# The relationship between coalescence times and population divergence times

NOAH A. ROSENBERG and MARCUS W. FELDMAN

Department of Biological Sciences
Stanford University
Stanford, CA
USA 94305

July 31, 2001

Chapter 9 in the forthcoming volume: M. SLATKIN, M. VEUILLE, eds. *Modern Developments in Theoretical Population Genetics.* Oxford: Oxford University Press, 2002.

**Corresponding Author:**
Noah A. Rosenberg
University of Southern California
1042 W. 36th Place
DRB 155
Los Angeles, CA 90089-1113
(213) 740-2416 (phone)
(213) 740-2437 (fax)
noahr@usc.edu

**Running head:**
Coalescence times and divergence times

## Abstract

The *divergence time* of two populations is the amount of time that has elapsed since the populations arose from an ancestral group, while the *coalescence time* of a set of copies of a gene is the amount of time that has elapsed since the most recent common ancestor of the gene copies lived. We briefly review the methods that have been used to infer divergence times and coalescence times from genetic data. We then consider the relationship between divergence times and coalescence times in a population genetic model that includes divergence followed by migration between two descendant populations, paying particular attention to the fact that migration can cause coalescence to occur more recently than divergence. Insights gained from the model and its special cases are applied to four examples: the divergences of humans and chimpanzees, modern humans and Neanderthals, Africans and non-Africans, and Native Americans and Asians. For each example, we discuss the connection between hypothesized divergence times and estimated coalescence times.

## 9.1 Introduction

Two important concepts in the study of the genetic history of organisms are the *divergence time* of populations, and the *coalescence time* of copies of a locus in a sample of individuals. Although divergence times and coalescence times have been considered extensively in separate mathematical contexts, fewer studies have utilized a framework that includes both concepts. Most of these unified studies have treated ancient divergences for which the numerical difference between divergence time and coalescence time is not very pronounced. Additionally, most approaches have not allowed migration to take place between populations after a divergence. The purpose of this article is to examine the relationship between divergence times and coalescence times, as well as to discuss the numerical difference between them in simple models that permit migration.

A unified perspective that includes both divergence times and coalescence times is warranted because divergence is a phenomenon of great evolutionary interest, while a coalescence-based approach is of great utility in the analysis of genetic data. Gene coalescence times can be naturally inferred from genetic data, but individually, they provide limited information about population histories. Population divergence times are often more helpful towards understanding evolutionary history, but the class of assumptions about demography and genetic data that allow the inference of divergence times is more restrictive. Thus, if we better understood the relationship between divergence times and coalescence times, we could more

sensibly interpret coalescence times inferred from genetic data in terms of their implications for population histories. The distribution of inferred coalescence times across neutral loci might even become a useful source of information for testing historical models of divergence (Ruvolo, 1996; Takahata *et al.*, 2001).

In Section 9.2, we introduce divergence times and coalescence times with a series of examples. Section 9.3 reviews how divergence times and coalescence times can be estimated from genotypic data under different demographic and mutational conditions. Section 9.4 considers the difference between divergence times and coalescence times for two-population models with no migration and with a constant rate of migration per generation. We conclude in Sections 9.5, 9.6, and 9.7 with a discussion of implications for the interpretation of studies of real populations.

We note that Takahata and Satta (Chapter 5) consider a companion issue to the topics studied here. As we will see, if no migration takes place after population divergence, then coalescence necessarily precedes divergence, and the magnitude of the difference between population divergence times and gene coalescence times depends on population size prior to divergence. Indirectly, this dependence underlies the procedures used by Takahata and Satta to estimate ancestral population size using genotypes taken from two or three related species. Here we consider the converse dependence, that is, how knowledge of current (and ancestral) population sizes leads to greater accuracy regarding the numerical relationship between divergence times and coalescence times.

## 9.2 Definitions

### 9.2.1 Definition of divergence time

The *divergence time* ($T_D$) or splitting time of two populations is the time that has elapsed since an ancestral group of individuals separated into the two descendant groups. We will assume divergences take place instantaneously, so that the divergence time is a property of a pair of populations. For any two populations in a collection, the divergence time is perfectly specified from a tree of populations. Thus, in Figure 9.1, although more than two populations are present, the divergence time for any pair of populations is uniquely determined: it equals the time since the two populations were part of the same ancestral group. The divergence time of two descendant populations is well-defined even if the existence of additional descendant populations is ignored. The divergence time is also well-defined if the ancestral populations are allowed to experience multifurcation events, that is, instantaneous separation into more than two descendant groups.

### 9.2.2 Definition of coalescence time

The *coalescence time* of a collection of copies of a locus, also known as the *time to the most recent common ancestor* ($T_{MRCA}$), is the time that has passed since the existence of the most recent common ancestor (MRCA) of a collection of copies of a locus. Unlike the divergence time, which is uniquely determined once two populations are selected, the coalescence time is uniquely determined only when a locus and a set of individuals are chosen. Coalescence times do vary across loci and across sets of individuals, as will become clear below.

Because the coalescence times of sets of copies of loci have some properties that are not immediately obvious, we consider a few examples. First, we illustrate how coalescence times can vary across loci.

**Example 1:** Suppose that after an extreme bottleneck, a population of mammals contains only five living individuals (individuals 16-20), who have the genealogy and genotypes shown in Figure 9.2. Suppose also that no mutations have occurred in this population in the last several generations. For the living individuals, 16-20, we decide to compute the coalescence times of four loci: the Y chromosome, the mitochondrial genome, and two autosomal loci, *aut1* and *aut2*, which lie on different chromosomes. Locus *aut1* has alleles $A$ and $B$, while locus *aut2* has alleles $C$ and $D$.

In mammals, only males have Y chromosomes, which are thus uniparentally transmitted from fathers to sons. Because the three living males (individuals 16, 17, and 19) share the same father, the Y chromosome of that father (individual 13) is the most recent common ancestor of all Y chromosomes in the population. The value of $T_{MRCA}$ for the Y chromosome is one generation.

Mitochondrial DNA, however, is uniparentally transmitted through the mother and it is present in individuals of both sexes. Because the five living individuals do not all have the same mother but they do all have the same maternal grandmother, the mitochondrial genome of that grandmother (individual 8) is the MRCA of all surviving mitochondrial genomes. She lived two generations in the past, so $T_{MRCA}$ for the mitochondrial genome is two generations.

All copies of the autosomal locus *aut1* can be traced back two generations to individual 8, but all copies of autosomal locus *aut2* must be traced back three generations to individual 3 in order to reach a common ancestor. Due to the stochastic nature of allelic transmission from parent to offspring, the two autosomal loci have different coalescence times in the population.

**Example 2:** In Example 1, we were able to observe the genotypes of all individuals in the population. Because complete observation of all individuals is rarely possible for most populations, a more typical situation uses a *sample* of individuals from a population. Consider the genealogy of the eight members of a haploid species in Figure 9.3. In haploid species, unless "horizontal transfer" of genetic material takes place between individuals, every locus has the same genealogy. For this example, assume that horizontal transfer and recombination do not take place in the species under consideration. Suppose we selected a sample of size two from this population of eight individuals. If our sample consisted of individuals 1 and 2, the coalescence time would be four generations. If we instead considered individuals 1 and 4, the coalescence time would be 12 generations. If we considered individuals 1 and 8, the coalescence time would be 20 generations. The coalescence time for a sample is at most the coalescence time for the whole population, and as more individuals are added to a sample, the sample coalescence time approaches the population coalescence time. For this example, a sample of size five is sufficient to guarantee that the sample coalescence time equals the population coalescence time.

In Example 1, coalescence times varied across loci, but because we genotyped the entire population, variation among samples was not an issue. In Example 2, coalescence times for a given locus varied across samples. Because we considered a haploid species with no recombination, however, for any sample, all loci had the same coalescence times.

More typical of realistic studies of diploid populations, the coalescence time varies across both loci and samples considered. Suppose that for Figure 9.2, we had only taken a sample of size two rather than the whole population. Had we sampled individuals 19 and 20, we would have inferred the coalescence times for the Y chromosome, the mitochondrial genome, *aut1* and *aut2* to be 0, 1, 2, and 2 generations, respectively. Had we sampled only individuals 16 and 17, we would have inferred the Y-chromosomal, mitochondrial, *aut1*, and *aut2* coalescence times to be 1, 1, 2, and 3 generations. Recall that the correct coalescence times are 1, 2, 2, and 3 generations.

We note, however, that the variation of coalescence time across *loci* is of much greater magnitude than the variation across *samples*. The variance of the coalescence time across a large number of independent loci in a sexual population of constant haploid size $N$ is approximately $1.16N^2$ (e.g. Tavaré *et al.*, 1997). In contrast, for the same population, the probability that the coalescence time for a sample of size $n$ equals the coalescence time for the whole population is approximately $(n-1)/(n+1)$ (Saunders *et al.*, 1984). Thus, many loci are needed in order to accurately obtain the values of statistics of the coalescence time

distribution for a population. For any given locus, however, a small sample is sufficient in principle to obtain the true population coalescence time at the locus.

One final complication with coalescence times is that when considered with respect to a population, $T_{MRCA}$ for a particular locus need not change linearly as time progresses (recall that divergences are treated as fixed events, so that divergence times only increase linearly with time as soon as the divergence takes place). Consider again Figure 9.2 as an example. Suppose that our observations of the population were taken one day after the death of individual 15. Had we taken our measurements of the entire population on the previous day, we would have found the coalescence times for the Y chromosome, the mitochondrial genome, *aut1*, and *aut2* to be 4, 3, 4, and 4 respectively - for all four genes the coalescence times are different from the situation in which only individuals 16-20 were considered. In a large population, however, there is little need to be concerned with the possibility that $T_{MRCA}$ for a given gene will change nonlinearly over short periods of time. Considering Figure 1 and Table 1 of Underhill *et al.*(2000), in which 1062 human Y chromosomes from around the world are partitioned into 116 haplotypes, the most recent common ancestor of human Y chromosomes could only be changed by a catastrophic demographic event affecting millions of dispersed possessors of haplotypes 1-8, or by a similar event affecting the even more numerous carriers of haplotypes 9-116. For a sample of moderate size from a population of constant haploid size $N$, the probability that the MRCA $t$ generations into the future will be different from the current population MRCA is approximately $1 - e^{-t/N}$ (Watterson 1982; Saunders *et al.*, 1984).

To summarize, using the examples in Figures 9.2 and 9.3, we have discussed how (1) for a given set of individuals, the coalescence time may vary across loci; (2) for a given locus, the coalescence time may vary across samples; (3) the coalescence time of a particular locus for a sample is at most equal to the coalescence time of the locus for the whole population; (4) for a given population, the coalescence time of a locus may change nonlinearly as time progresses. These are general properties not just for our examples, but for samples and populations of any size.

### 9.2.3 Comments on terminology

It may appear inconsistent that the concepts of population divergence and gene coalescence are labeled with directional language in opposite senses. Looking forward in time, populations diverge, and looking backward, lineages coalesce. We could just as well describe population divergence as the joining of populations backwards in time, or gene coalescence

as the radiation of lineages forward in time.

However, the different directional senses of these two concepts are rooted in the traditions from which the concepts derive; the reasons for the apparent linguistic inconsistency are largely historical. Population divergence is a type of event that takes place as evolution moves forward in time. Evolution naturally occurs in the same direction as time, so that in an evolutionary framework, populations diverge and species radiate. While population divergences arise from evolutionary considerations, gene coalescences derive from a genealogical perspective. In this framework (e.g. Hudson, 1990; Nordborg, 2001), we assume that information is known only about the present, and we treat the past as a random process proceeding backwards from the present. From this point of view, lineages coalesce backwards in time.

To avoid confusion, we employ "divergence" to refer only to population-level or species-level demographic splits and "coalescence" to describe only the ancestry of copies of a locus. We also employ "earlier," "older," and "before" to mean "more ancient." We take "later", "younger" and "after" to mean "more recent." The directional senses of other words that relate to time should be clear from their contexts. We use the term "migration," although for the purposes of this article, "horizontal gene transfer" or "lateral gene transfer" would describe the same phenomenon just as well.

## 9.3 Methods of inference

### 9.3.1 Inference of divergence time

The need to calculate population divergence times arises in many situations. For example, to understand the history of a species, one might wish to know how long populations have been separated from each other. Estimated population divergence times can serve to confirm the influence of geologic or climatic processes on evolution. Knowing the values of population divergence times can help to date migrations or adaptive innovations. In addition to their uses in identifying biogeographic patterns, divergence times also assist in framing the phylogenetic infrastructure upon which hypotheses about evolutionary processes can be tested. Methods of estimating divergence times have been reviewed recently (Nielsen *et al.*, 1998; Edwards and Beerli, 2000; see also Watterson, 1985). Here we only comment on assumptions about mutation and demography that have been made in the development of the various estimation tools.

