Lecture Notes on Gene Genealogies[1]

Alan R. Rogers

December 30, 2020

Contents

Lecture 1

Descriptive Statistics for DNA Sequences

1.1 DNA sequence data

Not until the 1980s did population geneticists begin the study of DNA sequence data. Until then, our measures of genetic variation were incomplete. We worked only with a small fraction of the genetic variation in our samples. With DNA sequence data we were finally able to study it all.

But this opportunity posed an immediate challenge. How should we *measure* that variation? Population geneticists were used to summarizing variation with statistics such as the sample heterozygosity: the probability that two random gene copies are copies of different alleles. But if the DNA sequences are long enough, it is unlikely that any two of them will be identical. The heterozygosity, in other words, is always 1. Clearly, new measures of variation are needed.

Table 1.1: Ten DNA sequences, each consisting of 40 sites. The sites are numbered across the top. The dots represent sites that are identical to the *reference sequence* at the top.

```
             0000000001 1111111112 2222222223 3333333334
             1234567890 1234567890 1234567890 1234567890

Sequence01   AATATGGCAC CTCCCAACCC TCTAGCATAT ACCACTTACA
Sequence02   .......T.. .C......TG C......C.. ..........
Sequence03   ..C....... .......... .......... ..........
Sequence04   .......T.. .C......TG C......... G.........
Sequence05   .......... .......... .......... ..........
Sequence06   .....A.... ........T. C......... G....C....
Sequence07   ..C....T.. .C......TG C......... G.........
Sequence08   .....A.T.. TC......TG C......... G.........
Sequence09   .......... .......... C......... ..........
Sequence10   .G...A.... ........T. C......C.. .T....C..G
Segregating:  ^^  ^ ^   ^^      ^^ ^      ^   ^^   ^^  ^
```

Table 1.1 presents 10 DNA sequences from some hypothetical species. Take a minute to study them. How many ways can you think of to summarize the variation in these data? This is precisely the problem that confronted population geneticists during the 1980s. The lecture that follows will summarize some of the ideas they came up with.

1.2 Statistics

Gene diversity (a.k.a. heterozygosity) is the probability that two random sequences are different. To calculate it, the straightforward approach is to examine all pairs and count the fraction of the pairs in which the two sequences are different from each other. It is often faster, however, to start by counting the number of copies of each type in the data. Let k_i denote the number of copies of type i, and $K = \sum k_i$ the number of gene copies in the sample. The the heterozygosity is estimated by

$$H = 1 - \sum_i \left(\frac{k_i}{K}\right)\left(\frac{k_i - 1}{K - 1}\right)$$

In the past, we have expressed heterozygosity as $2p(1-p)$ (for bi-allelic loci) or as $1 - \sum_i p_i^2$ (for loci with multiple alleles). These formulas are correct when p is the population allele frequency of the parents but contain a subtle bias when p is the allele frequency within a sample. The new formula corrects this bias.[1]

Number of segregating sites A "segregating site" is a site that is polymorphic in the data. The number of such sites is usually denoted by S.

Mean pairwise difference The average number, Π, of nucleotide site differences between pairs of sequences.

Mean pairwise difference per nucleotide If the sequences are L bases long, it is often useful to standardize Π by dividing it by L. The resulting statistic is

$$\pi = \Pi/L$$

Mismatch distribution A histogram whose ith entry is the number of pairs of sequences that differ by i sites. Here, i ranges from 0 through the maximal difference between pairs in the sample.

Site frequency spectrum A histogram whose ith entry is the number of polymorphic sites at which the mutant allele is present in i copies within the sample. Here, i ranges from 1 to $K - 1$.

Folded site frequency spectrum It is often impossible to tell which allele is the mutant and which is ancestral. In that case, we combine the entries for i and $K - i$, so the new i ranges from 1 through $K/2$.

[1] Imagine drawing two gene copies without replacement from a sample of size K. The first is a copy of allele A_i with probability k_i/K. Given this, the second is a copy of A_i with probability $(k_i - 1)/(K - 1)$. Thus, the sum of these quantities is the homozygosity and 1 minus this sum is the heterozygosity.

1.3 Data analysis

1.3.1 The number (S) of segregating sites

On the last line of Table 1.1, segregating (i.e. polymorphic) sites are indicated with a caret (ˆ). There are 15 such sites. Thus, the number of segregating sites is $S = 15$.

1.3.2 The mean pairwise difference (Π)

We want the average number of differences between pairs of individuals. There are two ways to do this calculation, the direct way and the easy way.

The direct way

Count the number of differences between each pair of sequences. For example, sequences 1 and 2 differ at 6 sites. Compare every pair of sequences, and write down the number of differences between each pair. If you do this (and I don't recommend it), you should end up with 45 numbers that sum to 248. The average is $\Pi = 248/45 = 5.511111$.

The easy way

The direct calculation involved two steps. Step 1 calculated the number (248) of pairwise differences, and then step 2 divided by the number (45) of pairs. The first of these numbers can be thought of as a sum over sites: the number of pairwise differences at site 1 plus that at site 2 and so on. The monomorphic sites make no contribution to this sum, so we need consider only the 15 polymorphic sites. And each site makes a contribution that is easy to calculate.

Suppose that at some site the sample contains only two nucleotides: x As and y Gs. Among pairs of sequences there will be some AA pairs, some AG pairs, and some GG pairs, but only the AG pairs will contribute a difference. The number of such pairs is $x \times y$, so this is the value that this particular site makes to the sum of pairwise differences.

For example, consider site 6 in the data above. There are 3 As and 7 Gs, so there are $3 \times 7 = 21$ AG pairs, and site 6 contributes 21 to the sum of pairwise differences. At site 2, on the other hand, there are 1 G and 9 As, so the site contributes $1 \times 9 = 9$ to the sum. Summing across the 15 polymorphic sites gives 248 as before.

There is also an easy way to find the number of pairs. In a sample of K sequences, there are $K(K-1)/2$ pairs. There are 10 sequences in the data above, so the formula gives $(10 \times 9)/2 = 45$ pairs.

1.3.3 Computer output

Here is the output of my seqstat program, which calculates descriptive statistics for DNA sequences:

```
%                                  seqstat
%                (descriptive statistics from sequence data)
%                             by Alan R. Rogers
%                                Version 5-1
%                                30 Jan 2000
%                        Type ‘seqstat -- ’ for help

% Cmd line: seqstat af10.seq

% Population 0
meanPairwiseDiff = 5.511111 ;
nsequences = 10 ;
nsites =  40 ;
mismatch = 1 5 3 2 2 6 8 8 5 2 2 1 ;
segregating = 15 ;
spectrum =  6 2 2 5 0 ;
% Count of minor allele at each polymorphic site:
%psite  site  count | psite  site  count
     1     2      1 |     9    21      3
     2     3      2 |    10    28      2
     3     6      3 |    11    31      4
     4     8      4 |    12    32      1
     5    11      1 |    13    36      1
     6    12      4 |    14    37      1
     7    19      4 |    15    40      1
     8    20      4 |
```

⋆ EXERCISE **1–1** Here is a set of 10 made-up DNA sequences, each with 10 nucleotide sites.

```
S01       AAACT GTCAT
S02       ..... A....
S03       ..G.. A....
S04       ..G.. A....
S05       ..... A....
S06       ..... AC...
S07       ..... A....
S08       ..... A....
S09       ..... A...C
S10       ..... A....
```

Calculate the mean pairwise difference, the number of segregating sites, the mismatch distribution and the site frequency spectrum.

Lecture 2

The Method of Maximum Likelihood

Before doing this exercise, please read *Using Likelihood*, which you can find at `http://www.anthro.utah.edu/~rogers/pubs/index.html`.

2.1 Maximum likelihood exercises with genetics problems

⋆EXERCISE **2–1** Suppose that we have data from a genetic system with two alleles, and that we observe N_1 individuals of genotype A_1A_1, N_2 of genotype A_1A_2, and N_3 of genotype A_2A_2. If the (unknown) genotype frequencies are P_1, P_2, and P_3, then the likelihood function is

$$L \propto P_1^{N_1} P_2^{N_2} (1 - P_1 - P_2)^{N_3}$$

I have used the symbol "$\propto$" (which stands for "is proportional to") rather than the equals sign because this expression ignores a proportional constant that will not affect the answer. The log likelihood is

$$\begin{aligned} \ln L &= \text{const.} + N_1 \ln P_1 + N_2 \ln P_2 \\ &\quad + N_3 \ln(1 - P_1 - P_2) \end{aligned}$$

Find the values of P_1, P_2, and P_3 that maximize the likelihood.

⋆EXERCISE **2–2** If we assume that the population is in Hardy-Weinberg equilibrium then the likelihood and log likelihood functions are

$$\begin{aligned} L &\propto [p^2]^{N_1} [2p(1-p)]^{N_2} [(1-p)^2]^{N_3} \\ \ln L &= \text{const.} + (2N_1 + N_2) \ln p \\ &\quad + (N_2 + 2N_3) \ln(1 - p) \end{aligned}$$

Find the value of p that maximizes the likelihood.

Lecture 3

Genetic Drift

3.1 The four causes of evolutionary change

1. mutation
2. selection
3. migration (a.k.a. gene flow)
4. genetic drift

3.2 What is genetic drift?

- It is everything that is left over after you account for the effects of mutation, selection, and migration.
- It consists of all the stochastic (random) effects on allele frequencies. These include everything from Mendelian segregation to the risk of accidentally walking in front of a bus.

How can one possibly model such an ill-defined hodgepodge?

3.3 The Wright-Fisher model

The population does not vary in size. In each generation, it consists of N individuals, each produced by the union of two randomly chosen gametes.

Generating gametes Each gamete is constructed by the following algorithm: (1) choose a parent at random from among the N individuals of the previous generation. (2) Choose a random half of that parent's DNA. (Don't worry about genetic linkage; we will be dealing here with one locus at a time.) If there are two alleles A_1 and A_1, segregating at some locus, what is the probability that the gamete that we construct carries a copy of A_1? If the parent was an A_1A_1 homozygote, then we are bound to get A_1 in the gamete. If the parent was an A_1A_2 heterozygote, then the gamete has a 50% chance of carrying A_1. Thus, the algorithm generates an A_1-bearing gamete with probability $p_1 = P_{11} + P_{12}/2$, where P_{11} and P_{12} are the frequencies of genotypes A_1A_1 and A_1A_2 within the parental generation. Note that the formula for p_1 is exactly the same as the formula for the frequency of A_1 among the parents. Conclusion: The Wright-Fisher algorithm for generating gametes is equivalent to drawing genes at random with replacement from the parental population. To clarify this idea, many authors have made use of the urn metaphor.

The urn metaphor In an urn full of balls, a fraction p of the balls are red and a fraction $1 - p$ are black. Each ball represents a gene. The red balls represent copies of one allele; the black ones copies of another. The fraction p represents the frequency of the red allele in the population. The urn will be used to produce a new generation in which there are N diploid individuals, or $2N$ genes. To produce the new generation, we perform the following operation $2N$ times: draw a random ball from the urn, write down its color, and then return the ball to the urn. The number of red balls drawn represents the number of copies of the red allele in the new generation, and similarly for the black balls. Both numbers are random variables. Their probability distribution was taken by Wright and Fisher as a model of the process of genetic drift.

The Wright-Fisher model is undoubtedly simpler than reality, but it has been remarkably successful at dealing with the stochastic variation in real populations. Let us be content with it, at least for the moment, and ask about its properties. In the urn, the frequency of red balls is p. Let p' denote the frequency of red balls among those drawn. The difference between p' and p represents the effect of genetic drift. How large is this difference likely to be?

If we repeated the urn experiment over and over, the average value of p' would get closer and closer to p. Another way to say this is to say that *the expected value of p' equals p*. In notation,

$$E[p'] = p$$

where the symbol E represents the "expectation," or average.

But unless N is extremely large, there will be some difference between p' and p, so we can write

$$p' = p + \epsilon$$

Here, ϵ (the greek letter "epsilon") represents the effect of genetic drift. Its expected value is 0, but its variance is[1]

$$V[\epsilon] = E[\epsilon^2] = \frac{p(1-p)}{2N}$$

Genetic drift is important when this variance is large; unimportant when it is small. The formula captures two influences:

1. Drift is unimportant when $p(1 - p)$ is near 0. This happens when $p \approx 0$ and also when $p \approx 1$.

2. Drift is unimportant when N is very large.

3. Drift is most important when $p \approx 1/2$ and N is small.

Show a plot of p against t.

[1] To see where this formula comes from, look up the binomial distribution in any text on probability and statistics.

Table 3.1: Average heterozygosity

Pop.	Bl. grp.[a]	Protein[b]	Classical[c]	RFLP[d]	RSP[e]	STR-4[f]	STR-2[g]	STR-3[h]
Africa	0.164	0.179	0.163	0.297	0.322	**0.769**	**0.807**	**0.850**
Asia	0.145	0.164	0.189	0.327	0.377	0.681	0.685	0.820
Europe	**0.179**	**0.186**	**0.202**	**0.379**	**0.432**	0.724	0.730	0.807

Note: Largest entry in each column is in boldface. Columns are in order of increasing European heterozygosity.

[a] 32 blood groups [22].

[b] 80 protein polymorphisms [22].

[c] 110 classical polymorphisms [1].

[d] 79 restriction fragment length polymorphisms [1].

[e] 30 RFLPs consisting solely of restriction site polymorphisms [13].

[f] 30 tetranucleotide STRs [13].

[g] 30 dinucleotide short tandem repeat polymorphisms (STRs). Difference between Africa and Europe is significant [1].

[h] 5 trinucleotide STRs [34].

3.4 Classical theory of homozygosity and heterozygosity

Let J_t represent the probability that two genes chosen at random from some population are copies of the same allele. If the population mates at random, then J will also be the homozygosity. The gene diversity (or heterozygosity) is $H = 1 - J$. Several gene diversity estimates are shown in table 3.1. These statistics are affected by several evolutionary forces:

1. Mutations reduce J because they are more likely to make identical genes less similar than to make different genes identical.

2. Genetic drift tends to move allele frequencies toward 0 and 1. Consequently $2pq$ gets smaller, heterozygosity declines, and homozygosity increases.

To measure the effects of these forces, we need a model. Let us begin with a model dealing only with the first force.

3.4.1 Drift only

Let J_t denote the probability that two genes drawn at random from the population of generation t are copies of the same allele. There is a simple model that relates J in one generation to its value in the generation before:

$$J_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) J_t$$

The first term on the right accounts for the possibility that the two genes in generation $t + 1$ may be copies of the same gene in generation t. Since there are $2N$ genes in the population, the two genes are copies of the same gene with probability $1/(2N)$ and are copies of distinct genes with probability $1 - 1/(2N)$. In the latter case, they are by definition copies of the same allele with probability J_t.

This model says that each generation's value of J is a weighted average of 1 and the previous value of J. Consequently, J converges toward 1. We are eventually left with no heterozygotes at all.

3.4.2 Drift plus mutation

To make the model interesting, we need to add in some other evolutionary force. Let us add in mutation. The easiest way to do this employs the model of "infinite alleles," which assumes that every mutation produces an allele that has never existed before. Thus, two genes can be identical only if there has been no mutation along the evolutionary path that connects them. In particular, there can have been no mutation during the past generation in the path leading to either of our two genes. If u is the mutation rate per generation, then $1 - u$ is the probability that no mutation occurs along a single evolutionary path during a generation, and $(1-u)^2$ is the probability that neither of our two genes has mutated in the past generation. Thus,

$$J_{t+1} = (1-u)^2 \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) J_t \right)$$

Population geneticists are not very accurate people and tend to ignore whatever they can. Consider the following:

u	$(1-u)^2$	$1-2u$
0.00100	0.9980010000	0.99800
0.00010	0.9998000100	0.99980
0.00001	0.9999800001	0.99998

The smaller the value of u, the less the difference between $(1-u)^2$ and $1-2u$. Since mutation rates are very small numbers, population geneticists never trouble themselves about the difference between $(1-u)^2$ and $1-2u$. In the present case, $1-2u$ makes the algebra simpler. Similarly, if u is small and N is large, $(1-2u)/(2N)$ is hardly different from $1/(2N)$. Applying both simplifications gives

$$J_{t+1} \approx \frac{1}{2N} + \left(1 - 2u - \frac{1}{2N}\right) J_t$$

This equation doesn't look very different from the one with drift only, but it leads to a very different conclusion. Rather than converging toward unity, this one levels out at a different equilibrium value, as shown in figure 3.1.

To find the equilibrium algebraically, set $J_{t+1} = J_t$ and solve the resulting equation. The result is

$$J = \frac{1}{4Nu + 1} \tag{3.1}$$

The gene diversity (or heterozygosity) is

$$H = 1 - J = \frac{4Nu}{4Nu + 1} \tag{3.2}$$

3.4.3 A simpler way: coalescent theory

Take a random pair of genes and peer backwards down their ancestries. So long as the two evolutionary paths remain distinct, two types of event may happen in any given generation:

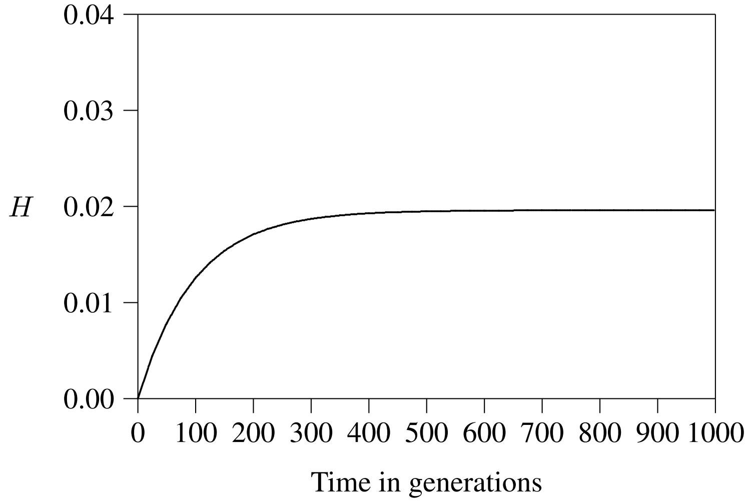

Figure 3.1: How heterozygosity changes over time. Assumes: $u = 0.005$, $N = 2500$, H(0)=0

A mutation A mutation may occur in either path with probability u. The combined probability in both paths is $2u$.[2]

A coalescent event The two paths will coalesce when we reach the most recent generation in which they share a common ancestor. This is an event with probability $1/(2N)$.

