A Comparison of Coalescent Estimation Software

Kristen Piggott Shepherd
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd
Part of the Statistics and Probability Commons

BYU ScholarsArchive Citation
https://scholarsarchive.byu.edu/etd/21

This Selected Project is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
A COMPARISON OF COALESCENT ESTIMATION SOFTWARE

by

Kristen P. Shepherd

A project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

December 2002
Copyright © 2002 Kristen P. Shepherd

All Rights Reserved
BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Kristen P. Shepherd

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date  David G. Whiting, Chair

Date  Bruce J. Collings

Date  Gilbert W. Fellingham
As chair of the candidate’s graduate committee, I have read the project of Kristen P. Shepherd in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

David G. Whiting
Chair, Graduate Committee

Accepted for the Department

Gilbert W. Fellingham
Graduate Coordinator

Accepted for the College

G. Rex Bryce
Associate Dean, College of Physical and Mathematical Sciences
ABSTRACT

A COMPARISON OF COALESCENT ESTIMATION SOFTWARE

Kristen P. Shepherd
Department of Statistics
Master of Science

Coalescent theory is a method often used by population geneticists in order to make inferences about evolutionary parameters. The coalescent is a stochastic model that approximates ancestral relationships among genes. An understanding of the coalescent pattern of a sample of sequences, along with some knowledge of the mutations that have occurred, provides information about the evolutionary forces that have acted on the population. Processes such as migration, recombination, variable population size, or natural selection are the forces that affect the genealogies and lead to genetic variability in a sample. Coalescent theory provides a statistical description of the variability in the sample, which in turn leads to inference about evolutionary parameters such as population size, population growth rates, and migration rates.

Several methods have been developed that model the coalescent under different sets of evolutionary assumptions. We have examined and compared three computer packages
that estimate parameters under the coalescent model: LAMARC, Genetree, and UPBLUE. These are not commercial computer programs, but have been developed by researchers for their own use and made available to others. Their performance in various areas has not been previously well established.

We compared the programs in the general areas of model assumptions, availability, usability, and results. No single program is superior to the others in all areas, but each have strengths and weakness. Program selection must be based on the researcher’s data, goals and preferences. This comparison should help population geneticists determine which program would be most applicable for their data and research.
ACKNOWLEDGMENTS

I would like to thank Dr. Whiting for his time and effort throughout this project and especially for helping me finish from long-distance. I’m also grateful for professors who have taught me and staff that have assisted me. Finally, I would like to thank Bryan and other family members for their constant support and encouragement.
# TABLE OF CONTENTS

1 Introduction ........................................................................................................ 1
2 Population Genetics ............................................................................................ 3
3 Basic Coalescent Theory ..................................................................................... 6
   3.1 An Illustrative Example: Coalescent Model from a Sample ....................... 7
   3.2 The Parameter $\theta$ .................................................................................. 11
   3.3 The Wright-Fisher Model ......................................................................... 12
   3.4 Kingman’s Coalescent ............................................................................. 14
4 The Statistical Model .......................................................................................... 16
   4.1 Coalescent Times ...................................................................................... 16
   4.2 Mutations ................................................................................................. 18
5 Estimation Methods ........................................................................................... 19
   5.1 History ....................................................................................................... 19
   5.2 Major Research Groups .......................................................................... 20
   5.3 Program Comparison ............................................................................... 23
6 Methodology ........................................................................................................ 25
   6.1 Description of Datasets .......................................................................... 25
      6.1.1 Nuu-Chah-Nulth Data ....................................................................... 26
      6.1.2 Africa Data ....................................................................................... 27
      6.1.3 Simulated Datasets .......................................................................... 27
   6.2 Platform Availability ................................................................................. 28
   6.3 Documentation .......................................................................................... 29
   6.4 Data Input Format ..................................................................................... 31
   6.5 Running Time ........................................................................................... 33
   6.6 Output ....................................................................................................... 35
   6.7 Ease of use ............................................................................................... 36
   6.8 Stability ..................................................................................................... 36
7 Detailed Program Evaluation ............................................................................. 38
   7.1 LAMARC Package ................................................................................... 38
      7.1.1 Search Strategy .................................................................................. 38
      7.1.2 Fluctuate ............................................................................................ 40
      7.1.3 Recombine ......................................................................................... 44
      7.1.4 Migrate .............................................................................................. 49
      7.1.5 Lamarc ............................................................................................... 51
   7.2 Genetree ..................................................................................................... 52
      7.2.1 Wright-Fisher Model ($\theta$ Only) ...................................................... 53
      7.2.2 Variable Population Size ................................................................. 54
      7.2.3 Migration ........................................................................................... 57
7.3 UPBLUE ........................................................................................................... 61
8 Discussion and Conclusions ............................................................................ 62
Appendix ............................................................................................................ 66
A.1 Tables ........................................................................................................... 67
A.2 Datasets ......................................................................................................... 71
  A.2.1 Nuu-Chah-Nulth Data .............................................................................. 71
  A.2.2 Africa data .............................................................................................. 79
  A.2.3 Simulated Datasets ................................................................................. 79
A.3 Data Input Format Examples ......................................................................... 80
  A.3.1 Fluctuate ................................................................................................ 80
  A.3.2 Recombine ............................................................................................. 81
  A.3.3 Migrate ................................................................................................ 81
  A.3.4 Lamarc ................................................................................................ 82
  A.3.5 Genetree ................................................................................................. 84
  A.3.6 UPBLUE ................................................................................................ 84
A.4 Computer Runs ............................................................................................. 85
  A.4.1 Parmfile for Fluctuate Results ................................................................. 85
  A.4.2 Parmfile for Recombine Results .............................................................. 86
  A.4.3 Parmfile for Migrate Results ................................................................. 87
  A.4.5 Example of Command Line for Genetree Results ................................. 88
A.5 Glossary ......................................................................................................... 89
Bibliography ..................................................................................................... 91
1 INTRODUCTION

Coalescent theory is a method often used by population geneticists in order to make inferences about evolutionary parameters. The coalescent is a stochastic model that approximates ancestral relationships among genes. The general idea is that, under a few basic assumptions, a sample of \( n \) DNA sequences can be traced backwards through time to a point where only \( n-1 \) sequences existed. At that point, two of the original \( n \) sequences are said to have “coalesced” to one. These \( n-1 \) sequences then coalesce to \( n-2 \), then \( n-3 \), and so on, until all \( n \) sequences in the original sample can be traced back to one common ancestor, referred to as the most recent common ancestor (MRCA). Measuring backwards, the number of ancestors \( t \) time units ago is a random death process (Kingman 1982). The distribution of coalescent times (i.e., times between coalescent events), as well as the relationships among the sequences, determined by which sequences coalesce, together define a gene genealogy.

A DNA sequence consists of a string of nucleotides (nitrogenous bases attached to sugar and phosphate molecules). Throughout time, mutations accumulate along fragments of DNA. A mutation is a change in the pattern of nucleotides (also called bases or sites) that most commonly occurs during DNA replication. An extra nucleotide may be inserted somewhere along the sequence, or one may be deleted, changing the length of the sequence. A substitution is another type of mutation, where a new (different) nucleotide takes the place of an old (original) one (Li 1997).

An understanding of the coalescent pattern of a sample of sequences, along with some knowledge of the mutations that have occurred, provides information about the genetic diversity of a population, and the evolutionary forces that have acted on it. Genetic
diversity refers to the extent of variability in the sequence pattern of a particular gene, i.e., a segment of DNA (Li 1997). Evolution is a process that results in heritable changes in a population spread over many generations (see, e.g., http://www.talkorigins.org/faqs/evolution-definition.html). Processes such as migration, recombination, or natural selection are some of the forces that may result in evolutionary changes taking place. In other words, these forces change the distribution of gene genealogies (i.e., branching pattern) and times between coalescent events in predictable ways. This predictability leads to a statistical description of the variability in the sample, which in turn leads to inferences about genetic diversity in the population and evolutionary parameters such as population size, population growth rates, and migration rates (Felsenstein 1999).

Several methods of estimating genetic diversity and evolutionary parameters under the coalescent model have been developed. Our research focuses on three main groups of individuals that are currently working on coalescent estimation techniques and have developed software for that purpose. We have compared and evaluated the performance of LAMARC (Kuhner et al. 1995, 1997, 1998; Beerli and Felsenstein 1999), Genetree (Bahlo and Griffiths, 2000), and UPBLUE (Fu 1994a, 1994b). These are not commercial programs, and their performance is not well established. A comparison of available software will help coalescent researchers in choosing a program to analyze their data.
2 POPULATION GENETICS

Coalescent theory is being used for a variety of research purposes including studies in zoology and botany. Coalescent theory provides information about historical processes that have led to the observed variation in a sample. Because species loss is generally related to a loss of genetic variation or diversity, an understanding of these processes obtained through genetic studies can be useful for conservation and species management strategies.

One example is a study performed on the California Gnatcatcher (*Polioptila californica*), a threatened bird once found throughout southern California coastal regions (Zink et al. 2000). It has been a major species involved in the economic and conservation dispute over the development of sage scrub habitat. Small populations of this species inhabit these controversial areas. The relevant question is whether these small populations, or subspecies, can be considered as “evolutionarily significant units”, a classification that is often considered as units to be conserved (U.S. Departments of the Interior and U.S. Department of Commerce 1996). The issue involves the distribution of genetic diversity within the species. Zink et al. (2000) incorporated coalescent theory in their study and found that there were no significant genetic divisions within the species; hence, the preservation of the sage scrub habitat could not be linked solely to this species. For other zoological studies relating to conservation, see Barrowclough et al. (1999) and Good et al. (1997).

Similar studies have been performed where the coalescent was applied to plant population genetics. It has been used to estimate coalescent times and effective population size (i.e., number of reproducing individuals in the population; refer to Section 3.3), to test
for natural selection, and to study and compare the diversity of different plant species. For examples, see Clegg (1997), Eyre et al. (1998), Matos (1998), Matos and Schaal (2000), and Carbone (2001).

Human population studies is another field where coalescent methods are commonly used. For example, there has been much debate over the hypothesis of a common African ancestor over 200,000 years ago (summarized by Stoneking 1993). The study of genetic data, including the use of coalescent theory, has been used to investigate this claim and other questions regarding historical human population sizes and growth rates, migration patterns, etc. Current research suggests models that are more complex than this common ancestor hypothesis (see, e.g., Hammer et al. (1998) and http://www.stats.ox.ac.uk/mathgen/evolve.html).

Sherry et al. (1997) used coalescent methods to estimate human effective population sizes throughout different historical time periods. They estimate an effective population size of approximately 18,000 humans during the last one to two million years.

Migration studies are another area in which coalescent theory has been applied. Archaeological and linguistic data have typically been used to investigate migration patterns. Coalescent theory provides one way of analyzing molecular evolutionary data to augment the amount of information used in these types of research efforts. For example, there has been considerable debate concerning the origin of New World natives. Griffiths and Tavaré (1994a) suggest the use of molecular data to further investigate the number, time and structure of native migrations, leading to inferences about the colonization of the Americas. In their paper describing the application of the coalescent for ancestral inference, for instance, they perform exploratory analysis on mitochondrial sequence data...
from the Nuu-Chah-Nulth North American Indian tribe. They were able to estimate and compare substitution rates across the DNA molecule, determine the most likely ancestral lineage of the sample, and estimate the mean and standard deviation of the distribution of the time to the MRCA of the sample (14,400 ± 4,680 years).

One vast area where coalescent methods are beginning to be applied is HIV research. As we have seen, these methods are typically applied to genes in plant, animal, or bacteria populations. However, the evolutionary processes that affect these genes are similar in HIV populations, and the high mutation rate provides sufficient genetic variation to make it a prime candidate for analysis by coalescent theory. Inferences can be made regarding evolutionary parameters such as mutation rate, generation time, and recombination rate.

Recombination is the process by which segments of DNA are exchanged between different molecules (Li 1997). Recombination has the potential to generate diversity in HIV sequences (Crandall 1999), and the estimation of the recombination rate can answer questions related to antiretroviral resistance. These inferences can be used to address other research questions about the infection, such as the relationship between viral diversity and disease progression (Rodrigo and Felsenstein 1999).

Some research has also been done on the theory and application of coalescent methods to samples taken from an infected individual over time, rather than sampling from several individuals at one point in time, as is done in typical coalescent studies. This is a common sampling method in HIV studies (Rodrigo and Felsenstein 1999).
3 BASIC COALESCENT THEORY

The ancestral relationships among a sample of genetic sequences can be depicted using a coalescent tree. Coalescent trees are constructed of nodes (vertices) and branches (see Figure 1). The root of the tree is the node from which a unique path leads to any other nodes, i.e., the MRCA. The overall branching pattern is referred to as the topology.

![Coalescent Tree](image)

**Figure 1. Coalescent tree for a sample of 5 genetic sequences**

Coalescent times (i.e., times between coalescent events) are depicted in Figure 1 as segments of the coalescent tree, divided by dashed lines ($t_2$-$t_5$). The random mutations that are ultimately responsible for the variability in the sample occur along branches of the tree. A reasonable model of the distribution of coalescent times and the ancestral relationships, as well as the number of substitutions or mutations along the branches, allows for a statistical description of the variability in the sample. This can then be used to estimate parameters and make population inferences. A major objective in population genetics, in its simplest sense, is to estimate the parameter $\theta$, which is related to the amount of genetic
diversity in the population (Rodrigo and Felsenstein 1999). Coalescent theory currently provides the most efficient estimators of $\theta$ (Felsenstein 1992a).

3.1 An Illustrative Example: Coalescent Model from a Sample

Often genetic data is in the format of several DNA sequences, coded according to their nucleotides ($A =$ adenine, $C =$ cytosine, $G =$ guanine, $T =$ thymine). The nucleotides are sometimes referred to as bases or sites. Consider the example of a sample of 7 sequences (labeled A through G) of 40 sites, shown in Table 1. This example follows the tree construction method of Griffiths and Tavaré (1995).

Note that there is variability among the sequences at 10 of the sites (shaded). In the population genetics literature, these are referred to as segregating sites (Li 1997), or as single nucleotide polymorphisms (SNPs) (Kuhner et al. 2000a). In one mutation model (known as the infinitely-many-sites model), only information within the segregating sites is used for the reconstruction of the ancestry of the sequences because variability in the data occurs only at these sites (Fu 1994a). Other models use information from the entire sequence (see, e.g., Kuhner et al. 1995). Using the former approach, we include the ten segregating sites from Table 1 in Table 2 below.

The pattern of segregating sites provides information about the mutations that have occurred in the history of the sample (Griffiths and Tavaré 1994a). A difference of nucleotides at a particular site indicates that a mutation has occurred sometime in the past. Thus, we can think of the sites as representing mutations.
Table 1. Seven sequences of 40 sites

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|

Table 2. Segregating sites

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>E</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>F</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>
The relationships among the sequences can be depicted using an unrooted genealogical tree based on the segregating sites (Griffiths and Tavaré 1994a). In this type of tree, the vertices represent sequences, with the mutations represented as dots along the edges. For example, the difference between sequences A and C occurs only at Site 1. In the unrooted genealogical tree in Figure 2, the dot representing Site 1 is placed between the circles representing sequences A and C. The differences between sequences B and E occur at Sites 4 and 7. We depict this relationship similarly in the unrooted tree. In Figure 2, we include relationships between sequences (labeled A-G) for all ten sites.

A rooted genealogical tree can be constructed from the unrooted tree by placing the root either at one of the sequences or between sites along the branches. For example, two possible rooted trees from the unrooted tree in Figure 2 are given in Figure 3. Note the difference in the shape of the tree when the root is changed from being between sites 2 and 6 to being placed at sequence E. If there are \( s \) segregating sites, then \( s+1 \) rooted trees are possible.

Figure 2. Unrooted genealogical tree
A coalescent tree is created when a time scale is imposed on a rooted genealogical tree. Even with the same branching topology, different time distributions produce different coalescent trees. Thus, there are obviously an infinite number of trees possible. An example of the same branching pattern with different time distributions is provided in Figure 4. The different distributions of coalescent times could have resulted from different evolutionary processes.

The set of all possible coalescent trees essentially provides a sampling distribution from which to estimate genetic diversity and evolutionary parameters.
3.2 The Parameter $\theta$

The parameter $\theta$ plays an important role in population genetics, as it appears in several theoretical formulas and equations (Li 1997). It is related to the accumulation of genetic variation in a population. For simplicity, we define

$$\theta = 2\kappa N \mu$$  \hspace{1cm} (1)

where $N$ is the effective population size (see Section 3.3), $\mu$ is the mutation rate, and

$$\kappa = \begin{cases} 1 & \text{for a haploid population} \\ 2 & \text{for a diploid population} \end{cases}$$

The relationship among the parameters in equation (1) is based upon the fact that populations with the same effective population size and mutation rate are defined to have the same expected level of diversity (Rodrigo and Felsenstein 1999). (Note that population geneticists don’t use the simplified form in equation (1). Rather, depending on the problem at hand, they use either $\theta = 2N\mu$ or $4N\mu$.)