The first methods of estimating divergence times assumed that no mutation occurred after divergence, that descendant populations had equal and constant size, and that the

genetic differences between populations were solely due to genetic drift. The divergence time was then estimated using theoretical results relating expected divergence times to summary statistics computed from allele frequency distributions at many loci in two populations (e.g. Watterson, 1985; Nei, 1987). For this type of allele frequency data, advances have included maximum likelihood estimation of divergence times (Nielsen *et al.*, 1998; Nielsen and Slatkin, 2000), incorporation of loci evolving by a stepwise mutation process (Goldstein *et al.*, 1995; Feldman *et al.*, 1999; Zhivotovsky, 2001), relaxed assumptions about the relative sizes of the two populations (Gaggiotti and Excoffier, 2000), and permission of gene flow after the divergence (Nielsen and Slatkin, 2000; Nielsen and Wakeley, 2001; Zhivotovsky, 2001).

A second class of methods has been designed for use with completely linked genetic markers. These methods can accommodate a variety of mutational models, including the infinitely many sites model for DNA sequence data (Wakeley and Hey, 1997; Nielsen, 1998), the stepwise mutation model for microsatellite evolution (Wilson *et al.*, 2001), and finitely many allele models (Takahata and Satta, 1997). Some of these methods assume deep divergences (Takahata and Satta, 1997; Edwards and Beerli, 2000), but they generally make no assumptions about how long ago divergence took place.

Most approaches have in common a demographic model in which a single ancestral population of constant size separates into two descendant populations of constant (and possibly different) size. A few recent methods have attempted to accommodate change in population sizes through time: for example, in a method applicable to completely linked haplotypes at unique event polymorphism and microsatellite loci, Wilson *et al.*(2001) allowed for the two descendant populations to be growing at the same exponential rate. They also studied a model in which populations are constant in size until a particular moment at which they begin to grow exponentially.

A more difficult issue is the estimation of divergence time when gene flow occurs between the descendant populations after the divergence. Gene flow after a divergence slows the rate at which two populations become genetically distinctive. Thus, because lineages are exchanged between populations, methods that assume no gene flow after divergence will underestimate the divergence time. Using a maximum likelihood method based on allele frequency data at many unlinked biallelic loci, Nielsen and Slatkin (2000) developed a joint estimator of divergence time and the symmetric migration rate for two populations. For DNA sequence data evolving under the infinitely many sites model, Nielsen and Wakeley (2001) developed a similar joint estimation procedure. Both methods rely on algorithms that allow the likelihood function for migration and divergence parameters, conditional on observed sample configurations, to be approximated through simulation.

The newest methods of estimating divergence times that have begun to incorporate more complex demography than constant size models are encouraging, in that estimation of population divergence times can now be performed with species for which a model of constant size is not tenable. Nevertheless, satisfactory inference of divergence times under complex demographic models will require considerable advances. Summary statistics often take on the same or similar values under drastically different demographic conditions (e.g. Wakeley, 1996a), so that new genetic distance statistics are unlikely to accurately estimate divergence time when migration is permitted. These statistics will likely be most helpful when a specific idealized demography can be assumed. Methods that jointly infer divergence times and migration rates seem to be more promising, though these approaches can be extremely computationally intensive (Nielsen and Slatkin, 2000; Nielsen and Wakeley, 2001).

### 9.3.2 Inference of coalescence time

Concurrent with the recent improvements in the estimation of divergence times, inference tools for coalescence times have also been developed using a variety of demographic models that are suitable for different kinds of genetic data. In general, methods for inferring coalescence times numerically compute posterior distributions of $T_{MRCA}$ given prior distributions of various mutational and demographic parameters and either a full set of individual genotypes (reviewed by Stephens and Donnelly, 2000; Stephens, 2001) or a collection of summary statistics (e.g. Fu and Li, 1997; Tavaré *et al.*, 1997; Fu and Li, 1999; Pritchard *et al.*, 1999; Stumpf and Goldstein, 2001). Methods of inferring $T_{MRCA}$ from summary statistics, such as the mean number of pairwise differences between sequences or the number of segregating sites in a sample, are potentially less accurate due to inefficient use of data. As a trade-off, the gain in speed of computation due to the summarizing of data can increase the complexity of demographic and mutational models under which $T_{MRCA}$ can be inferred (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999). As is true for divergence times, a mutational model and a demographic model are currently necessary for the estimation procedure, and the choice of model can have a large impact on the eventual estimates (e.g. Pritchard *et al.*, 1999; Thomson *et al.*, 2000; Tang *et al.*, 2001).

The coalescence time for a locus is usually inferred from completely linked haplotypes. Although methods based on summary statistics often directly estimate $T_{MRCA}$, most approaches do not treat $T_{MRCA}$ as a model parameter, and they instead infer $T_{MRCA}$ as a byproduct of a procedure that estimates demographic and mutational parameters. Models under which $T_{MRCA}$ can be inferred for appropriate data include the infinitely many sites

model for DNA sequence data (e.g. Griffiths and Tavaré, 1994a, Tavaré *et al.*, 1997) and stepwise mutation models for microsatellite evolution (Wilson and Balding, 1998; Pritchard *et al.*, 1999; Wilson *et al.*, 2001). An additional computational strategy for inferring demographic parameters (Kuhner *et al.*, 1995, 1997, 1998) can be modified to infer $T_{MRCA}$ under a finitely many allele model (Tang *et al.*, 2001).

As with divergence times, the inference of $T_{MRCA}$ is most advanced for panmictic models of constant-sized populations (e.g. Griffiths and Tavaré, 1994a; Fu and Li, 1997; Tavaré *et al.*, 1997). However, $T_{MRCA}$ has also been studied under models of continuous exponential growth (Griffiths and Tavaré, 1994b; Thomson *et al.*, 2000; Tang *et al.*, 2001), constant population size followed by continuous exponential growth (Pritchard *et al.*, 1999; Wilson *et al.*, 2001), and a single instantaneous burst of population growth (e.g. Stumpf and Goldstein, 2001).

Inference of $T_{MRCA}$ using models of population structure has proven more challenging. Schemes have been developed that allow inference of demographic parameters in island migration models (Beerli and Felsenstein, 1999; Bahlo and Griffiths, 2000), but greater computational efficiency is required for use with reasonably sized data sets. The method of Wilson *et al.*(2001) can infer $T_{MRCA}$ in a model of population divergence with no migration after the divergence.

In principle, the inference of $T_{MRCA}$ with a minimum of demographic assumptions would be desirable, because choosing an appropriate demographic model to describe data is difficult. For example, the observed mean number of pairwise site differences between two DNA sequences could be taken as a simple estimator of $T_{MRCA}$ (when normalized by twice the mutation rate). This crude estimator can be calculated from any DNA sequence data set (the analog for microsatellite loci evolving under the simple stepwise mutation model is the mean number of mutational step differences between haplotypes divided by twice the mutation rate), and it is computationally simple. Unless the sample has size two (see Walsh [2001] for estimation of $T_{MRCA}$ in this case), the pairwise difference estimator is problematic, however: even if pairwise coalescence times were known for all pairs of gene copies in a sample, this method would still underestimate the coalescence time. In Figure 9.3 the average pairwise coalescence time over all 28 pairs would produce an estimate of 14.4 generations, rather than the correct value of 20 generations.

Although the bias of the pairwise difference estimator may cause gross underestimation in the case of a population that has remained constant in size, the bias is less severe for an exponentially growing population. The effect of exponential growth on the shape of coalescent trees is to shrink the lengths of older tree branches compared to the lengths of

younger branches (Slatkin and Hudson, 1991; Donnelly and Tavaré, 1995). Thus, pairwise coalescence times in exponentially growing populations tend to be closer to the coalescence time of a sample than in constant size populations (consider Donnelly [1996, figures 1-3]). Consequently, the use of pairwise differences to estimate a sample coalescence time might be more acceptable in populations that are known to be exponentially growing (Stumpf and Goldstein, 2001).

In order to use this type of nonparametric estimator, the bias of pairwise difference methods must be understood under different demographic models. For a constant size model, the bias of the pairwise difference estimator is severe: on average only a third of all pairs of sequences have the same coalescence time as the whole population (using the result of Saunders *et al.*[1984] that was mentioned in Section 9.2.2). On the other hand, in the limiting case of a population that experienced a recent extremely rapid burst in size and that derived from a small ancestral population, most pairwise coalescence times are approximately equal to the sample coalescence time, and the bias of the pairwise difference estimator will be relatively small. With intermediate values, the dependence of the bias on the growth rate must be computed. An additional problem is that except in the idealized case in which all pairwise coalescence times equal the sample coalescence time, the calculation of confidence intervals of a pairwise difference estimate may be misleading. The variance of the estimator depends on generally unknown demographic parameters such as growth rates. Thus, because the bias and variance of the mean number of pairwise differences are difficult to determine unless a highly-idealized demography is assumed, the pairwise difference method does not offer a general escape from assuming a demographic model in order to estimate coalescence times.

At present, methods that compute posterior distributions of $T_{MRCA}$ using a set of individual genotypes and a particular model offer the most rigorous approaches to the inference of $T_{MRCA}$. Because these approaches generally infer $T_{MRCA}$ as a result of a more complex analysis, they also offer the additional benefit of producing estimates of demographic parameters, such as growth rates, migration rates, and population sizes, which are generally more useful than $T_{MRCA}$ for interpreting population histories. Additional work on efficient inference of coalescence times has focused on generalizing the genetic and demographic models that provide a framework for inference of population parameters. As these models expand to further incorporate phenomena such as ascertainment bias, asymmetric migration, inbreeding, mutational biases, population divergences, recombination, samples that derive from different points in time (by including both recent and ancient DNA), and selection, and as computational schemes become more efficent, the estimation of $T_{MRCA}$ using real data sets will be

much improved.

## 9.4 Relationship between coalescence times and divergence times

As we have described in Section 9.3, separate explorations of the properties of divergence times and coalescence times have usually been made. For the remainder of this article, we turn to the less well-studied relationship between divergence times and coalescence times. We consider a model in which two populations derive from an ancestral population and in which constant migration takes place between descendant populations after the divergence. The size of the ancestral population is the sum of the sizes of the descendant populations. The model is the same one that has been considered in a number of recent studies (e.g. Wakeley, 1996b). In special cases of the migration model considered here, the distribution of the random variable $T_{MRCA} - T_D$ has been previously studied, as discussed below.

This particular model is of interest because complete isolation between descendant populations is characteristic mainly of species-level divergences. After divergence, however, migration between descendant groups is a frequent occurrence for populations of the same species. Migration models may also be more appropriate in studies of bacteria for which occasional horizontal transfer of genes takes place between species after divergence events. Our approach is similar to other studies of $T_{MRCA}$ (e.g. Harding, 1996; Marjoram and Donnelly, 1997), in which certain genetic models are chosen, and the distribution of $T_{MRCA}$ (in our case, $T_{MRCA} - T_D$) is obtained numerically under the models.

### 9.4.1 A diverging population model

In principle, the relationship between divergence times and coalescence times can be studied with any model of population history that includes population divergences and in which the coalescence times of samples can be simulated. The simplest of such histories allows a single population to split at a specific point in time into two populations. We consider two haploid populations that have constant sizes $N_1$ and $N_2$ in the present. In population $i$ ($i = 1, 2$), the number of offspring produced by an arbitrary individual has variance $\sigma_i^2$. Migration occurs between the two populations at a constant rate, so that in each generation, a fraction $m_{ji}$ of the individuals in population $i$ are migrants from population $j$. At the divergence time, $T_D$ generations in the past, the two populations arose instantaneously from an ancestral population of size $N_1 + N_2$, in which the variance of the offspring distribution was $\sigma_{1+2}^2$. We refer to the "two-population" and "one-population" phases of the model to describe the periods that contain descendant and ancestral populations, respectively. We also use the

terms "descendant phase" and "ancestral phase."

For convenience, we make several assumptions that simplify the presentation. First, we assume that the variance of the offspring distribution is 1 in both the ancestral population and the two descendant populations. If these three variances were not all equal to 1, then all results that follow could be corrected by replacing population sizes with ratios of population sizes to the corresponding offspring distribution variances (e.g. Nordborg, 2001a).

We also assume that the duration of a generation is the same in both the ancestral and the two descendant populations, equal to $G$ years. If this assumption were violated, as might be true when considering species divergences or human populations with different reproductive behavior, then the two descendant populations would have experienced different numbers of generations between the divergence event and the present day. Under these circumstances, it would be appropriate to measure time in absolute units such as years, rather than in generations.

Lastly, we assume $N_1 = N_2 = N$. Because the amount of coalescence of lineages that takes place scales with the population size, if the two populations were of vastly different sizes, the numerical values of the results would be quite different from those presented here. In general, the behavior of the larger population would dominate, so that coalescence times would be determined by the size of the larger population. However, a smaller difference between population sizes would not seriously affect the qualitative nature of the results. Thus, we ignore this complication of unequal population size.

We wish to study the random variable $T_{MRCA} - T_D$, the difference between the random time to the most recent common ancestor of the sample and the fixed divergence time. In the case of two constant populations with no migration, we can compute the distribution of this random variable exactly. For more complex demographies, we can often obtain asymptotic results regarding $T_{MRCA} - T_D$. In any case, backwards simulations of genealogies of samples taken from diverged populations are straightforward to execute (e.g. Takahata, 1989; Takahata and Slatkin, 1990); as described in the next section, we have taken advantage of this fact.