The hazard of an event of either type is

$$2u + 1/(2N)$$

When an event does occur, it is a mutation with probability

$$\frac{2u}{2u + 1/(2N)} = \frac{4Nu}{4Nu + 1}$$

In this case, the two genes are copies of different alleles. The formula gives the probability that two random genes will be copies of different alleles—the gene diversity. Notice that it is identical to equation 3.2.

The simplicity of this approach is remarkable. It has led to profound changes in population genetics during the past decade or two. We will return to it in the section on gene genealogies.

$\star$ EXERCISE **3–1** For classical polymorphisms, human gene diversity is roughly 0.16. What does this imply about the quantity $4Nu$? (In the literature, $4Nu$ is often denoted by θ, the greek letter "theta").

$\star$ EXERCISE **3–2** If the mutation rate were 10^{-6}, what value of N would be needed to account for this level of heterozygosity?

[2] This is only an approximation. If I really wanted to be accurate, I would say that $(1-u)^2$ was the probability of no mutation along either of the two paths and $1-(1-u)^2$ the probability of at least one mutation. But when u is small this latter probability is indistinguishable from $2u$.

$\star$ EXERCISE **3–3** Using these same values for N and u, suppose that H were equal to 0.01 in generation 0. What would its value be in generation 20?

$\star$ EXERCISE **3–4** Using the same value for N, plot the variance of ϵ for values of p ranging from 0 through 1.

$\star$ EXERCISE **3–5** The square root of the variance is called the standard deviation and (in this case) provides an estimate of the magnitude of a typical value of ϵ. For what value of p is this standard deviation largest? How large is it at this value of p?

$\star$ EXERCISE **3–6** Figure 3.1 assumed that $u = 0.005$ and $N = 2,500$. Under these assumptions, what is the equilibrium value of H? Is it consistent with the figure?

Lecture 4

Gene Genealogies

The coalescent process [12, 17] describes the ancestry of a sample of genes. As we trace the ancestry of each modern gene backwards from ancestor to ancestor, we occasionally encounter common ancestors—genes whose descendants include more than one gene in the modern sample. Each time this happens, the number of ancestors shrinks in size. Eventually, we reach the gene that is ancestral to the entire modern sample, and the process ends.

Since the mid-1980s, this model has revolutionized our understanding of the effects of genetic drift and mutation. Many of the results obtained this way have been entirely new. Others have merely confirmed results that were obtained long before. Either way, the coalescent model provides a method of studying drift and migration that is far easier than the methods that geneticists used to use. We begin with a few mathematical tricks, which will be useful later.

4.1 Preliminaries

Trick 1 *If the hazard of death is h per day, then the expected life-span is $1/h$ days.*

For concreteness, suppose that we are talking about the life-span of a piece of kitchen glassware. Eventually, someone will drop it and it will break. Suppose that the hazard of breakage is h per day and its expected lifespan is T days. Trick 1 tells us that $T = 1/h$.

We are envisioning time here as a continuous variable and assuming that the glass may break at any instant. It makes no sense to talk about the probability of breakage at a particular instant, because that has to be zero. Instead, h is a *probability density*. (See *Just Enough Probability*.) Specifically, it is the density that the glass will break at a particular instant given that it has not broken already. This sort of density is often called a *hazard*, and we are assuming that the hazard does not change. This implies that the lifespan (t) is a random variable whose probability distribution is *exponential*. The mean of this distribution is $1/h$, as shown in appendix 4.B. In the exercise below, you will derive this formula for the case in which time is discrete.

$\star$ EXERCISE **4–1** Trick 1 refers to the case in which time is continuous, but the result also holds when time is discrete. For example, suppose you are tossing a glass into the air and then catching it. On each toss, there is a probability h that you will drop the glass and break it. How many times, on average, can you toss

```
X -----------------------|
                         A------
Y -----------------------|

|--------- t ------------|
```

Figure 4.1: Coalescence of a sample of two genes

the glass before dropping it? The answer is $1/h$, just as in Trick 1. In this exercise, you will derive this formula. To do so, consider what happens on the first toss. Either you drop it or you catch it. If you drop it (probability h), it breaks, and its lifespan is 1 toss. The first component of T is therefore $h \times 1$. If you catch it (probability $1 - h$), the expected lifespan is $1 + T$. Why? Because h doesn't change. Our 1-toss-old glass can expect to survive T additional tosses, so its expected lifespan is $1 + T$. The second component of T is therefore $(1 - h)(1 + T)$. Write down an equation saying that T is the sum of these two components, and then solve that equation for T.

Trick 2 *There are $k(k-1)/2$ ways to choose 2 items out of k.*

There are k ways to choose the first item. Having chosen the first, there are $k-1$ ways to choose the second, so there are $k(k-1)$ pairs. But this counts pair AB separately from pair BA. We are interested in unordered pairs, so the number is $k(k-1)/2$.

4.2 Coalescence time in a sample of two genes

The genealogy of two genes, X and Y, is shown in figure 4.1. Genes X and Y live in the present generation, and their common ancestor A lived t generations ago. Consequently, as we look backward from the present into the past, the two lines of descent remain distinct for t generations, at which time they *coalesce* into a single line of descent. In a given generation, the lines coalesce if the two genes in that generation are copies of a single parental gene in the generation before. Otherwise, the two lines remain distinct.

What can we say about the length of time, t, that they remain distinct? The problem is a lot like the one above involving kitchen glassware. If we knew the hazard, h, that the lines of descent will coalesce during a generation, then trick 1 would tell us immediately the mean number of generations until the two lineages coalesce.

Consider the tiny population shown in figure 4.2. Each row illustrates the population in a single generation, and within each row each character represents a gene. The generations are numbered backwards, so that generation 0 is the present and generation 1 contains the parents of generation 0. The vertical line indicates that gene Z is the parent of gene X. What is the probability that it is also the parent of gene Y? If each gene in generation 1 is equally likely to be Y's parent, then this probability is

$$h = 1/10$$

since there are 10 genes in generation 1. This answer would be the same no matter which gene in generation 1 had been X's parent.

```
                  _________Population___________
Generation 0:     0  Z  0  0  0  0  0  0  0  0
                     |
Generation 1:     0  X  0  Y  0  0  0  0  0  0
```

Figure 4.2: A sample of two genes (X and Y) in a population of size 10. Gene Z is the parent of gene X.

Trick 1 immediately tells us that the mean coalescence time is 10 generations. Of course, 10 is really the number of genes in the population. If there are $2N$ genes in the population, then

$$h = 1/2N \tag{4.1}$$

Now the answer becomes somewhat more interesting. The average pair of genes last shared a common ancestor $2N$ generations ago. This provides a connection between population size and the genealogy of genes. As we shall see in lecture 5, this connection lets us use genetics to study the history of population size.

In equation 4.1, the symbol N is a little confusing. If we are talking about an autosomal locus, then there are two genes for every person and N is the number of people in the population. The meaning of N is different, however, if we are talking about a mitochondrial locus. In that case, the gene is transmitted only through women, and locus is effectively haploid. Consequently, the number $(2N)$ of genes is the number of females in the population, and the symbol N represents half the number of females.

● EXAMPLE **4–1**

Suppose that we somehow knew that the average pair of mitochondrial genes last shared a common ancestor 100,000 years ago. What would this imply about population size? (Ignore the issue of statistical error.)

○ ANSWER

100,000 years is about 4000 generations, so the assumption implies that

$$2N = 4000$$

Since we are talking about a mitochondrial gene, this is really the number of women. If there are as many men as women, then the population would contain 8000 individuals. This is about the size of a large village or a very small town.

4.3 Coalescence times in a sample of K genes

Now consider a sample of K genes. (Figure 4.3 shows the case in which $K = 4$.) As we move backwards in time from the present, the first coalescent event that we encounter reduces our sample from K to $K - 1$, the second from $K - 1$ to $K - 2$, and so on. After $K - 1$ coalescent events, only a single lineage is left and no further coalescent events can occur. There are thus $K - 1$ intervals to consider. The first (i.e. the most recent) interval is the one in which there are K lines of descent. This interval is t_K generations. The next interval has $K - 1$ lines of descent and is t_{K-1} generations long. The last interval is the one with two lines of descent and is t_2 generations long. Since the length of each interval is independent of all the other lengths, we can consider them one at a time.

```
W --------|
          |------------|
X --------|            |
                       |--------------|
Y ---------------------|              |---
                                      |
Z ------------------------------------|

|-- t_4 --|---- t_3 ---|------ t_2 ---|
```

Figure 4.3: Coalescence of a sample of four genes

Consider a generation during which the sample has i genes. By trick 2 there are $i(i-1)/2$ pairs of genes. For any given pair, the probability that the two genes are copies of the same parental gene is equal to $1/2N$. Consequently, we might expect the probability of a coalescent event to be close to

$$h_i = \frac{i(i-1)}{4N} \tag{4.2}$$

per generation. Although this argument is loose, it turns out that the result is correct. (To find out why, see section 4.A.) The expected length of this interval is given by trick 1 and equals

$$1/h_i = \frac{4N}{i(i-1)} \tag{4.3}$$

For example, in a sample of size 5, the four coalescent intervals have hazards and expected lengths as follows:

	Coalescent	Expected
Interval	hazard	length
5	$h_5 = \frac{5 \times 4}{4N} = 10/2N$	$2N/10$
4	$h_4 = \frac{4 \times 3}{4N} = 6/2N$	$2N/6$
3	$h_3 = \frac{3 \times 2}{4N} = 3/2N$	$2N/3$
2	$h_2 = \frac{2 \times 1}{4N} = 1/2N$	$2N$

4.4 The depth of a gene tree

The *depth* of a gene tree is the time (usually in generations) since the *Last Common Ancestor* (LCA) of all the genes in the sample. The tree's depth is simply the sum of its coalescent intervals, and we already have a formula for the expected length of each interval. For example, in a sample 2 genes, the expected depth of

The coalescent interval containing i lineages has expected depth $1/h_i = 4N/i(i-1)$, so the total expected depth in a sample of K gene copies is

$$4N \sum_{i=2}^{K} 1/i(i-1)$$

This sum is easy to simplify once you notice that

$$\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$$

With this substitution, the series becomes

$$4N\left(\frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \cdots - \frac{1}{K-1} + \frac{1}{K-1} - \frac{1}{K}\right)$$

Adjacent terms cancel, and we are left with equation 4.4 [29, p. 132].

Box 1: The expected depth of a gene genealogy

the gene tree is $2N$ generations. In a sample of 3 genes it is $2N + 2N/3 = 8N/3$ generations. Here are a few more examples:

Sample size	Mean depth of tree
2	$1/h_2 = 2N$
3	$1/h_3 + 1/h_2 = 8N/3$
4	$1/h_4 + 1/h_3 + 1/h_2 = 3N$
5	$1/h_5 + 1/h_4 + 1/h_3 + 1/h_2 = 16N/5$

Notice that in each case, the mean tree depth is equal to

$$4N(1 - 1/K) \tag{4.4}$$

where K is the number of gene copies in the sample. This formula turns out to be true in general, as shown in Box 1 [29, p. 132]. In a large sample, the $1/K$ term is unimportant and the answer is even simpler: the average depth of a gene tree is approximately $4N$ generations.

⋆ EXERCISE **4–2** Box 1 uses the fact that

$$\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$$

Verify that this is true by deriving the left side from the right.

⋆ EXERCISE **4–3** Suppose that we draw a sample of size $K = 10$ from a population with $2N = 5000$ genes. What are the expected lengths (in generations) of all the coalescent intervals?

⋆ EXERCISE **4–4** In a sample of 10,000 genes, what is the expected age of the LCA? What fraction of this age is accounted for by the interval during which the tree contained only two lineages?

```
-|
 |--|
-|  |
    |---|
---||   |
   ||   |
---|    |---|
        |   |
--------|   |-------------------------------------------|
            |                                           |
------------|                                           |
                                                        |
------|                                                 |
      |                                                 |
---|  |-------|                                         |
   |--|       |                                         |---
---|          |------|                                  |
              |      |                                  |
-|            |      |                                  |
 |------------|      |                                  |
-|                   |                                  |
                     |                                  |
-------------|       |----------------------------------|
             |       |
-|           |-|     |
 |-----------| |     |
-|             |     |
               |     |
|              |     |
|---|          |-----|
|   |          |
    |-|        |
----| |        |
      |        |
-|    |--------|
 ||   |
-||   |
  |---|
--|
```

Figure 4.4: Coalescence of a sample of twenty genes

When the sample is large, $K(K-1)/2$ is a large number. Consequently, initial coalescent intervals tend to be short. In a sample of size 20, the most recent coalescent interval is (on average) 0.5 percent as long as the interval that ends with the root. Figure 4.4 shows an example. Note the short terminal branches and the deep basal branch.

4.A A more detailed treatment (optional)

4.A.1 Preliminaries

Before explaining the formula for the general case—that of a coalescent interval during which the sample has K genes—we need one additional mathematical trick.

Trick 3 *The sum of the numbers from 1 through k is $k(k+1)/2$.*

This trick was supposedly discovered by the mathematician Carl Friedrich Gauss when he was just six years old. According to the story, Gauss's teacher gave the class an assignment to keep it busy while he graded papers: Sum the numbers from 1 through 100. Two minutes later, Gauss walked to the front of the room with his answer. The answer was correct, but Gauss was punished for failing to do the work the hard way. Here is how he did it.

First he wrote down

$$1 + 2 + \cdots + 99 + 100$$

Then, being bored and discouraged, he wrote it out backwards just below:

$$\begin{array}{ccccccccc} 1 & + & 2 & + & \cdots & + & 99 & + & 100 \\ 100 & + & 99 & + & \cdots & + & 2 & + & 1 \end{array}$$

Then the insight struck—he noticed that each of the columns added to 101:

$$\begin{array}{ccccccccc} 1 & + & 2 & + & \cdots & + & 99 & + & 100 \\ 100 & + & 99 & + & \cdots & + & 2 & + & 1 \\ \hline 101 & + & 101 & + & \cdots & + & 101 & + & 101 \end{array}$$

Since there are 100 columns, the sum of all the numbers here is 100×101. And this is twice the sum that he was looking for. Thus,

$$1 + 2 + \cdots + 99 + 100 = \frac{100 \times 101}{2}$$

In the general case,

$$1 + 2 + \cdots + k = \frac{k(k+1)}{2}$$

Trick 4 *e^x is approximately $1 + x$ when x is small.*

Here e^x is the exponential function, and is also written $\exp(x)$. You can verify the trick with a calculator.

```
X --------------------|
                      B---------------|
Y --------------------|               |
                                      A---
Z ------------------------------------|


|--------- t_3 -------|------ t_2 ----|
```

Figure 4.5: Coalescence of a sample of three genes

4.A.2 Coalescence times in an interval with three genes

Figure 4.5 shows the genealogy of a sample of three genes. It has two coalescent events, one at node A (the root) and another at node B. The time between the present and the root can be broken into two intervals of length t_2 and t_3, where t_2 is the number of generations during which the genealogy has two lines of descent and t_3 is the number of generations during which it had three. What can we say about the lengths of these intervals?

The first point to notice is that the intervals are independent. As we ponder the length of one interval, we need not worry about the length of the other. And we already know the mean of t_2: the preceding section showed that the mean coalescence time for a sample of two genes is $2N$ generations.

This leaves us with only one question to answer: What is the mean time until the first coalescent event in a sample of three genes? We could answer this question using trick 1 if we knew the hazard of a coalescent event in a sample of that size. Let us therefore consider the probability that a coalescent event occurs during some given generation.

It is easier to calculate first the probability of the event that all three lines of descent remain distinct. This requires that

1. X and Y are copies of different parental genes. We already know that this event has probability $1 - 1/2N$.

2. Z is neither a copy of X's parent nor a copy of Y's parent. This event has probability $(2N - 2)/2N$. (Of the $2N$ genes that we can choose between, 2 produce a coalescent event and $2N - 2$ do not.) This probability can also be written as $1 - 2/2N$.

Thus, the probability that no coalescent event occurs in some particular generation is equal to

$$1 - h = (1 - 1/2N)(1 - 2/2N)$$

Now it is time to invoke trick 4. If the population is large, then $2N$ will be a large number and $1/2N$ and $2/2N$ will both be small. Trick 4 thus allows the probability above to be re-expressed as

$$1 - h \approx e^{-1/2N} e^{-2/2N} = e^{-3/2N}$$

Now invoke trick 4 once again to simplify the exponential:

$$\begin{aligned} 1-h &\approx 1-3/2N \\ h &\approx 3/2N \end{aligned}$$

Having found the hazard of a coalescent event in a sample of three genes, trick 1 now gives us the mean length of the part of the genealogy during which there were three lines of descent:

$$\text{mean of} \quad t_3 = 2N/3$$

In a sample of three genes, the hazard of a coalescent event is three times as large as the hazard in a sample of two. Consequently, the mean waiting time until the first coalescent event is only 1/3 as large. The expected depth of the tree is the expected sum of t_2 and t_3. It equals $2N + 2N/3$, or $8N/3$. Three quarters of this total is taken up by the portion of the genealogy during which there are only two lines of descent.

● EXAMPLE **4–2**

In a population of 10^7, what is the mean time in years until a sample of three mitochondrial genes coalesce to a single line of descent.

○ ANSWER

If there are 10^7 people, there will be about half that many females, so $2N = 5 \times 10^6$. The coalescence time is $8N/3 = 6.67 \times 10^6$ generations. If generations are 25 years long, this is 167×10^6 years. So the Last Common Ancestor (LCA) should have lived during the Jurassic period. Incidentally, this example is far-fetched for humans, because it implies far more mitochondrial variation than really exists.