Generally, neither $N$ nor $\mu$ can be estimated separately. However, if one is assumed to be known and $\theta$ can be estimated using coalescent methods, then the other can be inferred. The parameter $\mu$ most often refers to the mutation rate per gene per generation ($\mu_g$); however, it may also equivalently refer to mutation rate per locus per generation or per sequence per generation. In some instances, $\mu$ is measured in terms of mutations per site per generation ($\mu_s$) (Kuhner et al. 1995). A simple conversion can then be made for comparison purposes by multiplying $\mu_s$ by the sequence length (Tavaré et al. 1997).
3.3 The Wright-Fisher Model

The Wright-Fisher model (Wright 1931, Fisher 1930), a well-known description of a theoretical population, provides a framework for the development of simple coalescent theory (Kingman 1982). In its simplest version, the assumptions of the Wright-Fisher model are:

- Finite and constant population size $N$
- Non-overlapping generations
- Random mating
- No migration
- No recombination
- Selective neutrality.

The first assumption is a finite and constant population size of $N$ individuals. If the population is diploid (i.e., two copies of genetic material in the cells, such as plants and animals) there are really $2N$ copies of DNA under consideration. If the individuals are haploid (i.e., only one copy of genetic material in the cells, such as egg, sperm and pollen cells), $N$ stands alone as the population size. The only distinction needed is the appropriate designation of $\kappa$ as 1 or 2 for haploid or diploid, respectively.

The second assumption of the Wright-Fisher model is non-overlapping generations, which simply means that for each generation, only offspring of the preceding generation are present in the population (Rodrigo and Felsenstein 1999). The third assumption, random mating, implies no preference in mate selection with regard to ancestry or genotype (i.e., composition of alleles) (Levin 2000). Migration is a process in which populations exchange a certain proportion of individuals each generation (Rodrigo and Felsenstein 1999). As defined earlier, recombination is the process by which segments of DNA are exchanged between different molecules (Li 1997).
Finally, the assumption of selective neutrality states that natural selection has not acted on the population. Natural selection is an evolutionary process caused by varying degrees of reproductive success among individuals in the population. Genetic variation among individuals is a necessary precursor for natural selection. Those character traits that inhibit reproductive success are selected against (i.e., individuals possessing them are less likely to reproduce), thus the allele frequency that defines that trait decreases from generation to generation. This is called negative, or purifying, selection. Positive selection occurs when a mutant allele has a selective advantage and individuals possessing it are more likely to reproduce (Li 1997).

Natural selection is not the only process that leads to changes in allele frequencies over time (i.e., evolution). Random genetic drift is another such process, where the changes are due solely to chance effects, independent of genetic fitness. The neutral theory of molecular evolution, often used in population genetics, states that natural selection plays a minimal role in evolutionary changes relative to mutation and genetic drift (Clegg 1997). Several statistics have been developed to test the hypothesis of selective neutrality (summarized in Li 1997 and Rodrigo and Felsenstein 1999).

A final note needs to be made regarding the “effective population size”, $N_e$. It was originally developed by Wright (1938) to account for the discrepancies between the observed population size and the actual reproducing population size due to violations of the aforementioned assumptions. For example, some individuals in a population may be in their pre- or post-reproductive stages of life (overlapping generations), or situations exist where the number of reproducing males is not equal to the number of reproducing females, such as polygamous populations or those with a non-reproducing caste. Also, factors such
as environmental catastrophes or cyclical reproduction violate the assumption of a constant population size (Li 1997). \( N_e \) is generally smaller than \( N \), and is defined to more accurately represent the reproducing population size. Essentially, a population with effective size \( N_e \) experiences the same intensity of genetic drift as does an ideal population of the same size, and therefore acquires or loses diversity at the same rate (Beerli et al. 2001). The parameter \( \theta \), then, is actually defined as

\[
\theta = 2 \kappa N_e \mu.
\]

For the purposes of simplicity in this paper, however, \( N \) will be used to represent the effective population size, \( N_e \).

3.4 Kingman’s Coalescent

Kingman (1982) developed the coalescent as a Markov chain that can be used to approximate the ancestral structure of \( n \) individuals from a population. It was based upon the Wright-Fisher model, with a few modifications.

The coalescent can be applied to a wide variety of selectively neutral reproduction models, only one of which is the Wright-Fisher model (Griffiths and Tavaré 1997). According to Griffiths and Tavaré (1994a), reproductive neutrality is imposed by requiring offspring to be exchangeable in probability (see Koch 1982) and identical in different generations. Kingman (1982) showed that all exchangeable reproductive models can be approximated by the coalescent. Consequently, the assumption can be rephrased such that every individual must have the same opportunity or propensity to produce offspring, independently for each offspring (Rodrigo and Felsenstein 1999).

Recent research suggests that even the assumption of selective neutrality may also be relaxed. It appears that the genealogical trees used to create the coalescent model
remain close to those expected under neutral models of evolution, even when there is evidence of moderate levels of purifying selection. Research also suggests that even strong purifying selection may have a minimal influence on results based on coalescent theory (Golding 1997, Neuhauser and Krone 1997). How the coalescent model is affected by positive selection is not yet known (Rodrigo and Felsenstein 1999). Beerli et al. (2001) recommend that a test for selection be performed prior to applying the methods based on coalescent theory.

The other assumptions of the Wright-Fisher model, as noted above, are idealized and rarely found in nature. Departures from these assumptions provide insights into the evolutionary forces that “leave an indelible mark on the genealogy of genes sampled from the population” (Rodrigo and Felsenstein 1999, p. 240). The population geneticist’s interest lies not in the trees themselves, but in the parameters that generate the trees (Felsenstein 1999). Thus, methods have been developed that model the coalescent when certain assumptions are violated. This allows for the estimation of migration rates, recombination rates, or population growth rates, in addition to the mutation rate or effective population size which are incorporated into \( \theta \). In Table 3, we reference models that have been developed under conditions that violate basic Wright-Fisher model assumptions.

<table>
<thead>
<tr>
<th>Varying Population Size</th>
<th>Migration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuhner, Yamato, and Felsenstein 1998</td>
<td>Nath and Griffiths 1993</td>
</tr>
<tr>
<td></td>
<td>Beerli and Felsenstein 1999, 2001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recombination</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hudson and Kaplan 1985, 1988</td>
<td>Kaplan, Darden and Hudson 1988</td>
</tr>
<tr>
<td>Griffiths and Marjoram 1996</td>
<td>Krone and Neuhauser 1997</td>
</tr>
<tr>
<td>Hey and Wakeley 1997</td>
<td></td>
</tr>
</tbody>
</table>
4 THE STATISTICAL MODEL

The coalescent may be thought of as a random tree with respect to both the time scale and the topology. The topology, which determines who is related to whom is obtained by randomly merging pairs of individuals (Griffiths and Tavaré 1994a). Recall that \( t_i \) is the amount of time it takes for \( i \) sequences to coalesce to \( i - 1 \) (see Figure 1). Some authors (e.g., Rodrigo and Felsenstein 1999) measure time in generations, while others including Kingman (1982) and Tavaré et al. (1997), measure time in units of \( \kappa N \) generations.

4.1 Coalescent Times

The distribution of the coalescent times, \( t_i \), may be derived in two equivalent manners. Kingman (1982), the father of coalescent theory, used a stochastic process approach to derive the distribution of \( t_i \). Following Rodrigo and Felsenstein (1999), we appeal to basic probability arguments in our derivation.

Consider a sample of two from a population of \( \kappa N \) individuals (recall that \( \kappa = 1 \) for haploid cells and \( \kappa = 2 \) for diploid). The number of ways two individuals can arise from unique ancestors in the previous generation is \( \kappa N(\kappa N - 1) \). The number of ways two individuals can be selected from the previous generation is \( (\kappa N)^2 \) (i.e., sampling with replacement from a finite population). Therefore, the probability \( P(1) \) that two individuals are derived from different ancestors one generation in the past (where time is measured in discrete generations) is

\[
P(1) = \frac{\kappa N(\kappa N - 1)}{(\kappa N)^2} = 1 - \frac{1}{\kappa N}.
\]
Similarly, the probability that \( i \) individuals are descendants of \( i \) different individuals one generation in the past is

\[
P(1) = \prod_{j=1}^{i-1} \left(1 - \frac{j}{\kappa N}\right) \approx 1 - \frac{i(i-1)}{2\kappa N}.
\]  

(2)

The probability of a coalescent event \( t_i \) generations ago for \( i \) individuals (where \( t_i = 0 \) at the time of sampling) is the probability that no coalescent events occur in the first \( t_i - 1 \) generations, and that one coalescent event occurs in generation \( t_i \):

\[
P(t_i) = \left[1 - P(1)\right]P(1)^{t_i-1}.
\]

For discrete time, the \( t_i \) are thus distributed as geometric random variables.

Generalizing to continuous time, the distribution of the \( t_i \) can be approximated by an exponential distribution:

\[
P(t_i) \approx \frac{i(i-1)}{2\kappa N} \exp\left(-\frac{i(i-1)}{2\kappa N} t_i\right).
\]

(3)

(Notice that in the Kingman (1982) or Tavaré et al. (1997) derivation, \( 2\kappa N \) in equations (1) and (2) is replaced by the number 2.)

The expected value and the variance of the time to the MRCA, \( T_{\text{MRCA}} \), is easily found under this time scale (Rodrigo and Felsenstein 1999):

\[
E(T_{\text{MRCA}}) = \sum_{i=2}^{n} E(t_i) = \sum_{i=2}^{n} \frac{2\kappa N}{i(i-1)}
\]

\[
= 2\kappa N \left(1 - \frac{1}{n}\right),
\]

(4)

\[
V(T_{\text{MRCA}}) = (2\kappa N)^2 \sum_{i=2}^{n} \frac{1}{i(i-1)^2}.
\]
4.2 Mutations

Mutations are included in the coalescent model under the assumption of a *molecular clock*. This assumption states that the rate of molecular evolution for a specific locus (gene) is approximately constant in all evolutionary lineages (Li 1997). However, when more than one gene is considered, the mutation rate $\mu$ is allowed to vary across loci. This is currently modeled with a one-parameter gamma distribution (Beerli and Felsenstein 1999).

In coalescent theory, it is assumed that mutations accumulate on the branches of a genealogy according to some evolutionary model with a mean of $\mu t$, where $t$ represents the branch length in time. The most commonly used model to describe the molecular clock assumption is the Poisson process (Takahata 1991). However, this has been called into question by some who note the presence of overdispersion. Other models have been suggested, including a fractal-renewal process, a Lévy-stable process, a fractional-difference process, and a log-Brownian process (Takahata 1991).

There are also various ways to model the effects of mutations. The simplest is called the infinitely-many-sites model (see Watterson 1975). This model assumes that each mutation occurs at a new and different site of the sequence. It is an approximation to the evolution of a sample of sequences of finite length, when it can be assumed that recurrent mutations at a site are rare (Griffiths and Tavaré 1994a). Under this model, the mutation rate is typically $\mu_g$ (per gene) (Kuhner et al. 1995; see the discussion in Section 1.3.2). The finite-sites model allows for recurrent mutations at the same site, and the mutation rate is typically $\mu_s$ (per site). A third model is the infinite-allele model, which assumes that each mutation creates a new allele not currently existing in the population (Li 1997).
5 ESTIMATION METHODS

5.1 History

Several different methods have been developed to measure genetic diversity. Two summary statistics used to estimate $\theta$ that are commonly used and relatively easy to compute are Watterson’s (1975) estimator and Tajima’s (1983) estimator. Watterson’s estimator uses the number of segregating sites in a sample, denoted $K$. Recall that a segregating site is one that shows variation from other sequences in the sample; so at least one sequence has a nucleotide that is different from those in other sequences at that site (see Section 3.1).

Tajima’s estimator uses pairwise genetic distances (1983). The genetic distance between two genes is defined as the number of substitutions or mutations that have accumulated independently in each gene since the time of their divergence (Rodrigo and Felsenstein 1999). It is measured by the number of nucleotide differences between the pair (Li 1997).

Both of these estimators are based solely on information obtained from the observations at the time the sample was taken. The estimates are simple to compute, but the variances are large (Fu 1994a). In addition, the pairwise genetic distances in Tajima’s estimator are not independent, and the overall average genetic distance is dependant on the genealogy of the sample (Felsenstein 1992a). It became apparent that an even more efficient estimate of $\theta$ could be obtained by including information from genealogical relationships among the sequences in a sample.

Estimates that incorporate genealogical relationships are called phylogenetic estimates. Felsenstein (1992a) showed the inefficiency of Watterson’s and Tajima’s
estimators compared to phylogenetic estimators. Fu and Li (1993) studied the maximum amount of gain in efficiency that could be obtained by incorporating ancestral relationships into the estimate of $\theta$. They show that Watterson’s estimator is an asymptotically optimal estimate, but substantial improvement is possible for finite sequences, particularly with large values of $\theta$. They derive a lower bound for the variances of estimators using maximum likelihood methods. Many current researchers doing work in coalescent theory are essentially searching for this optimum, phylogenetic estimator (Li 1997).

5.2 Major Research Groups

Several methods of estimating genetic diversity and evolutionary parameters under the coalescent model have been developed. Our research focuses on three main groups of individuals that are currently working on coalescent estimation techniques and have written computer programs for that purpose. The three groups vary in certain aspects of the development of their model and estimation techniques. For an overview, refer to Table 4.

The major players of the first group (Felsenstein’s group) are Felsenstein, Kuhner, Yamato and Beerli (see Felsenstein (1981, 1992a, 1992b, 1999), Kuhner et al. (1995, 1997, 1998, 2000a, 2000b), Beerli and Felsenstein (1999) and Beerli et al. (2001)). The second group (Griffiths’ group) includes Griffiths, Tavaré, and their collaborators (see Griffiths and Tavaré (1993, 1994a, 1994b, 1995, 1997), Griffiths and Marjoram (1996), Tavaré et al. (1997), and Nath and Griffiths (1993, 1996)). These groups each use a maximum likelihood approach, which is evaluated by Markov Chain Monte Carlo methods. The third group (Fu’s group) uses a recursive least-squares approach that is claimed to be computationally simpler than those used by Felsenstein’s or Griffiths’ groups (Li 1997).
This group includes Fu, Li and Vasco (see Fu and Li (1993), Fu (1994a, 1994b), Li (1997), and Vasco et al. (2001)).

Table 4. Model differences between three main groups of researchers

<table>
<thead>
<tr>
<th></th>
<th>Felsenstein LAMARC</th>
<th>Griffiths Genetree</th>
<th>Fu UPBLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach</strong></td>
<td>Likelihood</td>
<td>Likelihood</td>
<td>UPBLUE</td>
</tr>
<tr>
<td><strong>Technique</strong></td>
<td>Metropolis-Hastings</td>
<td>MCMC</td>
<td>Least Squares</td>
</tr>
<tr>
<td><strong>Time Scale</strong></td>
<td>No. of generations</td>
<td>$\kappa N$ generations</td>
<td>No. of generations</td>
</tr>
<tr>
<td><strong>Mutation Model</strong></td>
<td>Not specified</td>
<td>Poisson process</td>
<td>Poisson process</td>
</tr>
<tr>
<td><strong>Mutation Effects</strong></td>
<td>Finite-sites</td>
<td>Infinitely-many-sites</td>
<td>Infinitely-many-sites</td>
</tr>
<tr>
<td><strong>Mutation Rate $\mu$</strong></td>
<td>Site/generation, $\mu_s$</td>
<td>Gene/generation, $\mu_g$</td>
<td>Gene/generation, $\mu_g$</td>
</tr>
</tbody>
</table>

Felsenstein’s group estimates $\theta$ and other evolutionary parameters using the likelihood approach (Rodrigo and Felsenstein 1999). In their distributinal assumptions, time is measured in terms of the number of generations. They assume a molecular clock but do not explicitly state which mutation model they employ. Rather, they require only that the expected number of mutations or substitutions that occur during a time period defined by $t$ generations in the past is $\mu t$ (Rodrigo and Felsenstein 1999). While the Poisson process is typically assumed, it is not clear whether the model used by Felsenstein’s group is Poisson or more general, e.g., also allowing for overdispersion. Felsenstein’s group assumes the finite-sites model for mutations (Beerli and Felsenstein 1999), with $\mu_s$ as the mutation rate (Kuhner et al. 1995). They use a Metropolis-Hastings algorithm to evaluate the likelihood and estimate parameters. In addition to the parameter $\theta$, they have developed models that include an exponential population growth rate, $g$,
recombination rate, $r$, and migration rate, $m$. The software package they have produced, LAMARC, estimates each of these parameters.