### 9.4.2 The two-population divergence model: simulations

To explore the relationship between divergence times and coalescence times in our model, we performed simulations in two steps. First, for the descendant phase, we simulated the coalescence times of $n_1$ sampled lineages from population 1 and $n_2$ sampled lineages from population 2 using a two-population model with constant population size and constant mi-

gration. Because a large amount of migration increases the simulation time dramatically (e.g. Nordborg, 2001a), we did not use very large migration rates. All simulations assumed constant population size.

We implemented this step using methods similar to those of Hudson (1990). Looking backwards in time, one of four types of events can occur: (a) a coalescence of two lineages in population 1; (b) a coalescence of two lineages in population 2; (c) the migration of a lineage in population 1 from population 2; (d) the migration of a lineage in population 2 from population 1. The distribution of the waiting time $W_1$ to the occurence of an event of type (a) can be simulated according to

$$W_a = \frac{-2N_1}{n_1^*(n_1^* - 1)} \ln(U) \tag{9.1}$$

where $U$ is a uniform(0,1) random variable and $n_1^*$ is the current number of lineages in population 1. The waiting time until an event of type (b) is analogous.

The distribution of the waiting time $W$ until an event of type (c) is simulated according to

$$W_c = \frac{-1}{n_1^* m_{21}} \ln(U) \tag{9.2}$$

where $U$ is a uniform(0,1) random variable, $n_1^*$ is the current number of lineages in population 1, and $m_{21}$ is the fraction of population 1 that derives from population 2 in every generation (Hudson, 1990). The waiting time until an event of type (d) is analogous.

To determine which event took place, we simulated values of $W_a$, $W_b$, $W_c$, and $W_d$. The event corresponding to the smallest waiting time was then allowed to occur. After updating the values of the total elapsed time and the numbers of lineages in the two populations ($n_1^*$ and $n_2^*$) we simulated new values for the four waiting times and we allowed the event of smallest waiting time to occur. This process was repeated until either all lineages coalesced or until the pre-specified divergence time $T_D$ was reached, whichever happened first. This implementation is similar to the standard simulation strategy (Hudson, 1990), in which the total rate of events of all types is computed, a random number is chosen to determine the time of an event, and a second random number is chosen to determine the type of the event. It is straightforward to show that our procedure and the usual one are equivalent, though ours requires more random numbers to be simulated. With our method, the distribution of the waiting time to the most recent event is the minimum of four exponential random variables specifying the waiting times to the four types of events: this waiting time is exponentially distributed with rate equal to the sum of the rates of occurrence of the four types of events (as in the standard method). Additionally, the distribution of the type of the event determined

by the minimum of four exponential random variables is given as in the standard simulation method: the probability that the event is, say, of type (a), equals the ratio of the rate of occurence events of type (a) and the total rate of occurrence of all events.

In the second step, which we performed only if all lineages had not coalesced during the first step, we counted the number of ancestral lineages $n_{1+2}$ that were present at the divergence time $T_D$. Treating all lineages as part of a single panmictic population, we then simulated coalescence times in the ancestral phase similarly to above (all events were coalescences), until all lineages coalesced to a single lineage.

### 9.4.3 Constant population size with no migration

We return to the case of no migration, for which a combination of analytical results and simulations are used to understand the properties of $T_{MRCA} - T_D$. For this case, the theory derives largely from Takahata and Nei (1985), who considered very similar problems. The main theoretical result is that the distribution of $T_{MRCA} - T_D$ can be represented as

$$T_{MRCA} - T_D = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Pr(r_1 = i|N_1, T_D)\, Pr(r_2 = j|N_2, T_D)\, T_{i+j}(N_1 + N_2), \tag{9.3}$$

where $r_1$ and $r_2$ represent the number of ancestral lineages in populations 1 and 2 at the time of divergence. The conditional probability $Pr(r = j|N, T)$ that a sample of size $n$ taken from a population of size $N$ has $j$ distinct ancestors at time $T$ generations has been calculated by Tavaré (1984, equations 6.1 and 6.2); in Tavaré's (1984) notation, $Pr(r = j|N, T)$ would be represented as $g_{nj}(T/N)$. The expression $T_{i+j}(N_1 + N_2)$ is the coalescence time for a sample of size $i + j$ from a population of size $N_1 + N_2$, and properties of its distribution are well-known (e.g. Hudson, 1990; Tavaré *et al.*, 1997):

$$T_k(N) = \sum_{l=2}^{k} W_l(N) \tag{9.4}$$

where $W_l(N)$ is exponentially distributed with mean $\frac{2N}{l(l-1)}$.

Because the time it takes for coalescence to occur within a population is proportional to population size (see equation 9.1), scaling population sizes and divergence times by the same constant will leave the random variable $T_{MRCA} - T_D$ unchanged relative to population size. If $N_1$ and $N_2$ are say, doubled, and if $T_D$ is correspondingly doubled, then the same amount of coalescence takes place in the two-population phase as with the original values of $N_1$ and $N_2$, although this phase lasts twice as long due to the doubling of $T_D$. Because the population size in the ancestral phase is twice as large, the length of the ancestral phase is

also doubled. Thus, with two populations of equal size $N$ and equal samples of size $n$, the parameter $T_D/N$ and the value of $n$ determine the time to coalescence.

The distribution of $T_{MRCA} - T_D$ is somewhat unwieldy, and we have explored its properties using the simulations previously described. A discussion of special cases, as well as calculations of the mean and certain quantiles, suffice to describe the main properties of the distribution. Because the distribution of $T_{MRCA} - T_D$ is influenced by the sample size $n$ and the scaled divergence time $T_D/N$, the special cases to consider are (1) very small values of $T_D/N$; (2) very large values of $T_D/N$; (3) samples of size 1 in each population.

**Small values of $T_D/N$:** When $T_D$ is small compared to the total population size $N_1 + N_2$, our model approximates a one-population model with population size $N_1 + N_2$ and sample size $n_1 + n_2$. At the time of divergence, most lineages have not coalesced, so that most coalescences take place in the ancestral phase (Figure 9.4A). Consider $T_D$ approaching 0 for fixed sample sizes. Then $T_{MRCA} - T_D \approx T_{MRCA}$, and we can obtain (Tavaré *et al.*, 1997, equations 2 and 3):

$$E[T_{MRCA} - T_D] \approx 2\left(1 - \frac{1}{n_1 + n_2}\right)(N_1 + N_2), \tag{9.5}$$

$$Var[T_{MRCA} - T_D] \approx \left[8\left(\sum_{j=2}^{n_1+n_2} \frac{1}{j^2}\right) - 4\left(1 - \frac{1}{n_1 + n_2}\right)^2\right](N_1 + N_2)^2. \tag{9.6}$$

If we then consider the large-sample limit after letting $T_D$ become small, $E[T_{MRCA} - T_D]$ is near $2(N_1+N_2)$ and $Var[T_{MRCA}-T_D]$ is near $1.16(N_1+N_2)^2$. Note that in the one-population model that is approximated for small values of $T_D/N$, sample size has only a small effect on $E[T_{MRCA}]$ and an even smaller effect on $Var[T_{MRCA}]$. As sample size increases to 10 per population, the expected coalescence time is 95% of the large-sample limiting expected coalescence time. With samples of size 50 per population, the corresponding fraction is 99%. In our simulations we use samples of size 50 per population. This sample size is similar to reasonable values for applications, and the simulated values of $T_{MRCA} - T_D$ are then extremely close to the large-sample limit.

**Large values of $T_D/N$:** For values of $T_D$ that are very large compared to $N_1 + N_2$, all lineages in each population usually coalesce to a single lineage during the two-population phase. Thus, at the time of divergence, each population will almost always be represented by only a single lineage (Figure 9.4B). Thus, $T_{MRCA} - T_D \approx T_2$, where $T_2$ is the time to the coalescence of two lineages in a single population of size $N_1+N_2$. Because $T_2$ has expectation $N_1+N_2$, the difference between $T_{MRCA}$ and $T_D$ will be very small compared to the divergence

time.

**Samples of size 1:** If samples have size 1 in both populations, then at the time of divergence each sample necessarily has a single ancestral lineage. Thus, the time to coalescence is $T_D$ plus the exponentially distributed time to the coalescence of two sequences, and the distribution of $T_{MRCA} - T_D$ is independent of the value of $T_D/N$. The mean value of $T_{MRCA} - T_D$ is therefore $N_1 + N_2$ and the variance is $(N_1 + N_2)^2$. This case has often been studied from a different point of view: because the expected number of pairwise site differences between two DNA sequences sampled one from each of two populations is proportional to $T_{MRCA}$ for the sequences, results about coalescence times of two sequences are often hidden in results on mean numbers of pairwise sequence differences.

**General case:** For intermediate values of $T_D/N$, more than one lineage may be represented in each population at the time of divergence. Because the mean coalescence time of a sample increases (slowly) with the number of ancestral lineages at the time of divergence, and because this number of ancestral lineages decreases with time, the value of $T_{MRCA} - T_D$ decreases with increasing $T_D/N$ (Figure 9.5). The initial and final values of this monotonic are $4N$ and $2N$ respectively, consistent with our discussion of the special cases of small and large values of $T_D/N$ (recall that $N_1 = N_2 = N$ in the simulations).

Similarly, because the variability of the coalescence time increases (also slowly) with the number of lineages, the range of the inner 95% of the distribution of $T_{MRCA} - T_D$ decreases slowly as $T_D/N$ increases. This range eventually reaches the asymptotic interval given by the quantiles of the exponential distribution with mean $2N$. These asymptotic values are $(-2N \ln 0.975, -2N \ln 0.025)$, or $(0.051N, 7.38N)$. Had we used samples of size 1 for the plots instead of 50, the 2.5th percentile, mean, and 97.5th percentile would have been $0.051N$, $2N$, and $7.38N$, respectively, for all values of the scaled divergence time $T_D/N$.

Note that because we performed only 10,000 simulations at each value of $T_D/N$, the upper 97.5th percentile did not actually decline monotonically. Although the theory predicts a monotonic decline, this choppiness showcases the fact that the $T_{MRCA} - T_D$ distribution has a long right-hand tail. The difference between the 97.5th percentile and the mean is considerably higher than the corresponding difference of the 2.5th percentile and the mean.

When the ratio of $T_{MRCA} - T_D$ to divergence time $T_D$ is considered, the excess $T_{MRCA} - T_D$ dwarfs the divergence time for small values of $T_D/N$, eventually declining to a negligible fraction of $T_D$ (Figure 9.6 and see also Figure 2 of Edwards and Beerli, 2000). This observation has important consequences for the interpretation of inferred $T_{MRCA}$ values. Consider diver-

gence times typical of ancient species divergences. For ancient divergences, the divergence time likely exceeds the effective population size by a large amount. Under these conditions, the relative value of $T_{MRCA} - T_D$ to $T_D$ is very small, and numerical values of the coalescence time can be used as good surrogates for the divergence time. On the other hand, for very recent divergences, $T_{MRCA} - T_D$ is quite large, and coalescence happens much longer ago in the past than population divergence. In this case, divergence times are very different from coalescence times. Thus, for any inference about population history from coalescence times, *it is essential to know whether the data fall into the regime of large, intermediate, or small divergence time compared to population size.* We will return to this point further using examples in Section 9.5.

### 9.4.4 Constant population size with migration

The behavior of $T_{MRCA} - T_D$ is more complex when migration takes place after the population divergence, because it is possible for $T_{MRCA} - T_D$ to be negative, as we will see. The distribution of $T_{MRCA} - T_D$ is difficult to determine in models with migration; however, a slight adaptation of the calculations of Wakeley (1996b) gives the distribution for the case in which samples have size 1 in both populations. Wakeley (1996b) describes a three-state Markov chain for this situation, in which the three states describe the positions of the two lineages ancestral to the sampled lineages. At any given time in the past, the two ancestral lineages can be in the same population, in opposite populations, or they can have coalesced at a common ancestor. The probabililty that coalescence has occurred more recently than a given point in time, that is, the probability that the Markov chain has been absorbed more recently than that time point, can be computed by inserting the quantities in the appendix of Wakeley (1996b) into his equation 3 and simplifying considerably. If $-T_D \leq t \leq 0$ is measured in generations and $\gamma = Nm$, then

$$\begin{aligned} Pr(T_{MRCA} - T_D \leq t) &= 1 - \frac{e^{-(1+4\gamma)\frac{T_D+t}{2N}}}{\sqrt{1+16\gamma^2}} \left[ \sqrt{1+16\gamma^2} \cosh\left( \frac{T_D+t}{2N} \sqrt{1+16\gamma^2} \right) + \right. \\ &\quad \left. (1+4\gamma) \sinh\left( \frac{T_D+t}{2N} \sqrt{1+16\gamma^2} \right) \right]. \end{aligned} \tag{9.7}$$

When $t = 0$, equation 9.7 gives the probability that coalescence occurs in the two-population phase. For $t > 0$, the time to coalescence is simply the familiar time to coalescence of two lineages in the one-population phase. If coalescence occurs in the one-population phase, recall that the time to coalescence during this phase is exponentially distributed with mean $2N$. Therefore, for $t > 0$, we obtain:

$$Pr(T_{MRCA} - T_D \leq t) = Pr(T_{MRCA} - T_D \leq 0) +$$

$$[1 - Pr(T_{MRCA} - T_D \leq 0)][1 - e^{-t/(2N)}]. \qquad (9.8)$$

It is clear that in the case of $\gamma = 0$, equations 9.7 and 9.8 reduce to the familiar exponential distribution of waiting time to coalescence characteristic of the model with no migration.