4.A.3 Coalescence times in an interval with i genes

As in the case of three lines of descent, it is easiest to calculate $1 - h$, the probability that no coalescent event occurs during some particular generation. When there are i genes in the sample, this requires

Event	Probability
Gene 2 and gene 1 have different parents	$1 - 1/2N$
Gene 3's parent differs from the preceding 2 parents	$1 - 2/2N$
Gene 4's parent differs from the preceding 3 parents	$1 - 3/2N$
....	
Gene i's parent differs from the preceding $i - 1$ parents	$1 - (i - 1)/2N$

The probability that a coalescent event does *not* occur is

$$\begin{aligned} 1-h &= (1-1/2N)(1-2/2N)\cdots(1-(i-1)/2N) \\ &\approx e^{-1/2N}e^{-2/2N}\cdots e^{-(i-1)/2N} && \text{(trick 4)} \\ &= \exp\left[-\tfrac{1}{2N}(1+2+\cdots+(i-1))\right] \\ &= \exp\left[-\tfrac{i(i-1)}{4N}\right] && \text{(trick 3)} \\ &\approx 1-\tfrac{i(i-1)}{4N} && \text{(trick 4)} \end{aligned}$$

Thus,

$$h \approx \frac{i(i-1)}{4N} \tag{4.5}$$

4.B The mean of an exponential random variable (optional)

If t is an exponential random variable, then its density function is he^{-ht}. The mean of this distribution is:

$$E[t] = \int_0^\infty hte^{-ht}dt$$

Substituting $x = ht$ turns this into

$$E[t] = h^{-1}\int_0^\infty xe^{-x}dx$$

On integrating by parts, the integral on the right becomes

$$\int_0^\infty xe^{-x}dx = -xe^{-x}|_0^\infty - \int_0^\infty (-e^{-x})dx$$

The first term on the right is 0 and the second is $-e^{-x}|_0^\infty = 1$. Thus, $E[t] = 1/h$.

Lecture 5

Relating Gene Genealogies to Genetics

This course began with a section on probability theory. Then came material on genetic variation and genetic drift. In the last lecture, we discussed gene genealogies. These may have seemed like disconnected threads. This lecture will tie them all together.

We will make extensive use of the theory introduced last time about genealogical relationships among genes. Although that theory is elegant, it is also limited, for gene genealogies cannot be observed. They describe obscure events that happened many thousands of years ago. We can never *know* the genealogy of a sample of genes. We can *estimate* it from genetic data, but that requires a theory that relates gene genealogies to observable genetic data.

This lecture begins by adding mutations to gene genealogies, and then relates these to two genetic statistics: the number S of segregating sites, and the mean pairwise difference π between nucleotide sequences. Finally, it will consider two ways to estimate the parameter θ.

5.1 The number of mutations on a gene genealogy

Consider the gene genealogy below:

```
--x----x---
           |
           |------x-
           |        |
------xx---         |
                    |-----x----
                    |
                    |
---x-x----------x---

|---t -----|---t ---|
     3          2
```

Each "x" represents a different mutation, and I'll assume that each mutation is at a different nucleotide site. Although there are 9 mutations, the variation within the sample results only from the 8 that are "downstream" of the genealogy's root. The 9th mutation would produce an identical effect on all members of the sample and is therefore of no interest to us. For our purposes, this is a genealogy with 8 mutations.

How many mutations would we expect to see in a sample of 3 genes? The answer will depend in part on the number of nucleotide sites being examined. We expect more mutations per generation on an entire chromosome than at a single nucleotide site. Although this effect is large, we can avoid dealing with it directly by using a flexible definition of the mutation rate, u. If we are studying a single nucleotide site, then u will represent the expected number of mutations per generation per site. If we are studying a larger region—say an entire gene or chromosome—then u will represent the expected number of mutations per generation in this entire region.

Unless we are studying a large genomic region, u will be very small and can be interpreted not only as the expected number of mutations per generation, but also as the probability of a single mutation. This is because we don't lose much, when mutations are very rare, by ignoring the remote possibility that several of them happen at once.

The expected number of mutations depends not only on the mutation rate, u, but also on the total length of the gene genealogy. In a sample of 3 genes, this length is

$$L = 3t_3 + 2t_2 \tag{5.1}$$

where t_3 is the length of the coalescent interval during which the genealogy had 3 lines of descent, and t_2 is the length of the other interval. If there are u mutations per generation, then we would expect this tree to have uL mutations—if we knew the value of L.

But since the value of L is ordinarily unknown,

$$E[\text{\# of mutations}] = E[uL] = uE[L]$$

To calculate this expected value, we need the expectation of L. The direct approach would involve inspecting thousands of gene genealogies, each generated by the coalescent process described in the last lecture. You cannot, of course, look at even one real gene genealogy, let alone thousands of them. You could write a computer program to simulate the process, but we can do the same job more easily using the theory from the previous lecture.

The E stands for "expectation," and in the present context it refers to an average over genealogies. For example, $E[t_2]$ is the expectation (that is, average) of t_2 over a very large number of genealogies. We learned in the last lecture that $E[t_2] = 2N$ and $E[t_3] = 2N/3$. Thus, equation 5.1 implies that the expectation of L is

$$E[L] = 3E[t_3] + 2E[t_2] = 2N + 4N$$

In general, the expected length of the coalescent interval during which there are i lines of descent is (see equation 4.3)

$$E[t_i] = \frac{4N}{i(i-1)}$$

and the contribution of this interval to the expected length (including all branches) of the tree is

$$iE[t_i] = \frac{4N}{i-1}$$

The total expected length of the tree is

$$E[L] = \sum_{i=2}^{K} iE[t_i] = 4N \sum_{i=1}^{K-1} \frac{1}{i} \tag{5.2}$$

where K is the number of genes in the sample. The expected number of mutations on the gene genealogy is thus

$$\begin{aligned} E[\text{\# of mutations}] &= uE[L] \\ &= 4Nu \sum_{i=1}^{K-1} \frac{1}{i} \\ &= \theta \sum_{i=1}^{K-1} \frac{1}{i} \end{aligned} \tag{5.3}$$

where u is the mutation rate per generation, and $\theta = 4Nu$, a quantity that appears often in the formulas of population genetics. It equals twice the number of mutations that occur each generation in the population as a whole. Like u its magnitude depends on the size of the genomic region under study (see p. 30). Below, we will consider the problem of estimating θ from genetic data.

5.2 The model of infinite sites

We can now calculate the expected number of mutations in a gene genealogy of any size. The next challenge is to connect this result to data—to genetic differences between individuals. The easiest approach involves what is known as the "model of infinite sites." This model assumes mutation never strikes the same nucleotide site twice. This model is never really correct, but it is often a good approximation, especially in intra-specific data sets where the genetic differences between individuals are small. It makes sense to use it when the mutation rate per nucleotide site is low enough that only a small fraction of nucleotide sites will mutate twice in any given gene genealogy.

This model, of course, is only an approximation. In the real world, nucleotide sites may mutate more than once. Appendix section 5.B (p. 35) shows that these violations occur, on average, at a fraction $(ut)^2/2$ of sites, where u is the mutation rate per site per generation and t is the number of generations.

For example, suppose that some branch of the gene genealogy is $t = 10^4$ generations long and that $u = 10^{-8}$. Along this branch, the expected number of mutations at a single nucleotide site is $ut = 10^{-4}$. In an entire human genome, the number of sites is about 3×10^9. The expected number of sites that violate the infinite sites model is therefore

$$3 \times 10^9 \times 10^{-8}/2 = 15$$

The model of infinite sites is expected to fail only at 15 sites out of 3 billion. This analysis should not be taken too literally, because the mutation rate is not really constant across the genome, and we may be interested in much longer time intervals. Nonetheless, it does show that the model of infinite sites works well when the product ut is small.

5.3 The number of segregating sites

If mutation never strikes the same site twice, then the number S of segregating (i.e. polymorphic) sites in a data set is the same as the number of mutations in its gene genealogy, as given in equation 5.3. The expected number of segregating sites is [36]

$$E[S] = \theta\{1 + 1/2 + 1/3 + \cdots + 1/(K-1)\} \tag{5.4}$$

Finally, we have arrived at a statistic that can be calculated from data. The expected number of segregating sites is equal to θ times some number that increases with sample size. Thus, we expect more segregating sites in a large sample. But the effect of sample size is not pronounced, because the sum in the expression above doesn't increase very fast. Here are a few example values:

K	$\sum_{i=1}^{K-1} 1/i$
2	1.00
3	1.50
5	2.08
10	2.82
100	5.17
1000	7.48

In a sample of 100, we expect only about 5 times as many segregating sites as in a sample of 2.

The effect of the population size, on the other hand, is pronounced since $E[S]$ is proportional to θ, and θ is proportional to the population's size. In a population twice as large, we expect twice as many segregating sites.

5.4 The mean pairwise difference

Given any pair of DNA sequences, it is a simple matter to count the number of nucleotide positions at which they differ. Given a sample of size K, there are $K(K-1)/2$ pairwise comparisons that can be made and we can count the number of nucleotide differences between each pair. Averaging these numbers gives a statistic that is called the "mean pairwise nucleotide difference" and is generally denoted by the symbol π.[1]

What is the expected value of π? The number of nucleotide site differences between a pair of sequences is the same as the number of segregating sites in a sample of size 2. Thus, equation 5.4 tells us that the average pair of sequences differs at θ sites. Averaging over all the pairs in a sample doesn't change this expectation, so

$$E[\pi] = \theta \tag{5.5}$$

This gives us the expected value of a second statistic that can be estimated from genetic data, and this time the formula is especially simple. As in the case of S, we can expect the value of π to be large if the population is large, small if the population is small.

[1] Some authors [20] use the capital letter (Π) to denote the mean pairwise differences per sequence and the lower-case letter (π) to refer to the mean pairwise difference per site. I use the lower-case letter for both purposes.

⋆EXERCISE **5–1** Just above, I said that if the expected difference between each pair of sequences is θ, then the expectation of π is also θ. Prove that this is so.

⋆EXERCISE **5–2** For the following questions, assume that the population mates at random, has constant size $2N = 1000$, that there is no selection, and that the mutation rate is $u = 1/2000$ per sequence per generation. Assume that you are working with a sample of $K = 5$ DNA sequences, and that mutations obey the model of infinite sites.

1. What is the expected depth of the gene tree? (In other words, the expected number of generations since the last common ancestor.)

2. What is the expected length of the tree? (In other words, the expected sum of the lengths of all branches in the tree.)

3. What is the expected number of mutations on the tree?

4. What is the expected number of mutational differences between each pair of sequences?

5.5 Theta and Two Ways to Estimate It

In this lecture, we have twice run into the parameter θ, which is proportional to the product of mutation rate and population size. This parameter appears often in population genetics, and it is useful to have a way to estimate it. The results above suggest two ways. Equation 5.4 suggests

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{K-1} \frac{1}{i}} \tag{5.6}$$

and equation 5.5 suggests.

$$\hat{\theta}_\pi = \pi \tag{5.7}$$

Here $\hat{\theta}$ is read "theta hat." The "hat" indicates that these formulas are intended to estimate the parameter θ.

To make sure that these formulas estimate the same parameter, it is important to be consistent. S usually refers to the number of segregating sites within some larger DNA sequence. To make π comparable, we interpret it here as the mean pairwise difference *per sequence* rather than that *per site*. We also need to interpret u as the mutation rate per sequence when we define $\theta = 4Nu$.

With these consistent definitions, $\hat{\theta}_S$ and $\hat{\theta}_\pi$ estimate the same parameter. It seems natural to suppose that their values would be similar in real data. Let's have a look at some human mitochondrial DNA sequence data.

5.6 Example

Jorde et al (ref) published sequence data from the control region of human mitochondrial DNA. The example described here uses 430 nucleotide positions from HVS1 (the first hypervariable region). Jorde et al sequenced DNAs from all three major human racial groups, but this example will deal only with the 77 Asian and 72 African sequences. In these data:

	Asian	African
S	82	63
$\sum_{i=1}^{K-1} 1/i$	4.915	4.847
$\hat{\theta}_S$ (per sequence)	16.685	12.998
π (per sequence)	6.231	9.208

The theory above says that π and $\hat{\theta}_S$ are both estimates of the parameter θ, so we have every reason to expect their values to be similar. Yet $\hat{\theta}_S$ is half again as large as π in the African data and nearly three times as large in the Asian. Why are these numbers so different?

There are at least four possibilities worth considering:

Sampling error To figure out whether these discrepancies are large enough to worry about, we need a theory of errors.

Natural selection The theory we have used assumes neutral evolution. If selection has been at work, then we have no reason to think that π and $\hat{\theta}_S$ will be equal. In fact, the difference between π and $\hat{\theta}_S$ is often used to test the hypothesis of selective neutrality. (Look up Tajima's D in any textbook on population genetics.)

Variation in population size Our theory also assumes that the population has been constant in size. We need to investigate how π and $\hat{\theta}_S$ respond to changes in population size.

Failure of the infinite sites model Our theory assumes that mutation never strikes the same site twice.

5.A The probability that a nucleotide site is polymorphic within a sample

In comparisons between pairs of haploid human genomes, about one nucleotide site in a thousand is polymorphic. In larger samples, of course, the polymorphic fraction is larger. What is the fraction (Q_K) that is expected to be polymorphic in a sample of size K? It is easier to work with the monomorphic fraction, $1 - Q_K$. The gene genealogy is monomorphic only if no mutation occur in any coalescent interval. Let us consider first the coalescent interval during which there were k ancestors. As we trace time backwards across this interval, we might encounter either of two types of event: a mutation or a coalescent event. We encounter a coalescent event first if (and only if) the interval is free of mutations.

During this interval, coalescent events happen with hazard $\lambda_2 = k(k-1)/4N$ per generation, and mutations happen with hazard $\lambda_1 = ku$, where u is the mutation rate per site per generation. Once an event does occur, it is a coalescent event with probability

$$z_k = \frac{\lambda_2}{\lambda_1 + \lambda_2} = 1 - \frac{\theta}{\theta + k - 1}$$

where $\theta = 4Nu$ [see reference 32, pp. 48–49]. This is the probability that the kth coalescent interval is free of mutations. When θ is small, z_k is approximately

$$z_k \approx 1 - \frac{\theta}{k-1} \approx e^{-\theta/(k-1)}$$

Because the mutations that occur in different coalescent intervals are independent, these probabilities multiply. The entire gene genealogy is free of mutations with probability

$$\begin{aligned} 1 - Q_K &= z_2 z_3 z_4 \cdots z_K \\ &\approx \exp[-\theta\{1 + 1/2 + 1/3 + ... + 1/(K-1)\}] \end{aligned} \tag{5.8}$$

The expected fraction of polymorphic sites is Q_K. For example, if $\theta = 1/1000$, the fraction of polymorphic sites should be 0.001 in a sample of size 2, 0.003 in a sample of 10, and 0.005 in a sample of 100. The polymorphic fraction increases with K, but not very fast.

5.B When you assume the model of infinite sites, how wrong are you likely to be? (optional)

The model of infinite sites assumes that mutation never strikes the same site twice. Clearly, this is only an approximation, and when we use this model we are bound to introduce errors. The question is, how large are these errors likely to be? What fraction of the sites in our data can be expected to mutate more than once?

To find out, let us consider the mutations that occur at some nucleotide site along a single branch of a gene genealogy. If the branch is t generations long, then the number, X, of mutations is a Poisson-distributed random variable with mean $\lambda = ut$, where u is the mutation rate per generation.

Consider the probability, P, that $X < 2$. This is the probability that our site conforms to the model of infinite sites. Because X is Poisson,

$$P = e^{-\lambda} + \lambda e^{-\lambda}$$

If λ is small, $e^{-\lambda} \approx 1 - \lambda + \lambda^2/2$, ignoring terms of order λ^3. (This is from the series expansion of the exponential function.) To this standard of approximation,

$$\begin{aligned} P &\approx 1 - \lambda + \lambda^2/2 \\ &\quad + \lambda - \lambda^2 + \lambda^3/2 \\ &\approx 1 - \lambda^2/2 \end{aligned}$$

The fraction of sites that *violate* the infinite sites model is approximately $1 - P = \lambda^2/2$—a very small number.

Lecture 6

The Site Frequency Spectrum

6.1 The empirical site frequency spectrum

In a sample of K genes, a polymorphic site can divide the sample into 1 mutant and $K - 1$ non-mutants, into 2 mutants and $K - 2$ non-mutants, and so on. There may be at most $K - 1$ copies of the mutant if the site is to be polymorphic. In many cases we can't tell which allele is the mutant, so category i gets conflated with category $K - i$. Such spectra are called "folded." I will call a site a "singleton" if the mutant is present in a single copy, a "doubleton" if it is present in two copies, and so on.

6.1.1 An unfolded spectrum

Consider the set of DNA sequence data below:

```
               123456
HumanSequence1 AATAGC
HumanSequence2 ..AC..
HumanSequence3 .TACT.
HumanSequence4 ..ACT.
---------------------
ChimpSequence1 AAAATC
ChimpSequence2 AAAATC
```

There are 4 human sequences and 2 chimpanzee sequences. There are 6 sites of which 4 are polymorphic (segregating) within the human sample. We calculate the empirical spectrum by considering the sites one at a time.

Site 1 is fixed and therefore does not contribute to the spectrum.

Site 2 has both an A and a T within the human sample but has only an A within the chimpanzee sample. The odds are that the ancestor of humans and chimps had an A at this site, so we can infer that T is the mutant allele. Since there is only one copy of T in the human sample, site 2 is a singleton. So far, our spectrum looks like this:

```
Singletons : 1
Doubletons :
Tripletons :
```

Site 3 is like site 2. The human sample has a T and 3 As, and the chimp sample has only As. We infer that T is the mutant allele and count this site as another singleton. The spectrum now looks like this:

```
Singletons : 2
Doubletons :
Tripletons :
```

Site 4 has an A and 3 Cs, but it appears that A was the ancestral allele. We count this site as a tripleton, so the spectrum becomes

```
Singletons : 2
Doubletons :
Tripletons : 1
```

Site 5 has 2 Gs and 2 Ts. It does not matter which of these is ancestral. Either way, the site is a doubleton. The spectrum becomes

```
Singletons : 2
Doubletons : 1
Tripletons : 1
```

Site 6 does not contribute to the spectrum.

We are done. The empirical spectrum has 2 singletons, 1 doubleton, and 1 tripleton.