Griffiths’ group also uses a likelihood approach. In the development of their model, however, time is measured in terms of $\kappa N$ generations (Tavaré et al. 1997). Griffiths’ group assumes the Poisson process for accumulation of mutations. Most of their work deals with the infinitely-many-sites model for the effects of mutations, but they have also developed methods that assume the infinite-allele model (Griffiths and Tavaré 1994a, 1994b). They follow the convention of using $\mu_g$ when the infinitely-many-sites model is assumed (Kuhner et al. 1995). Griffiths’ group has also developed models that include the population growth rate and migration rate, and use an MCMC approach to estimate these parameters. Their program is called Genetree.

Fu’s group uses a least squares approach to estimate $\theta$. They measure time in terms of the number of generations (Vasco et al. 2001), and assume the infinitely-many-sites model for mutations, with $\mu_g$ as the mutation rate (Fu 1994a). They also assume a Poisson process for mutations. Fu’s group has developed a model that allows recombination and migration in the population, but their current software only estimates $\theta$ (Fu 1994b). They call this method UPBLUE because they use an UPGMA tree (unweighted pair-group method with arithmetic mean), and produce the best linear unbiased estimator of $\theta$.

These three groups have expanded upon and referenced earlier work done by a fourth group that includes Hudson and Kaplan (see Hudson (1982, 1987, 1990, 1993), Hudson and Kaplan (1985, 1988), and Kaplan et al. (1988)). This group had extended the coalescent model to include selection and recombination. We have not been successful in finding any software programs developed by them.
5.3 Program Comparison

Table 5 contains the extant programs used to estimate coalescent parameters. Note that the first 7 programs (through EVE) are those that can be classified into the three research groups aforementioned. The remaining coalescent programs were developed independently. At this point we have been unable to obtain these miscellaneous programs. Table 5 also includes the parameters estimated by the programs and the evolutionary forces included in the model used for estimation. Table 6 contains an explanation of the symbols used.

A more expanded version of Table 5, including additional comments and information (e.g., software version and source), is labeled Table A.1 in the Appendix. Note that there are other software programs that implement coalescent theory, but do not estimate any parameters (e.g., simulation programs). We have included these in Table A.2.
Table 5. Program comparison of function and evolutionary forces

<table>
<thead>
<tr>
<th>Group</th>
<th>Program</th>
<th>Estimation / Function</th>
<th>Evolutionary Forces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felsenstein</td>
<td>LAMARC Fluctuate</td>
<td>$\theta = 2 \kappa N \mu_s$ Growth Rate ($g = l / \mu_s$)</td>
<td>Exponential growth</td>
</tr>
<tr>
<td></td>
<td>LAMARC Recombine</td>
<td>$\theta = 2 \kappa N \mu_s$ Recombination Rate ($r = \rho / \mu_s$)</td>
<td>Recombination</td>
</tr>
<tr>
<td></td>
<td>LAMARC Migrate</td>
<td>$\theta_i = 2 \kappa N_i \mu_s$ ($i^{th}$ population) Migration Rate ($4N_i m_{ij}$)</td>
<td>Migration</td>
</tr>
<tr>
<td></td>
<td>LAMARC Lamar</td>
<td>$\theta = 2 \kappa N \mu_s$ Recombination Rate ($r$) Migration Rate ($M_{ij} = m_{ij} / \mu_s$)</td>
<td>Recombination Migration</td>
</tr>
<tr>
<td>Griffiths</td>
<td>Genetree</td>
<td>$\theta = 2 \kappa N \mu_s$ Migration Rate ($2N m_{ij}$)</td>
<td>Exponential growth Migration</td>
</tr>
<tr>
<td></td>
<td>UPBLUE (Fu and Li)</td>
<td>$\theta = 2 \kappa N \mu_s$</td>
<td>Exponential growth</td>
</tr>
<tr>
<td></td>
<td>Eve (Vasco)</td>
<td>$\theta = 2 \kappa N \mu_s$ Growth rate ($r$)</td>
<td>Exponential growth</td>
</tr>
<tr>
<td>Miscellaneou</td>
<td>Sweep_bott (Galtier)</td>
<td>Detects bottlenecks and selective-sweeps</td>
<td>Bottlenecks$^a$ Selective sweeps$^b$</td>
</tr>
<tr>
<td></td>
<td>SIMCOAL (Excoffier, Novembre, Schneider)</td>
<td>Simulates molecular data in interconnected populations with arbitrary demography</td>
<td>Variable population size Migration</td>
</tr>
<tr>
<td></td>
<td>MICSAT (Wilson and Balding)</td>
<td>MCMC sampling for microsatellite loci</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BATWING (Wilson and Balding)</td>
<td>$\theta = 2 \kappa N \mu_s$ Exp. growth rate TMRCA</td>
<td>Exponential growth Subdivision</td>
</tr>
<tr>
<td></td>
<td>PAL (Drummer, Buckland, Skinner)</td>
<td>Simulates coalescence intervals and estimation of demographic parameters</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Explanation of symbols for Table 5

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_s$</td>
<td>Mutation rate per site per generation</td>
</tr>
<tr>
<td>$\mu_g$</td>
<td>Mutation rate per gene per generation</td>
</tr>
<tr>
<td>$l$</td>
<td>Exponential growth rate</td>
</tr>
<tr>
<td>$g$</td>
<td>Scaled growth rate, $\theta = \theta_e e^{gt}$, the value of $\theta$ at time $t$, given its value now ($\theta_e$)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Recombination rate per inter-site link per generation</td>
</tr>
<tr>
<td>$r$</td>
<td>Ratio of per-site chance of recombination to per-site chance of mutation</td>
</tr>
<tr>
<td>$4N_i m_{ij}$</td>
<td>Immigration from population $j$ into population $i$</td>
</tr>
<tr>
<td>$M_{ij} = 4N_i m_{ij} / \theta$</td>
<td>Number of migrants moving into the $i^{th}$ population per unit time</td>
</tr>
<tr>
<td>$T_{MRCA}$</td>
<td>Time to the most recent common ancestor</td>
</tr>
</tbody>
</table>

$^a$Sudden reduction in population size

$^b$Rapid fixation of allele through directional selection
The goal of our research was to evaluate and compare the software programs currently available that perform coalescent estimation. Specifically, we compared the programs in the LAMARC package (Fluctuate, Recombine, Migrate, and Lamarc), developed by Felsenstein’s group, Genetree, written by Griffiths, and one program written by Fu’s group, called UPBLUE. (EVE, also developed by Fu’s group, is currently unavailable). We compared the programs in general areas of availability, usability, and results. The rest of this chapter follows the outline form:

6.1 Description of datasets  
6.2 Platform availability  
6.3 Documentation  
6.4 Data input format  
6.5 Running time  
6.6 Output  
6.7 Ease of use  
6.8 Stability

A discussion of each of the points of comparison will be found in its respective section in this chapter. For a detailed evaluation of program results see chapter 7.

Six other programs were included in Table 5 that relate to coalescent theory (Sweep_bott, Eve, SIMCOAL, MICSTAT, BATWING, and PAL). These miscellaneous programs were either unavailable or did not perform the same functions and were not comparable.

6.1 Description of Datasets

Ideally the same dataset would be used to compare each of these programs. However, we discovered that differences in the models and assumptions used by the programs prevented us from using only one dataset, although we tried to be as consistent as
possible. For example, Migrate estimates migration rates between two or more subpopulations, but assumes no recombination. A dataset that is appropriate for this program would not be appropriate for Recombine for instance, which expects a dataset with recombination. Furthermore, two different models of the effects of mutations were used by the authors of these three packages. Felsenstein’s group uses the finite-sites model while Griffiths’ and Fu’s groups use the more restrictive infinitely-many-sites model. Data that naturally fit the infinitely-many-sites model (e.g. long sequences with few segregating sites) and are therefore appropriate for Genetree and UPBLUE are also appropriate for the LAMARC programs. Data that show evidence of recurrent mutations at a given site do not fit the infinitely-many-sites model. These datasets may be analyzed by the LAMARC programs, but are not suitable for Genetree and UPBLUE. Sometimes data may be adjusted to fit the more restrictive model (e.g. Griffiths and Tavaré 1994a, 1994b), but in general it is not a natural adaptation and may be difficult to do.

To compare these package we used three datasets: the Nuu-Chah-Nulth data, the Africa data, and simulated datasets. The datasets and simulation programs and parameters are in Appendix A.2.

6.1.1 Nuu-Chah-Nulth Data

This dataset consists of mtDNA (mitochondrial DNA) from 63 individuals from the Nuu-Chah-Nulth North American Indian tribe (Ward et al. 1991). There are 28 unique sequences, where each sequence consists of 360 base pairs (sites). For the data to fit the infinitely-many-sites model, we removed 8 sites and 4 of the 28 unique sequences because of recurrent mutations. Ten of the unique sequences became duplicates after these removals, and were therefore deleted. The resulting dataset that corresponds with the
infinitely-many-sites model consists of 55 original sequences (14 unique sequences) with 352 sites.

6.1.2 Africa Data

This dataset consists of 22 DNA sequences from the Biaka Pygmy population from the Central African Republic (CAR) and 42 sequences from the Luo population of Kenya (KEN). They are part of a larger dataset from several populations that has been analyzed extensively by Harding et al. (1997). For comparing these programs we chose to use only these two populations. The sequences consist of 2670 base pairs encompassing the $\beta$-globin gene. The segregating sites fit the infinitely-many-sites model without any adjustments, and we used only these sites.

6.1.3 Simulated Datasets

We simulated several different datasets for comparing some of the programs. The first group was created using the Hudson Simulator (available from http://www.daimi.au.dk/~compbio/hudson/hudson.html), which simulates nucleotides evolving under the coalescent model (Hudson 1983). The recombination rate $R$ is scaled as $4N\rho$, where $\rho$ is the recombination rate for a single sequence per generation. The first simulation had recombination rate $R = 0.1$ and the second was simulated with $R = 1.0$. Both datasets consist of 40 sequences of 750 base pairs. These datasets were analyzed by the Recombine program. Only one dataset for each value of $R$ was simulated because the random seed generator could not be changed. (Note: we have changed the symbols from those used by the Hudson Simulator in order to compare symbols later. Originally, $R = \rho$ and $\rho = r$).
The second group of datasets was generated using SIMCOAL, a coalescent simulation program for haploid molecular data in interconnected populations (http://cmpg.unibe.ch/software/simcoal/). (Written by Excoffier, L., Computational and Molecular Population Genetics Lab, University of Bern, 12/11/2001). Each of the 10 datasets in this group consists of 2 populations, 20 sequences in each with 500 base pairs. The mutation rate per generation was specified to be 0.0005 and the population size for both populations was 20,000 (haploid genes). See Appendix A.2.3 for other parameter specifications. The migration matrix used for the 10 SIMCOAL datasets was:

\[
\begin{pmatrix}
0.0 & 0.0005 \\
0.0005 & 0.0
\end{pmatrix}
\]

This matrix represents symmetric migration, so the migration rate from population 1 to population 2 is 0.00005, which is the same as the migration rate from population 2 to population 1. Entries in the matrix are the migration rates per gene per generation backwards in time.

6.2 Platform Availability

The LAMARC package is quite versatile in terms of platform availability (see Table 7). The source code (written in C or C++) is available for all four programs, and claims to compile on most workstations. The programs are also available as executables for MS Windows, Apple Macintosh, and some Unix operating systems. We have downloaded and successfully run the Windows executable versions for each of the programs in the LAMARC package.
The Genetree program is not quite as flexible in terms of platform. The source code (written in C) is available and claims to compile under gcc, and a precompiled version is also available for Windows. No executables are available for Unix or Mac.

The UPBLUE program was written in FORTRAN code. It is available for users to compile themselves, and also runs simply by entering the location of the dataset (in the right format) directly onto the website http://www.hgc.sph.uth.tmc.edu/cgi-bin/upblue.pl. Access to the website, however, is not consistently available.

<table>
<thead>
<tr>
<th>Source</th>
<th>Unix</th>
<th>Windows</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluctuate</td>
<td>C</td>
<td>B</td>
<td>B (multiple)</td>
</tr>
<tr>
<td>Recombine</td>
<td>C</td>
<td>B (multiple)</td>
<td>B (multiple)</td>
</tr>
<tr>
<td>Migrate</td>
<td>C++</td>
<td>B</td>
<td>B (multiple)</td>
</tr>
<tr>
<td>Lamarc</td>
<td>C++</td>
<td>B</td>
<td>B (multiple)</td>
</tr>
<tr>
<td>Genetree</td>
<td>C</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>UPBLUE</td>
<td>Fortran</td>
<td>Access through the internet</td>
<td></td>
</tr>
</tbody>
</table>

B indicates pre-compiled binaries available

6.3 Documentation

In general, the documentation for the programs within the LAMARC package are thorough and quite readable. The manuals for Fluctuate (9 pages) and Recombine (17 pages) are similar in content and format, even word for word in some places. They include a table of contents, a fairly complete description of the input files, output files and menu options, and some examples. The manual for Migrate is even more complete (61 pages), including graphics, more examples, FAQs, errors and warnings, troubleshooting, and a more complete description of the output. It is also available as an HTML document. The help source for Lamarc is only available as an HTML document. It is similar in content to Migrate, although it could use more examples. In general the documentation for these
programs were suitable for application, although there are some exceptions (see e.g. 7.1.3 Recombine).

Examples of input and output files are included in the distribution for Fluctuate, Migrate, and Lamarc. The distribution for the current version of Recombine (v. 1.4.1, Aug 2002) does not include any example files, although previous versions did include them.

The manual for Genetree (25 pages) includes the basic elements listed before, but consistently provides more examples for each section. It also contains a section on hypothesis testing which aids the user somewhat with application and interpretation. Another section included in the document is on graphical output. Although the manual is very readable, it is not thorough enough for the complex strategy of estimating parameters. No example files are included in the distribution.

The help source for UPBLUE consists of a few comments within the FORTRAN code. They explain changes the user must make to the program depending on sample size, the memory required for different sample sizes, the code to initiate the program, and a brief description and example on the data input format. Overall, it is not very descriptive, nor is it smooth reading. It seems sufficient to run the program however. An example input file is included in the distribution.

Besides Migrate, no information is provided on citation for the documentation for these programs. Hereafter we will reference them as Fluctuate doc. 1998 (v. 1.3), Recombine doc. 2000 (v. 1.40 alpha), Lamarc doc. 2002, and Genetree doc. 2002. The reference for Migrate is Beerli (1997-2002). All are available from the websites listed in Table A.1 in the Appendix.
6.4 Data Input Format

Our main function of interest in these programs is coalescent estimation using sequence data (DNA/RNA). For some programs, the sequence data may be in the form of ordered single nucleotide polymorphisms (SNPs). Each of the programs under comparison analyzes sequence data, but some of them also analyze other types of data, including ordered enzyme electrophoretic data and microsatellite data. See Table 8 for a comparison among the programs. We will only compare the data input format for the programs with respect to sequence or SNP data. See Appendix A.3 for examples for each program.

<table>
<thead>
<tr>
<th>Fluctuate</th>
<th>Recombine</th>
<th>Migrate</th>
<th>Lamarc</th>
<th>Genetree</th>
<th>UPBLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>Sequence</td>
<td>Sequence</td>
<td>Sequence</td>
<td>SNPs</td>
<td>Pairwise nucleotide differences</td>
</tr>
<tr>
<td>SNPs</td>
<td>SNPs</td>
<td>SNPs</td>
<td>SNPs</td>
<td>Microsatellite</td>
<td>(Sequence)</td>
</tr>
<tr>
<td>Electrophoretic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsatellite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The format for sequence data to be used by the programs in the LAMARC package must be fully aligned and in one of two formats: interleaved (first line of all sequences, second line of all sequences, etc.) or sequential (all of sequence 1, all of sequence 2, etc.). The sequence for each individual is listed with a corresponding name (regardless of duplicates). The format is very similar among Fluctuate, Recombine and Migrate, with only slight changes in the content or position of header lines (numbers of populations, loci, sequences and sites, titles and subtitles). The data format required for the Lamarc program is called XML format, which is characterized by <tags> that identify pieces of information. A converter program is included in the distribution that converts Phylip, Recombine, and Migrate files to the XML format.
Recombine, Migrate, and Lamarc also take sequence data in the form of SNPs. The format is the same as full sequence data, with the invariable sites removed.

Genetree only accepts SNPs and requires the user to indicate at each polymorphic (segregating) site which bases are ancestral and which are mutant. This specification corresponds to a specific rooted genealogical tree (Griffiths and Tavaré 1995). Results may be based on this specific genealogy, or an unrooted genealogical tree where the exact knowledge of ancestry is not required. Also, each allele is listed only once along with its frequency. This varies from the LAMARC programs, which list the data for each individual, regardless of duplicates.