A generalization of Wakeley's (1996b) solution to include larger sample sizes appears to involve rather tedious calculations (because the number of states in the Markov chain grows with the square of the number of sampled lineages), so we rely exclusively on special cases and a series of simulations to explore the properties of $T_{MRCA} - T_D$. For this more general model including migration, the parameters that influence $T_{MRCA} - T_D$ are the scaled divergence time ($T_D/N$), the total amount of migration in each direction per generation ($Nm$, where $m$ is the fraction of each population that migrates to the other population in each generation: $m_{12} = m_{21} = m$), and the sample size in each population ($n$). Using these three independent variables, we again consider special cases.

**Small values of $T_D/N$:** When $T_D = 0$, the model simplifies to the classical one-population model with constant size, discussed in Section 9.4.3 and by many authors (e.g. Tajima, 1983; Hudson, 1990; Tavaré *et al.*, 1997; Nordborg, 2001a). The expected value of $T_{MRCA} - T_D$ is then $2(1 - \frac{1}{n_1+n_2})(N_1 + N_2)$, as in equation 9.5. For small but nonzero values of $T_D/N$, this result holds approximately.

**Large values of $T_D/N$ for a fixed nonzero value of $Nm$:** If we hold $Nm$ constant and nonzero and we let $T_D/N$ increase without bound, the model approaches the classical island migration model (e.g. Nath and Griffiths, 1993; Wakeley, 1998). In this limit, two populations have been separated for all time but they exchange migrants in each generation (Figure 9.7A). The behavior of $T_{MRCA} - T_D$ is very different in this case from the behavior in the case with no migration. Because $T_D/N$ is extremely large, it is almost certain that coalescence happens in the two-population phase. Thus, the fact that the two populations derive from an ancestral population is irrelevant to the coalescence of sampled lineages.

Although it is difficult to compute the distribution of $T_{MRCA}$ (and hence $T_{MRCA} - T_D$) in the island model for a general sample size, the computation of $E[T_{MRCA} - T_D]$ is feasible for small samples (e.g. Notohara, 1990; Nath and Griffiths, 1993; Wakeley, 1998). For two lineages sampled from the same population, the result is independent of the (nonzero) migration rate: $E(T_{MRCA}) = 2N$. If one lineage is sampled from each population, then $E(T_{MRCA}) = (2 + \frac{1}{Nm})N$. Note that a discontinuity occurs at $Nm = 0$, where the mean waiting time to coalescence is $N$ for two lineages in the same population and $\infty$ for two lineages in opposite populations (e.g. Nath and Griffiths, 1993).

**Small values of $Nm$ for a fixed value of $T_D/N$:** Of course, when $Nm = 0$ this reduces to the no-migration model of divergence previously discussed in Section 9.4.3. For nonzero but small $Nm$, the no-migration model is essentially accurate unless $T_D/N$ is exceedingly large (then the previous case applies).

**Large values of $Nm$ for a fixed value of $T_D/N$:** As is true in the case of extremely large $T_D/N$ for fixed nonzero $Nm$, for large amounts of total migration ($Nm$), coalescence generally occurs in the two-population phase, after the divergence (Figure 9.7B). The classical island model applies, except if $T_D/N$ is very small (see the case of small $T_D/N$).

**General case - mean value of $T_{MRCA} - T_D$:** The mean value of $T_{MRCA} - T_D$ decreases monotonically as migration rate increases and as $T_D/N$ increases (Figure 9.8; see also Figure 2 of Wakeley [1996b]). This contrasts with the model with no migration, in which this expectation approaches the constant $2N$.

The reasons for this behavior are clear. First, with samples of size 1 and a migration rate of 0, $E[T_{MRCA} - T_D] = 2N$ (the analogous result is $4N$ for large samples), as in Section 9.4.3. As migration increases for a fixed divergence time, more and more coalescences take place in the two-population phase, so that $T_{MRCA} < T_D$ in more and more cases. Asymptotically, in the extreme case of complete panmixia, $E[T_{MRCA} - T_D] = 2N - T_D$ (analogously, $4N - T_D$ for large samples).

Now as $T_D$ increases, there is more time for all lineages to migrate into the same population and for coalescence to take place during the two-population phase. Eventually, for large values of $T_D$, the results from the classical island model are applicable because coalescence always takes place in the two-population phase. Thus, we have (for samples of size 1) $E[T_{MRCA} - T_D] = (2 + \frac{1}{Nm})N - T_D$ at large values of the divergence time. For sufficiently large migration rates, the mean is the same order of magnitude as the divergence time itself, but negative, so that the coalescence time is a horrible estimator of $T_D$. Of course, with a *very* large amount of migration, the descendant populations cannot really be considered "diverged" and it may not make sense to discuss divergence times at all.

Thus, with two exceptions, traveling along any trajectory of increasing $Nm$ or $T_D/N$, the distribution of $T_{MRCA} - T_D$ shifts downward. The exceptions are the case of $Nm = 0$, for which $E[T_{MRCA} - T_D]$ asymptotically equals $2N$ with increasing $T_D/N$ (it always equals $2N$ if sample sizes are 1), and the line $T_D/N = 0$, along which the mean is $4N(1 - \frac{1}{2n})$ independent of the value of $Nm$.

The downward shift in the mean of $T_{MRCA} - T_D$ reflects the increasing probability that

coalescence takes place more recently than divergence (Figure 9.9). With the same exceptions of the $Nm = 0$ and $T_D/N = 0$ curves, $Pr(T_{MRCA} - T_D < 0)$ increases on any trajectory of increasing $Nm$ or $T_D$. For the six trajectories of increasing $T_D$ with $Nm > 0$ (Figure 9.9), all graphs eventually reach a probability of 1. We note that this probability can help determine if the asymptotic results from the classical island model are appropriate. If $Nm$ and $T_D/N$ are sufficiently large that all coalescences happen in the two-population phase (so that $Pr(T_{MRCA} - T_D < 0)$ is near one), then it is irrelevant that the populations had an ancestral population. The expected time to coalescence for a sample of size 1 from each population is then $(2 + \frac{1}{Nm})N$.

We note that the effect of sample size on the mean $T_{MRCA} - T_D$ is largely unimportant, but not negligible. Looking backwards in time, small samples require only a single migration event followed by a coalescence in order to reach the common ancestor. However, loosely speaking, large samples need an initial period for all the lineages to coalesce separately in each population (possibly including some migrations), and an eventual migration followed by coalescence of the last two lineages. Thus, the initial mean $T_{MRCA} - T_D$ is higher for the large sample (consider corresponding curves in Figures 9.8A and 9.8B) and the initial $Pr(T_{MRCA} < T_D)$ is smaller (Figures 9.9A and 9.9B).

**General case - shape of the $T_{MRCA} - T_D$ distribution:** So far we have found one major difference from the model with no migration: instead of reaching a constant value of $2N$ (for samples of size 1) as the scaled divergence time increases, the mean value of $T_{MRCA} - T_D$ declines to be comparable with $-T_D$. We now turn to a second major difference in the shape of the distribution of $T_{MRCA} - T_D$.

Suppose $T_D/N$ is fairly large (to some extent, the effect we are about to describe still happens for smaller $T_D/N$). We have already considered the cases of a small amount of migration and a large amount of migration. In the case of little migration, the ancestral lineages almost always coalesce in the one-population phase, with most coalescences happening very quickly as soon as the ancestral phase is entered. In the case of a lot of migration, coalescences happen quickly in the two-population phase. Thus, there must be an intermediate amount of migration for which sizeable fractions of the coalescences happen in each of the two phases. In this situation, the distribution of coalescence times is bimodal, with the smaller mode being the modal time of the two-population phase coalescences, and the larger mode equaling the modal time for one-population phase coalescences.

In fact, this qualitative shift in the distribution of $T_{MRCA} - T_D$ was observed in our simulations (Figure 9.10). For Figures 9.10A-C, which used samples of size 1, the fractions of

coalescences that happened in the two-population phase were 2%, 77%, and 99%, respectively; for Figures 9.10D-F, which used samples of size 50, the corresponding values were 1%, 68%, and 98%. Figures 9.10B and 9.10E showcase the bimodalilty of the distribution: some coalescences take place in each phase; if coalescence is not complete by the end of the two-population phase, it usually happens quickly as soon as the one-population phase begins.

Interestingly, for larger samples, the effect is not as pronounced: all other parameters being equal, the difference between the two modes of the bimodal $T_{MRCA} - T_D$ distribution is smaller with a large sample. This is again due to the fact that while coalescence can happen in a small sample very quickly, it takes more time for coalescence to occur in a large sample. Thus, the descendant phase mode in the larger sample is closer to the divergence time. The cases of small migration ($Nm = 0.001$) are essentially identical for small and large samples, because the large sample has coalesced to a small number of lineages by the time of divergence. The cases of intermediate migration ($Nm = 0.1$) and large migration ($Nm = 10$) are affected by sample size due to the sizeable waiting time to coalescence required by the large sample but not by the small sample.

Because corresponding bimodal distributions of $T_{MRCA} - T_D$ are more spread out for small samples than for large samples, it makes sense that the variance of $T_{MRCA} - T_D$ achieves higher values for a sample of size 1 than for a sample of size 50 (Figure 9.11 and see also Figure 2 of Wakeley [1996b]). This contrasts with the model with no migration, in which the variance of $T_{MRCA} - T_D$ increases (slowly) with sample size (recall equation 9.6). However, because the distribution of $T_{MRCA} - T_D$ does have the unusual bimodality for some values of $Nm$, we suggest that the variance of the distribution is perhaps not the most appropriate summary statistic for this random variable.

To summarize, the new phenomena introduced by permitting migration after the divergence include: (1) coalescence happens later than divergence some of the time, and for sufficiently large migration rates and divergence times, coalescence *nearly always* happens after divergence; (2) for extreme values of the migration rate and scaled divergence time, the distribution of $T_{MRCA} - T_D$ is the same as in the classical island model, the divergence model with no migration of Section 9.4.3, or the classical one-population model; (3) for certain intermediate values of the migration rate, the distribution of $T_{MRCA} - T_D$ exhibits an emergent bimodality due to a combination of behaviors from the classical island model and the two-population divergence model with no migration; (4) this bimodality is more pronounced with smaller samples. With these added phenomena, relating empirically obtained coalescence times to divergence times involves an additional level of complexity when

migration is allowed.

## 9.5 Studies of human populations

Of course, a main goal of this work is to understand the relationship between $T_{MRCA}$ and $T_D$ not only in abstract population genetic models, but also in real populations. Thus, if we select a demographic model and a specific population divergence, and if we have some limited knowledge about the parameters of the model, then we should be able to approximate the degree to which $T_{MRCA}$ differs from $T_D$. We should also be able to decide if this excess is too great for $T_{MRCA}$ of a gene to be a useful quantity in describing particular past demographic events.

We now consider some examples from human populations. Because coalescence times for many human genes have been studied, and because many divergences have been reasonably well-characterized, these examples provide the best demonstration of the utility of a joint consideration of $T_{MRCA}$ and $T_D$. In each case, the model considered is clearly a crude oversimplification of the actual population histories, especially because genetic substructure within each member of a population pair would certainly affect $T_{MRCA} - T_D$. However, because we are mainly interested in a qualitative understanding of $T_{MRCA} - T_D$ rather than careful statistical inference, this crude model is still appropriate.

### 9.5.1 The divergence of humans and chimpanzees

There is considerable uncertainty about the specific time of the human-chimpanzee divergence, the ancestral population size, and the appropriate population models that describe both species since the time of divergence (e.g. Ruvolo, 1997; Satta *et al.*, 2000; Chen and Li, 2001). Because the event was a speciation, a divergence model with no migration accurately represents the human-chimpanzee split. Here we make the simplifying assumption that both species have been constant in size since the separation. Suppose that the divergence took place about 4 to 7 million years ago, and that an effective population size of 5,000 to 100,000 in each species describes most human and chimpanzee genes. Assuming a generation time of 15 to 25 years, we obtain a value between 1.6 and 93.3 for $T_D/N$. Using Figures 9.5 and 9.6 and these values, we expect 95% of values of $T_{MRCA} - T_D$ to be between about $0.05N$ and $8N$. Plugging in the lower bounds for the population size and generation times, we get a range of 3,750 to 20,000,000 years for the value of $T_{MRCA} - T_D$. Using values of 20,000 for the population size, 20 for the generation time, and $2N$ for the mean of $T_{MRCA} - T_D$, we should expect the mode of the distribution to be around 800,000 years: for comparisons of

humans and chimpanzees, most coalescence dates should be near 800,000 years earlier than the time of divergence.