6.1.2 A folded spectrum

In the preceding section, the chimpanzee sequences were used at each site to infer which nucleotide was ancestral and which was the mutant. Let us now pretend that we have no chimpanzee sequences and therefore cannot tell the the ancestral allele from the mutant. Instead of counting mutants, we will count the rarest (sometimes called the minor) allele at each site. This time, however, I will omit the invariant sites (1 and 6), which do not contribute to the spectrum.

Site 2 The rare allele, T, is present in a single copy, so this site contributes to the singleton category just as it did for the unfolded spectrum.

Site 3 Ditto: another singleton

Site 4 The rare allele, A, is present in a single copy, so this site is a singleton. Recall that it was a tripleton in the unfolded spectrum.

Site 5 A doubleton

The folded spectrum looks like this:

```
Singletons : 3
Doubletons : 1
```

The only difference is that site 4, which was a tripleton in the unfolded spectrum, becomes a singleton in the folded spectrum.

In general, the ith category of the folded spectrum contains not only category i of the unfolded spectrum, but also category $K - i$, where K is the number of DNA sequences in the sample.

6.2 The expected spectrum under neutrality and constant population size

This section deals with the special case of selective neutrality and constant population size. I will assume initially that we can tell mutants from ancestral alleles so that our spectrum will be unfolded.

6.2.1 A site's position in the spectrum depends on its position in the gene tree

Consider the following gene tree:

```
     ----A------
                |
                |----B---
                |        |
     -----------         |
                         |-----------
                         |
                         |
     ------C-------------
```

Mutations A and C are singletons, whereas B is a doubleton. A mutation that occurs in the most recent coalescent interval can only be a singleton. One that occurs in the next most recent interval can be either a singleton or a doubleton. One in the interval before that can be a singleton, a doubleton, or a tripleton. And so on.

6.2.2 A tree with two leaves has nothing but singletons

To get a sense of how the process works, it helps to start with a tree with just two leaves:

```
 -------------------
                    |
                    |
                    |-----------
                    |
                    |
 -------------------

|- 2N generations --|
```

Since the hazard is $1/2N$, the mean depth of this tree is $2N$ generations and the total length is $4N$. We expect $4Nu = \theta$ mutations, all of which will be singletons.

6.2.3 A tree with three leaves has (on average) the same number of singletons and half that number of doubletons

Now consider a tree with three leaves:

```
---------
         |
         |--------------------
         |                    |
---------                     |
                              |--------
                              |
                              |
------------------------------

|--2N/3--|-------- 2N --------|

         Generations
```

If we could look at the spectrum just before the most recent coalescent event, it would look just like that of the tree with two leaves: θ singletons and no doubletons. At the time of the coalescent event, half of these mutations (the ones on the upper branch) become doubletons. There is no further change in the number of doubletons, so the expected number of doubletons in a 3-leaf gene tree is $\theta/2$. (We don't need to worry that mutation will turn any of our doubletons back into singletons because, under the infinite sites model, mutation never strikes the same site twice.)

We start this latest coalescent interval with $\theta/2$ singletons, but then more singletons are added because of new mutations. How many new mutations should we expect to see? The interval's expected length is $2N/3$ generations (see section 4.3), and it contains 3 lines of descent, so the sum of the branch lengths within this interval is (on average) $2N$ generations. We therefore expect $2Nu = \theta/2$ new singleton mutations.

The number of singletons that is added is precisely equal to the number that was lost. Thus, the new spectrum has θ singletons and $\theta/2$ doubletons.

6.2.4 The theoretical spectrum for an arbitrary number of leaves

The argument that I used above gets progressively more tedious as leaves are added. It is better to use a different argument. I will skip the details here, but the results look like this [6]:

Sample size	Theoretical spectrum (singletons, doubletons, ...)			
2	θ			
3	θ,	$\theta/2$		
4	θ,	$\theta/2$,	$\theta/3$	
5	θ,	$\theta/2$,	$\theta/3$,	$\theta/4$
		Etcetera		

It is remarkable that as we increase sample size, the number of mutants in each category doesn't change. We merely add a new category at the right side of the spectrum.

To use the theoretical formula with data, we need to substitute some estimate of θ. We might use the mean pairwise difference, π, or the estimator

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{K-1} \frac{1}{i}}$$

where K is the number of DNA sequences in the sample. Either of these estimators might work, since both of them estimate θ under the stationary neutral model (see the discussion of equation 5.6 on page 33). To choose between these estimators, we need some additional criterion.

The sum of the observed spectrum is equal to the number S of segregating sites. It would be useful if the theoretical spectrum summed to the same value. This turns out to be so only if $\hat{\theta}_S$ is used to estimate θ.

6.2.5 Folded theoretical spectra

When the spectrum is folded, we cannot distinguish category i from category $K - i$. Consequently, the expected number in category i in the folded spectrum is the sum θ/i and $\theta/(K - i)$, the expected numbers in the two corresponding categories in the unfolded spectrum.

This works so long as i and $K - i$ are not the same number. If they *are* the same number, then the expected spectrum is simply θ/i.

6.3 Human site frequency spectra

Figure 6.1 shows all of the human site frequency spectra that I was able to cull from the literature in the year 2000. In each plot, the empirical spectrum is shown as a histogram, and the expected values under neutral evolution with constant population size are shown as bold dots. The top row shows three systems in which there is an excess of singletons, compared with the stationary neutral model. The middle row shows three systems that seem to fit the neutral model, and the bottom row shows three systems in which there is a deficit of singletons and an excess of sites at intermediate frequency.

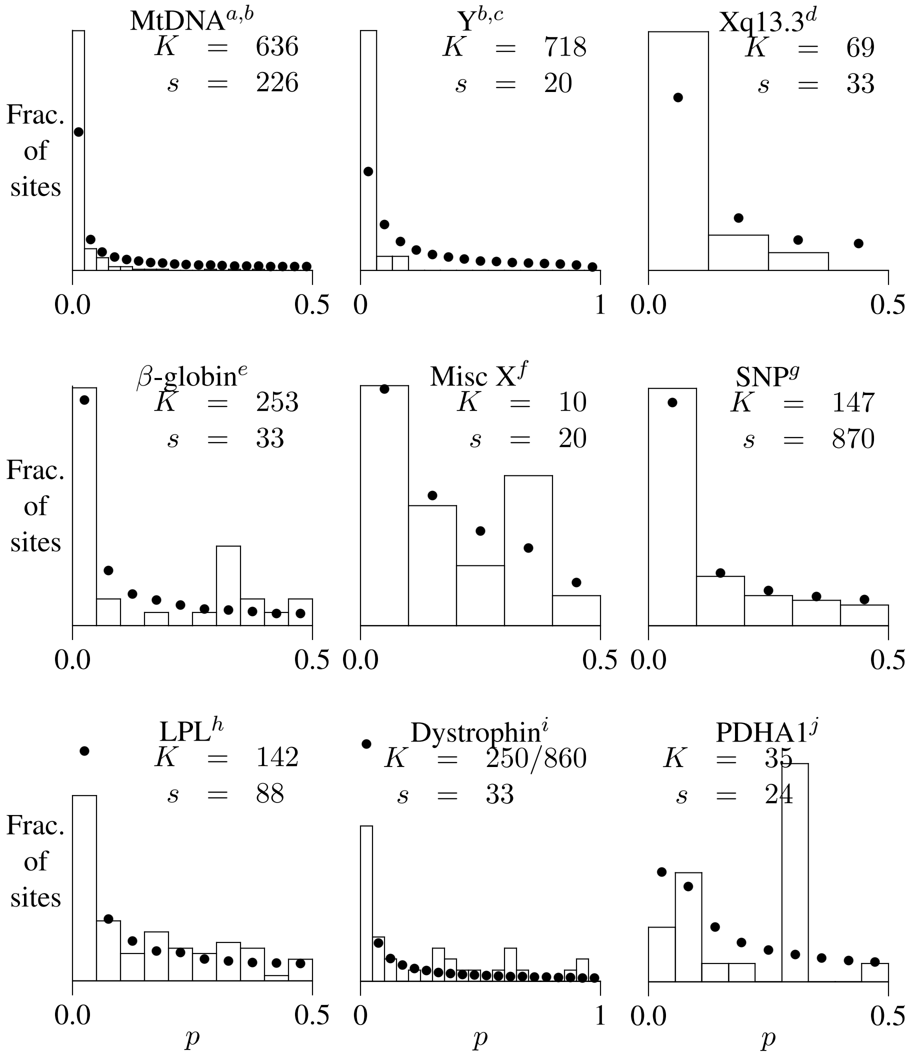

Figure 6.1: Site frequency spectra. The open rectangles in each panel show observed spectra; the bold dots show the spectra expected under the infinite sites model with no selection and constant population size. K is the number of chromosomes sampled, s is the number of segregating sites, and p is the frequency of the mutant allele (where ancestral state could be determined) or the rarest allele (where ancestral state is unknown). Sources: [a][13], [b][9], [c][30], [d][15], [e][8], [f][21], [g][7], [h][4], [i][39], [j][10]

6.4 Exercises

$\star$ EXERCISE **6–1** For this exercise, use the toy data set in section 1.1, on p. 5. (1) Use S to estimate θ, (2) from this value, calculate the number of sites expected in each frequency category, (3) fold the resulting theoretical spectrum by summing values for i and $K-i$. (4) Compare the result with the empirical spectrum that we calculated earlier, in section 1.3.3.

Lecture 7

The Mismatch Distribution

7.1 The observed mismatch distribution

Count the number of site differences between each pair of sequences in a sample, and use the resulting counts to build a histogram. You end up with a "mismatch distribution." The ith entry of the mismatch distribution is the number of pairs of sequences that differ by i sites.

For example, consider this data set:

S01	AAACT GTCAT
S02	. . . . . A . T . .
S03	. . G . . A . . . .
S04	. . G . . A . T . .
S05	. . . . . A . . . .

To calculate the mismatch distribution, we need to count the differences between every pair of sequences. Here are my counts:

	Pairwise
Pair	differences
1×2	2
1×3	2
1×4	3
1×5	1
2×3	2
2×4	1
2×5	1
3×4	1
3×5	1
4×5	2

There are five 1s, four 2s, and one 3. Thus, the mismatch distribution is

	Number
Pairwise	of
differences	pairs
0	0
1	5
2	4
3	1

Here, the right column gives the number of pairs that exhibit each level of difference. We often re-express these as fractions of the number of pairs of sequences. Since there are 10 pairs in our data set, this gives:

	Fraction
Pairwise	of
differences	pairs
0	0.0
1	0.5
2	0.4
3	0.1

Now the numbers in the right column sum to 1—they are relative frequencies. This is the *observed* or *empirical* mismatch distribution.

7.2 The expected mismatch distribution under neutral evolution with constant population size

The previous section concerned the observed mismatch distribution, which we calculate from genetic data. Each entry in this distribution is a random variable, so it is natural to wonder about its expected value. This is easy to calculate, under a model of constant size and selective neutrality: a random pair of sequences differs by i sites with probability [36]

$$F_i = \left(\frac{1}{\theta+1}\right)\left(\frac{\theta}{\theta+1}\right)^i, \qquad (i = 0, 1, 2, \ldots) \tag{7.1}$$

where $\theta = 4Nu$, u is the mutation rate per generation, and $2N$ is the number of genes in the population. This formula is graphed in figure 7.1 along with an empirical mismatch distribution from human mtDNA. (The empirical distribution, shown as open circles, is analogous to the one calculated in section 7.1.)

The poor fit between the observed and expected curves is striking. As usual, there are several hypotheses to consider:

Sampling error Perhaps the poor fit is an artifact attributable to sampling error. This possibility is especially important here because the pairs of sequences in this analysis are not independent: They are correlated both because each sequence participates in several pairs and also because of the genealogical relationships among sequences.

Selection More on this later

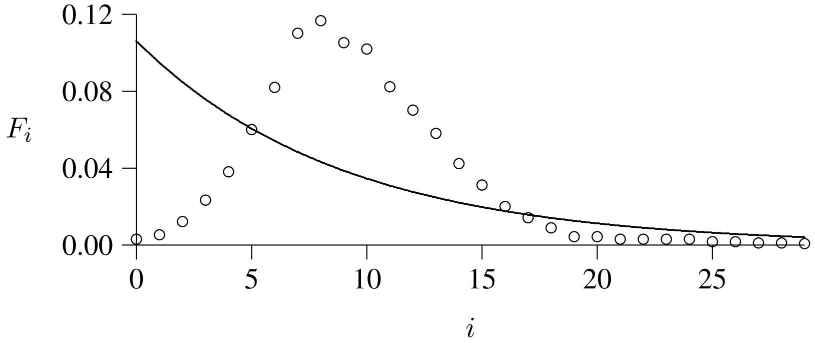

Figure 7.1: The poor fit of the equilibrium distribution
The open circles show the empirical pairwise difference distribution of Cann et al. [3], based on their figure 1. The solid line is an equilibrium distribution with the same mean.

Failure of infinite sites hypothesis

Non-random mating

Variation in population size

Work has been done on all of these possibilities, but we will consider only the last.

7.3 Coalescent theory in a population of varying size

The principles of coalescent theory still hold in a population of varying size. At any given time, t, the hazard of a coalescent event is

$$h_i(t) = \frac{i(i-1)}{4N(t)}$$

where i is the number of distinct lines of descent in the gene genealogy at time t and $2N(t)$ is the number of genes in the population size at time t.

It is no longer true, however, that the mean waiting time until a coalescent event is $1/h_i$. That only works in populations of constant size. Some theoretical results are still possible, but the more complex the model the more we are forced to base inferences on computer simulations.

7.4 The coalescent as an algorithm for computer simulations

Fortunately, coalescent theory is just as useful in computer simulations as in theoretical work. The style of reasoning is the same, except that coalescent intervals are generated from random numbers. As in the case of constant population size, intervals tend to be long in those parts of the tree where there are only a few lines of descent.

Now, however, there is an additional factor to consider: Coalescent intervals also tend to be long when the population size is large. It is not hard to understand why this should be so. Two random genes drawn from a large population are likely to be distantly related and thus separated by a long genealogical path. In a small population, a pair of genes is likely to be separated by a shorter path.

Table 7.1: Hypothetical two-epoch population history

	$2N_i$	t_i
Epoch	(genes)	(generations)
0	10^6	2000
1	10^4	∞

Computer simulations are most useful if they deal with only a few parameters at a time. Thus, we need some economical way to describe the history of a population that changes in size.

7.5 Stepwise models of population history

Population growth is ordinarily regulated by density-dependent mechanisms that allow small populations to grow and cause large ones to shrink. Consequently, most biologists think that the sizes of most populations are roughly constant most of the time. Now and then, something happens to disturb this equilibrium, and the population either grows or shrinks until a new equilibrium is attained. Over a long time scale, the periods of change would look like relatively sudden changes in the size of a population that was otherwise roughly constant.

This is the conventional view of demographic history, and it motivates the stepwise model of population history that is used here. I will assume that history can be divided into a series of epochs during which the population does not change. Epochs are separated by episodes of rapid change, which I treat as instantaneous. They are numbered backwards from the most recent (epoch 0) to the most ancient.[1]

Table 7.1 shows the parameters of a hypothetical population history with two epochs. The table describes a population that was small (10,000 people) early in time, grew suddenly 2000 generations ago, and has been large (1,000,000 people) ever since. The history of this hypothetical population is graphed in figure 7.2.

With genetic data is it not possible to estimate N_i and t_i directly. We can, however, estimate the related

[1] Rogers and Harpending [25] numbered them in the opposite direction.

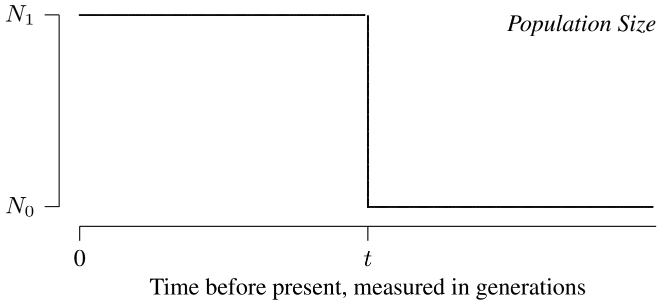

Figure 7.2: At t generations before the present, the population expands from female size N_1 to female size N_0. The model's 3 parameters are $\theta_0 \equiv 2uN_0$, $\theta_1 \equiv 2uN_1$, and $\tau \equiv 2ut$, where u is the aggregate mutation rate over all sites.

Table 7.2: Alternate parameterization of the same population history

Epoch	θ_i	τ_i
0	2000	4
1	20	∞

parameters

$$\theta_i = 4N_i u \tag{7.2}$$

$$\tau_i = 2ut_i \tag{7.3}$$

where u is the mutation rate. θ_i measures the size of the population during epoch i in units of $1/(2u)$ genes; τ_i measures the length of this epoch in units of $1/(2u)$ generations. The population history of table 7.1 can be re-expressed in terms of θ and τ by multiplying each parameter by $2u$. For example, if the mutation rate is 0.001 (per sequence per generation), then we obtain table 7.2.

7.6 Simulations of stationary populations

Figures 7.3–7.6, on pages 50–53, each show a simulation of a population with constant size. In each figure, the upper panel shows the time path of population size, which is simply a horizontal line in these first four figures. Proceding down each page, you will find the gene genealogy, the mismatch distribution and the spectrum. The expected mismatch distribution is shown as a solid line; the expected spectrum is shown as a series of filled circles. In both cases, the expected values refer to a model of neutral evolution in a population of constant size.

These figures make use of what I will call the "mutational time scale." A unit of mutational time is the amount of time it takes, on average, for one mutation to accumulate along the genealogical path separating two sequences. Thus, if two sequences have been separate for 5 units of mutational time, we expect them to differ at about 5 nucleotide sites. Each unit of mutational time is equal to $1/(2u)$ generations.

● EXAMPLE **7–1**

In human mitochondrial D-loop sequence, it has been estimated that the mutation rate is 4.1×10^{-6} per nucleotide per generation [33]. In the Jorde et al HVSI data there are 430 nucleotides, so $u = 430 \times 4.1 \times 10^{-6} = 0.0018$ per generation. Each unit of mutational time is $1/(2u) = 278$ generations, which would correspond roughly to 6950 years.