The data for Genetree may be in one of three formats. The first format lists just the polymorphic sites of only the unique sequences, along with the frequency of occurrence. An accompanying file must indicate ancestor bases at each site. In the second format the data are listed as “binary sequences”. In this form the ancestors are replaced with zeros and the mutants are replaced with ones. Finally, in the last format the data are represented as mutation paths forming a gene tree, as described by Griffiths (1987, 1994b, 1995). The program actually acts on the mutation paths, and a conversion program is included in the distribution to change the data in the other two formats into gene trees. The conversion program also reports on the inconsistencies with regards to the infinitely-many-sites model.

The data input format for UPBLUE is different from LAMARC and Genetree in that the nucleotides are not even needed. The first line simply lists the number of sequences. The following lines list the number of pairwise nucleotide differences between sequences. The first two numbers on each line indicate which sequences, and the last element indicates the number of nucleotide differences between those sequences. For
example, 1 4 1.0 means there is one nucleotide difference between sequence 1 and 4. Given
the sequence data, there are programs that will calculate nucleotide differences to achieve
this form (e.g. PAUP: http://paup.csit.fsu.edu/). The input file for UPBLUE must be in
plain text format.

The input format is different for each of these programs; there is no consistent or
standard form. The format of the researcher’s data will determine the ranking of these
programs with regard to ease of data input.

6.5 Running Time

There are several factors that determine the running time of these programs. For
LAMARC programs, the most obvious factors are the number and length of “chains” used
to search the parameter space. Other search strategies such as “heating” (in Recombine,
Migrate and Lamarc) and “replication” (in Lamarc) also increase running time. (See 7.1.1
Search Strategy for further discussion on these options). Other options for Migrate that
increase running time include writing genealogies to an outfile, evaluating likelihood
profiles at percentiles, and allowing the mutation rate to vary over loci. An option in
Lamarc allows using genotype data rather than haplotype data, which also increases the
running time. All of these programs also allow different categories of substitution rates,
but the run time increases proportionally to the number of rate categories. Run time also
increases with the number of sequences and the number of sites (less than linearly).

The LAMARC programs require initial parameter values to start the simulations.
For Fluctuate, run time is longer when high negative values of the initial growth rate
parameter are used, compared to positive values (see, e.g., Table 10). Likewise, for
Recombine, high values of the initial recombination rate (greater than 1.0) may lead to very slow progress in the initial chains.

For Genetree the number of replications is the driving force behind program running time. Another influential factor is whether the results are based on the specific unrooted genealogy as input by the user or on the summed probabilities from the unrooted genealogy. If the unrooted genealogy is used, run time increases proportional to the number of segregating sites. The simplest run where only $\theta$ is estimated requires the least amount of time. Estimation of additional parameters such as migration rates and growth rates slows down the program somewhat, and running time varies with the magnitude of the initial values (e.g. slower for higher values of initial growth parameter). Finally, the general procedure of finding MLEs in Genetree requires running the program several different times. This process contributes greatly to total estimation time.

A direct comparison of running time between the LAMARC programs and Genetree is impossible because of the many different options and the necessity of using different datasets. Individual runs of the LAMARC programs varied from 8 seconds to 127 minutes, depending on the dataset, chain settings, parameters to be estimated and their initial values. Individual runs of the Genetree program for 100,000 replications averaged around 50 minutes. However, the estimation procedure for Genetree requires several runs of the program to calculate a likelihood surface. For example, the combined running time for estimating $\theta$ and the population growth parameter for the Nuu-Chah-Nulth data in Genetree was $\approx 17$ hours. The average run time for Fluctuate on the same data was 29 minutes (averaged over runs with different initial growth parameters).
UPBLUE is by far the fastest among the three coalescent estimation programs here considered. There are no options to select that could influence run time, and results are practically immediate. In general, the speed of these programs (from fastest to slowest) is UPBLUE < LAMARC < Genetree.

6.6 Output

An important characteristic of computer programs that requires comparison is the usefulness, readability and interpretability of the output. A program that performs many functions but whose output is confusing is not very desirable.

The output files for all of the programs in the LAMARC package are similar in content and format. The main elements of output include data summaries, point estimates, log-likelihood tables for the parameters, and log-likelihood surface graphs. The format of the output is one main text file divided into sections, defined by headings, and usually set apart by some sort of border. The labeling and comments are adequate for readability. The graphs are easy to read and accompanied by legends, and the tables are sufficiently labeled. Table A.3 in the Appendix lists the specific portions of the output for each program.

The output from Genetree is a little less convenient than LAMARC. The basic results are output directly to the command screen, or may be redirected to a separate file. If the results are to be based on the unrooted tree, then all possible rooted trees must first be created, each in its own file. The unrooted probabilities are summed over all rooted trees. To incorporate variable population size or migration, separate files must be created that contain the initial exponential growth rate and migration matrix. Then, if population parameters are to be estimated, a surface file is created. When obtaining results from the unrooted tree, a separate surface file is created for each rooted tree, as well as for the
unrooted tree. Reading results from several different files is not as convenient as seeing all results in one file.

The output from UPBLUE is short and simple. The Internet version prints the results directly on the Internet page, in about 16 lines. The estimate of $\theta$ is given (as well as intermediate estimates), the estimated variance, and Watterson’s and Tajima’s estimates.

6.7 Ease of use

The LAMARC programs are all menu driven. If the required and/or optional input files are in the same directory as the program, then the program will automatically bring up a menu of options. Alphabetic letters or symbols change default options or bring up submenus. If the user has created a parmfile (parameter file) in the same directory, the options listed in the parmfile show up as defaults in the menu, but may still be changed.

Genetree is command line driven. The user may view the options that are available by typing “genetree” at the command prompt, or he/she may reference them from the documentation. The minimum requirements include the name of the input file, an initial value of $\theta$, number of replications, and a random seed. When listed on the command line after “genetree”, these specifications, as well as any combination of options, will initiate one run of the program.

Running UPBLUE from the Internet is practically effortless. The user simply gives the full path of the input file, and clicks the “Start Estimation” button.

6.8 Stability

Various factors can contribute to a crash for any of the programs. The most obvious one is an input file with incorrect format. None of the programs will run in this
case, usually without any indication of why. The LAMARC programs will sometimes print an error, but do not indicate the cause, leaving the task of discovering the cause of the crash up to the user. Other factors that may cause a crash in the LAMARC programs include a transition/transversion ratio of 0.5 (see A.5 Glossary), large values of the initial parameter, a star-shaped phylogeny, and insufficient memory. Recombine crashed more often when trying to analyze SNP data, and when heating was used (discussed in 7.1.1 Search Strategy). Genetree sometimes crashes when the program tries to take the square root of a negative number. What initiates this phenomenon is unknown. The solution we have found is to slightly vary different input options. In general, when the data is in the right format and directions are followed, the programs are pretty stable.
7  DETAILED PROGRAM EVALUATION

7.1  LAMARC Package

7.1.1  Search Strategy

The programs in the LAMARC package have six options that control the search strategy for the maximum likelihood estimates. The program begins with a genealogy (either arbitrarily created or input by the user) and sequentially makes modifications to it. Each modification is referred to as a “step”. Each change is either accepted or rejected based on the data. A continuous series of steps is called a “chain”. The genealogies along the chain are sampled at intervals, and a likelihood curve is constructed from the sampled genealogies (Fluctuate doc. 1998).

The LAMARC programs are sensitive to the initial values of the estimates. To overcome this, the programs run several short chains to acquire good initial estimates, then run a few long chains to refine them. The final estimates are based on the long chains. The search options in each of the programs control the number of short and long chains, the length of the chains (in number of steps), and the sampling increment (how often to sample the chains).

To investigate the effects of these chain settings we performed a $2^6 \times 2$ fractional factorial designed experiment using the Fluctuate program. We used the Nuu-Chah-Nulth data and only estimated $\theta$ (no growth parameter). The factors and levels were the number of short chains (5, 20), number of short increments (10, 40), number of short steps (200, 600), number of long chains (2, 6), number of long increments (10, 40), and number of long steps (1,000, 20,000). We added three center points (10, 25, 400, 4, 25, 10500) for an estimate of variability and to check for curvature. A different random seed was used for
each run. We allowed the program to use Watterson’s estimator of $\theta$ as the initial value to start the runs. The factor levels and runs performed are in Table 9, along with the results. (We shall hereafter use the notation $10/25/400/4/25/10500$ to refer to the chain settings in the order: number of short chains/ short increments/ short steps/ number of long chains/ long increments/ and long steps).

Table 9. Experimental runs performed on Fluctuate

<table>
<thead>
<tr>
<th>s_chains</th>
<th>s_inc</th>
<th>s_steps</th>
<th>l_chains</th>
<th>l_inc</th>
<th>l_steps</th>
<th>$\hat{\theta}$</th>
<th>seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>40</td>
<td>600</td>
<td>2</td>
<td>10</td>
<td>20000</td>
<td>0.0136</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>600</td>
<td>6</td>
<td>10</td>
<td>1000</td>
<td>0.0154</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>200</td>
<td>6</td>
<td>10</td>
<td>20000</td>
<td>0.0145</td>
<td>210</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>600</td>
<td>6</td>
<td>40</td>
<td>20000</td>
<td>0.0148</td>
<td>190</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>200</td>
<td>2</td>
<td>40</td>
<td>20000</td>
<td>0.0131</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>600</td>
<td>6</td>
<td>40</td>
<td>20000</td>
<td>0.0141</td>
<td>175</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>600</td>
<td>2</td>
<td>40</td>
<td>1000</td>
<td>0.0137</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>200</td>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>0.0163</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>200</td>
<td>6</td>
<td>40</td>
<td>1000</td>
<td>0.0145</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>600</td>
<td>6</td>
<td>10</td>
<td>1000</td>
<td>0.0107</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>400</td>
<td>4</td>
<td>25</td>
<td>10500</td>
<td>0.0133</td>
<td>62</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>400</td>
<td>4</td>
<td>25</td>
<td>10500</td>
<td>0.0141</td>
<td>62</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>200</td>
<td>2</td>
<td>40</td>
<td>20000</td>
<td>0.0151</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>200</td>
<td>2</td>
<td>10</td>
<td>1000</td>
<td>0.0168</td>
<td>59</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>400</td>
<td>4</td>
<td>25</td>
<td>10500</td>
<td>0.0142</td>
<td>62</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>600</td>
<td>2</td>
<td>40</td>
<td>1000</td>
<td>0.0167</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>600</td>
<td>2</td>
<td>10</td>
<td>20000</td>
<td>0.0237</td>
<td>67</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>200</td>
<td>6</td>
<td>10</td>
<td>20000</td>
<td>0.0135</td>
<td>210</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>200</td>
<td>6</td>
<td>40</td>
<td>1000</td>
<td>0.0097</td>
<td>9</td>
</tr>
</tbody>
</table>

The average estimate ($\hat{\theta}$) was 0.0146, with a standard deviation of 0.0028. The average number of seconds for each run was 70.2, with a standard deviation of 71.4. ANOVA results showed no factors had a significant effect on $\hat{\theta}$ at the 0.05 level. The significant effects on running time were the number of long chains (p-value = 0.0207) and the number of long steps (p-value = 0.0012). As the value of these settings increases so
does the running time. This analysis was specific to this dataset on Fluctuate, but we have assumed throughout our evaluation of the other LAMARC programs that chain settings have no effect on $\hat{\theta}$.

Another search option available in Recombine, Migrate, and Lamarc is called “heating”, which improves the program’s ability to search a good range of possible trees. This involves combining efforts of several chains, with different chances of accepting changes. It is beneficial when the search remains near the initial value, or when the acceptance rate for each chain is low. The drawback is increased running time.

A third option to improve the parameter search is called “replication”, available only in Lamarc. This involves repeating an entire set of chains many times with different starting genealogies and combining the results. Replication may help when separate runs of the program produce inconsistent results or when estimates vary wildly from chain to chain. Again, the drawback is increased running time (proportional to the number of replicates).

7.1.2 Fluctuate

Many options are available for the programs in the LAMARC package. To aid comparison, we used the simplest options possible, and the specific settings are included in parmfiles in Appendix A.4. For each of the programs, we used Watterson’s estimate of $\theta$ for the initial value and the program-generated initial tree. Results may be improved by using a better estimated starting tree input by the user (Fluctuate doc. 1998).

Fluctuate can be used under the Wright-Fisher model or assuming exponential growth. See Table 9 for the results of Fluctuate when $\theta$ was estimated for the Nuu-Chah-
Nulth data under the Wright-Fisher model. No estimate of variance is given by Fluctuate when the constant population size model is used.

Using the center point values for the chain settings from the experiment in section 7.1.1 (10/25/400/4/25/10500), we also ran the program to estimate $\theta$ and the exponential growth parameter $g$ for the Nuu-Chah-Nulth data. There is no built-in estimate for the initial value $g_0$ like Watterson’s estimate of $\theta$, so the user must make an educated guess. The documentation (Fluctuate doc. 1998) states that setting $g_0 = 0.0$ will cause the program to crash, and very large values of $g_0$ (especially negative ones) may also cause a crash. It gives no indication of what “large” means however. We selected values for $g_0$ of ±5.0, ±2.0, ±1.0, ±0.5, and ±0.00001. We ran a final run $g_0$ set to the average of the first 10 runs (285.0). The results are in Table 10.

<table>
<thead>
<tr>
<th>$g_0$</th>
<th>$\hat{\theta}$</th>
<th>s.d. ($\hat{\theta}$)</th>
<th>$\hat{g}$</th>
<th>s.d. ($\hat{g}$)</th>
<th>$\text{cov} (\hat{\theta}, \hat{g})$</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.0</td>
<td>0.0198</td>
<td>0.0026</td>
<td>241.32</td>
<td>83.79</td>
<td>-7.0756e-4</td>
<td>63:58</td>
</tr>
<tr>
<td>-2.0</td>
<td>0.0211</td>
<td>0.0028</td>
<td>227.14</td>
<td>95.64</td>
<td>-1.7050e-3</td>
<td>68:17</td>
</tr>
<tr>
<td>-1.0</td>
<td>0.0222</td>
<td>0.0027</td>
<td>234.35</td>
<td>89.39</td>
<td>4.8491e-3</td>
<td>40:57</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.0322</td>
<td>0.0030</td>
<td>447.60</td>
<td>54.08</td>
<td>-1.6297e-5</td>
<td>47:33</td>
</tr>
<tr>
<td>-0.00001</td>
<td>0.0302</td>
<td>0.0043</td>
<td>593.52</td>
<td>92.20</td>
<td>-6.4165e-8</td>
<td>33:10</td>
</tr>
<tr>
<td>0.00001</td>
<td>0.0162</td>
<td>0.0023</td>
<td>212.50</td>
<td>81.54</td>
<td>8.0687e-4</td>
<td>31:25</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0131</td>
<td>0.0017</td>
<td>126.20</td>
<td>72.32</td>
<td>2.9639e-6</td>
<td>05:00</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0156</td>
<td>0.0035</td>
<td>220.28</td>
<td>116.59</td>
<td>-3.7052e-3</td>
<td>02:20</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0228</td>
<td>0.0025</td>
<td>252.18</td>
<td>57.62</td>
<td>1.1679e-6</td>
<td>02:11</td>
</tr>
<tr>
<td>5.0</td>
<td>0.0189</td>
<td>0.0027</td>
<td>304.42</td>
<td>94.09</td>
<td>-2.3905e-4</td>
<td>02:51</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.0214</strong></td>
<td><strong>0.0028</strong></td>
<td><strong>285.95</strong></td>
<td><strong>83.73</strong></td>
<td><strong>-7.1306e-5</strong></td>
<td><strong>29:46</strong></td>
</tr>
<tr>
<td>285.0</td>
<td>0.0148</td>
<td>0.0016</td>
<td>299.88</td>
<td>123.80</td>
<td>-1.5487e-5</td>
<td>01:35</td>
</tr>
</tbody>
</table>

We have not discovered any obvious method of selecting the appropriate estimates. Comparison of log-likelihood values is not valid because the reported likelihoods are relative and not absolute. Figures 2 and 3 plot the approximate 95% confidence intervals
for $\hat{\theta}$ and $\hat{g}$ and show how variable they are ($\hat{\theta} \pm 1.96 \times \text{s.d.}$). However, the standard deviations for $\hat{\theta}$ and $\hat{g}$ included in the output are very approximate. The documentation states that the assumption is made that the likelihood curve is more nearly normal than it generally is in practice. They aid in interpreting the parameter estimates and how flat the likelihood surface is and how seriously to take the results. The standard deviations and the estimates improve if more loci are analyzed, and increasing the number and length of chains may sometimes improve standard deviations (Fluctuate doc. 1998).

![Graph showing confidence intervals for $\hat{\theta}$ from Fluctuate](image)

**Figure 2.** Confidence intervals for $\hat{\theta}$ from Fluctuate
The estimates of $\theta$ are most useful when either the population size or the mutation rate is known, and the other can be calculated using the relationship $\theta = 2\kappa N \mu$. When interpreting results from Fluctuate, it is important to remember that the mutation rate ($\mu_s$) is the rate per site per generation. For the Nuu-Chah-Nulth data, $\kappa = 2$. So for example, the average value of $\hat{\theta}$ from Table 10 was 0.0214. If we assume a mutation rate of $\mu_s = 1.0 \times 10^{-6}$, then the current effective population size is $N = 5350$. For human populations $\mu$ is usually estimated using other methods and $N$ is calculated from $\theta$ (Genetree doc. 2002).