Thus, most genes should coalesce to a common ancestor not long before the human-chimpanzee split. However, it should not be surprising that some small number of genes will have ancient coalescence times well before the separation. A substantial fraction of genes, about 40%, have such deep coalescences that either chimpanzee or human sequences coalesce with gorilla sequences more recently than the coalescences of human and chimpanzee sequences coalesce together (e.g. Satta *et al.*, 2000; Chen and Li, 2001). Although orangutan sequences are generally treated as outgroups in genetic comparisons of humans, chimpanzees, and gorillas, for 7 of 53 genomic segments, at least one of three pairwise distances involving orangutans was smaller than the three distances not involving orangutans (Table 1 of Chen and Li [2001]). Because the separation of the orangutan lineage likely occurred only 5 to 12 million years before the human-chimpanzee split (using the estimate of 12-16 million years for the divergence time of orangutans and the human-chimpanzee-gorilla clade, as was done by Chen and Li [2001]), and because these values are well within our rough range for $T_{MRCA} - T_D$, it seems likely that many genes will eventually be discovered for which human and orangutan sequences or chimpanzee and orangutan sequences coalesce more recently than do human and chimpanzee sequences.

### 9.5.2 The divergence of modern humans and Neanderthals

For this example, we also employ a divergence model with no migration after the divergence of modern human and Neanderthal populations. This model is certainly debatable, and others might easily be considered (Wall, 2000; Nordborg, 2001b). We postulate a divergence time of 250,000 to 1,000,000 years, equal effective population sizes for modern humans and Neanderthals of 5,000 to 20,000, and a generation time of 20 to 25 years. Note that this approach treats Neanderthals as still in existence; this assumption is acceptable because the time since their extinction (assume for now that modern humans have no Neanderthal ancestry) is small compared to the divergence time of Neanderthals and modern humans. With these values, we then obtain for $T_D/N$ a value of 0.5 to 10. Figure 9.5 suggests a value of $T_{MRCA} - T_D$ between $0.05N$ and about $9N$, corresponding to a range of 5,000 to 4,500,000 for the value of $T_{MRCA} - T_D$. A guess for the center of the range might be about $2.5N$ for $T_{MRCA} - T_D$, a generation time of 20 years, and an effective population size of 10,000. These values suggest that coalescence times of autosomal loci should be about 500,000 years more than the divergence time for humans and Neanderthals.

Unlike the case of the human-chimpanzee split, for which most values of $T_{MRCA} - T_D$ should be fairly small compared to the divergence time of about 4 to 7 million years, the calculations suggest that most values of Neanderthal-modern human $T_{MRCA} - T_D$ should be comparable to the divergence time of Neanderthals and modern humans. Although great effort has been expended to obtain coalescence times for modern human and Neanderthal mitochondrial DNA, this coalescence time alone, estimated at 465,000 years (Krings *et al.*, 1999), will be unable to resolve controversies regarding the time of the divergence between modern humans and Neanderthals. Although the mitochondrial genome has a smaller population size than do autosomal loci and therefore its coalescence time should be closer to population divergence times than those of autosomal loci by a factor of four, the 465,000 figure is consistent with a recent divergence of 250,000 years ago. It could potentially be consistent with an ancient divergence at 1,000,000 years ago if ancient admixture occurred between humans and Neanderthals. For example, consider the results of the migration model in Figure 9.8, and take the lower estimate of 5,000 for $N$. Then with a divergence time of 1,000,000 years and a generation time of 20, $T_D/N$ equals 10. If 465,000 years is taken as the expected value of $T_{MRCA}$, then the scaled value of $E[T_{MRCA} - T_D]$ is -5.35. With $Nm = 1$, the expected value of $T_{MRCA} - T_D$ at $T_D/N = 10$ is about -5.5 (Figure 9.8B; a smaller migration rate between $Nm = 0.1$ and $Nm = 1$ would give an analogous result for the plot with samples of size 1 in Figure 9.8A). Thus, with the limited genetic data currently available, it is not yet possible to rule out a situation of ancient divergence followed by some admixture between Neanderthals and modern humans.

### 9.5.3 The divergence of African and non-African populations

Regardless of the degree to which ancient human populations interbred and the timing of modern human origins, the fact that ancient hominid fossils derive from Africa make it likely that the most ancient divergence among the ancestors of modern humans took place between ancestors of some African populations and ancestors of all other populations (for more details, see Cavalli-Sforza *et al.*, 1994; Takahata, 1995; Jorde *et al.*, 1998; Mountain, 1998). First we consider a model of constant-sized populations without migration. Supposing a time of divergence of 50,000 to 100,000 years ago between African and non-African populations, an effective population size for each of 5,000 to 20,000, and a generation time of 20 to 25 years. We obtain $T_D/N$ between .1 and 1. Then $T_{MRCA} - T_D$ should be between about $N$ and $9.5N$, centered near $3N$ (Figure 9.5). We then might guess that $T_{MRCA} - T_D$ would be between 100,000 and 4,750,000 years, with a mode near 600,000 years (using 10,000 and 20 for the

central effective population size and generation time, respectively). Thus, we should expect coalescence dates to be hundreds of thousands of years before the divergence of African and non-African populations. For the Y chromosome and the mitochondrial genome, which each have only a quarter the population size of an average autosomal gene, the value of $T_{MRCA} - T_D$ should be about a quarter as large as for autosomal genes.

In fact, these predictions are somewhat consistent with observations. First, many estimated $T_{MRCA}$ values are close to 650,000-700,000 years (Takahata *et al.*, 2001), the approximate predicted average value of $T_{MRCA}$, while the coalescence times for the Y-chromosome and the mitochondrial genome are much smaller (Tang *et al.*, 2001). The model suggests that for nearly all loci, all humans will have coalesced more recently than the human-chimpanzee divergence. Thus, it is no surprise that in studies of many genes, human sequences are always more closely related to each other than any human sequence is to a chimpanzee sequence. However, the high end of the predicted human coalescence time distribution is very close to the range of possible human-chimpanzee divergence times. Thus, when comprehensive analyses of human and chimpanzee genomes have been completed, regions may be found for which some humans coalesce with some chimpanzees more recently than all humans coalesce together.

If we consider the model with migration, the results are similar. A recent study that inferred migration rates between African and non-African populations using an island model obtained $Nm$ values between 1 and 7 for the Y chromosome and the mitochondrial genome (Tang *et al.*, 2001). This study used the polymorphism in approximately 5 kilobases of Y chromosome sequence in about 100 geographically diverse males and about 1.5 kilobases of mitochondrial DNA sequence in about 180 geographically diverse individuals, together with the Monte Carlo likelihood procedure of Beerli and Felsenstein (1999), in order to estimate migration rates. With the same range for $T_D/N$ of 0.1 to 1, however, the divergence time is too small for this level migration to seriously affect the value of $T_{MRCA} - T_D$ when a large sample is used (Figure 9.8B). With these values of $T_D/N$, it is highly unlikely that coalescence is more recent than divergence (Figure 9.9B).

### 9.5.4 The divergence of Native Americans and Asian populations

Though the timing and number of distinct migration events are uncertain, most theories about the arrival of Asian ancestors in the America place the first crossing of people into the New World between 12,000 and 40,000 years ago (e.g. Ward *et al.*, 1991; Horai *et al.*, 1993; Cavalli-Sforza *et al.*, 1994; Crawford, 1998; Ward, 1999). If we suppose an American-Asian

divergence time of 12,000 to 40,000 years, an effective population size of 1,000 to 10,000 for Asians and Americans, and a generation time of 20 to 25 years, we obtain $T_D/N$ between 0.05 and 2 and $T_{MRCA} - T_D$ between $0.05N$ and $9.5N$, with mode near $3N$. This suggests a range for $T_{MRCA} - T_D$ between 2000 and 2,375,000 years with mode near 120,000 years. As in the previous case, migration has no major effect on $T_{MRCA} - T_D$ for scaled divergence times less than 2 (Figure 9.8A), so the predicted coalescence times under the model without migration are likely to be reasonable.

As in the other examples, there is no reason to expect coalescence times to be remotely close to the divergence time. For example, ancient estimated coalescence times for Native Americans need not imply an older divergence time than the classical estimates of around 15,000 years ago, though they are certainly consistent with older divergences. Calculations of coalescence times at many loci will be needed in order to determine the divergence of Americans and Asians. Because the number of migratory events that contributed to the Native American population is not known, a model that includes a variable number of migration pulses may be more appropriate than the constant migration model used here.

## 9.6 Extensions to three or more populations: gene trees and species trees

We note that the relationship of divergence times and coalescence times has also arisen in the context of comparing gene trees and species trees. It has long been known that genealogical history of a particular unduplicated gene in diploid species need not match the history of species divergences (e.g. Pamilo and Nei, 1988; Takahata, 1989; Maddison, 1997; Nordborg, 2001a). When considering three or more populations, this discordance arises because ancestral lineages from the two most recently diverged populations coalesce so far back in time that they have the opportunity to coalesce with lineages from additional populations (as described by Takahata and Satta in Chapter 5, the phenomenon of discordance due to "deep coalescence" can be harnessed to provide estimates of ancestral population sizes). Alternatively, migration between two highly diverged populations may cause the coalescence of their lineages to be more recent than coalescences of less diverged populations that have not exchanged migrants.

Thus, the addition of more populations into the divergence framework will further complicate the relationships between $T_{MRCA}$ and each of the pairwise values of $T_D$. However, simulations similar to ours could be performed with three or more populations, using any of a variety of population models. Divergence times could be held constant as independent

variables, and we could explore the effect of divergence times, population sizes, and other parameters such as migration rates and population growth rates on both $T_{MRCA} - T_D$ and on the probability that the gene tree is concordant with the species tree. For a three-population model with constant-size populations and no migration among descendant populations, this problem has been studied in detail (e.g. Pamilo and Nei, 1988; Takahata, 1989). Takahata (1995) also considered the value of $T_{MRCA}$ using an ancestral population that experienced a single multifurcation event, after which migration occurred among descendant populations.

It would be interesting to consider the effect of migration on $T_{MRCA} - T_D$ and on the gene tree/population tree concordance probability in a three-population model. As with the two-population model, migration will decrease $T_{MRCA} - T_D$. Depending on relative migration rates between pairs of populations, the concordance probability will be increased or decreased. For example, if migration has occurred only between the two most recently diverged populations, the gene tree is more likely to be concordant to the population tree than in a model without migration. However, if migration has occurred only between two anciently diverged populations and not between recently diverged populations, the gene tree and population tree are less likely to be concordant than in the case of no migration. A simulation framework such as that used in Section 9.4, or the analytical methods of Wakeley (1996b), could potentially explore this issue in greater detail.

## 9.7 Conclusions

In the four examples in Section 9.5, we observed several different qualitative regimes. For the human-chimpanzee divergence, $T_{MRCA} - T_D$ is small compared to $T_D$, so that coalescence takes place just a little earlier than divergence. For the modern human-Neanderthal divergence $T_{MRCA} - T_D$ and $T_D$ are expected to be comparable in magnitude. For the separation of African and non-African groups and for the American-Asian divergence, $T_{MRCA} - T_D$ is expected to be *much* larger in general than $T_D$. If one is interested in studying any of these four example problems in greater detail, one might choose a more sophisticated model that includes changes in population size (e.g. Marjoram and Donnelly, 1997) or different kinds of migration and mixture after divergence (e.g. Nordborg, 2001b).

In any case, for sufficiently large divergence times compared to population sizes, if migration between populations is sufficiently small, coalescence times are useful approximations for divergence times. Otherwise, they are not highly informative. We suggest (based on Figures 9.5 and 9.6) that for values of $T_D/N$ larger than 5, coalescence times and divergence times will generally be reasonably close in constant-size models without migration. Even so,

the variance of $T_{MRCA} - T_D$ is sufficiently large that individual genes may exhibit coalescence times much larger than the divergence time. For smaller values of $T_D/N$, coalescence times must be very carefully interpreted if they are to provide evidence regarding demographic hypotheses.

In models with migration, coalescence times are even less informative. Coalescence can be more ancient *or* more recent than divergence, so that if migration is suspected of occurring, it is not appropriate to treat coalescence times as the upper bounds for divergence times. A large amount of migration will even allow $T_{MRCA} - T_D$ to be negative with high probability.

Interestingly, migration may also cause the distribution of $T_{MRCA} - T_D$ to be bimodal. Intuitively, this makes sense: by chance, for some genes, all gene copies in one population will descend from a migrant from the other population; for others, not all copies of the gene will be traceable to migrants, and coalescence will occur among ancestors that predate population divergence. However, by considering many genes, we can use this bimodality to our advantage: if we detect such a bimodal distribution of coalescence times across many genes, we will have evidence for migration between two descendant populations. So far, too few genes have been studied with respect to any particular population divergence in order to test for bimodality of the $T_{MRCA} - T_D$ distribution. As more data become available, however, the bimodal distribution of $T_{MRCA} - T_D$ across genes may offer a basis for a solution of the problem of detecting ancient gene transfer between populations.