Notice that neither the mismatch distributions nor the spectra of these equilibrium populations look much like the theoretical formulas would predict. Far from showing the smooth decline seen in the theoretical curves, the simulated mismatch distributions tend to be ragged, with multiple peaks. The site frequency spectra also exhibit pronounced departures from their expected values. In order to get answers that look like the theory, we would need to run the simulations many times and average the results. This is bad news, since in real data analysis we cannot rewind the evolutionary process and look at it again and again. The situation is not hopeless, but it is clear that merely inspecting graphs such as these is not going to give us dependable answers. We are going to need more sophisticated statistical methods.

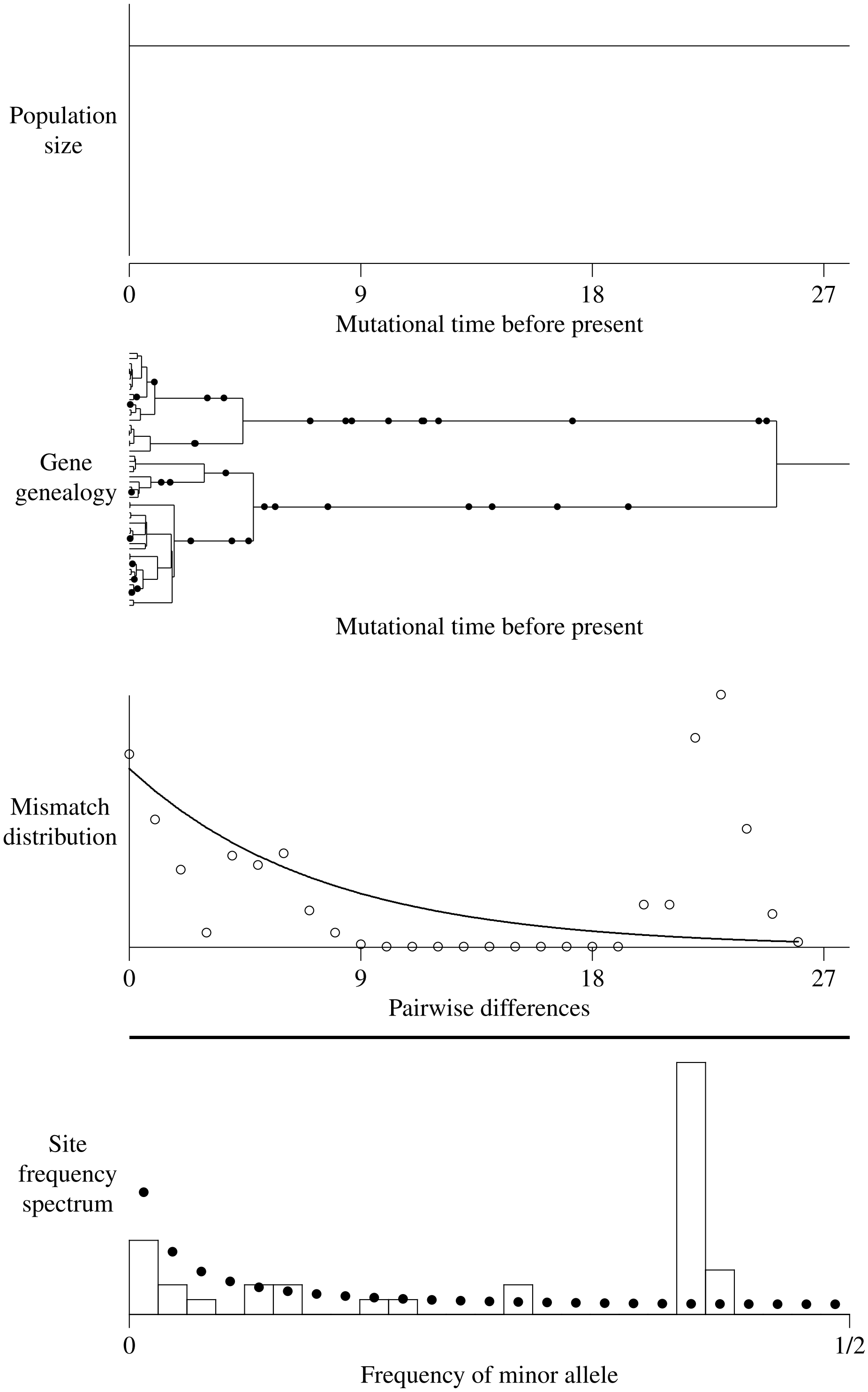

Figure 7.3: An equilibrium population with $\theta = 7$

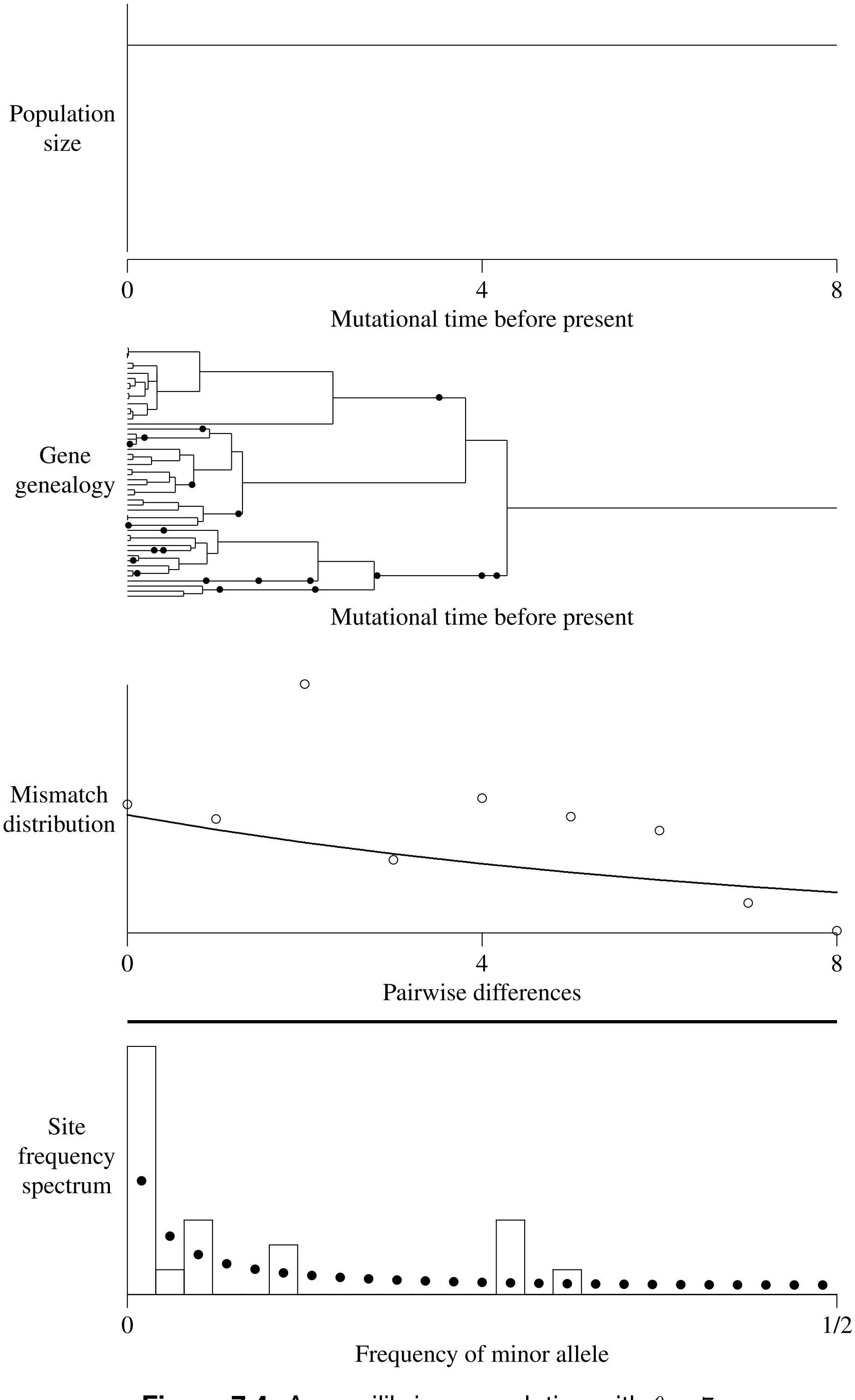

Figure 7.4: An equilibrium population with $\theta = 7$

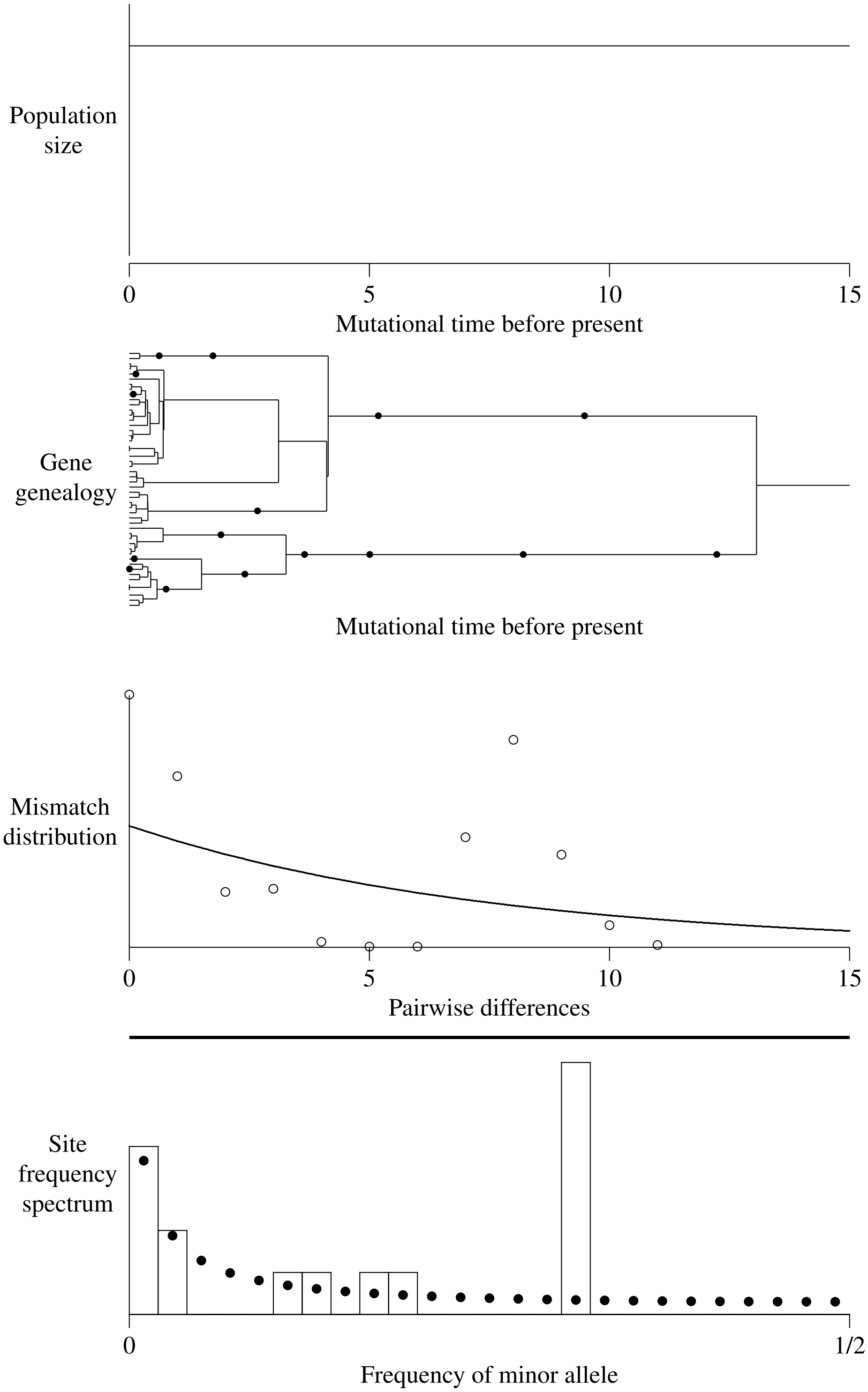

Figure 7.5: An equilibrium population with $\theta = 7$

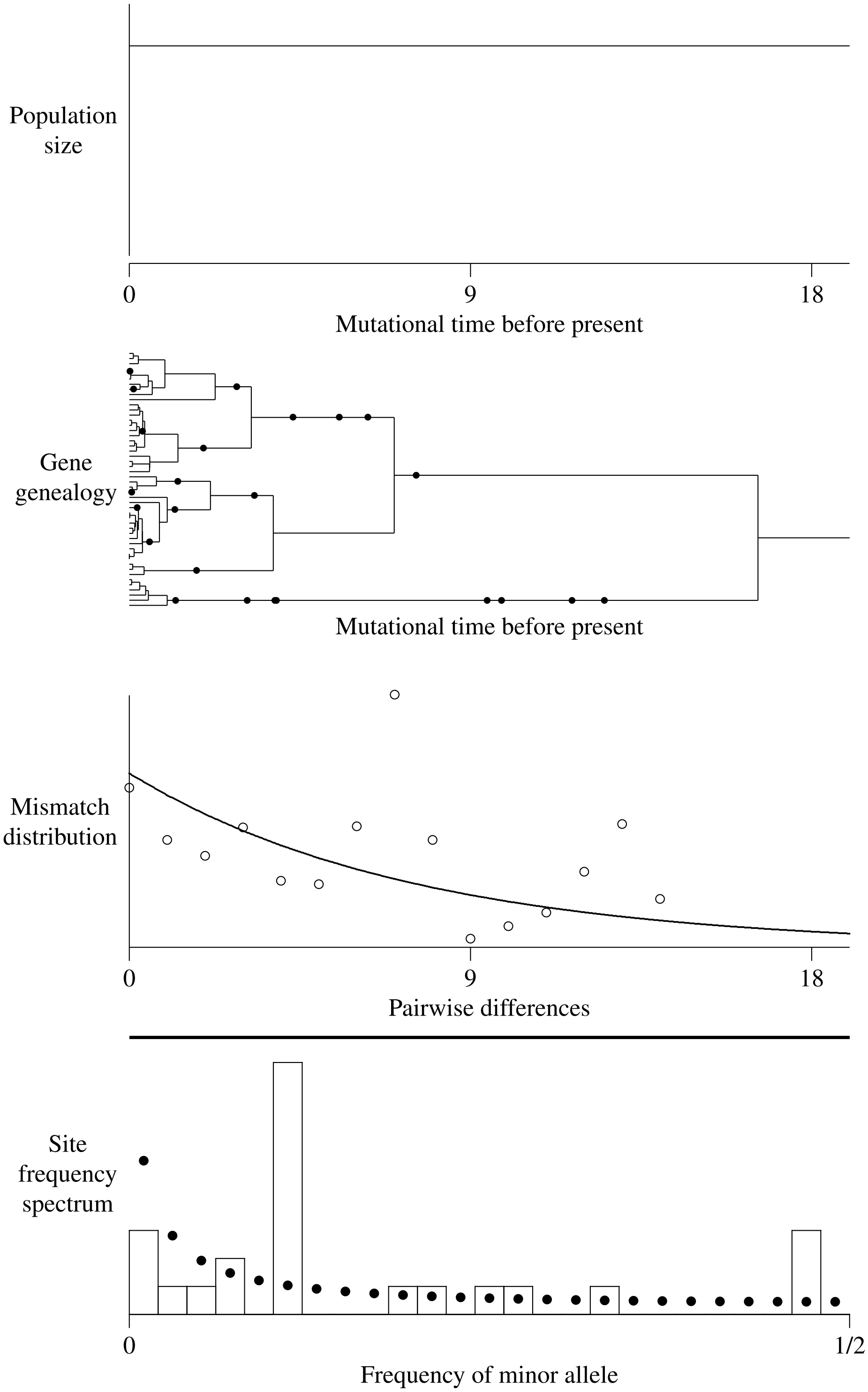

Figure 7.6: An equilibrium population with $\theta = 7$

7.7 Simulations of expanded populations

Figures 7.7–7.10, on pages 55–58, show simulations of populations that grew suddenly by 100-fold at 7 units of mutational time before the present. In these graphs, the solid lines drawn for mismatch distributions refer to the expectation under the model of population history that was used in the simulations. On the other hand, the filled circles shown with the site frequency spectra refer as before to a model of constant population size.

The gene genealogies of these expanded populations look different. Coalescent events occur only rarely during the period when the population was large, but occur rapidly in the earlier period when the population was small. This gives the gene genealogies a comb-like shape, which seldom appears in the genealogies of stationary populations. Many pairs of individuals differ by just over 7 units of mutational time.

The mismatch distributions in these simulations are all unimodal, with peaks just a little before 7. This reflects the fact that many pairs of individuals differ by just over 7 units of mutational time.

The spectra in these simulations exhibit an excess of singletons. This is because the terminal branches in the gene genealogies are long and attract a disproportionate number of the mutations. The mutations that fall on these long terminal branches are all singletons. Methods for calculating the expected spectrum are introduced by Harpending et al. [9] and by Wooding and Rogers [38].