The exponential growth rate $g$ is in units of $1/\mu_s$ per generation. Positive values indicate population growth and negative values indicate decline. The average value of $\hat{g}$ from Table 10 is 285.95. Assuming the same mutation rate ($\mu_s = 1.0 \times 10^{-6}$), the unscaled exponential growth rate for this data is 0.0003 per generation.
The parameters $\theta$ and $g$ are not independent. For a higher growth rate, the current population size is expected to be larger compared to its average size. Fluctuate produces a table that translates the parameters into population terms, given the maximum likelihood estimates $\hat{\theta}$ and $\hat{g}$, and various mutation rates and time frames. For the last run from Table 10, this portion of the output is given in Table 11.

Table 11. Translation of parameters into population terms for ML Estimates

\[
\theta = 0.0148 \text{ and } g = 299.9; \text{ from Fluctuate output}
\]

<table>
<thead>
<tr>
<th>Mutation Rate</th>
<th>Population Sizes (generations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1000</td>
<td>1.0000e-010 3.6914e+007 3.6914e+007 3.6914e+007 3.6914e+007</td>
</tr>
<tr>
<td>-100</td>
<td>1.0000e-008 3.6903e+005 3.6913e+005 3.6914e+005 3.6925e+005</td>
</tr>
<tr>
<td>100</td>
<td>1.0000e-007 3.6803e+004 3.6903e+004 3.6914e+004 3.6925e+004</td>
</tr>
<tr>
<td>1000</td>
<td>1.0000e-006 3.5823e+003 3.6803e+003 3.6914e+003 3.7025e+003</td>
</tr>
<tr>
<td>-1000</td>
<td>1.0000e-005 2.7350e+002 3.5823e+002 3.6914e+002 3.8038e+002</td>
</tr>
<tr>
<td>-100</td>
<td>1.0000e-003 3.4965e-013 1.8401e-001 3.6914e+000 3.8972e+013</td>
</tr>
</tbody>
</table>

Kuhner et al. (1998) discovered an upwards bias in $\hat{g}$ from Fluctuate. They claim that the bias is an inherent property of the estimator and should be expected from any method of estimation from genealogy structure. They also report a smaller upward bias in $\hat{\theta}$ due to the correlation between the two parameters. According to Kuhner et al. (1998), sampling additional unlinked loci is more effective in reducing the bias than increasing the number of chains or the length of the chains from the same locus.

7.1.3 Recombine

Recombine is appropriate for data where the assumption of no recombination is violated. It estimates $\theta$ and the recombination rate $r$. The program may still be used to analyze data with no recombination (such as mitochondrial DNA) by fixing $r = 0$. We used...
the Nuu-Chah-Nulth data (without recombination) and the Hudson simulations (with recombination) to investigate the characteristics of Recombine.

For the Nuu-Chah-Nulth data with chain settings 10/20/5000/4/20/5000 and $r = 0$ fixed, Recombine estimated $\hat{\theta} = 0.0144$, with a 95% confidence interval of (7.77e-03, 2.23e-02). This is very close to the average results from Fluctuate ($\bar{\theta} = 0.0146$) when $g = 0$.

Recombine will also accept sequence data as ordered SNPs with information about distances between sites. Kuhner et al. (2000a) investigate the usefulness of SNP data for estimating population parameters and describe two likelihood methods implemented in their programs. The “conditional likelihood” method should be used when only the SNPs are available, and should not be used when estimating a recombination rate. This method modifies the likelihood by conditioning on the site being a SNP. The “reconstituted DNA” basically tries to “reconstruct the original DNA sequence” (Kuhner et al. 200a). This method should be used if the number of unobserved sites is known, and may be used whether estimating recombination or not.

The menu options and documentation of Recombine, however, are ambiguous about how to implement these methods. According to the documentation, the menu option “SNPs with recombination” indicates which method will be used (No = conditional likelihood, Yes = reconstituted DNA). A “spacefile” may be created that contains the distances between SNPs, and affects the results, regardless of which option is selected. There is also a parmfile option, “full-snp”, that is not described in the documentation and its purpose is unknown.

Using the Nuu-Chah-Nulth data in SNP format to estimate $\theta$ with Recombine produced confusing results. For example, the conditional likelihood method does not need
information on other sites, yet the results when the spacefile was used are more similar to those when the full DNA sequences were used (see Table 12). For another confusing example, we fixed $r$ at zero so it would not be estimated, yet the reconstituted DNA method is chosen by selecting “Yes” to “SNPs with Recombination”. This method requires knowledge of invariable sites, yet functions without a spacefile. It is not clear what the program is doing under either method when the spacefile is or is not included.

Table 12. Recombine results using Nuu-Chah-Nulth SNP data

<table>
<thead>
<tr>
<th></th>
<th>SNPs with Recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td><strong>θ</strong> 95% CI</td>
<td></td>
</tr>
<tr>
<td>Spacefile</td>
<td>0.0123 (8.07e-3, 2.85e-2)</td>
</tr>
<tr>
<td>No</td>
<td>0.0207 (1.07e-2, 3.18e-2)</td>
</tr>
</tbody>
</table>

Simulation results by Kuhner et al. (2000a) show (for finite data) accurate estimates from SNPs when $\theta$ is relatively high (0.1), though not as efficient compared to the full DNA model. When $\theta$ is lower (0.01) estimates based on SNPs from finite data will be inaccurate. In such cases, Kuhner et al. (2000a) suggest using methods that rely on the infinitely-many-sites model to analyze SNP data. Estimates are expected to be consistent based on SNPs from an infinite amount of data.

To investigate the ability of Recombine to estimate recombination rates we used the Hudson simulations, with chain settings 10/20/5000/2/20/50000. Other options were set as simple as possible. For an example of parameter specifications see the parmfile in Appendix A.4.2.
In Recombine the parameter $r$ is scaled in terms of the mutation rate. Specifically, $r = \rho_s/\mu$, where $\rho_s$ is the recombination rate per site per generation. The parameter can be interpreted as the ratio of the per-site chance of recombination to the per-site chance of mutation (Recombine doc. 2000). The recombination rate used to simulate the data from the Hudson simulator is scaled as $R = 4N\rho_g$, where $\rho_g$ is the recombination rate per gene per generation. The different scaling makes it difficult to compare, since there are unknown values involved ($N$, $\mu$), and different interpretations of $\rho$. The parameters $N$, $\mu$ and $\theta$ aren’t known from the simulated data.

Like Fluctuate, there is no built-in estimate for the initial value $r_0$ in Recombine, so the user must make an educated guess. Setting $r_0$ to 0.0 will waste the first chain, and very large values (greater than 1.0) may slow down or crash the program (Recombine doc. 2000). If no recombination is expected, $r_0$ should be set to a very low non-zero value. We selected values for $r_0$ of 0.01 and 0.1 to be used with the Hudson simulations ($R = 0.1$ and $R = 1.0$). The results are in Table 13. Approximate 75%, 95%, and 99% confidence intervals are also produced by Recombine for both $\hat{\theta}$ and $\hat{r}$. We have included the 95% confidence intervals.

<table>
<thead>
<tr>
<th>$r_0$</th>
<th>$\hat{\theta}$</th>
<th>95% CI ($\hat{\theta}$)</th>
<th>$\hat{r}$</th>
<th>95% CI ($\hat{r}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.2104</td>
<td>(1.42e-01, 3.28e-01)</td>
<td>0.00000</td>
<td>(0.0, 5.29e-03)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2111</td>
<td>(1.44e-01, 3.26e-01)</td>
<td>0.00198</td>
<td>(3.72e-05, 1.13e-02)</td>
</tr>
<tr>
<td></td>
<td>(R = 1.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.2603</td>
<td>(1.83e-01, 3.88e-01)</td>
<td>0.00976</td>
<td>(2.50e-03, 2.60e-02)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2356</td>
<td>(1.68e-01, 3.40e-01)</td>
<td>0.02600</td>
<td>(2.07e-02, 3.26e-02)</td>
</tr>
</tbody>
</table>

Table 13. Results from Recombine with 2 datasets, varying $r_0$.
Comparing the values of \( \hat{r} \) from Recombine with \( R \) from the Hudson simulator depends on the values of \( N \) and \( \mu \). For example, the estimate from the simulation with \( R = 0.1 \) and \( r_0 = 0.1 \) was \( \hat{r} = 0.00198 \). If we assume \( \mu_s = 1.0 \times 10^{-6} \), then \( \rho_s = r \times \mu_s = 1.98 \times 10^{-9} \), or \( \rho_g = 1.485 \times 10^{-6} \) (750 sites/sequence). If \( N \approx 16,800 \) then the values are comparable (\( \hat{R} = 4N\rho_g = 0.0998 \)). In general, the estimates of \( r \) from the simulation with \( R = 0.1 \) were lower than from the simulation with \( R = 1.0 \). Similarly, the estimates of \( r \) when started from the \( r_0 = 0.01 \) were lower than the estimates from \( r_0 = 0.1 \). Kuhner et al. (2000a) conclude from simulation studies that Recombine cannot accurately estimate \( r \) with sequences shorter than 500 base pairs.

When \( r_0 = 0.01 \) for both datasets there were many of the chains that had a low acceptance rate, meaning a small percentage of trees were accepted from each chain. The documentation indicates that an acceptance rate of 5% or lower indicates the sampler is having trouble moving around the parameter space, and heating may improve the results. However, when attempting to use heating for these datasets the program crashed with no explanation.

Another option to help the acceptance rate is the choice between two tree-rearrangement strategies. The runs in Table 13 used the “normal” option. The “final-coalescence” strategy is slower and intended for highly recombinant data. Using the final-coalescence strategy on the \( R = 0.1 \) simulated data with \( r_0 = 0.01 \) improved the acceptance rates somewhat, but they were still below the 5% level for some chains.
7.1.4 Migrate

The current version of Migrate (v. 1.6.8) allows estimation of parameters for an \( n \)-population model. The estimated parameters are \( \theta = 2\kappa N^0 \mu \) and \( M_{ji} = \frac{m_{ji}}{\mu} \), where \( m_{ji} \) is the immigration rate per generation from population \( j \) into \( i \). The migration rate estimate may also be expressed as \( \theta_i M_{ji} = 2\kappa N^0 m_{ji} \).

Initial values of the parameters to start the chains may be user-input values or \( F_{ST} \) calculations. \( F_{ST} \) calculations are based on mean differences in populations compared to mean differences between populations (Smith 1970). According to the Lamarc documentation the \( F_{ST} \) estimates are not always sensible, particularly when within-population variability is greater than between-population variability (Lamarc doc. 2002). On the other hand, simulation studies by Beerli and Felsenstein (2001) showed the program converging even when unfavorable starting parameters are chosen.

Migrate is flexible with regards to the migration model. It allows different subpopulation sizes, symmetric or asymmetric migration rates between subpopulations, and arbitrary migration scenarios that may be known by the researcher (e.g. fixed population sizes, equal migration rates, zero migration rates, etc.). It can also perform likelihood ratio tests of hierarchical migration scenarios.

We used Migrate to analyze the SIMCOAL simulations with the chain settings 10/25/400/2/25/4000. In Migrate the menu option “Number of steps along chains” refers to the trees recorded (in this case 400 and 4,000), while in Fluctuate and Recombine it refers to the total trees sampled (25*400 = 10,000). In other words, the chain settings 10/25/400/2/25/4000 in Migrate are equivalent to 10/25/10000/2/25/100000 in Fluctuate.
and Recombine. This can be confusing when going from one program to another. For specific parameter options see the example of the parmfile in Appendix A.4.3.

The SIMCOAL datasets were created with a symmetric migration matrix with rate $m_{ji} = 0.00005$ and $N = 20,000$ for haploid data, therefore $\theta_i M_{ji} = 2.0$ from population 1 to population 2 and vice versa. The true value of $\theta$ from the simulated data is not known. The point estimates and 95% confidence intervals from Migrate are in Table 14. In only one dataset the confidence intervals for both migration rates included the true parameter. In 7 datasets only one confidence interval included the true parameter, and in 2 datasets neither interval contained the true value.

Table 14. Point Estimates and 95% confidence intervals from Migrate
from 10 SIMCOAL datasets with $\theta_i M_{ji} = 2.0$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\theta}_1$</th>
<th>$\hat{\theta}_2$</th>
<th>$\theta_i M_{ji}$</th>
<th>$\theta_2 M_{1i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.036</td>
<td>0.040</td>
<td>5.025</td>
<td>1.27e-07</td>
</tr>
<tr>
<td></td>
<td>(0.023, 0.062)</td>
<td>(0.017, 0.081)</td>
<td>(2.259, 9.270)</td>
<td>(9.56e-8, 3.19e-4)</td>
</tr>
<tr>
<td>2</td>
<td>0.020</td>
<td>0.035</td>
<td>1.044e-07</td>
<td>1.386</td>
</tr>
<tr>
<td></td>
<td>(0.011, 0.039)</td>
<td>(0.021, 0.061)</td>
<td>(7.83e-8, 2.61e-4)</td>
<td>(0.485, 4.002)</td>
</tr>
<tr>
<td>3</td>
<td>0.029</td>
<td>0.040</td>
<td>0.765</td>
<td>1.161e-07</td>
</tr>
<tr>
<td></td>
<td>(0.018, 0.050)</td>
<td>(0.023, 0.073)</td>
<td>(0.190, 2.058)</td>
<td>(8.71e-8, 2.90e-4)</td>
</tr>
<tr>
<td>4</td>
<td>0.060</td>
<td>0.040</td>
<td>1.026</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>(0.037, 0.1)</td>
<td>(0.024, 0.076)</td>
<td>(0.28, 2.405)</td>
<td>(0.021, 1.326)</td>
</tr>
<tr>
<td>5</td>
<td>0.035</td>
<td>0.048</td>
<td>2.039</td>
<td>1.522</td>
</tr>
<tr>
<td></td>
<td>(0.021, 0.062)</td>
<td>(0.029, 0.087)</td>
<td>(0.391, 4.028)</td>
<td>(0.529, 3.213)</td>
</tr>
<tr>
<td>6</td>
<td>0.021</td>
<td>0.057</td>
<td>3.633e-08</td>
<td>5.566</td>
</tr>
<tr>
<td></td>
<td>(0.011, 0.045)</td>
<td>(0.034, 0.094)</td>
<td>(2.72e-8, 0.626)</td>
<td>(0.610, 4.357)</td>
</tr>
<tr>
<td>7</td>
<td>0.027</td>
<td>0.027</td>
<td>1.348e-09</td>
<td>2.813</td>
</tr>
<tr>
<td></td>
<td>(0.014, 0.053)</td>
<td>(0.016, 0.047)</td>
<td>(1.01e-9, 0.384)</td>
<td>(0.782, 6.147)</td>
</tr>
<tr>
<td>8</td>
<td>0.017</td>
<td>0.07</td>
<td>1.523</td>
<td>2.953</td>
</tr>
<tr>
<td></td>
<td>(0.009, 0.033)</td>
<td>(0.046, 0.112)</td>
<td>(0.306, 3.344)</td>
<td>(1.5, 5.261)</td>
</tr>
<tr>
<td>9</td>
<td>0.029</td>
<td>0.065</td>
<td>3.748</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>(0.015, 0.055)</td>
<td>(0.042, 0.109)</td>
<td>(1.863, 5.143)</td>
<td>(4.786, 9.422)</td>
</tr>
<tr>
<td>10</td>
<td>0.058</td>
<td>0.052</td>
<td>1.714e-07</td>
<td>5.278</td>
</tr>
<tr>
<td></td>
<td>(0.034, 0.11)</td>
<td>(0.036, 0.08)</td>
<td>(1.29e-7, 4.29e-4)</td>
<td>(2.287, 10.024)</td>
</tr>
</tbody>
</table>

Shading indicates intervals that include the true parameter

*Estimated scaled migration rate from population 1 to population 2
bEstimated scaled migration rate from population 2 to population 1
To investigate the inaccuracy from these simulations we changed some of the options and repeated the two runs (datasets 1 and 10) in which neither confidence interval included the true parameter. Using the same random seed for each dataset we doubled the length of the chains to 10/25/800/2/25/8000. This had no effect on the dataset 1, while one interval from dataset 10 now included the true parameter. With the longer chains we also input initial migration rates equal to the true parameter, rather than using the $F_{ST}$ estimates. This still had no effect on the dataset 1, while both intervals from dataset 10 now included the true parameter. Finally, we kept the original chain lengths and $F_{ST}$ estimates, but changed the random number seed. In this last run one interval from dataset 1 now included the true parameter, while both intervals from dataset 10 included the true parameter. Depending on the data, each of these options seems to have an effect on the estimates.