Finally, the distribution of $T_{MRCA} - T_D$ under the assumption that a gene is neutral can be used as a null distribution in order to test specific genes for causal connections to population or species divergences. Consider as an example the ancient human domestication of a wild plant. Genotypes at a random gene would only have become different in wild and cultivated varieties slowly, by genetic drift, as humans performed artificial selection. If ancient humans selected for specific single-gene traits, then the differentiation of these causal genes across wild and cultivated plants would have occurred rather rapidly as soon as selection began. Thus, due to selection pressure, the coalescence times across wild and cultivated varieties for the genes causally related to the domestication event would likely be quite close to the divergence time of wild and cultivated populations. The coalescence times of random genes, on the other hand, would likely be much older, because their differentiation was only due to genetic drift. Similar reasoning applies to divergences of species (see Ting *et al.*, 2000): since their differentiation coincides with species divergences, "speciation genes" should have concordant gene trees with the species tree more often than do random genes. Thus, an approach for studying causal factors linked to speciation, domestication, or population divergences might be based on distributions of $T_{MRCA} - T_D$ calculated across entire genomes.

## Acknowledgments

We thank Russell Thomson for many useful suggestions and for reading an earlier draft of the paper. We also thank Marissa Baskett, Aaron Hirsh, Joanna Mountain, Rasmus Nielsen and Hua Tang for comments, and Dylan Schwilk for helpful conversations. N.A.R. is supported by a Program in Mathematics and Molecular Biology Graduate Fellowship. This work was supported in part by NIH grant GM28016 to M.W.F.

Bahlo M and Griffiths RC (2000). Inferences from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79-95.

Beerli P and Felsenstein J (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763-773.

Cavalli-Sforza LL, Menozzi P, and Piazza A (1994). *The history and geography of human genes*. Princeton University Press, Princeton.

Chen F-C and Li W-H (2001). Genomic divergences between humans and other Hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, **68**, 444-456.

Crawford MH (1998). *The origins of Native Americans*. Cambridge University Press, Cambridge.

Donnelly P (1996). Interpreting genetic variability: the effects of shared evolutionary history. In *Variation in the human genome*, pp. 25-40. Wiley, Chichester, UK.

Donnelly P and Tavaré S (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, **29**, 401-421.

Edwards SV and Beerli P (2000). Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839-1854.

Feldman MW, Kumm J, and Pritchard JK (1999). Mutation and migration in models of microsatellite evolution. In DB Goldstein and C Schlötterer, eds. *Microsatellites: evolution and applications*, pp. 98-115. Oxford University Press, Oxford.

Fu Y-X and Li W-H (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution*, **14**, 195-199.

Fu Y-X and Li W-H (1999). Coalescing into the 21st Century: an overview and prospects of coalescent theory. *Theoretical Population Biology*, **56**, 1-10.

Gaggiotti OE and Excoffier L (2000). A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proceedings of the Royal Society of London Series B - Biological Sciences*, **267**, 81-87.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, and Feldman MW (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences USA*, **92**, 6723-6727.

Griffiths RC and Tavaré S (1994a). Ancestral inference in population genetics. *Statistical Science*, **9**, 307-319.

Griffiths RC and Tavaré S (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, **344**, 403-410.

Harding RM (1996). Using the coalescent to interepret gene trees. In AJ Boyce and CGN Mascie-Taylor, eds. *Molecular biology and human diversity*, pp. 63-80. Cambridge University Press, Cambridge.

Horai S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, and Tajima K (1993). Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Molecular Biology and Evolution*, **10**, 23-47.

Hudson RR (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1-44.

Jorde LB, Bamshad M, and Rogers AR (1998). Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *BioEssays*, **20**, 126-136.

Krings M, Geisert H, Schmitz RW, Krainitzki H, and Pääbo S (1999). DNA sequence of the mitochondrial hypervariable region II from the Neanderthal type specimen. *Proceedings of the National Academy of Sciences USA*, **96**, 5581-5585.

Kuhner MK, Yamato J, and Felsenstein J (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421-1430.

Kuhner MK, Yamato J, and Felsenstein J (1997). Applications of Metropolis-Hastings genealogy sampling. In P Donnelly and S Tavaré, eds. *Progress in population genetics and human evolution*, pp. 183-192. Springer-Verlag, New York.

Kuhner MK, Yamato J, and Felsenstein J (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-434.

Maddison WP (1997). Gene trees in species trees. *Systematic Biology*, **46**, 523-536.

Marjoram P and Donnelly P (1997). Human demography and the sime since mitochondrial Eve. In P Donnelly and S Tavaré, eds. *Progress in population genetics and human evolution*, pp. 107-131. Springer-Verlag, New York.

Mountain JL (1998). Molecular evolution and modern human origins. *Evolutionary Anthropology*, **7**, 21-37.

Nath HB and Griffiths RC (1993). The coalescent in two colonies with symmetric migration. *Journal of Mathematical Biology*, **31**, 841-852.

Nei M (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Nielsen R (1998). Maximum likelihood estimation of population divergence times and population phylogenies under the infinites sites model. *Theoretical Population Biology*, **53**, 143-151.

Nielsen R, Mountain JL, Huelsenbeck JP, and Slatkin M (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*, **52**, 669-677.

Nielsen R and Slatkin M (2000). Likelihood analysis of ongoing gene flow and historical association. *Evolution*, **54**, 44-50.

Nielsen R and Wakeley J (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885-896.

Nordborg M (2001a). Coalescent theory. In DJ Balding, C Cannings, and M Bishop, eds. *Handbook of statistical genetics*, pp. 179-212. Wiley, Chichester, UK.

Nordborg M (2001b). On detecting ancient admixture. In P Donnelly, ed. *Genes, fossils, and behaviour: an integrated approach to human evolution.* IOS Press, Amsterdam.

Notohara M (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, **29**, 59-75.

Pamilo P and Nei M (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, **5**, 568-583.

Pritchard JK, Seielstad MT, Pérez-Lezaun A, and Feldman MW (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791-1798.

Ruvolo M (1996). A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. *Molecular Phylogenetics and Evolution*, **5**, 202-219.

Ruvolo M (1997). Molecular phylogeny of the Hominoids: inferences from multiple independent DNA sequence data sets. *Molecular Biology and Evolution*, **14**, 248-265.

Satta Y, Klein J, and Takahata N (2000). DNA archives and our nearest relative: the trichotomy problem revisited. *Molecular Phylogenetics and Evolution*, **14**, 259-275.

Saunders IW, Tavaré S, and Watterson GA (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability*, **16**, 471-491.

Slatkin M and Hudson RR (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555-562.

Stephens M (2001). Inference under the coalescent. In DJ Balding, C Cannings, and M Bishop, eds. *Handbook of statistical genetics*, pp. 213-238.

Stephens M and Donnelly P (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, **62**, 605-635.

Stumpf MPH and Goldstein DB (2001). Genealogical and evolutionary inference with the human Y chromosome. *Science*, **291**, 1738-1742.

Tajima F (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Takahata N (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957-966.

Takahata N (1995). A genetic perspective on the origin and history of humans. *Annual Review of Ecology and Systematics*, **26**, 343-372.

Takahata N, Lee S-H, and Satta Y (2001). Testing multiregionality of modern human origins. *Molecular Biology and Evolution*, **18**, 172-183.

Takahata N and Nei M (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325-344.

Takahata N and Satta Y (1997). Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proceedings of the National Academy of Sciences USA*, **94**, 4811-4815.

Takahata N and Slatkin M (1990). Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology*, **38**, 331-350.

Tang H, Thomson R, Cavalli-Sforza LL, Shen P, Oefner P, and Feldman MW (2001). Sex differences in demographic histories of humans. In preparation.

Tavaré S (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, **26**, 119-164.

Tavaré S, Balding DJ, Griffiths RC, and Donnelly P (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505-518.

Thomson R, Pritchard JK, Shen P, Oefner PJ, and Feldman MW (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proceedings of the National Academy of Sciences USA*, **97**, 7360-7365.

Ting C-T, Tsaur S-C, and Wu C-I (2000). The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of Sciences USA*, **97**, 5313-5316.

Underhill PA, Shen P, Lin AA, *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genetics*, **26**, 358-361.

Wakeley J (1996a). Distinguishing migration from isolation using the variation of pairwise differences. *Theoretical Population Biology*, **49**, 369-386.

Wakeley J (1996b). Pairwise differences under a general model of population subdivision. *Journal of Genetics*, **75**, 81-89.

Wakeley J (1998). Segregating sites in Wright's island model. *Theoretical Population Biology*, **53**, 166-174.

Wakeley J and Hey J (1997). Estimating ancestral population parameters. *Genetics*, **145**, 847-855.

Wall JD (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, **154**, 1271-1279.

Walsh B (2001). Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics*, **158**, 897-912.

Ward R (1999). Language and genes in the Americas. In B Sykes, ed. *The human inheritance: genes, language, and evolution*, pp. 135-157. Oxford University Press, Oxford.

Ward RH, Frazier BL, Dew-Jager K, and Pääbo S (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences USA*, **88**, 8720-8724.

Watterson GA (1982). Mutant substitutions at linked nucleotide sites. *Advances in Applied Probability*, **14**, 206-224.

Watterson GA (1985). Estimating species divergence times using multi-locus data. In T Ohta and K Aoki, eds. *Population genetics and molecular evolution*, pp. 163-183. Japan Scientific Society Press, Tokyo.

Wilson IJ and Balding DJ (1998). Genealogical inference from microsatellite data. *Genetics*, **150**, 499-510.

Wilson IJ, Weale ME, and Balding DJ (2001). Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. *Journal of the Royal Statistical Society Series A - Statistics in Society*, in press.

Zhivotovsky LA (2001). Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Molecular Biology and Evolution*, **18**, 700-709.

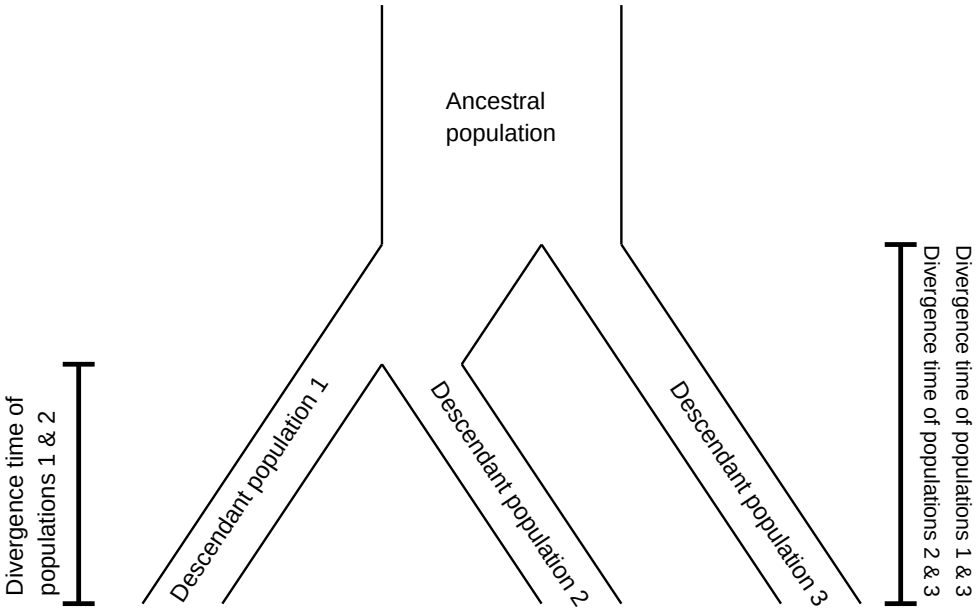



Figure 9.1: A divergence scheme that has resulted in three descendant populations.

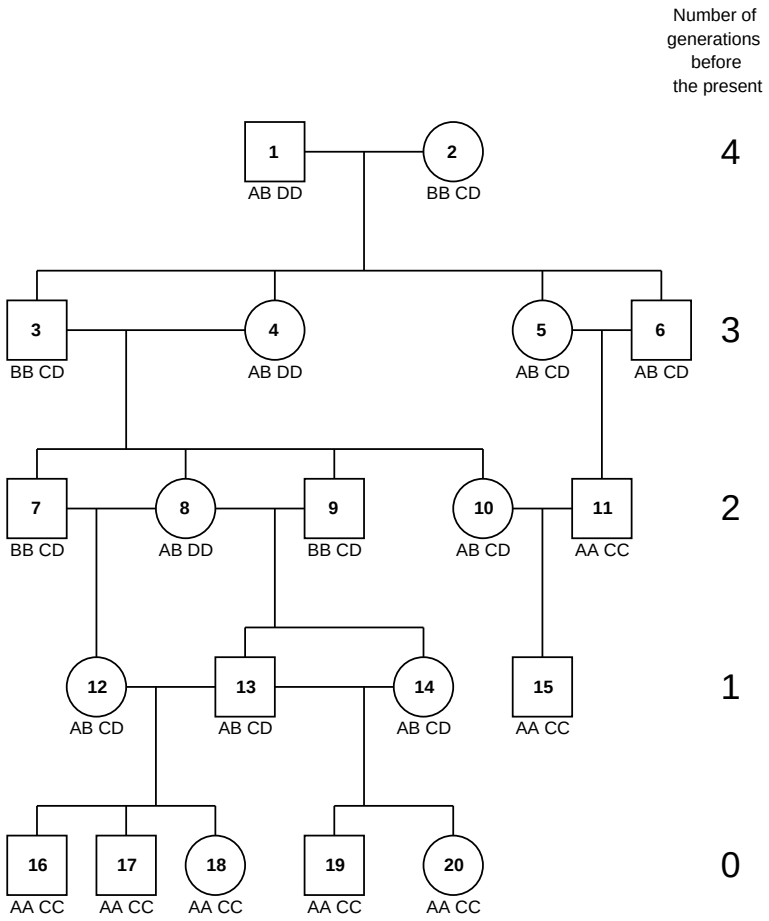



Figure 9.2: Genealogy of a diploid population. Numbers in the centers of circles (females) and squares (males) are individual identifiers. Genotypes at two autosomal loci are labeled beneath each individual. Autosomal locus 1 has alleles A and B; autosomal locus 2 has alleles C and D.