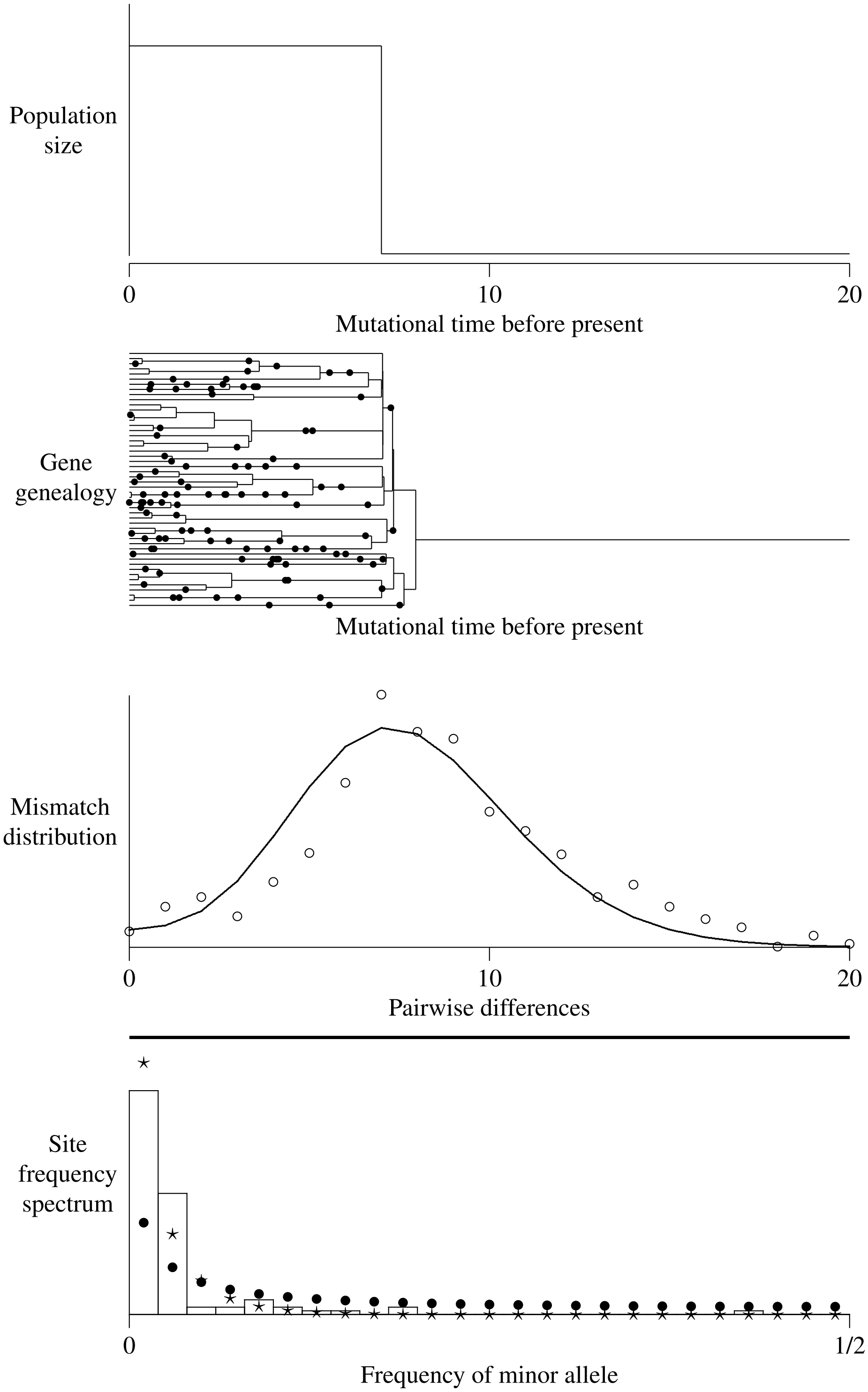

Figure 7.7: A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$. $\star$: expected; $\bullet$: expected if population size is constant.

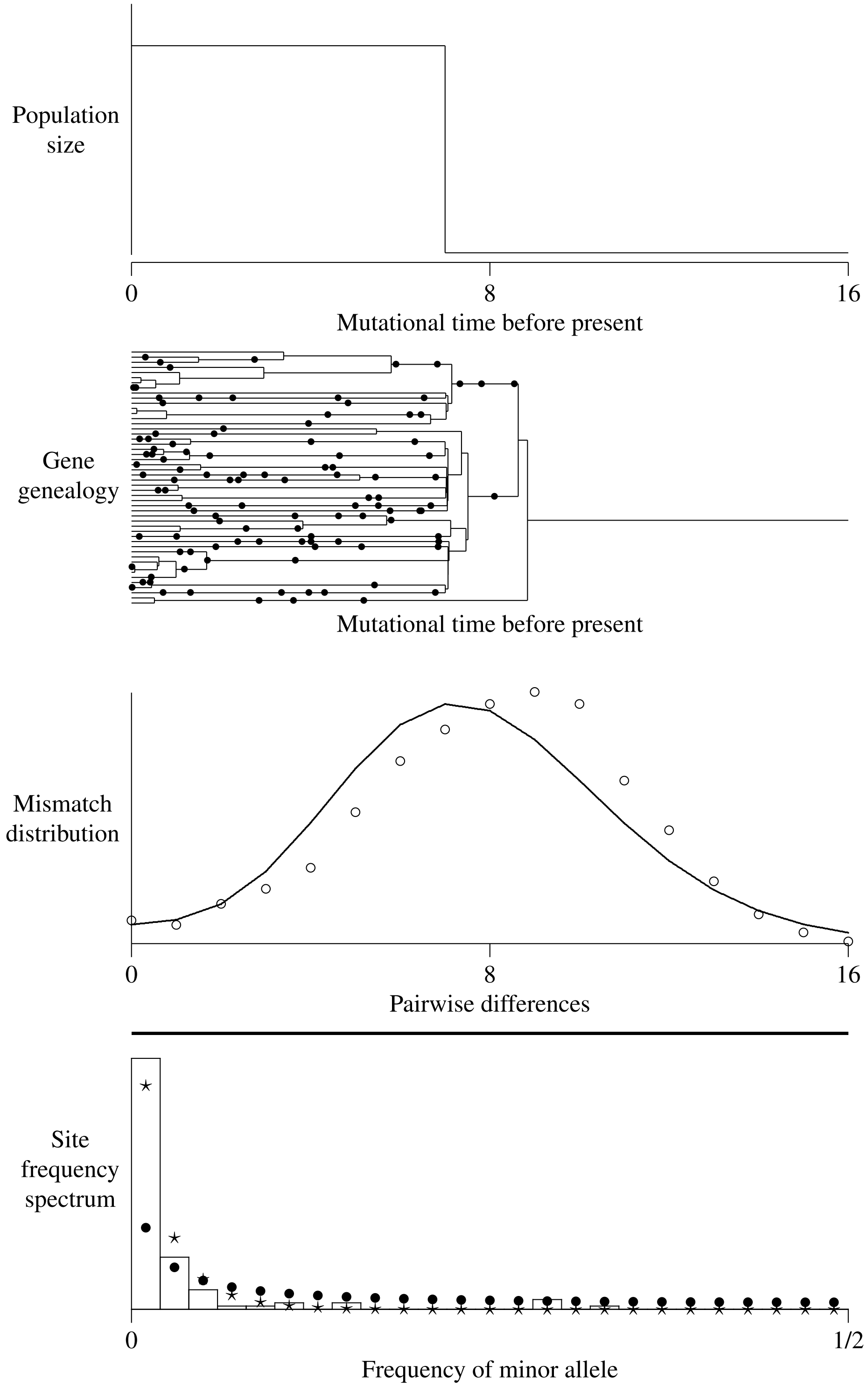

Figure 7.8: A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$. $\star$: expected; •: expected if population size is constant.

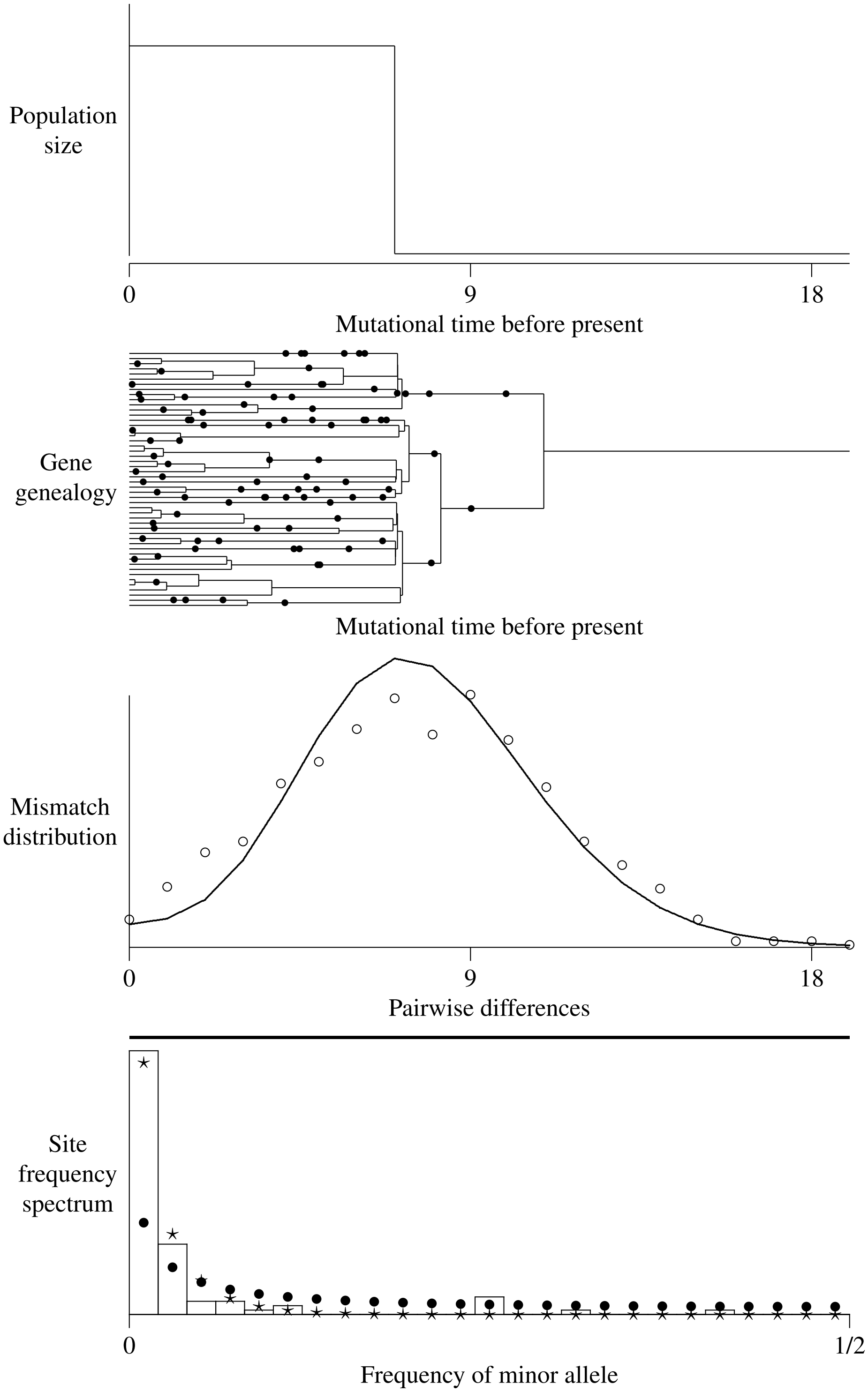

Figure 7.9: A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$. $\star$: expected; $\bullet$: expected if population size is constant.

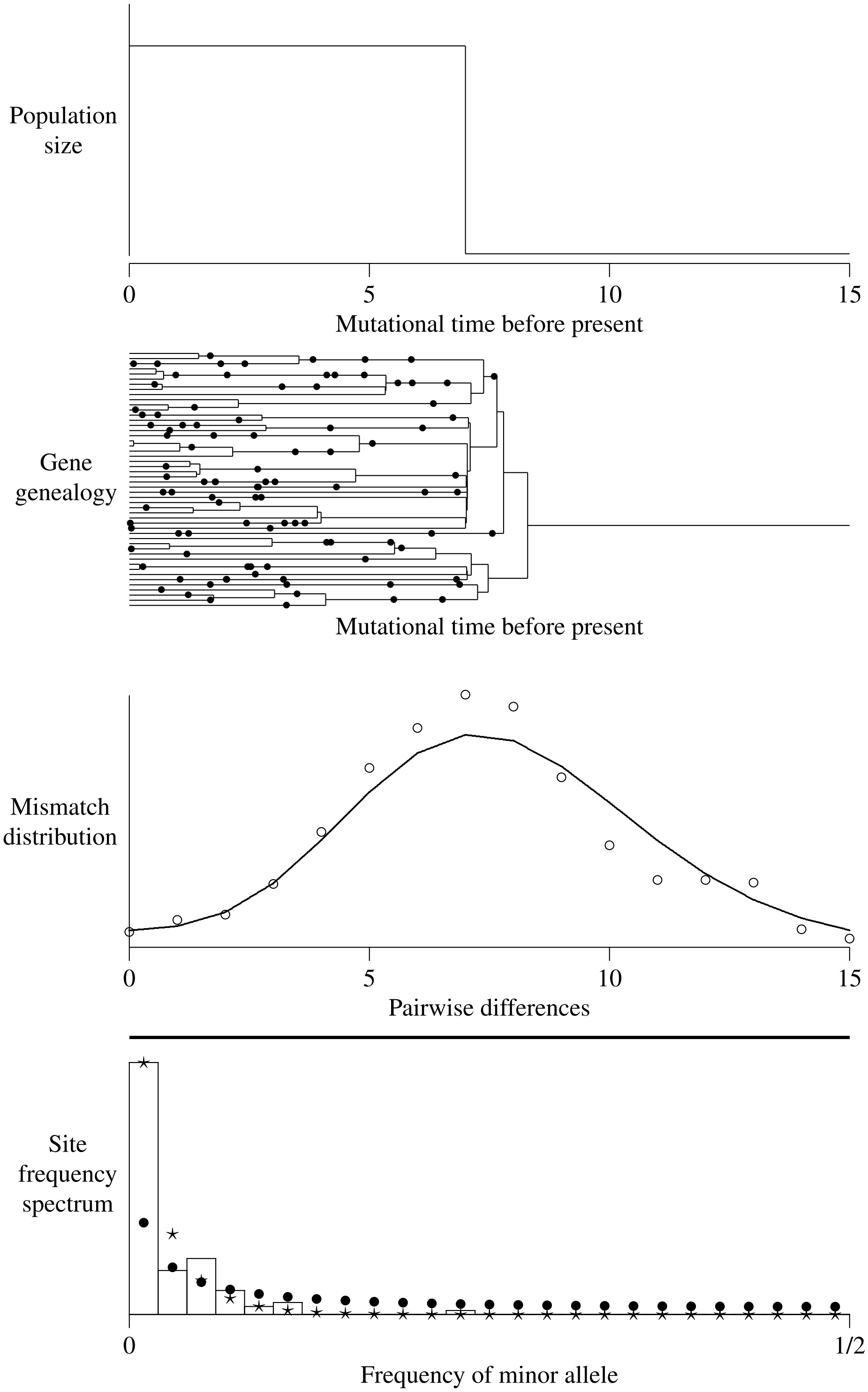

Figure 7.10: A coalescent simulation with growth: $\theta_0 = 1$, $\theta_1 = 100$, $\tau = 7$. $\star$: expected; $\bullet$: expected if population size is constant.

7.A Point estimators for expanded populations (optional)

If the expansion has been dramatic, the mismatch distribution will be a smooth wave with a single mode. In that case, the time of the expansion and the size of the pre-expansion population can be estimated using the following statistics [23]:

$$\hat{\theta}_1 = \sqrt{v - \pi} \tag{7.4}$$

$$\hat{\tau}_0 = m - \hat{\theta}_1 \tag{7.5}$$

where π is the mean pairwise difference per sequence within the sample, m is the mean of pairwise differences, and v is the variance.

7.B Statistical properties of point estimates (optional)

To determine the statistical properties of $\hat{\theta}_1$ and $\hat{\tau}_0$, I used the coalescent algorithm [12] to generate 1,000 simulated data sets at each of a wide variety of parameter values. In order to allow for changes in population size, I used a modified version of the coalescent algorithm, which is described elsewhere [24]. I estimated θ_1 and τ_0 from each simulated data set, thus obtaining an estimate of the sampling distribution of the estimators for each set of parameter values.

Figure 7.11 shows how the sampling distribution of $\hat{\tau}_0$ changes in response to variation in the underlying parameter τ_0. If $\hat{\tau}_0$ is in fact an estimator of τ_0, we would expect the median of $\hat{\tau}_0$ (shown as a solid line in the figure) to increase in response to increases in τ_0. This is indeed the case. An ideal estimator should also have a relatively narrow distribution at each value of τ_0. The dashed and dotted lines show that $\hat{\tau}_0$ also satisfies this test. The dashed lines enclose the central 50% of the distribution, and the dotted lines the central 95%. Both sets of lines enclose a relatively narrow interval about the median. In all of these respects, $\hat{\tau}_0$ behaves as an estimator of τ_0.

Figure 7.12 performs a similar analysis on $\hat{\theta}_1$, and shows it to perform well as an estimator when $\theta_1 > 1$. The distribution is tightly centered about the bold dots, showing that $\hat{\theta}$ is rich in information and nearly unbiased when $\hat{\theta}_1 > 1$. But when $\theta_1 < 1$, the upper quantiles of $\log_{10} \hat{\theta}_1$ are horizontal, while the median and lower quantiles of $\hat{\theta}_1$ equal 0. Thus an estimate of $\hat{\theta}_1 \approx 1$ is equally consistent with the hypotheses that $\theta_1 = 1$ and that $\theta_1 = 0$. Although $\hat{\theta}_1$ will always allow us to place an upper bound on θ_1, it can provide no lower bound unless $\hat{\theta}_1$ is much greater than 1. This is no serious problem; it means only that when the estimate is near unity, the confidence interval will reach all the way to 0.

(Need some prose to go with figure 7.13.)

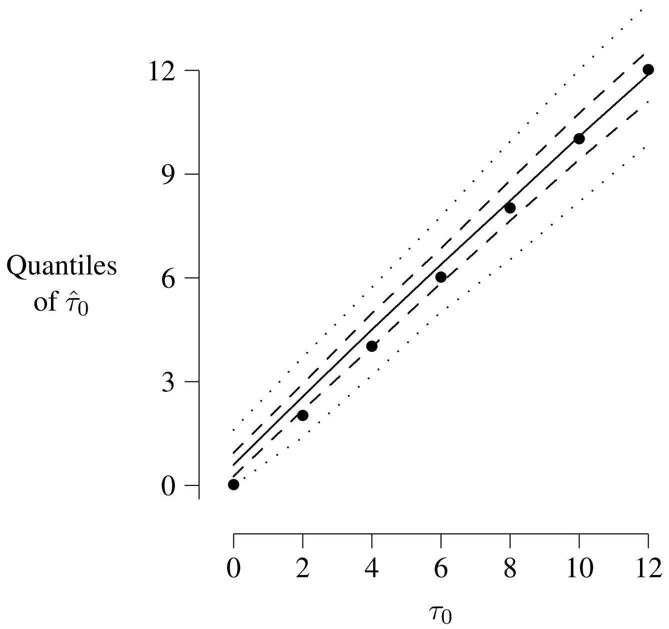

Figure 7.11: Quantiles of $\hat{\tau}_0$. 1000 data sets were simulated at each of several values of τ_0, and each was used to estimate the model's parameters. The bold dots indicate points at which $\hat{\tau}_0 = \tau_0$. The solid line is the median, the dashed lines enclose the central 50% of the distribution, and the dotted lines the central 95%. Each simulated data set was generated using the coalescent algorithm with $\theta_0 = 500$, $\theta_1 = 1$, and $N = 147$.

