evidence for a population bottleneck, as seen in the Italian, Chinese, and Seattle SNPs European data, is also supported by the fossil record. The earliest modern skeletal remains outside Africa, dating from 115 kya, have been found in Israel (Skhul and Qafzeh caves in the Levant region) (Stringer 2003). However, it is hypothesized that these anatomically modern populations that originally dispersed from Africa were replaced with more technologically and behaviorally advanced populations which dispersed from Africa approximately 60,000 years ago (Mellars 2006).

Such a dispersal could be reflected by the MLE parameters corresponding to a population bottleneck estimated for the non-African populations examined in this dissertation. In the combining p-values analyses, both the Italian and Chinese data sets are shown to be compatible with a range of bottleneck scenarios, including those beginning 3200 generations ago, which corresponds to 64 kya if a generation time of 20 years is assumed. Additionally, the Seattle SNPs European data set is also shown to be most compatible with a bottleneck beginning 3200 generations ago, although the simple bottleneck model can not be accepted as a complete explanation of the Seattle SNPs European data.

6.5 Compatibility of MLEs with observed data

In addition to noting the general consistency of the demographic estimates mentioned above, both with each other and with the fossil record, it is also important to consider whether the inferred parameter values are compatible with the observed patterns of variation. The demographic models discussed in the preceding chapters are simple growth or bottleneck models, and it is interesting to examine whether these simple models resolve any incompatibility with the standard neutral model without invoking any additional features such as selection or population structure.

In Chapters 2 and 4, a χ^2-like goodness of fit test was performed to determine whether the frequency spectra expected under the inferred demographic parameters are significantly different from the observed frequency spectra. In the case of the Hausa data, both the MLE parameters and the parameters corresponding to constant population size were compatible with the Hausa frequency spectrum, with the MLE growth parameters producing a slightly improved fit.

In contrast, the confidence regions for the Italian data set (using either the combined p-values or joint frequencies method) do not include parameters corresponding to constant population size, instead including parameters that represent a bottleneck scenario. While the Italian data set is incompatible with the constant-size model, the

MLE parameters identified by the joint frequencies method provide a good fit to the data, as application of the goodness-of-fit test described above indicates that the frequency spectrum expected under the MLE bottleneck parameters is not significantly different from the observed Italian frequency spectrum.

However, similar goodness-of-fit analyses of the Seattle SNPs data sets do not indicate compatibility with simple demographic scenarios. In Chapter 2, the composite likelihood MLE obtained for the Seattle SNPs African American data set was rejected by the goodness-of-fit test, even though the MLE growth model represented an improvement over the constant population size model. For the Seattle SNPs European data set first described in Chapter 2, the frequency spectrum expected under the composite likelihood MLE was compatible with the observed data. However, in Chapter 4, a larger amount of data for the Seattle SNPs European data set was available. Using the expanded data set, it was shown that the joint frequencies MLE parameters produce an expected frequency spectrum that is significantly different from the observed Seattle SNPs frequency spectrum.

6.6 Effect of data on demographic inference

The observation that the Hausa and Italian observed frequency spectra are consistent with their respective MLEs while those of the Seattle SNPs are not underscores the importance of data sets specifically tailored to demographic inference. All of the data sets examined in the preceding chapters are a result of full resequencing efforts. Such data are critical for demographic analyses, which rely on the accurate capture of low frequency variants. Additionally, the Hausa and Italian data sets are obtained from clearly defined, relatively homogeneous population samples and consist of noncoding regions that are far from known or predicted genes and regions conserved between human and mouse, reducing the potential influence of selection on the observed patterns of variation.

The Seattle SNPs data, on the other hand, are not ideal for demographic studies. First, these data consist of coding loci that have been implicated in inflammatory response. Despite the removal of coding SNPs and several putative selected loci from the analyses, it is possible that selection is a confounding factor in the Seattle SNPs demographic analyses. Additional features such as population structure could also contribute to the poor fit of the Seattle SNPs data to the frequency spectra predicted by the MLEs, particularly for the African American data set. If meaningful conclusions are to be drawn from data sets such as the Seattle SNPs, then more complex demographic models which incorporate features such as population structure, migration, or selection are likely to be required.

6.7 Demography and ρ

A final development reported in this dissertation is a method of incorporating demographic information into estimation of recombination rate, as described in Chapter 5. Simulation analyses confirm that deviations from the standard constant size model result in bias of ρ estimates when the standard model is assumed. Specifically, scenarios of growth lead to overestimates of $\hat{\rho}$ and population bottlenecks result in underestimates of $\hat{\rho}$. However, when the appropriate demographic history is considered, the bias in $\hat{\rho}$ is eliminated.

The ratio of $\frac{\hat{\theta}_W}{\hat{\rho}}$ is also considered, as it is often assumed to represent the ratio of the mutation rate to the recombination rate, even under changing population size scenarios. The results of Chapter 5, however, illustrate that estimates of θ and ρ made under the assumption of constant population size are not biased to the same degree under alternate demographic scenarios, and, as a result, $\frac{\hat{\theta}_w}{\hat{\rho}}$ is often not an accurate representation of the mutation to recombination rate ratio when the population size is not constant. Since many studies, including those detailed in this dissertation, have illustrated that the constant population size model is not consistent with many data sets, these results suggest that it is appropriate to incorporate information regarding demographic history into estimates of ρ.