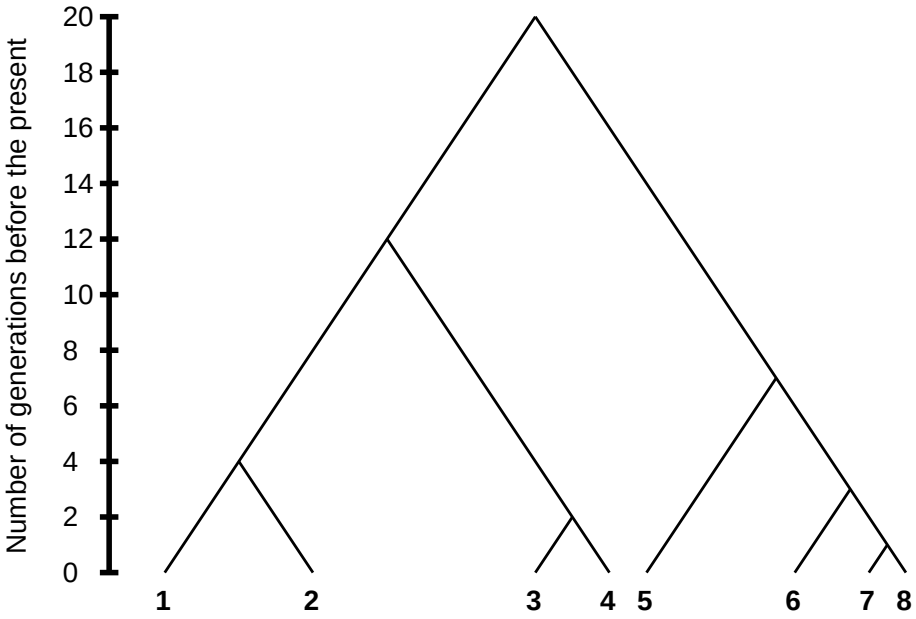



Figure 9.3: Genealogy of a haploid population. Individual identifiers are located at branch tips.

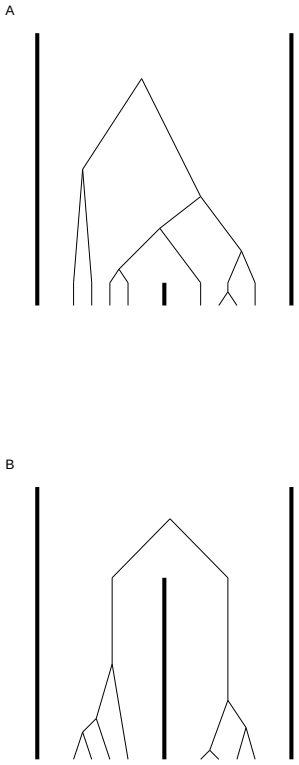



Figure 9.4: Diagram of divergence and coalescence in a model with no migration after the divergence. The center line separates the two descendant populations. The coalescent tree is drawn for samples of eight lineages, four from each descendant population. (A) Recent divergence. (B) Ancient divergence.

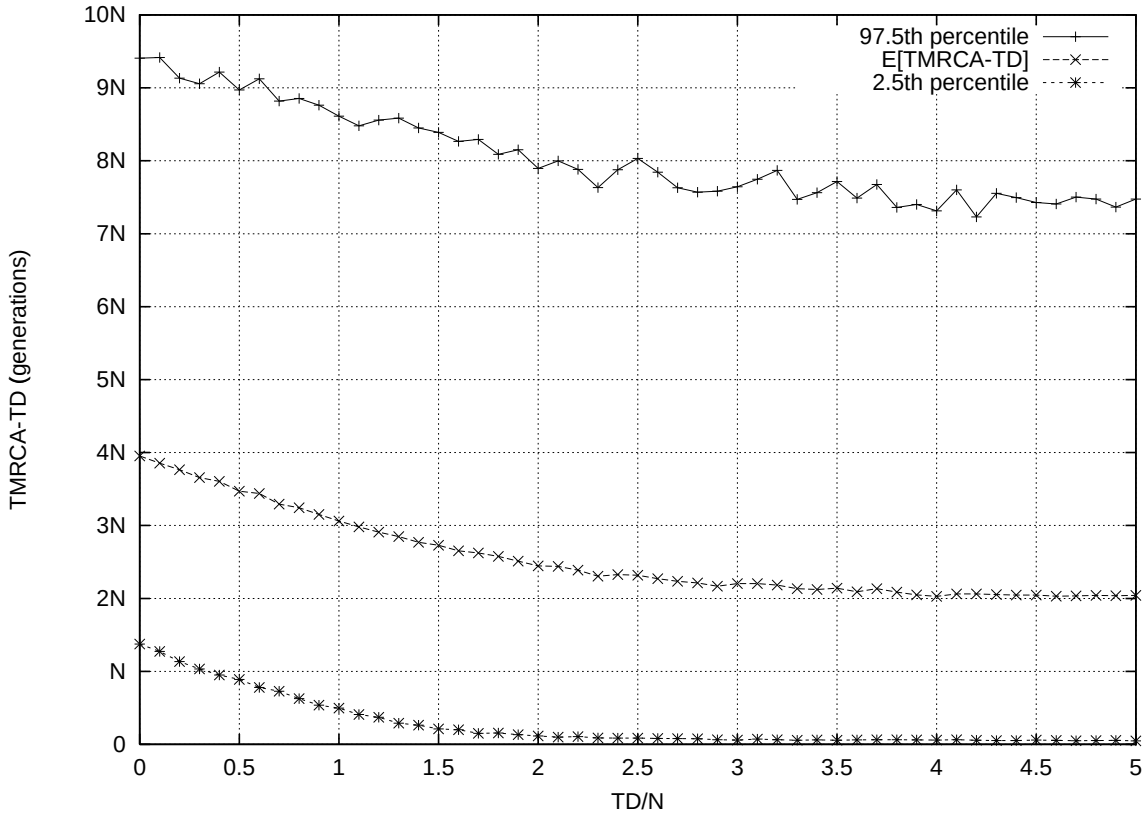



Figure 9.5: $T_{MRCA} - T_D$ vs. $T_D/N$ for a two-population divergence model with no migration after the divergence. The mean, the lower 2.5th percentile, and the upper 97.5th percentile are shown (based on 10,000 simulated coalescent trees for each value of $T_D/N$). Samples of size 50 were used in each population, and the descendant populations both had size $N$.

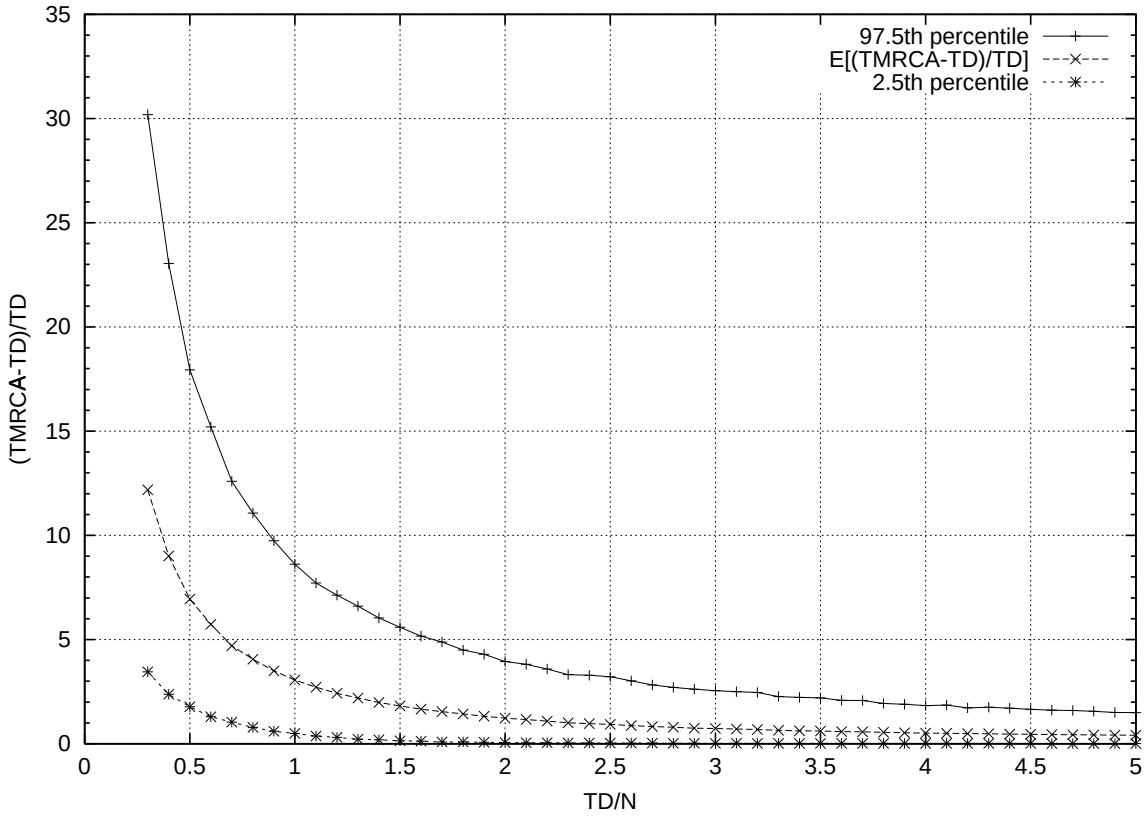



Figure 9.6: $\frac{T_{MRCA}-T_D}{T_D}$ vs. $T_D/N$ for a two-population divergence model with no migration after the divergence. The mean, the lower 2.5th percentile, and the upper 97.5th percentile are shown (based on 10,000 simulated coalescent trees for each value of $T_D/N$). Samples of size 50 were used in each population, and the two descendant populations both had size $N$.

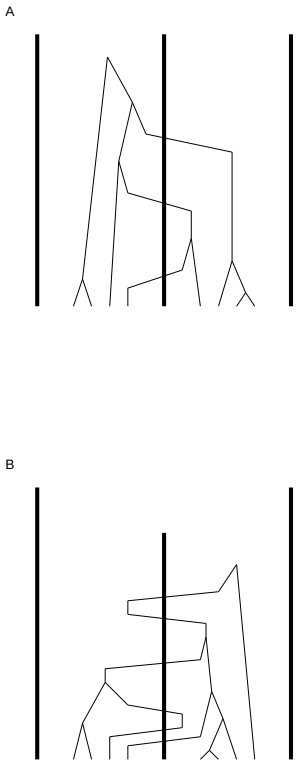



Figure 9.7: Diagram of divergence and coalescence in models with migration. The center line separates the two descendant populations. The coalescent tree is drawn for samples of eight lineages, four from each descendant population. (A) Two populations that have always been separated - a typical equilibrium island migration model. (B) Two populations that have had considerable migration after a divergence. In this example, the most recent common ancestor of the sample lived after the divergence.