Another possible reason for the inaccuracy of Migrate from these simulations may be lack of data. The documentation notes that in general, single locus data is not adequate for estimation of migration rates and produces large confidence intervals (Beerli 1997-2002). Simulation studies by Beerli and Felsenstein (2001) show fairly accurate results with 10-locus data sets, although as the complexity of the migration model increases the estimates are biased upwards and confidence intervals widen.

7.1.5 Lamarc

Lamarc, the most recent program added to the LAMARC package, performs coalescent estimation while incorporating both migration and recombination in the model. This type of data is common in nature (e.g. see Harding et al. 1997 and Kuhner et al. 2000b). However, we were not able to acquire full sequence data that met these
assumptions. A program called Treevolve (http://evolve.zoo.ox.ac.uk/software/Treevolve/Treevolve.html) claims to simulate data evolving under the coalescent model with migration and recombination. Unfortunately we could not get this program to work.

Without the appropriate data we were unable to evaluate the accuracy and consistency of results from Lamarc under the combined migration and recombination model. However, as a test run we used Lamarc to analyze one of the SIMCOAL simulations (dataset 1), which incorporates migration but not recombination. Using the chain settings 10/20/500/2/20/10000 the program took 127 minutes (Lamarc follows the method of Migrate for interpreting the chain settings). Both 95% confidence intervals from Lamarc included the true migration rate parameters, while neither interval from Migrate included the true parameters. The estimated recombination rate was $\hat{r} = 0.0$, with a 95% confidence interval of (0.0, 0.018). The estimates $\hat{\theta}_1 = 0.027$, $\hat{\theta}_2 = 0.041$ from Lamarc were similar to those from Migrate ($\hat{\theta}_1 = 0.036$, $\hat{\theta}_2 = 0.04$). See Appendix A.3.4 for the XML file, which contains the specific options.

### 7.2 Genetree

The simplest results from the Genetree program, given an input value of $\theta_0$, include the estimated likelihood of the input tree and the mean time to the MRCA (TMRCA), both with their standard errors. The user may also specify an exponential growth rate or a migration matrix to see the effect on the tree likelihood and TMRCA. To estimate $\theta$, the growth rate, or migration parameters the user must input initial values, give a range of values over which to calculate the likelihood surface (centered around the initial value) and indicate the number of replications or runs. We used the Nuu-Chah-Nulth dataset to
investigate characteristics of the Genetree program under the Wright-Fisher model and the variable population size model and the Africa dataset to examine the program under the migration model. Both datasets fit the infinitely-many-sites model.

7.2.1 Wright-Fisher Model ($\theta$ Only)

Watterson’s estimate for the Nuu-Chah-Nulth tribe is $\theta_W = 3.93$. Using this value as the initial $\theta_0$, the likelihood surface around this value (range 3.4 to 4.4) was monotonically increasing. By sequentially increasing $\theta_0$ to 4.9 we found the MLE, $\hat{\theta} = 4.8$. An example of the specific command options for the program runs is illustrated in Appendix A.4.5.

The value of $\hat{\theta}$ from Genetree is based on the mutation rate per gene per generation, $\mu_g$. To compare with Fluctuate, a simple conversion can be made by multiplying the estimate by the sequence length (352 sites for the Nuu-Chah-Nulth data). For example, the average value from Fluctuate under the constant population size model was $\hat{\theta}_F = 0.0146 \times 352 = 5.14$, compared to the Genetree estimate $\hat{\theta}_G = 4.8$.

To better understand the variability involved in Genetree with different random number seeds we ran three duplicate runs, estimating $\hat{\theta}$ under the Wright-Fisher model with $\theta_0 = 4.9$ and different seeds for each run. The value of $\hat{\theta}$ was consistently 4.8, the likelihood values ranged from 1.52e-19 to 1.95e-19, and the standard errors of the likelihood ranged from 1.03e-20 to 2.28e-20. The estimates are consistent given different seeds. (The specific results are in Table A.4 in the Appendix.)

The likelihood surfaces produced by Genetree are only locally accurate in the vicinity of the generating parameters (Genetree doc. 2002). The simulated likelihoods are
an average of independent estimates and are therefore approximately normally distributed. A 95% confidence interval of the likelihood may be calculated as $\pm 1.96*SE$. The documentation notes that the variance is often big and larger than the mean, so a large number of simulations is essential. Unless otherwise indicated, we ran all simulations with 100,000 repetitions, using a different random seed for each run. It may be possible to reduce the variances by using a larger number of replications.

7.2.2 Variable Population Size

When estimating both $\theta$ and the growth parameter ($\beta$) using Genetree, the documentation suggests finding $\hat{\theta}$ for various fixed values of $\beta_0$, then selecting the $\beta_0$ with the highest likelihood among the MLEs, then searching in the vicinity of this $\beta_0$, fixing the corresponding $\hat{\theta}$ value as $\theta_0$. The results from the first step of this recommended search strategy are in Table 15. We chose initial values from $\beta_0 = -2.0$ to $\beta_0 = 4.0$. (The documentation claims that runs get very long when $\beta_0$ is very much larger than 3.0.) Each row in Table 15 contains the results from one run of the program, given $\beta_0$ and $\theta_0$. Each run produces a likelihood surface around $\theta_0$, and $\hat{\theta}$ is chosen as the value with the highest likelihood.

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$\theta_0$</th>
<th>$\hat{\theta}$</th>
<th>Likelihood</th>
<th>S.E.</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.0</td>
<td>35.3</td>
<td>35.4</td>
<td>1.98e-19</td>
<td>2.48e-20</td>
<td>59:27</td>
</tr>
<tr>
<td>-1.0</td>
<td>13.4</td>
<td>13.0</td>
<td>1.67e-19</td>
<td>1.51e-20</td>
<td>54:51</td>
</tr>
<tr>
<td>0.0</td>
<td>4.9</td>
<td>4.8</td>
<td>1.95e-19</td>
<td>2.28e-20</td>
<td>18:56</td>
</tr>
<tr>
<td>1.0</td>
<td>5.7</td>
<td>5.75</td>
<td>3.48e-19</td>
<td>2.70e-20</td>
<td>51:23</td>
</tr>
<tr>
<td>2.0</td>
<td>6.6</td>
<td>6.5</td>
<td>5.12e-19</td>
<td>8.60e-20</td>
<td>51:10</td>
</tr>
<tr>
<td>3.0</td>
<td>7.2</td>
<td>7.25</td>
<td>5.11e-19</td>
<td>5.85e-20</td>
<td>51:32</td>
</tr>
<tr>
<td>4.0</td>
<td>7.9</td>
<td>7.85</td>
<td>3.17e-19</td>
<td>1.93e-20</td>
<td>52:55</td>
</tr>
</tbody>
</table>
As $\beta_0$ increases or decreases away from 0.0, $\hat{\beta}$ increases. The shaded cells in Table 15 indicate the results with the highest likelihood of all the maximized likelihoods. Fixing $\theta_0 = 6.5$ we searched for $\hat{\beta}$ in the vicinity of 2.0. The results are in Table 16. Instead of estimating $\hat{\beta}$ directly, Genetree creates a likelihood surface for a multiplier ($k$) of the initial value $\beta_0$. The shaded cells in Table 16 indicate $\hat{\beta} = 1.0 \times 2.5 = 2.5$.

This process can be repeated back and forth until the joint maxima of $\theta$ and $\beta$ are found. After fixing $\beta_0$ and $\theta_0$ each one more time, our final maximum likelihood estimates were $\hat{\theta} = 6.95$ and $\hat{\beta} = 2.625$ (likelihood = 4.7109e-19, s.e. = 4.9009e-20). (The results from fixing $\beta_0 = 2.5$ are in Table A.5 in the Appendix.)

This search process of fixing one parameter while searching for the other is complex and time consuming. It is possible in Genetree to estimate both $\theta$ and $\beta$ simultaneously. Using 10,000 repetitions, $\theta_0 = 6.95$, and $\beta_0 = 2.625$ we found $\hat{\theta} = 6.85$ and $\hat{\beta} = 2.49$, (likelihood: 3.8e-19, standard error: 6.27e-20). However, the documentation states that “running a two-dimensional surface for $\beta$ and $\theta$ is generally unsuccessful

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$k$</th>
<th>$\hat{\beta}$</th>
<th>Likelihood</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Increasing Likelihood Surface</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.35</td>
<td>2.025</td>
<td>5.46e-19</td>
<td>7.61e-20</td>
</tr>
<tr>
<td>2.0</td>
<td>1.2</td>
<td>2.4</td>
<td>5.47e-19</td>
<td>1.72e-19</td>
</tr>
<tr>
<td>2.5</td>
<td>1.0</td>
<td>2.5</td>
<td>7.18e-19</td>
<td>3.40e-19</td>
</tr>
<tr>
<td>3.0</td>
<td>0.8</td>
<td>2.4</td>
<td>4.21e-19</td>
<td>3.15e-20</td>
</tr>
<tr>
<td>3.5</td>
<td>0.7</td>
<td>2.45</td>
<td>4.07e-19</td>
<td>4.04e-20</td>
</tr>
</tbody>
</table>

This search process of fixing one parameter while searching for the other is complex and time consuming. It is possible in Genetree to estimate both $\theta$ and $\beta$ simultaneously. Using 10,000 repetitions, $\theta_0 = 6.95$, and $\beta_0 = 2.625$ we found $\hat{\theta} = 6.85$ and $\hat{\beta} = 2.49$, (likelihood: 3.8e-19, standard error: 6.27e-20). However, the documentation states that “running a two-dimensional surface for $\beta$ and $\theta$ is generally unsuccessful
because it is not locally very accurate and the likelihood surface may have a maximal ridge
where $\beta$ and $\theta$ increase together … It tends to maximize the likelihood on the boundary of
any likelihood grid” (Genetree doc. 2002). In this case the results were fairly similar for
the Nuu-Chah-Nulth data.

Under the variable population size model the estimate $\hat{\theta}_g = 6.96$ from Genetree is
similar to the average value $\hat{\theta}_F = 0.0214$ from Fluctuate, after making the appropriate
conversion ($\hat{\theta}_F = 0.0214 \times 352 = 7.53$). Assuming the same mutation rate as before, $\mu_s =
1.0 \times 10^{-6}$, then from Genetree, $N = 4936$, compared to $N = 5350$ from Fluctuate.

The value of $\beta$ from Genetree is scaled by the population size, $\beta = 2Nl$, where $l$ is
the exponential growth rate per gene per generation. Direct comparison between the
Fluctuate parameter $g$ and the Genetree parameter $\beta$ is not straightforward because
Felsenstein’s group scale time in units of $1/\mu_s$ generations (Kuhner 1998), while Griffiths
and Tavaré (1994b) scale time in units of $2N$ generations. The growth rate estimate from
Genetree was $\hat{\beta} = 2.625$. If we assume $\mu_s = 1.0 \times 10^{-6}$ and $N = 4936$, then $l = 0.0003$, which
was the same estimate as from Fluctuate.

Griffiths et al. (Genetree doc. 2002) emphasize that the results from Genetree are
more effective with hypothesis testing than with estimation. A null hypothesis of no
exponential population growth versus the alternative of positive exponential growth can be
tested using a likelihood ratio test. The log-likelihood from our MLE under the constant
population assumption is $-18.7090$. The log-likelihood corresponding to $\hat{\beta} = 2.625$ was
$-18.3269$. The value of the likelihood ratio statistic for this test is 1.02, suggesting little
evidence of exponential population growth.
7.2.3 Migration

To investigate the method of estimating migration rates in Genetree we used the Africa dataset. Harding et al. (1997) determined the root of the tree as haplotype B2 (see Appendix A.2.2). Rather than summing the probabilities over all unrooted trees, we made the same assumption and based our results on this rooted tree.

Watterson’s estimate \( \hat{\theta}_W \) for the panmictic population (no subdivision) is 2.96 (s.e. 1.09). (For the CAR population \( \hat{\theta}_W = 1.65 \) (s.e. 0.83), and for KEN \( \hat{\theta}_W = 3.25 \) (s.e. 1.24)). Under the constant population size assumption, \( \hat{\theta} = 3.15 \), following the procedure described in the preceding sections. Under the variable population size model, the maximum likelihood estimators were \( \hat{\theta} = 3.8 \) and \( \hat{\beta} = 0.56 \). The likelihood statistic for this data is 1.005, providing little evidence of population growth assuming a panmictic population.

The advantage of Genetree over Migrate is that migration rates and population growth rates can both be incorporated into the model and estimated, though in a somewhat limited way. None of the LAMARC programs have combined these parameters yet. These parameters are not independent, e.g. in a subdivided population the trees are longer while \( \theta \) and \( \beta \) are smaller (Genetree doc. 2002). The procedure for estimating migration rates in Genetree is the same as for estimating the population growth rate, i.e. searching for one parameter while holding others fixed. The drawback to this procedure is that the complexity grows as more parameters are added to the model, increasing the number of possibilities to explore. It is primarily for this reason that it is suggested that analyses are constrained to two populations (or a migration matrix with only 2 or 3 free parameters),
with as many locations as possible combined. Another motivation is that the increase of data in combined populations aids parameter estimation.

Using the Africa dataset with two populations, we first assumed a constant population size while estimating $\theta$. The migration matrix ($M$) is specified in a separate file, with zeros on the diagonals and migration rates elsewhere. For example, the matrix

$$\begin{pmatrix}
0.0 & 0.5 \\
3.0 & 0.0
\end{pmatrix}$$

indicates that the rate of migration from population 0 to 1 (CAR to KEN) is 0.5, and from population 1 to 0 (KEN to CAR) is 3.0. The $(i, j)^{th}$ value in the matrix is $m_{ij} = 2N_e q_{ij}$, where $q_{ij}$ is the migration rate per gene per generation, backwards in time, from population $i$ to population $j$. Note the differences from the parameter estimated by Migrate: $4N_e (i) m_{ji}$. Genetree uses the multiplier 2 rather than 4. The default in Genetree assumes the same effective population size for all populations, although this can be changed if the relative sizes are known. For example, a separate file with the values 0.25 0.75 indicates that population 1 is a third of the size of population 2. The $q_{ij}$ from Genetree is equivalent to the $m_{ji}$ from Migrate. (Note: we will use the notation $M = \{0.0, 0.5, 3.0, 0.0\}$ for migration matrices from here on).

For the Africa data we specified several different migration matrices in the range 0.5 to 5.0 for each parameter. Migration rates of about 5.0 in magnitude indicate that the population is essentially panmictic, or that the two populations behave as one with migration unrestricted (Genetree doc. 2002). For each migration matrix we estimated $\hat{\theta}$. Unlike Migrate, Genetree does not estimate separate values of $\hat{\theta}$ for each population, but assumes they are the same. The value of $\hat{\theta}$ with the highest likelihood was 2.45, which
occurred under $M = \{0.0, 5.0, 0.0, 3.0\}$. Fixing $\theta_0 = 2.45$ we searched in the vicinity of this $M$. Likelihood estimation for migration rates can be performed in two ways. First, a likelihood surface can be created for a multiplier ($h$) of the entire matrix $M$. The multiplier applies to each value of the matrix. The second option allows estimation of individual entries of $M$. We used the second option and found the likelihood surface increasing towards values greater than 5.0 for both parameters. We conclude from this that the data suggests essentially unrestricted migration between CAR and KEN.

Estimating population growth rates and migration rates for two populations simultaneously is more complex and is not necessarily an intended function of Genetree. The program allows different growth rates for each population, though it is recommended to use the same growth rate for all populations unless prior information is available. As noted in section 7.2.2 Variable Population Size for one population, Genetree calculates a likelihood surface for a multiplier ($k$) of $\beta_0$. When two growth rates are specified, $\beta_{01}$ and $\beta_{02}$, Genetree still estimates only one value of $k$ that is applied to both populations. Hence, different growth rates for two populations cannot be estimated separately. If the migration matrix is assumed known as well as the growth rates relative to each other, then their magnitude can be estimated while holding the same relative relationship. We did not test this with the African data since it is essentially from a panmictic population and no migration matrix applies.

The same procedure follows if growth rates are assumed known and migration rates are estimated. We previously determined that the African data suggests no population growth. However for purposes of example, we fixed $\beta_{01} = \beta_{02} = 0.56$ (the MLE) and
estimated the migration matrix. Again, the likelihood surface indicated the maximum occurred at values greater than 5.0 for both migration rates.

It is possible in Genetree to perform a likelihood ratio test to test the hypothesis of no population subdivision (Genetree does not do automatic LRTs). This test involves a combinatorial factor as explained by Bahlo (2000). A simple description and example is given in the documentation.