REFERENCES

Adams, A. M. and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168**: 1699–1712.

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. **2**: e286.

Ardlie, K. G., L. Krugliak, and M. Seilstad, 2002a Patterns of linkage disequilibrium in the human genome. Nature Rev. Genet. **3**: 299–309.

Ardlie, K. G., L. Krugliak, and M. Seilstad, 2002b Patterns of linkage disequilibrium in the human genome. Nature Rev. Genet. **3**: 566.

Aris-Brosou, S. and L. Excoffier, 1996 The impact of population expansion and mutation rate heterogeneity on dna sequence polymorphism. Mol. Biol. Evol. **13**: 494–504.

Beerli, P. and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98**: 4563–4568.

Cargill, M., D. Altshuler, J. Ireland, P. Skyler, and K. Ardlie, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22**: 231–238.

Clark, J. D., Y. Beyene, G. WoldeGabriel, W. K. Hart, P. R. Renne, H. Gilbert, A. Defleur, G. Suwa, S. Katoh, K. R. Ludwig, J.-R. Boisserie, B. Asfaw, and T. D. White, 2003 Stratigraphic, chronological and behavioral contexts of pleistocene *Homo sapiens* from middle awash, ethiopia. Nature **423**: 747–752.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill, M. L. Petzl-Erler, G. K. Haines, and D. H. Barch, 1998 Branching pattern in the evolutionary tree for human mitochondrial DNA. Genetics **148**: 1269–1284.

Di Rienzo, A. and A. C. Wilson, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci U S A **88**: 1597–1601.

Eswaran, V., H. Harpending, and A. R. Rogers, 2005 Genomics refutes an exclusively African origin of humans. J Hum Evol **49**: 1–18.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

Fay, J. C., J. Wyckoff, and C.-I. Wu, 2001 Positive and negative selection on the human genome. Genetics **158**: 1227–1234.

Fearnhead, P. and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates (with discussion). JRSS, series B **64**: 657–680.

Fearnhead, P. and N. G. Smith, 2005 A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. Am J Hum Genet **77**: 781–794.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. Wall, J. Donfack, and A. Di Rienzo, 2001 Gene conversion and different populations histories may explain explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69**: 831–843.

Fu, Y.-X. and W.-H. Li, 1993 Statistical test of neutrality of mutations. Genetics **133**: 693–709.

Goldstein, D. B. and L. Chikhi, 2002 Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet **3**: 129–152.

Griffiths, R. C. and S. Tavare, 1998 The age of a mutation in the general coalescent tree. Stoc. Models **14**: 273–295.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res **15**: 790–799.

Harding, R. M. and G. McVean, 2004 A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev **14**: 667–674, Historical Article.

Harpending, H. and A. Rogers, 2000 Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet **1**: 361–385, Historical Article.

HARPENDING, H. C., M. A. BATZER, M. GURVEN, L. B. JORDE, A. R. ROGERS, and S. T. SHERRY, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**: 1961–1967.

HEY, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol **14**: 166–172.

HINDS, D. A., L. L. STUVE, G. B. NILSEN, E. HALPERIN, E. ESKIN, D. G. BALLINGER, K. A. FRAZER, and D. R. COX, 2005 Whole-genome patterns of common DNA variation in three human populations. Science **307**: 1072–1079.

HUDSON, R. R., 1983 Properties of the neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**: 183–201.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7**: 1–44.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159**: 1805–1817.

HUDSON, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI, and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136**: 1329–1340.

JEFFREYS, A. J. and C. A. MAY, 2004 Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. **36**: 151–156.

JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO, and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics **170**: 1401–1410.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS, and L. B. JORDE, 1998 Signatures of population expansion in microsatellite repeat data. Genetics **148**: 1921–1930.

KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON, B. RICHARDSSON, S. SIGURDARDOTTIR, J. BARNARD, B. HALLBECK, G. MASSON, A. SHLIEN, S. T. PALSSON, M. L. FRIGGE, T. E. THORGEIRSSON, J. R. GULCHER, and K. STEFANSSON, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31**: 241–247.

KREITMAN, M. and A. DI RIENZO, 2004 Balancing claims for balancing selection. Trends Genet **20**: 300–304.

Kuhner, M. K., J. Yamato, and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149**: 429–434.

Lahr, M. M. and R. A. Foley, 1998 Toward a theory of modern human origins: geography, demography, and diversity in recent human evolution. Yb. Phys. Anthropol. **41**: 137–176.