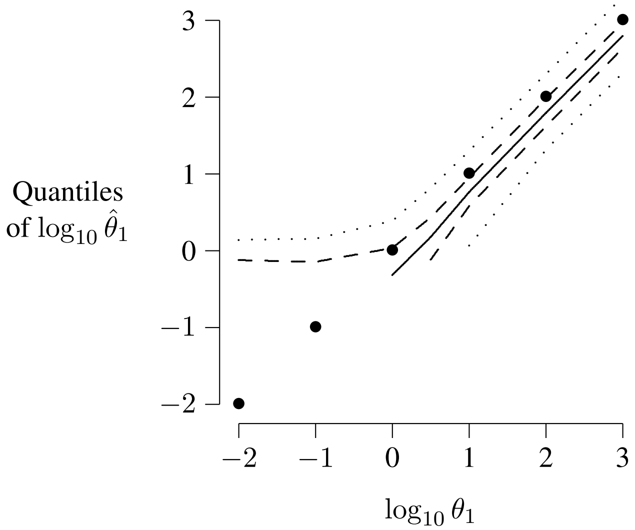

Figure 7.12: Quantiles of $\hat{\theta}_1$. 1,000 data sets were simulated at each of several values of θ_1, and each was used to estimate the model's three parameters. In each run, $\theta_0 = 1000$, $\tau = 7$, and $N = 147$. The lines and bold dots are interpreted as in figure 7.11.

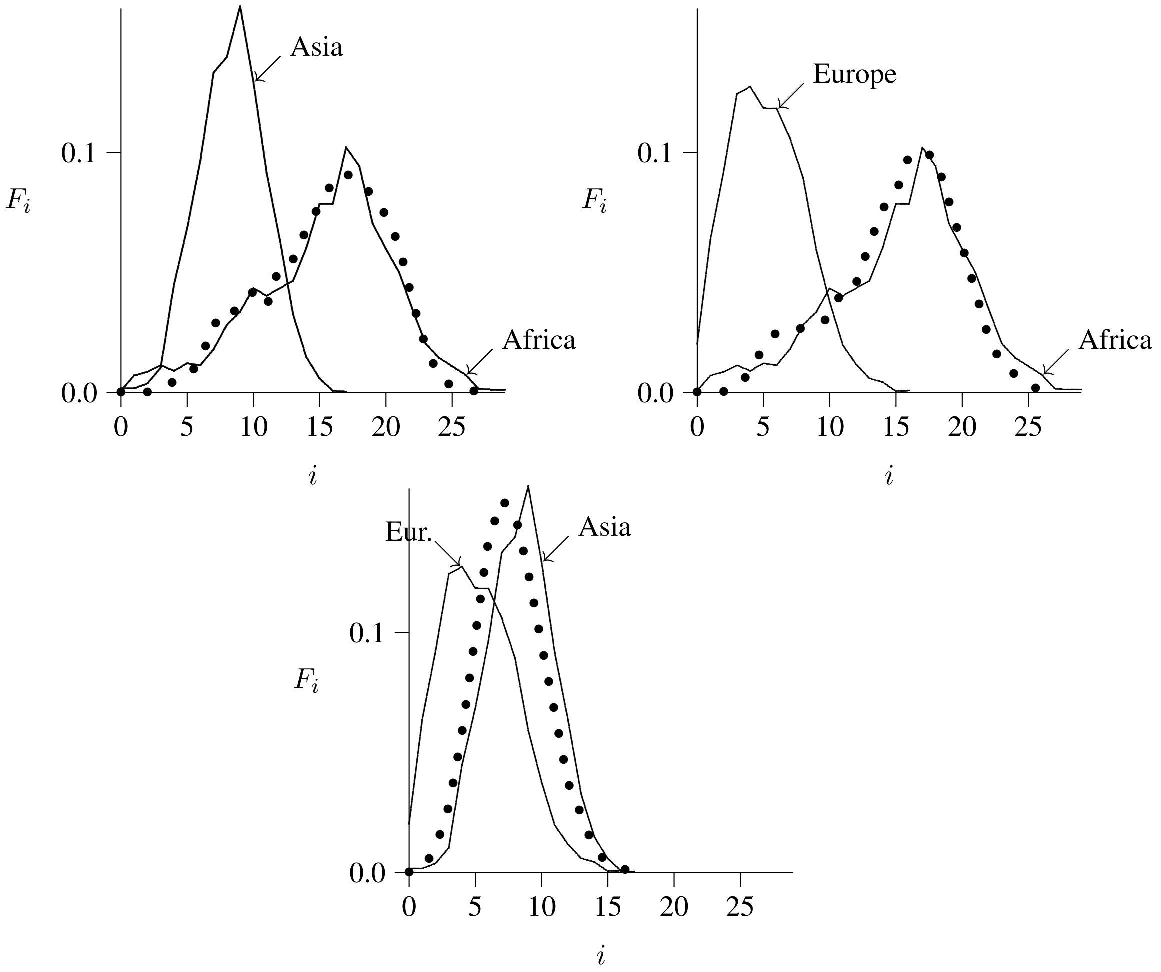

Figure 7.13: Mitochondrial Mismatch Distributions

In each panel, the solid lines show mismatch distributions for within-population comparisons, and the dotted lines show the analogous between-population comparisons. The data comprise 72 Africans, 77 Asians, and 89 Europeans [13].

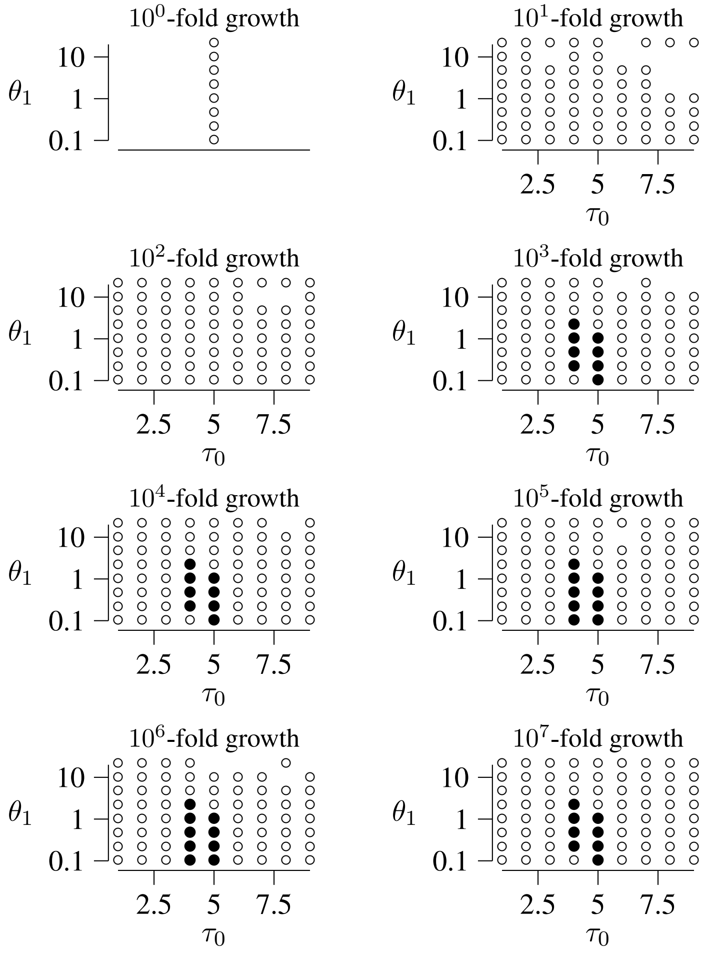

Figure 7.14: An experiment designed to test the method for generating confidence regions. Rather than beginning with a real data set, this experiment began with a simulated data set that was generated under the assumptions that $N = 147$, and $(\tau, \theta_0, \theta_1) = (4, 1, 500)$. A confidence region was then generated as described in the text. Open circles ($\circ$) represent points outside the 95% confidence region; closed circles ($\bullet$) represent points within. Note that the correct hypothesis $\{(\tau, \theta_0, \theta_1) = (4, 1, 500)\}$ falls inside the confidence region.

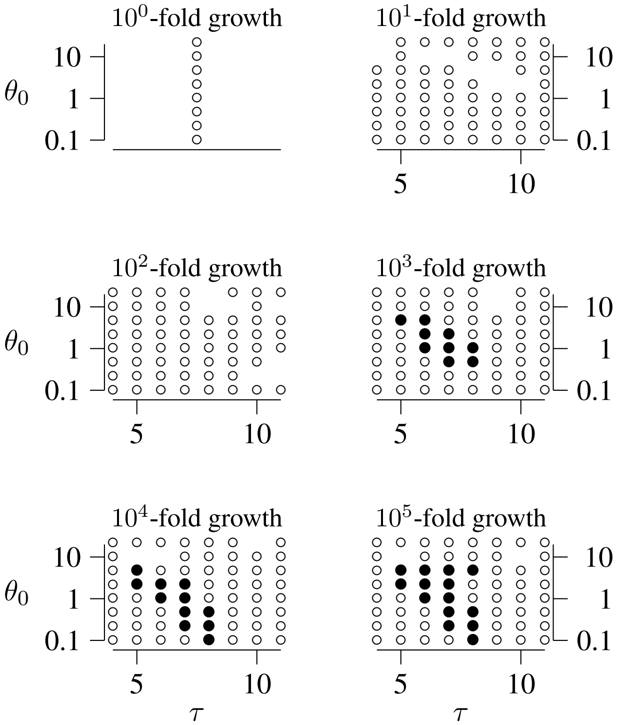

Figure 7.15: 95% confidence region for the Cann-Stoneking-Wilson data

Open circles ($\circ$) represent points outside the 95% confidence region; closed circles ($\bullet$) represent points within. The marginal confidence regions are τ : $[5, 8]$, θ_0 : $[0, 4.6]$, and θ_1/θ_0 : $(100, \infty]$. If the mutation rate is $u = 0.001$ per sequence per generation, and generations are 25 years, these intervals become t : $[62, 100]$ thousand years, and N_0 : $[0, 2320]$ females.

Lecture 8

Microsatellites

8.1 Repeat polymorphisms: Nomenclature

tandem repeat polymorphisms A particular sequence is repeated again and again. For example, if the repeated sequence is AT, then one chromosome might have a sting of three repeats (ATATAT) and another might have a string of five (ATATATATAT). They are called "tandem" repeats because the repeated units are stuck together end-to-end. There are several categories of repeat polymorphism:

Minisatellite the repeated unit is fairly large

Microsatellite the repeated unit is small, commonly 2–6 nucleotides.

Short tandem repeat (STR) a synonym for "microsatellite"

dispersed repeat polymorphisms These also consist of a sequence that is repeated again and again, but in these polymorphisms the repeated units are adjacent to one another. They are not even close together. They are widely dispersed within the genome. The *Alu polymorphisms* are an example of this class of polymorphism.

8.2 Properties relevant to statistical analysis

Rapid evolution Mutation rate in the neighborhood of 1/1000 per generation [37].

Di- and tetra-nucleotide repeats are likely to be neutral A tetranucleotide repeat doesn't work very well within a reading frame, because each mutation produces a frame-shift mutation, which probably has catastrophic consequences. Most of the di- and tetra-nucleotide repeats that we find are outside of reading frames.

Tri-nucleotide repeats are known within reading frames, and some cause genetic disease.

There are lots of microsatellite loci

Mutational process poorly understood Statistical analysis usually assumes that mutations add or subtract one repeat unit with equal probability. The is almost certainly too simple.

8.3 A sample of 60 STR loci

Figure 8.1 shows the frequency distributions of repeat counts at a sample of 60 tetra-nucleotide repeat loci described by Jorde and his colleagues [13, 14]. Each horizontal axis there shows the number of copies of the repeat unit, and each vertical axis shows the number of chromosomes in the sample that exhibited that number of copies.

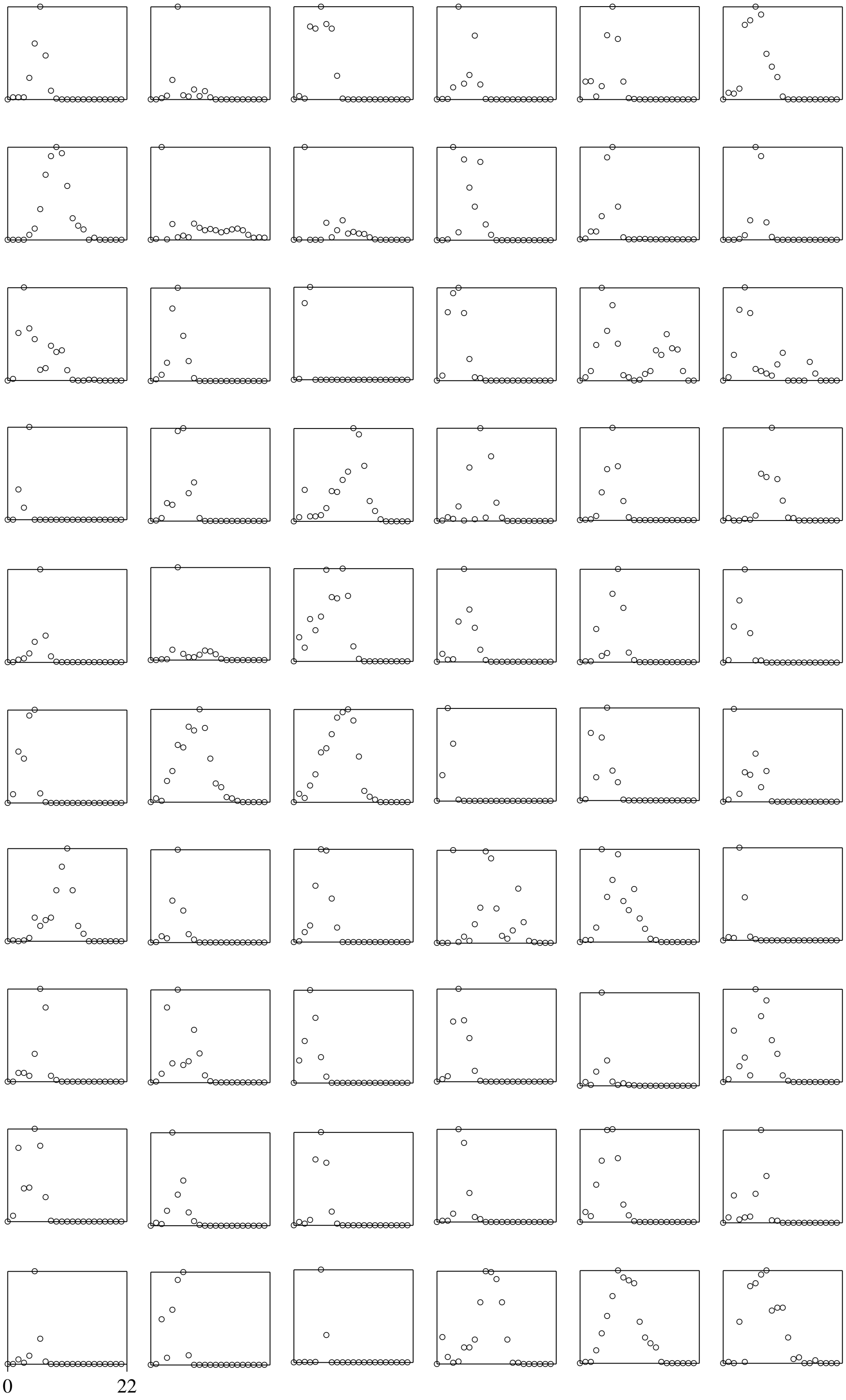

Figure 8.1: Frequency distributions of 60 STR loci

8.4 Remark about statistical methods

A variety of statistical methods have been introduced for analysis of STRs. Many of these seem useful, but none of them have yet become widely used. It is not obvious which should be included in an introductory account.

8.5 Descriptive statistics

Several of the statistical methods used with STRs involve the so-called "moments" of the distribution of repeat counts.

The mean For any given locus, the mean is the average number of repeat counts in the sample. For example, suppose we had data on 4 haploid genomes with repeat counts 4, 4, 5, and 6. The mean would be

$$m = (4 + 4 + 5 + 6)/4 = 4.75$$

The mean is also known as the "first moment" of the distribution.

The variance is the mean squared deviation from the mean. With the data above, this is[1]

$$V = \Big((4 - 4.75)^2 + (4 - 4.75)^2 + (5 - 4.75)^2 + (6 - 4.75)^2\Big)/4 = 0.6875$$

The variance is also called the "second moment" of the distribution.

The fourth moment is the mean of the fourth powers of deviations from the mean. With the example data,

$$m_4 = \Big((4 - 4.75)^4 + (4 - 4.75)^4 + (5 - 4.75)^4 + (6 - 4.75)^4\Big)/4 = 0.7695$$

Gene identity is defined as usual: it is the probability that a random pair of genes is identical. In the example data there are 4 genes, so there are $(4 \times 3)/2 = 6$ pairs. Only one of these pairs is identical, so the gene identity is

$$J = 1/6;$$

This is *not* a moment of the distribution.

8.6 The method of Kimmel et al [16]

Kimmel et al show that, at equilibrium between mutation and drift,

$$\begin{aligned} V &= \theta \\ J &= 1\Big/\sqrt{1 + 2\theta} \end{aligned}$$

[1] In a statistics course, you would learn to divide by 3 rather than 4 here in order to obtain an unbiased estimate. I am ignoring such niceties.