The technique of estimating $\theta$, $\beta$, and $m_{ij}$ is quite involved. The procedure of fixing parameters while searching for others is convoluted and requires many runs of the program. The documentation contains some guidelines on how to proceed, but we have found in general that these guidelines are not thorough enough. As stated by Griffiths (Genetree doc. 2002), “Analysis is not an automated process with a clear path and requires reasoning to decide which parameters are best. It is not as simple as finding the MLE of each of the parameters. Trying to find a MLE for the migration rate from one population to another with a great deal of accuracy is not possible”.

There are many other options available in Genetree that we did not cover. These include simulation of the empirical distribution of the TMRCA, simulation of ages of the mutations in the sample and graphical output of the gene tree. Another option allows Genetree to stop searching when there are only two individuals left and calculate the probability exactly.

Genetree does not currently incorporate recombination into its model. A program called recom, which does include recombination, is referenced in the literature (Griffiths and Marjoram 1996), but we have not been able to acquire it.
7.3 **UPBLUE**

UPBLUE does not require any parameter specifications, as do the other programs. All that is necessary is a dataset in the right format (see 6.4 Data Input Format). The program uses the pairwise nucleotide distances to calculate branch lengths of an UPGMA tree. It then performs an iterative procedure to estimate \( \theta \), using Watterson’s estimate to start the procedure, and cycling until the process converges. For the Nuu-Chah-Nulth data, the output lists the estimates of \( \theta \) at the end of each of 3 cycles: 3.48364, 3.48227, 3.48228. UPBLUE then performs a bias correction, which compensates for using an estimated genealogy rather than the true one. This correction is based on a regression equation from simulation studies (Fu 1994a). The final value of \( \hat{\theta} \) for the Nuu-Chah-Nulth data is 4.436 with estimated variance 1.761. This estimate is similar to \( \hat{\theta}_F = 5.14 \) from Fluctuate and \( \hat{\theta}_G = 4.8 \) from Genetree.

Variance estimates from UPBLUE were shown in simulation studies by Fu (1994a) to be quite close to the theoretical minimum variance. Tajima and Watterson estimates are also given (3.296 and 3.251 respectively).

Fu (1994b) has extended his estimation method to include the assumptions of recombination and migration, but these models have not been incorporated into the current UPBLUE program. Programs written in C for these models are available from the author, but they do not estimate recombination rates or migration rates.
LAMARC, Genetree and UPBLUE have strengths and weaknesses in different areas, which makes it impossible to recommend one program overall. LAMARC surpasses the others in availability, with executables for UNIX, Windows, and Mac. An executable for Genetree is available for Windows, while UPBLUE is accessible on the web. The source code is available for each of the programs. LAMARC and Genetree rank equal in the area of documentation, while UPBLUE falls far short. A comparison of the data input format depends on the data possessed by the user. LAMARC ranks first with full sequence data that includes all nucleotides, while Genetree is preferred if only the segregating sites are available. However, data does not always fit the more restrictive infinitely-many-sites model required by Genetree. The Nuu-Chah-Nulth data had to be adjusted, while the Africa data met the assumption. LAMARC also accepts other types of data, including electrophoretic and microsatellite data. UPBLUE requires nucleotide differences, which must be obtained from sequence data through another program (or by self-programming).

UPBLUE is by far the fastest estimation program, followed by LAMARC, then Genetree. This also depends on the options selected. The method suggested by Genetree of searching for one parameter while fixing all others is very time consuming. The default in the LAMARC programs is simultaneous estimation of multiple parameters, but there is the option to fix parameters, which follows the same method as Genetree.

The output of LAMARC and UPBLUE are both printed to a single file and are easy to read. UPBLUE has the advantage of being more simple and direct, while LAMARC is more complete and offers additional information to evaluate the results. The multiple file output from Genetree is more complicated, and does not provide as much information as
LAMARC. The internet version of UPBLUE is the simplest program to use, requiring only a click of a button. Genetree and LAMARC rate similarly, depending on the user’s preference of menu-driven or command-line driven. LAMARC has more options available that expand possibilities, but also add to the complexity. All of the programs are fairly stable, given the data is in the right format.

Our results from these programs could only be compared in a few instances. We analyzed the Nuu-Chah-Nulth data with each of the packages under the basic Wright-Fisher model (constant population size, no migration, and no recombination), and \( \hat{\theta} \) differed by 0.7 among the packages (≈15% of the average): \( \hat{\theta}_F = 5.14 \) (Fluctuate), \( \hat{\theta}_G = 4.8 \) (Genetree), and \( \hat{\theta}_U = 4.436 \) (UPBLUE). If we assume \( \mu = 1.0 \times 10^{-6} \) (or \( \mu_g = 3.52 \times 10^{-4} \)) then the estimates of the population sizes are \( N_F = 3650, N_G = 3409, \) and \( N_U = 3151 \). These estimates varied by 499 (≈15% of the average). We also compared the estimates from Fluctuate and Genetree from the Nuu-Chah-Nulth data under the variable population size model: \( \hat{\theta}_F = 7.53 \) and \( \hat{\theta}_G = 6.96 \). The un-scaled estimated exponential growth rate per generation from both programs for this data was 0.0003. Finally, we compared the results from the Migrate and Lamarc programs within the LAMARC package on a simulated dataset (migration, no recombination). Lamarc performed better than Migrate on the estimation of migration rates, Lamarc accurately estimated the recombination rate to be zero, and the \( \hat{\theta} \) estimates were similar between the two programs.

Other comparisons between the programs were not possible in our study because of the different datasets used, the different models implemented by the packages, and the different options unique to the programs. Simulation studies done by Beerli and
Felsenstein (2001) compare Migrate with the migration function of Genetree. With 100 loci there is much uncertainty in both programs, but they include the true parameter values in their 50% confidence regions. They also produce similarly shaped likelihood surfaces. LAMARC incorporates several chains to reduce the sensitivity to the initial parameter values. In Genetree, the user must find appropriate starting values, which may require several runs of the program, adding to the total time. Both programs may still produce biased results when the starting values are far from reality.

Another aspect to consider is the scaling of the estimates. LAMARC considers the mutation rate per site per generation, and scales the evolutionary estimates by this mutation rate. This scaling is rather unique, while Genetree and UPBLUE follow the more standard method in the literature. They consider the mutation rate per sequence per generation, and scale the evolutionary estimates by the population size $N$.

A major criticism of UPBLUE is that it relies on a single genealogy, while LAMARC and Genetree sample from multiple trees. Fu (1994a) analyzed 100 trees and concluded that $\hat{\theta}$ and its variance differed little among the trees.

In general, if a researcher desires a fast estimate of $\theta$ and has little time to figure out a program then UPBLUE may be the best choice. However, this program relies on the assumptions of the Wright-Fisher model. If the researcher expects deviations from any of these assumptions then LAMARC or Genetree should be used. LAMARC has the unique ability to allow recombination and simultaneous migration and recombination, while Genetree allows simultaneous migration and variable population size. Another important consideration is the type of data. If only the segregating sites are available then Genetree is
more appropriate, provided the data fit the infinitely-many-sites model. However, if the user has electrophoretic or microsatellite data then LAMARC is the only choice.

The coalescent model is a useful tool that has been implemented in many areas of research in population genetics. It provides an efficient way to estimate genetic diversity and to make inferences about the evolutionary history of a population. We have compared the performance of three programs (LAMARC, Genetree, and UPBLUE) written by major groups of coalescent researchers. These are not commercial computer programs, but have been developed by researchers for their own use and made available to others. Their performance in various areas has not been previously well established. Our comparison will hopefully help population geneticists determine which program would be most applicable for their data and research.
APPENDIX

A.1 Tables
A.2 Datasets
A.3 Data input formats
A.4 Program specifications
A.5 Glossary of genetics terms
### A.1 Tables

Table A.1 Coalescent simulation and analysis programs with comments

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimation / Function</th>
<th>Evolutionary Forces</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluctuate</td>
<td>$\theta = 4N\mu_s$ Growth Rate ($g$)</td>
<td>Recombination</td>
<td>Version 1.4 alpha, 9/2000</td>
</tr>
<tr>
<td>(Felsenstein’s group)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recombine</td>
<td>$\theta = 4N\mu_s$ Recombination Rate ($r = \rho/\mu$)</td>
<td>Migration</td>
<td>Version 1.2.4, 7/2001</td>
</tr>
<tr>
<td>(Felsenstein’s group)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migrate</td>
<td>$\theta_i = 4N_i\mu_i$ ($i$th population) Migration Rate ($4N_i m_{ij}$)</td>
<td>Migration</td>
<td>Version 1.0 alpha, 12/2001</td>
</tr>
<tr>
<td>(Felsenstein’s group)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamarc</td>
<td>$\theta = 2N\mu_s$ Recombination Rate ($r$) Migration Rate ($M_{ij} = m_{ij}/\mu$)</td>
<td>Recombination Migration</td>
<td></td>
</tr>
<tr>
<td>(Felsenstein’s group)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetree</td>
<td>$\theta = 4N\mu$ Migration Rate ($2N m_{ij}$) Growth Rate ($\beta = 2N g$)</td>
<td>Exponential growth Migration Infinitely-many-sites model</td>
<td>Version 9.0, 7/2000 <a href="http://www.stats.ox.ac.uk/~griff/software.html">http://www.stats.ox.ac.uk/~griff/software.html</a></td>
</tr>
<tr>
<td>(Griffiths’ group)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweep_bott</td>
<td>Detection of bottlenecks$^a$ and selective-sweeps$^b$ Bottlenecks Selective sweeps</td>
<td>Implements Griffiths and Tavaré’s algorithm. (Have to download in UNIX) <a href="http://helios.bto.ed.ac.uk/evolgen/galtier/galtier.html">http://helios.bto.ed.ac.uk/evolgen/galtier/galtier.html</a></td>
<td></td>
</tr>
<tr>
<td>UPBLUE</td>
<td>$\theta = 4N\mu$ Infinitely-many-sites model Recombination Migration</td>
<td>2/1996 <a href="http://www.hgc.sph.uth.tmc.edu/fu/">http://www.hgc.sph.uth.tmc.edu/fu/</a></td>
<td></td>
</tr>
<tr>
<td>(Fu and Li)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eve</td>
<td>$\theta = 2N\mu$ Growth rate ($r$) Exponential growth</td>
<td>Follows the method of Fu and Li Currently unavailable. (Vasco et al. 2001) <a href="http://zoology.byu.edu/crandall_lab/Vasco/eve.htm">http://zoology.byu.edu/crandall_lab/Vasco/eve.htm</a></td>
<td></td>
</tr>
<tr>
<td>(Vasco)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMCOAL</td>
<td>Simulates molecular data in interconnected populations with arbitrary demography Variable population size Migration Finite-sites model</td>
<td>Follows the method of Hudson.</td>
<td></td>
</tr>
<tr>
<td>(Excoffier, Novembre, Schneider)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Estimation / Function</td>
<td>Evolutionary Forces</td>
<td>Comments</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>MICSAT (Wilson and Balding)</td>
<td>MCMC sampling of coalescent trees for microsatellite loci</td>
<td>?</td>
<td>(Can’t get to work) <a href="http://maths.abdn.ac.uk/~ijw/">http://maths.abdn.ac.uk/~ijw/</a></td>
</tr>
<tr>
<td>BATWING (Wilson and Balding)</td>
<td>$\theta = 2N\mu$ Exponential growth rate (starting at time $\beta$ in the past TMRCA)</td>
<td>Exponential growth Population subdivision Infinitely-many-sites model <a href="http://maths.abdn.ac.uk/~ijw/">http://maths.abdn.ac.uk/~ijw/</a> (Unable to download)</td>
<td></td>
</tr>
<tr>
<td>PAL (Drummer, Buckland, Skinner)</td>
<td>Simulates coalescence intervals and estimation of demographic parameters</td>
<td>?</td>
<td>(Unable to download) <a href="http://www.pal-project.org/">http://www.pal-project.org/</a></td>
</tr>
</tbody>
</table>

*Sudden reduction in population size  
*bRapid fixation of allele through directional selection*  

**Table A.2 Other coalescent programs**

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimation / Function</th>
<th>Evolutionary Forces</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SkylinePlotter (Skinner)</td>
<td>Coalescent-based skyline plot to infer demographic history</td>
<td>?</td>
<td>Can’t figure it out [<a href="http://www.stat.uni-muenchen.de/~strimmer/research.html">http://www.stat.uni-muenchen.de/~strimmer/research.html</a> - coalescent](<a href="http://www.stat.uni-muenchen.de/~strimmer/research.html">http://www.stat.uni-muenchen.de/~strimmer/research.html</a> - coalescent)</td>
</tr>
<tr>
<td>Treevolve (Grassly and Rambaut)</td>
<td>Simulates the evolution of DNA sequences under a coalescent model</td>
<td>Exponential growth Subdivision/Migration Recombination</td>
<td><a href="http://evolve.zoo.ox.ac.uk/software/Treevolve/Treevolve.html">http://evolve.zoo.ox.ac.uk/software/Treevolve/Treevolve.html</a> Haven’t downloaded yet</td>
</tr>
</tbody>
</table>
Table A.3. Content of output files of LAMARC programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Output Content</th>
</tr>
</thead>
</table>
| Fluctuate | Summary of Data  
|          | Log-likelihood table for various values of $\theta$ and $g$  
|          | Point estimates  
|          | Log-likelihood surface graph  
|          | Covariance matrix and approximate standard deviations  
|          | Translation of parameters into population terms |
| Recombine | Summary of Data  
|          | Log-likelihood table for various values of $\theta$ and $r$  
|          | Point estimates  
|          | Log-likelihood surface graph  
|          | Values of $\theta$ and $r$ at intermediate chains  
|          | Approximate confidence intervals (75%, 95%, 99%)  
|          | Histogram of recombination events along the sequence |
| Migrate  | Summary of Data  
|          | Profile likelihood tables for each parameter  
|          | MCMC estimates  
|          | Log-likelihood surface graphs  
|          | Options in use  
|          | Summary of profile likelihood percentiles of all parameters |
| Lamarc   | ML estimates and percentiles  
|          | Profile likelihood tables for each parameter  
|          | Regional profile likelihood tables  
|          | Data, region and model summaries  
|          | Run reports by region |

Table A.4. Likelihood values for $\hat{\theta}$ from Genetree, given various seds

<table>
<thead>
<tr>
<th>Seed</th>
<th>$\hat{\theta}$</th>
<th>Likelihood</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>46250</td>
<td>4.8</td>
<td>1.9544e-19</td>
<td>2.2753e-20</td>
</tr>
<tr>
<td>33435</td>
<td>4.8</td>
<td>1.5164e-19</td>
<td>1.0348e-20</td>
</tr>
<tr>
<td>66200</td>
<td>4.8</td>
<td>1.7575e-19</td>
<td>1.9442e-20</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.0</td>
<td>2.1937e-20</td>
<td>6.4232e-20</td>
</tr>
</tbody>
</table>
Table A.5. MLE ($\hat{\theta}$) at various $\theta_0$ values with fixed $\beta_0=2.5$; from Genetree

<table>
<thead>
<tr>
<th>$\theta_0$</th>
<th>$\hat{\theta}$</th>
<th>Likelihood</th>
<th>S.E.(Likelihood)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.25</td>
<td>Increasing Likelihood Surface</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.50</td>
<td>6.9</td>
<td>4.2399e-19</td>
<td>3.4732e-20</td>
</tr>
<tr>
<td>6.75</td>
<td>6.95</td>
<td>5.2770e-19</td>
<td>5.9696e-20</td>
</tr>
<tr>
<td>7.00</td>
<td>6.9</td>
<td>5.1133e-19</td>
<td>6.4060e-20</td>
</tr>
<tr>
<td>7.25</td>
<td>6.9</td>
<td>4.8907e-19</td>
<td>4.8004e-20</td>
</tr>
<tr>
<td>7.50</td>
<td>Decreasing Likelihood Surface</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shaded row indicates MLE with highest likelihood
A.2 Datasets

A.2.1 Nuu-Chah-Nulth Data

2557_HV1  TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----
CCCCCCCCATGCTACACATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----

2559_HV1a TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----

2559_HV1b TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----

2559_HV1c TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----

2559_HV1d TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----

2560_HV1a TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTCGTACATTAC-TGCCAGCCACCATGGAATATTGTACCTGTT ACCATAAAAAACC-AATCC--ACATCAAAA----
CCCTCACCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2566_HV1b TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2570_HV1 TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2571_HV1a TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2571_HV1b TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2571_HV1c TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2571_HV1d TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA
2571_HV1e TTCTTTCATGGGGAAGCAGATTGGTGACACCCCAAGATGGATTGACTTCACCCCATCAACAACCCGCATATGGTACATTAC-TGACACCGCTCCATAGGATACAAACAAACCTATCCACCC-TTAACAGTACATAGTACATAAAACCATTACGTACATAGCACATTACAGTCAAAATCCCTTCTCGCCCCC-ATGGATGACCCTCCCTCA