Li, N. and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**: 2213–2233.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala, D. Church, S. Wheelan, J. Baker, M. Ward, M. Kholodov, L. Phan, E. Czabarka, J. Murvai, D. Cutler, S. Wooding, A. Rogers, A. Chakravarti, H. Harpending, P. Kwok, and S. Sherry, 2003 Sequence variantions in the public human genome data reflect a bottlenecked population history. Proc. Natl. Acad. Sci. USA **100**: 376–381.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166**: 351–372.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bently, and P. J. Donnelly, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

Mellars, P., 2006 Why did modern human populations disperse from Africa *ca.* 60,000 years ago? A new model. Proc. Natl. Acad. Sci. USA **103**: 9381–9386.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310**: 321–324.

Nickerson, D. A., V. O. Tobe, and S. L. Taylor, 1997 Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. **25**: 2745–2751.

Nielsen, R., 1999 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**: 931–942.

Nielsen, R., 2004 Population genetic analysis of ascertained snp data. Hum Genomics **1**: 218–224.

Nordborg, M. and S. Tavare, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18**: 83–90.

Parra, E. J., A. Marcini, J. Akey, and J. Martinson, 1998 Estimating african american admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63**: 1839–1851.

Pluzhinikov, A., A. Di Rienzo, and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161**: 1209–1218.

Polanski, A. and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics **165**: 427–436.

Pritchard, J. K. and M. Przeworski, 2001 Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. **69**: 1–14.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999 Population growth of human y chromosomes: a study of y chromosome microsatellites. Mol. Biol. Evol. **16**: 1791–1798.

Przeworski, M., R. R. Hudson, and A. D. Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16**: 296–302.

Ptak, S. E. and M. Przeworski, 2002 Evidence for population growth in humans is confounded by fine-scale population structure. Trends Genet. **18**: 559–563.

Ptak, S. E., K. Voelpel, and M. Przeworski, 2004 Insights into recombination from patterns of linkage disequilibrium in humans. Genetics **167**: 387–397.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander, 2001 Linkage disequilibrium in the human genome. Nature **411**: 199–203.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet **32**: 135–142.

Rogers, A. R. and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**: 552–569.

Rogers, A. R. and L. B. Jorde, 1995 Genetic evidence on modern human origins. Hum. Biol. **1**: 1–36.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, 2002 Genetic structure of human populations. Science **298**: 2381–2385.

S Tavare, R. C. G., D J Balding and P. Donnelly, 1997 Inferring coalescence times from dna sequence data. Genetics **145**: 505–518.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

Sherry, S. T., A. R. Rogers, H. Harpending, H. Soodyall, T. Jenkins, and M. Stoneking, 1994 Mismatch distributions of mtDNA reveal recent human population expansions. Hum Biol **66**: 761–775.

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

Smith, C. N. and P. Fearnhead, 2005 A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. Genetics **171**: 2051–62.

Soldevila, M., F. Calafell, A. Helgason, K. Stefansson, and J. Bertranpetit, 2005 Assessing the signatures of selection in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. Trends Genet **21**: 389–391, Comment.

Stajich, J. E. and M. W. Hahn, 2005 Disentangling the effects of demography and selection in human history. Mol Biol Evol **22**: 63–73.

Stephens, M., 2001 *Handbook of Statistical Genetics*, chapter Inference Under the Coalescent, pp. 213–238. Wiley and Sons.

Stringer, C., 2003 Out of ethiopia. Nature **423**: 692–695.

Tajima, F., 1989a The effect of change in population size on dna polymorphism. Genetics **123**: 597–601.

Tajima, F., 1989b Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics **123**: 585–595.

The International HapMap Consortium, 2003 The International HapMap Project. Nature **426**: 789–796.

Thornton, K. and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics **172**: 1607–1619.

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonne-Tamir, A. S. Santachiara-Benerecetti, P. Moral, and M. Krings, 1996 Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. Science **271**: 1380–1387.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Quian, R. R. Hudson, and A. Di Rienzo, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Nat. Acad. Sci. **102**: 18508–18513.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero, and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms–and inferences about human demographic history. Am. J. Hum. Genet. **69**: 1332–1347.

Wakely, J. and S. Lessard, 2003 Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics **164**: 1043–1053.

Wall, J. and M. Przeworski, 2000 When did the human population size start increasing? Genetics **155**: 1865–1874.

Wall, J. D., L. A. Frisse, R. R. Hudson, and A. D. Rienzo, 2003 Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am. J. Hum. Genet. **74**: 1330–1340.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theo. Popul. Biol. **7**: 256–276.

Weiss, G. and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149**: 1539–1546.

White, T. D., B. Asfaw, D. DeGusta, H. Gilbert, G. D. Richards, G. Suwa, and F. C. Howell, 2003 Pleistocene *Homo sapiens* from middle awash, ethiopia. Nature **423**: 742–747.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, R. Nielsen, and C. D. Bustamante, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Nat. Acad. Sci. USA **102**: 7882–7887.

Wilson, I. J. and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150**: 499–510.

Wooding, S. and A. Rogers, 2002 The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics **161**: 1641–1650.

Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber, 2001 Comparison of human genetic and sequence-based physical maps. Nature **409**: 951–953.

Zhu, L. and C. D. Bustamante, 2005 A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics **170**: 1411–1421.