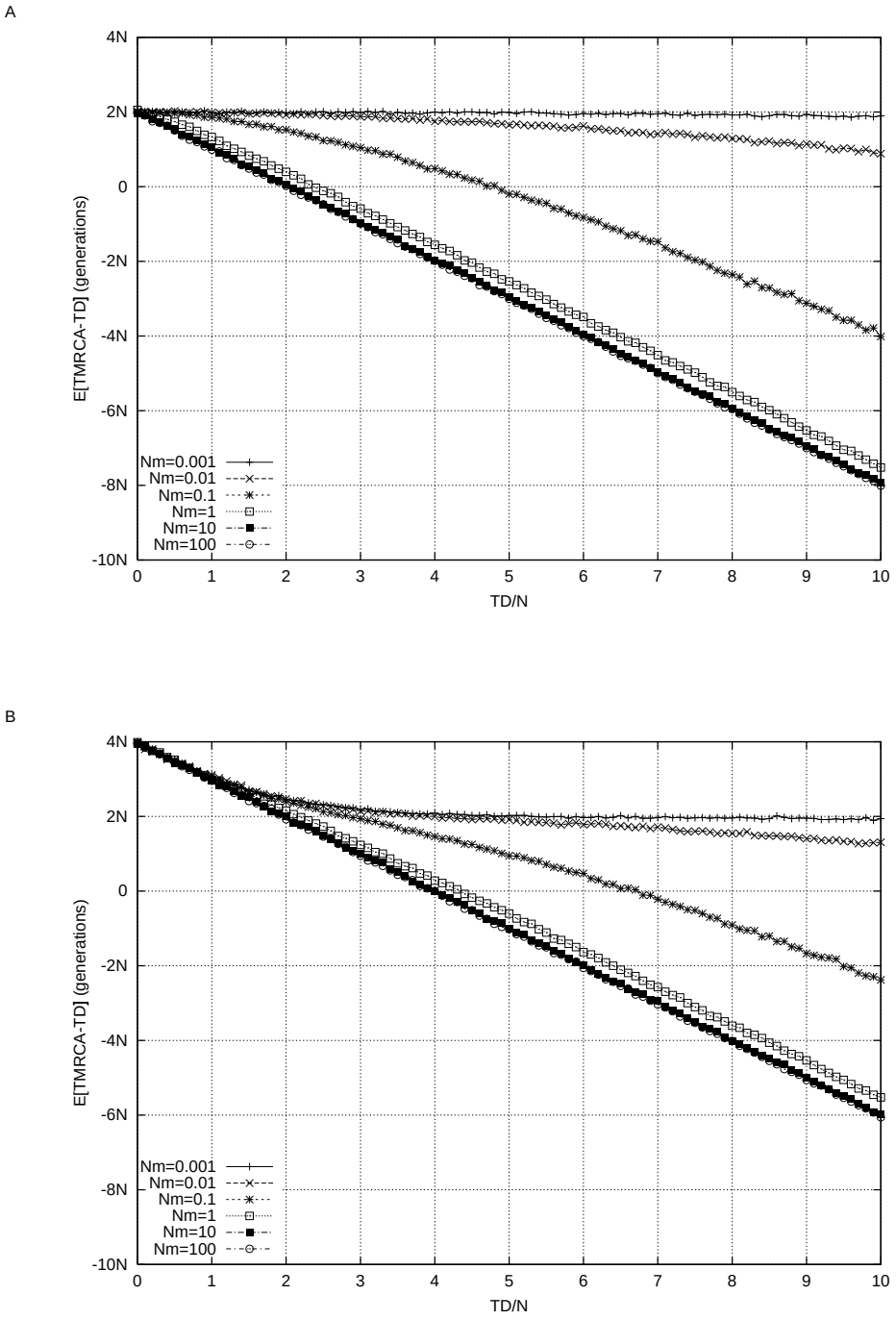



Figure 9.8: $E[T_{MRCA} - T_D]$ vs. $T_D/N$ for the two-population divergence model with constant migration after the divergence. Each point is based on 10,000 simulated coalescent trees. The descendant populations both had size $N$. (A) Samples of size 1 in each population. (B) Samples of size 50 in each population.

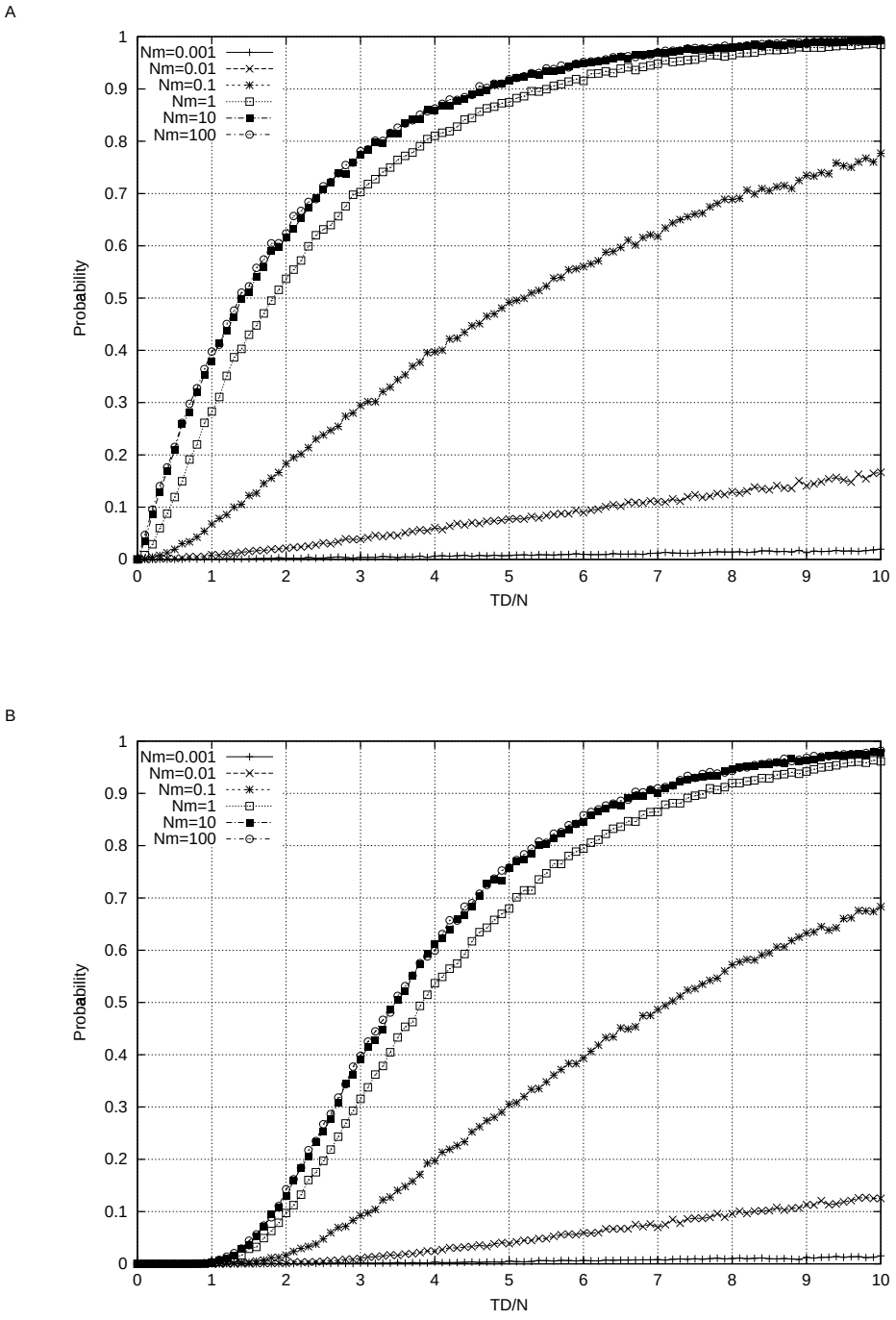



Figure 9.9: Probability that coalescence is *more recent* than divergence vs. $T_D/N$ for the two-population model with constant migration after the divergence. Each point is based on 10,000 simulated coalescent trees. The descendant populations both had size $N$. (A) Samples of size 1 in each population. (B) Samples of size 50 in each population.

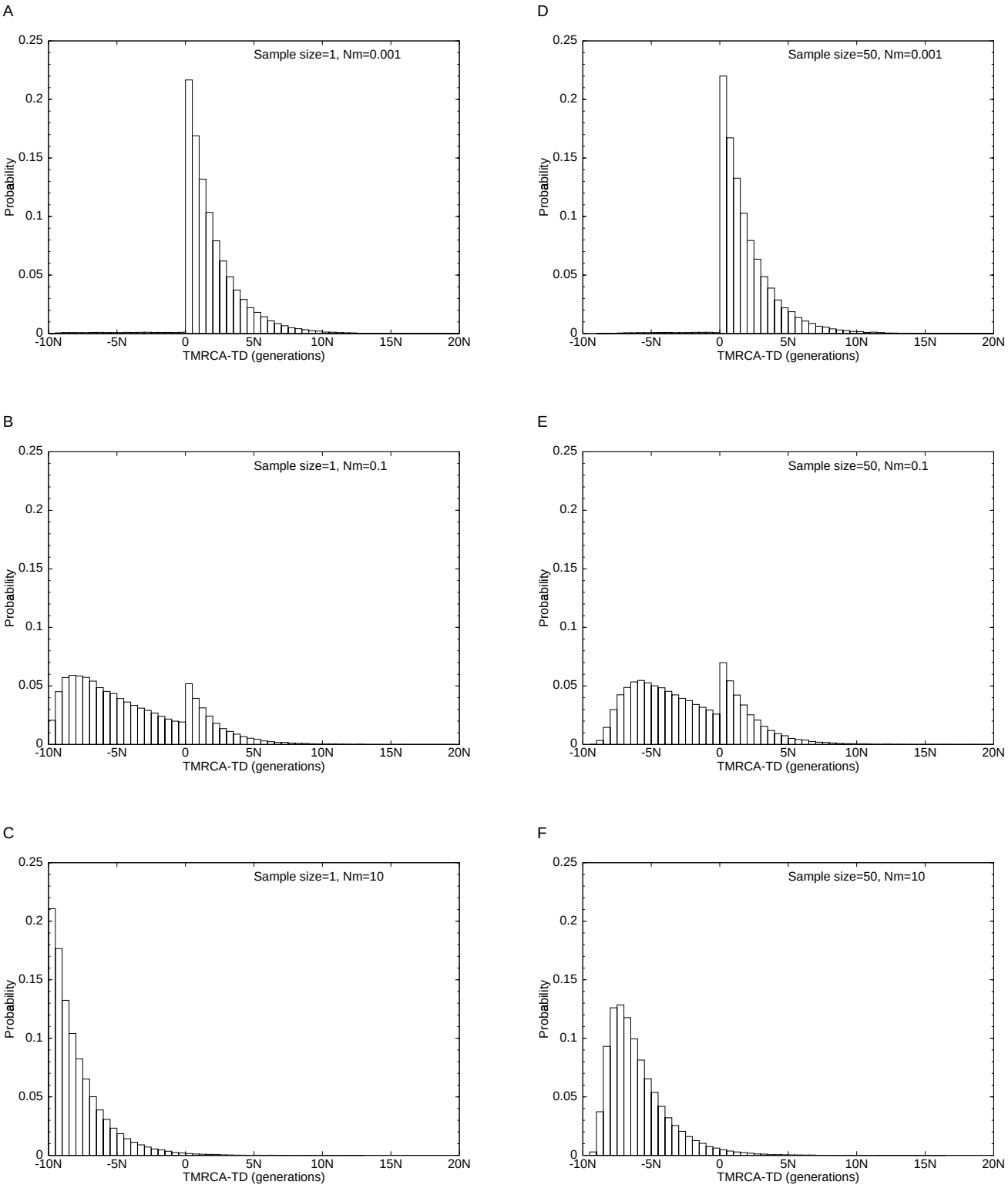
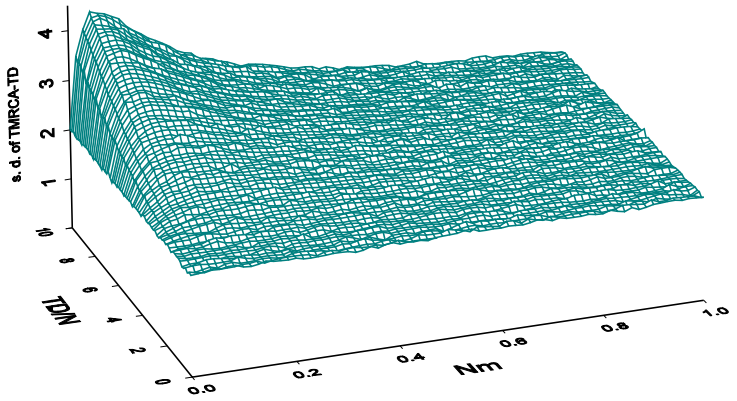



Figure 9.10: Probability distributions of $T_{MRCA} - T_D$ for the two-population divergence model with constant migration after a divergence $10N$ generations ago. Each distribution is based on 100,000 simulated coalescent trees. The descendant populations both had size $N$.

A




B

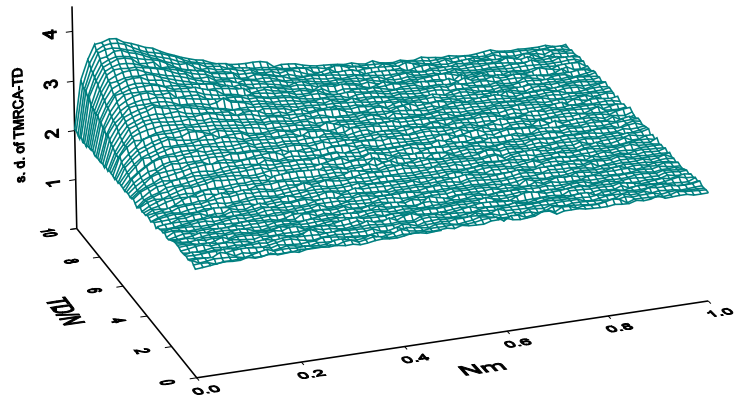



Figure 9.11: Standard deviation of $T_{MRCA} - T_D$ for the two-population divergence model with constant migration after the divergence, measured in units of $N$ generations. Each point is based on 10,000 simulated coalescent trees. (A) Samples of size 1 in each population. (B) Samples of size 50 in each population.