73
A.2.2 Africa data

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>508</th>
<th>906</th>
<th>1358</th>
<th>1416</th>
<th>1423</th>
<th>2008</th>
<th>2503</th>
<th>2554</th>
<th>2636</th>
<th>2792</th>
<th>2876</th>
<th>2924</th>
<th>2945</th>
<th>CAR</th>
<th>KEN</th>
<th>TOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>A3</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>B11</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B3</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>B2</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>B1</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>B9</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D1</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

aCAR: Central African Republic
bKEN: Kenya
cPolymorphic sites indicated by a number 1-3000

A.2.3 Simulated Datasets

Hudson Simulations (available from [http://www.daimi.au.dk/~compbio/hudson/hudson.html](http://www.daimi.au.dk/~compbio/hudson/hudson.html)) parameter inputs:

1st Simulation: n = 750; Kimura 2-parameter model; alpha = 0.05; beta = 0.025; no rate heterogeneity; recombination rate = 0.1; Phylip format; Hudson algorithm

2nd Simulation: n = 750; Kimura 2-parameter model; alpha = 0.05; beta = 0.025; no rate heterogeneity; recombination rate = 1.0; Phylip format; Hudson algorithm
SIMCOAL Simulations (available from http://cmpg.unibe.ch/software/simcoal/) parameter file:

//Input parameters for simcoal
2 samples to simulate
//Deme sizes
20000
20000
//Sample sizes
20
20
//Growth rates
0
0
//Number of migration matrices
1
//Migration rates matrix 0
0.0 0.00005
0.00005 0.0
//Historical events
0 historical events
//Mutation rate per generation
0.0005
//Sequence length
500
//Data type
DNA .66666666 (Corresponding to a transition/transversion ratio of 2)
//Mutation rates
0.0

A.3 Data Input Format Examples

A.3.1 Fluctuate

Nuu-Chah-Nulth data:

1
55 365
2557_HV1
TTCTTTCATGAGGAAGGAGATTTTGGTGACCGGATAAGGTATGTGACCTACCCACACAACCGCTATGTATTC
GTACATTAC-TGACCAGCCATGAAATTTGTACGCTACCTCCATAAACCC-AATCC---ACATCA-A---
CCACCTCCCATGCTTACAAGGATAAGCTACAACAATCCAAACTACACATCAAACCTCCACCCACCC-
TTAACGACATAGACATAAAAGCCATACGTACAGCAGAATTATCCCTTCTCGTCCACCCATGAGAC
80
A.3.2 Recombine

Hudson $R = 0.1$ dataset:

2559_HV1a

TTCTTTCTATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTC

GTACATTAC-TGSCAGGACCACCATTATTGTAGGTTACATAAA-AACCCC--ACATCAAAA----

CCCCCCTCCCCATGCTTTAAAGCAATAGCAATTCACACGTACCACCACCACC-

CTATCCACAATCATACTGCAACATCCAAAGGCCCAC-CCTCACCACAGTACCACAACACCTACCACCACC-

TTAACATAGGATAGATACAAAGCCATACGATAGCACCATTACAAATCCCTTCTGTCCCC--

ATGGATAGCCATCCTCA

......etc.

A.3.3 Migrate

SIMCOAL dataset:

2 1 Migration matrix 0, .00005, .00005, 0

500

20 Population0

0.0

GGTTTCTATAGCAGTTTACCCCGGATTTGGAATCCACATTAGACTACTATTTAATTAAATCCACTATTTCCCGGA
ATCCCCCCACCAATCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCG
GAGTGCTGGAGCCGGATCTCGCTCCGGAGCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG 0_1
GGGTTCTATAGCAATCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCG
GAGTGCTGGAGCCGGATCTCGCTCCGGAGCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG 1_1
GGGTTCTATAGCAATCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCGT
GAGTGCTAGAGCCGGAGTTACAGGTCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG 0_1
GGGTCTATAGCCAGTTTACCCGGATTTACCCGAAATCCACATTTAGACTACTATAATATTACCTCCACTATTCCCCGGGA
ATCCCCCACCACCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCGT
GAGTGCTAGAGCCGGAGTTACAGGTCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG 1_1
GGGTCTATAGCCAGTTTACCCGGATTTACCCGAAATCCACATTTAGACTACTATAATATTACCTCCACTATTCCCCGGGA
ATCCCCCACCACCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCGT
GAGTGCTAGAGCCGGAGTTACAGGTCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG

......etc.
20  Population1
1 0
GGGTCTATAGCCAGTTTACCCGGATTTACCCGAAATCCACATTTAGACTACTATAATATTACCTCCACTATTCCCCGGGA
ATCCCCCACCACCTGAGGATTACTCAGTAGAGGTCTTTTCGCTTTCTTACTCGCTAATATCGTATGATA
TGCAGTAAATGTCCGCCCGTATATTATCGACCCTCTGTCTGGTTATCTTTTTCCGGCTCCCGGGCTACCGA
TTGTAAGTCCTACCGAGACCCGGTGCGTTTGTGGTTAAGAGCAATTACGCTCCCTAGATATACGAGACCGA
TTATTCCAGAGTCTAATGTTGGAACTTATAGGTGGTTCCTCCGGGCCGGGAGTACGTCGATAGGGCCGCGT
GAGTGCTAGAGCCGGAGTTACAGGTCTCAACATGCCACGGGCCAGAAAGTTTCGTGAGACGGTGCCACAAT
ACAGCGAGTCTGGCGATCTCGGGGAAACCCCATCGTCGCCGGCAGTTCTCCCTCCTGAGCAGACCGAGATAGG
TTG

......etc.

A.3.4 Lamarc

<lamarc>
<!-- Created from the LamarcDS DataStore -->
<forces>
  <coalescence>
    <start-values> 0.01 0.01 </start-values>
    <method> USER USER </method>
    <max-events> 1000 </max-events>
  </coalescence>
  <migration>
    <start-values> 0.0 100 100 0.0 </start-values>
    <method> FST FST FST </method>
    <max-events> 10000 </max-events>
  </migration>
</forces>

82
<chains>
  <replicates>1</replicates>
  <heating>
    <temperatures>1</temperatures>
    <swap-interval>1</swap-interval>
  </heating>
  <strategy>
    <resimulating>1</resimulating>
  </strategy>
  <initial>
    <number>10</number>
    <samples>500</samples>
    <discard>1000</discard>
    <interval>20</interval>
  </initial>
  <final>
    <number>2</number>
    <samples>10000</samples>
    <discard>1000</discard>
    <interval>20</interval>
  </final>
</chains>

<format>
  <verbosity>verbose</ verbosity>
  <progress-reports>verbose</ progress-reports>
  <echo>true</echo>
  <plotting>
    <profile>false</profile>
    <posterior>false</posterior>
  </plotting>
  <seed>1005</seed>
  <parameter-file>parmfile</parameter-file>
  <results-file>outfile</results-file>
  <summary-file>sumfile</summary-file>
</format>

<data>
  <region name="Region1">
    <population name="Population0">
      <individual name="0 0">
        <sample name="0 0">
          <datablock type="DNA">
            GGGTCTATACGATTTTACCCGGATTGAATCACCAGATTGACTACTATTAAATTCCACTATTGCC
            CCGGAATCCCCCAACCAATCTGAGATTACCTAGTAGGTTTTTTCTTACTCGCTATATATGT
            ATGATAGCAGTAAATGTCCGCCGGTATATTATCGACCTCTGTGTTATACCTTTTTCCGCTCCCGGC
            TACCGATTGACCTAATTCCGGACGATCCTTTGTTCTGACGACGTACGCTCCCGTGTGACTAC
            GATGTTATTCCAGGCTAATTCCGAACTCTTAAGGTAACTCTAATTTGTTCTCCTCCGAGGTATC
            GTCTCCTCGCTGGGCTCGCTGGGGAACCGCATCGTCCGCCAGCTCGAGCCAGTACGAGATAGGTTG
          </datablock>
        </sample>
      </individual>
    </population>
  </region>
</data>
A.3.5 Genetree

Africa data:

Sequence file for 2 populations:
0 9 : ATCCGCTGGCAGTG
1 12 : ATCCGCTGGCAGTG
0 1 : ATCCGCTAGCAGTG
1 5 : ATCCGCTAGCAGTG
1 1 : CTCTCTGGCAACT
0 1 : ATCTCTGGCAAGTT
1 6 : ATCTCTGGCAAGTT
0 4 : CTCTCTGGCAAGTT
1 8 : CTCTCTGGCAAGTT
0 6 : ATCTCTGGGTGTGTT
1 9 : ATCTCTGGGTGTGTT
0 1 : ATCTCTGGGAAAGTT
1 1 : ACTGTTCGCAAGTT

Ancestor file:
CTCTCTGGCAAGTT

A.3.6 UPBLUE

Nuu-Chah-Nulth data:

55
1  2  1
1  3  1
1  4  1
1  5  1
1  6  2
1  7  2
1  8  2
1  9  2
1 10  2
1 11  7
1 12  7
1 13  6
1 14  6
1 15  6
1 16  5
1 17  5
1 18  5
1 19  5
1 20  5
1 21  5
A.4 Computer Runs

A.4.1 Parmfile for Fluctuate Results

interleaved=false
printdata=false
progress=true
print-trees=false
freqs-from-data=true
categories=false
watterson=true
usertree=false
ttratio=2.0
short-chains=20
short-inc=25
short-steps=400
long-chains=4
long-inc=25
long-steps=10500
growth-rate=-1.0
end

*Change these parameters when estimate $\theta$ only to investigate MCMC options
**Change this parameter for runs estimating $\theta$ and $g$

A.4.2 Parmfile for Recombine Results

interleaved=false
printdata=false
progress=true
print-trees=false
freqs-from-data=true
categories=false
watterson=true
usertree=false
autocorrelation=false:1.000000
newdata=true
same-ne=true
interactive=true
mhmcsave=false
panel=false
map=false
final-coalescence=false
full-snp=false
haplotyping=false
profile-like=true
norecsnp=false
ttratio=2.000000
short-chains=10
short-steps=5000
short-inc=20
long-chains=2
long-inc=20
long-steps=50000
rec-rate=0.100000
holding=none
mu-ratio=1.000000
trait-ratio=100.000000
dis-freq=0.500000
hapdrop=0
heating=1;1;
end
*Change this parameter for $r_0$

A.4.3 Parmfile for Migrate Results

```plaintext
# Parmfile for Migrate 1.6.7
# generated automatically on
# Monday October 07 11:22:45 2002
# please report problems to Peter beerli
# email: beerli@genetics.washington.edu
# http://evolution.genetics.washington.edu/lamarc.html

# General options ---------------------------------------
nmlength=10

# data options ------------------------------------------
datatype=SequenceData
ttratio=2.000000
freqs-from-data=YES
categories=1
rates=1:1.000000
prob-rates=1:1.000000
autocorrelation=NO
weights=NO
interleaved=NO
fast-likelihood=YES
usertree=NO
distfile=NO

# input/output options ----------------------------------
menu=YES
infile=infile
random-seed=AUTO #OWN:258497371
progress=YES
logfile=NONE
print-data=NO
outfile=outfile
plot=YES:OUTFILE:LOG:{0.000100,100.000000,0.000100,100.000000}:N36
profile=ALL:PRECISE
write-summary=NO

# likelihood-ratio test and AIC
1-ratio=MLE:m, m, m, m
aic-modeltest=NO

# parameter options ------------------------------------
theta=FST
migration=FST
mutation=NOGAMMA
fst-type=THETA
custom-migration={**}
geofile=NO
```

# search strategies ------------------------------------
short-chains=10
short-inc=25
short-sample=400
long-chains=2
long-inc=25
long-sample=4000
#obscure options
burn-in=10000
mig-histogram=NO
heating=NO
moving-steps=NO
long-chain-epsilon=INFINITY
gelman-convergence=No
replicate=No:LastChains
end

A.4.5 Example of Command Line for Genetree Results

Prompt > Genetree treefile_dat.# 7.25 100000 23844 –f surf35.# –e exp –g 6.75 7.75 21 –i .5 1.5 21 >Out35.txt

Commands in order:

- **Genetree**: Calls the program
treefile_dat.#: Instructs the program to use all rooted trees, each in a different file that were created previously with the –z option (treefile_dat.0 – treefile_dat.18)

- **7.25**: \( \theta \), Number of replications
- **100000**: Random seed
- **23844**: Indicates likelihood surface files will be created
- **-f**: Creates separate likelihood surface files for, one for each rooted tree and the unrooted tree (surf35.0 – surf35.18, surf35.all)
- **-e**: Indicates exponential growth
- **exp**: Name of the file containing \( \beta_0 \)
- **-g**: Allows likelihood estimation of \( \theta \)
- **6.75**: Lower range for the likelihood surface of \( \theta \)
- **7.75**: Upper range for the likelihood surface of \( \theta \)
- **21**: Number of surface points to calculate between lower and upper range for \( \theta \)
- **-i**: Allows likelihood estimation of \( \beta \)
- **.5**: Lower range for likelihood surface for multiplier of \( \beta \)
- **1.5**: Upper range for likelihood surface for multiplier of \( \beta \)
- **21**: Number of surface points to calculate between lower and upper range for \( \beta \)

>Out35.txt: Redirects output to this file name
## A.5 Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele</td>
<td>Alternative form of a gene.</td>
</tr>
<tr>
<td>Bottlenecks</td>
<td>Sudden reduction in population size.</td>
</tr>
<tr>
<td>Diploid</td>
<td>Two copies of genetic material in each cell.</td>
</tr>
<tr>
<td>DNA</td>
<td>(Deoxyribonucleic acid). The chemical inside the nucleus of a cell that carries genetic instructions for making living organisms.</td>
</tr>
<tr>
<td>Effective population size</td>
<td>Size of an idealized population that would have the same effect of random sampling on gene frequency as in the actual population (i.e. the size of the reproducing population).</td>
</tr>
<tr>
<td>Evolution</td>
<td>A change in the frequency of alleles in a population from generation to generation.</td>
</tr>
<tr>
<td>Gametes</td>
<td>Haploid reproductive cells.</td>
</tr>
<tr>
<td>Gene</td>
<td>Sequence of DNA that performs a specific function. Gene and locus are often used interchangeably.</td>
</tr>
<tr>
<td>Genetic diversity</td>
<td>The extent of polymorphism in a population.</td>
</tr>
<tr>
<td>Genetic drift</td>
<td>Changes in the frequencies of alleles in a population that occur by chance, rather than natural selection.</td>
</tr>
<tr>
<td>Haploid</td>
<td>One copy of genetic material in each cell.</td>
</tr>
<tr>
<td>Infinitely-many-sites model</td>
<td>A mutation model that assumes each mutation occurs at a new site of the sequence.</td>
</tr>
<tr>
<td>Locus</td>
<td>The physical position of a gene on a chromosome. Gene and locus are often used interchangeably.</td>
</tr>
<tr>
<td>Migration</td>
<td>The process by which geographically separate populations exchange a certain proportion of individuals each generation (e.g., a proportion $p_{12}$ of population 1 moves to join population 2, and $p_{21}$ of population 2 moves to join population 1.</td>
</tr>
<tr>
<td>MRCA</td>
<td>Most recent common ancestor.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mutation</td>
<td>Heritable change in genetic material. A change from one allele type to another, usually resulting from an error in replication of DNA.</td>
</tr>
<tr>
<td>Natural selection</td>
<td>Differential survival and reproduction of classes of organisms that differ from one another in one or more usually heritable characteristics.</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>A nitrogen base attached to sugar and phosphate molecules.</td>
</tr>
<tr>
<td>Phylogeny</td>
<td>The study of ancestral relations among species.</td>
</tr>
<tr>
<td>Phylogenetic tree</td>
<td>A branching diagram showing the ancestral relations among species.</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>When two or more alleles co-exist in a population. A common variation in the DNA sequence among individuals.</td>
</tr>
<tr>
<td>Population genetics</td>
<td>The study of processes influencing gene frequencies.</td>
</tr>
<tr>
<td>Random mating</td>
<td>No preference for one phenotype or the other in choosing a mate.</td>
</tr>
<tr>
<td>Recombination</td>
<td>Process by which segments of DNA are exchanged between different molecules</td>
</tr>
<tr>
<td>Segregating site</td>
<td>A site that shows variation among the sequences.</td>
</tr>
<tr>
<td>Selective neutrality</td>
<td>No natural selection.</td>
</tr>
<tr>
<td>Selective sweep</td>
<td>Rapid fixation of allele through directional selection</td>
</tr>
<tr>
<td>Transition</td>
<td>Change of a pyrimidine nucleotide (C or T) into another pyrimidine, or a purine (A or G) into another purine.</td>
</tr>
<tr>
<td>Transversion</td>
<td>Change of a pyrimidine nucleotide (C or T) into a purine nucleotide (A or G), or vice versa.</td>
</tr>
<tr>
<td>Wright-Fisher model</td>
<td>A widely-used model that describes a finite population under a set of assumptions.</td>
</tr>
</tbody>
</table>