Thus, θ can be estimated either by V or by $(1/J^2 - 1)/2$. The ratio

$$\beta = \frac{V}{(1/J^2 - 1)/2} \tag{8.1}$$

ought to equal unity at migration-drift equilibrium.

After a population increase, θ becomes larger so the numerator and denominator of this ratio will both begin to increase. The denominator, however, increases faster. Thus, $\beta < 1$ for some time after a population expansion. For the 60 tetranucleotide repeats shown in figure 8.1,

Population	β
Asia	1.8221
Europe	1.3364
Africa	1.1163

This pattern does not match that predicted by a population expansion. On the other hand, Kimmel et al show that these numbers do depart significantly from unity, so the hypothesis of mutation-drift equilibrium doesn't hold either. What *is* going on?

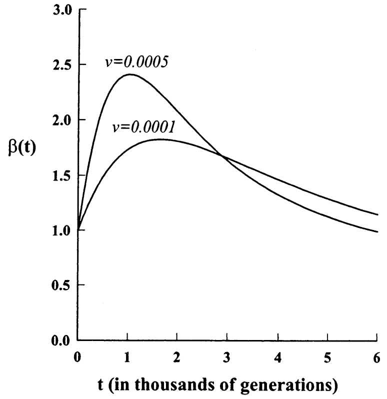

Figure 8.2: From [16, fig. 3]

When population size decreases, β is elevated above unity and then slowly decreases toward its equilibrium value of unity. Taken at face value, these data seem to imply a population collapse in the late Pleistocene—precisely the opposite of the conclusions reached from archaeology and mtDNA.

In an effort to reconcile these contradictory findings, Kimmel et al consider the effect of a "bottleneck," or temporary reduction in population size. At the start of the bottleneck, θ decreases so β is elevated above unity. After the population grows large again, this process will reverse and β will eventually converge to its equilibrium value of unity. However, the increase in β happens faster than the subsequent decrease. Consequently, β remains elevated for a long time after the bottleneck has ended. Kimmel et al show that the elevation would still be apparent for several thousand generations after the end of the bottleneck.

8.7 A method-of-moments estimate of τ

8.7.1 Method

Consider a model of population history in which the population was constant in size except for one instantaneous change, which occurred t generations ago. The number of genes was N_0 before the change and N_1 after. The effect of this history on genetic data can be fully described by three parameters:

$$\begin{aligned} \theta_0 &\equiv 2uN_0 \\ \theta_1 &\equiv 2uN_1 \\ \tau &\equiv 2ut \end{aligned}$$

My goal is to estimate these parameters.

Figure 8.3 shows the mismatch distributions from our sample of 60 STR loci. In each panel there, the horizontal axis shows the magnitude of the pairwise difference, and the vertical axis shows the numbers of pairs of chromosomes in the sample that exhibited such a difference.

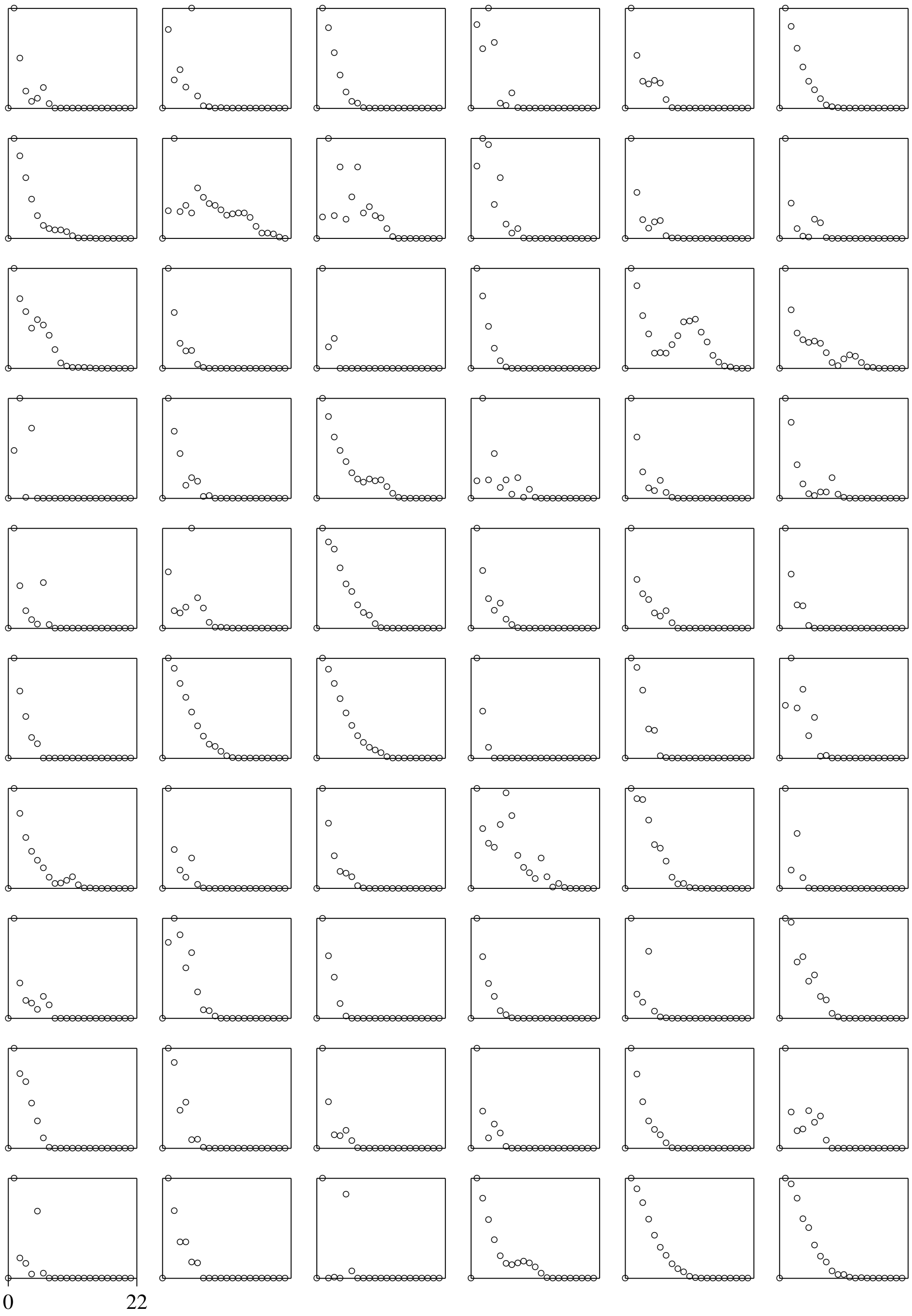

Figure 8.3: Mismatch distributions of 60 STR loci

Doing theory with samples of size 2 Rather than attempt a statistical theory for mismatch distributions from samples of size n, I work instead with theory for samples of size 2. This simpler theory should provide some guidance, since the expected values of the components of a mismatch distribution do not depend on sample size [24, footnote 4, p. 61]. Nonetheless, other statistical properties *do* depend on sample size, so caution is in order. The simple theory of random pairs will provide insight concerning which statistics are likely to be useful. Before these insights can be trusted, however, they must be tested by computer simulation.

The stepwise mutation model I use a very simple model of mutation, which assumes that each mutation either adds or subtracts exactly one repeat unit, the two alternatives having equal probability.

We already have a theory describing the effect of population history on mutational differences [19, 25], and the stepwise mutational assumption makes it easy to calculate the probability that two chromosomes differ by y repeat units given that they are separated by x mutations.

Figure 8.4 shows the effect of two population histories. The distribution of mutational changes shows a pronounced wave whose position reflects the time of the population expansion. This principle has proved useful in estimating expansion times from mitochondrial sequence data, where most mutations can be detected in comparisons between sequences [23]. But the lower panel shows that the distribution of stepwise differences is far less sensitive to differences in population history. Each mutation has an even chance of erasing the effect of the preceding mutation, and changes in the distribution are subtle. On the other hand, our sample of 60 STR polymorphisms provides information about 60 essentially independent gene genealogies, whereas the mtDNA provide information about just one. It is not yet clear which will be more informative.

The different histories shown in figure 8.4 leave the mode of the stepwise mismatch distribution unchanged but do affect its width. Thus, it seems natural to look for look for effects on the moments of the distribution (defined above). One can obtain estimators of θ_0 and τ by equating the moments of the sample with theoretical moments in a sample of size 2. This gives

$$\hat{\theta}_0 = \sqrt{(m_4 - V)/3 - V^2} \tag{8.2}$$

$$\hat{\tau} = V - \hat{\theta}_0 \tag{8.3}$$

These formulas are analogous to the two-parameter method of moments estimators introduced in lecture 7. There is no estimator for θ_1 here because the formulas assume that θ_1 is extremely large. They are not likely to be useful unless the population has in fact expanded.

To verify that these formulas are useful as estimators, I examined their behavior with simulated data. At each of a series of values of τ, I generated 1000 10-locus data sets and calculated $\hat{\tau}$ using each data set. The distribution of the results, shown in figure 8.5, is gratifying. The sampling distribution of the estimator is relatively narrow and its central tendency increases in response to increases in τ. There is a modest upward bias, but since this bias can be measured, it can be corrected. The behavior of $\hat{\theta}_0$, shown in figure 8.6, is less encouraging. The sampling distribution is wide and its central tendency varies little in response to changes in θ_0. Thus, $\hat{\theta}_0$ does not appear to be a useful estimator of θ_0.

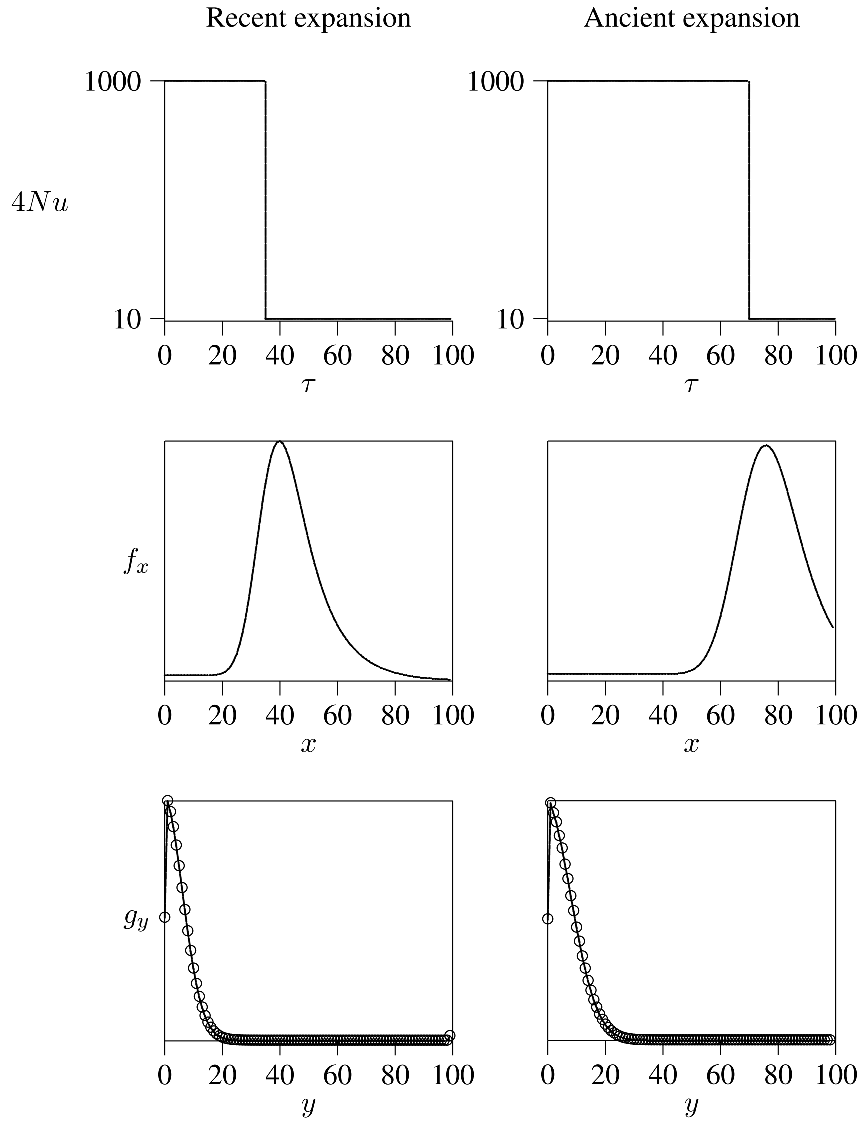

Figure 8.4: Effect of history on mutational differences and stepwise differences. In each column, the upper panel shows the history of population size on a time scale in which each unit equals $1/(2u)$ generations. The middle panel shows the probability f_x that a random pair of individuals differ by x mutational changes and the lower panel shows the probability g_y that they differ by y steps under the stepwise mutational process.

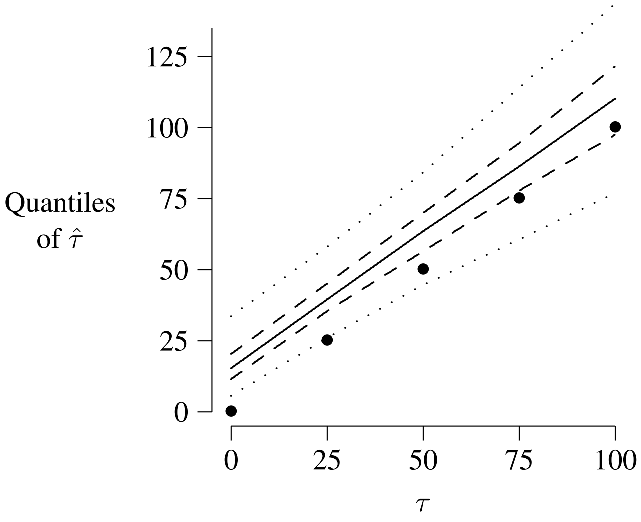

Figure 8.5: Quantiles of $\hat{\tau}$ from STR simulations. 1000 10-locus data sets were simulated at each of several values of τ, and each was used to estimate the model's parameters. The bold dots indicate points at which $\hat{\tau} = \tau$. The solid line is the median, the dashed lines enclose the central 50% of the distribution, and the dotted lines the central 95%. Each simulated data set was generated using the coalescent algorithm with $\theta_0 = 10$, $\theta_1 = 1000$, and $N = 100$.

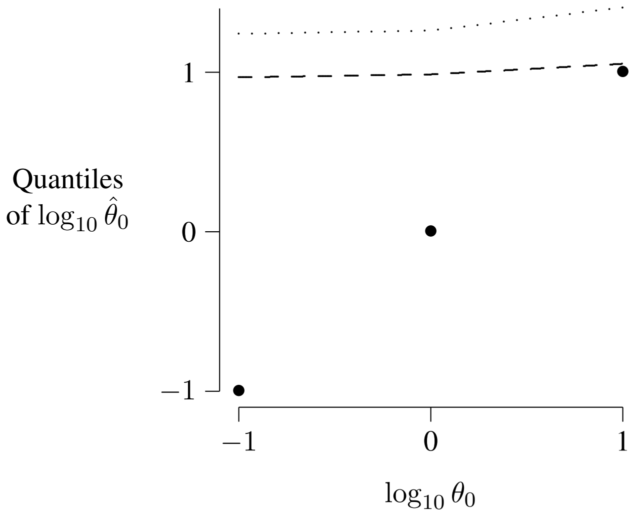

Figure 8.6: Quantiles of $\log_{10}\hat{\theta}_0$. 1,000 data sets were simulated at each of several values of θ_0, and each was used to estimate the model's three parameters. In each run, $\theta_1 = 10000$, $\tau = 70$, and $N = 100$. 97.5% of the simulated estimates fell below the dotted line; 75% fell below the dashed line. The median of $\hat{\theta}_0$ was 0 for all three values of θ_0.

Table 8.1: Estimates of τ

	60 STR loci†			mtDNA‡		
Population	L*	$\bar{\hat{\tau}}$	U**	L*	$\hat{\tau}$	U**
Africa	3.83	5.75	8.11	2.5	12.0	27.5
Asia	2.92	4.80	7.19	3.5	8.4	12.5
Europe	3.04	5.37	8.51	—	3.6	—

†Bootstrapped confidence intervals

‡Simulation-based confidence intervals

*Lower end of 95% confidence interval

**Upper end of confidence interval.

8.7.2 Results

The present method provides a useful estimate of τ but not of θ_0. This makes it useful for estimating the time of an expansion but useless for establishing that an expansion has in fact occurred. Fortunately, the study of Shriver et al. [27] has shown that the very data under study here *do* imply a population expansion. Thus, I address myself here to the task of estimating the time of this expansion.

To estimate τ from STR data, I applied equation 8.2 to each of the 60 loci. The averages of these estimates are shown in table 8.1 under the column labeled $\bar{\hat{\tau}}$. To measure the precision of the STR estimates, I repeated the calculations with each of 10000 bootstrap replicates. Each bootstrap data set contains 60 STR loci, which were selected by sampling with replacement from the original 60 loci. The STR and mitochondrial estimates agree in suggesting an earlier expansion in Africa than in Europe or Asia.

To convert τ from mutational time into years, we must divide by $2u$ and multiply by the generation time, say 25 years. Weber and Wong [37] suggest that mutation rates at many STR loci are near 1/1000 per generation. This estimate is undoubtedly rough, and it ignores variation across loci, but we set these issues aside in order to obtain the estimates of $\hat{t}$ in the STR column of table 8.2. The corresponding estimates in the mitochondrial column assume a per-nucleotide divergence rate of 14% per million years [31, p. 1506], or about $u = 1/1000$ per generation for 630 nucleotides.

For Africa and Asia, the mitochondrial and STR estimates differ by roughly a factor of two—well within

Table 8.2: Estimates of t

	60 STR loci†			mtDNA‡		
Population	L*	$\bar{\hat{t}}$	U**	L*	$\hat{t}$	U**
Africa	48,000	72,000	101,000	31,000	150,000	344,000
Asia	37,000	60,000	90,000	44,000	105,000	156,000
Europe	38,000	67,000	106,000	—	45,000	—

†Bootstrapped confidence intervals

‡Simulation-based confidence intervals

*Lower end of 95% confidence interval

**Upper end of confidence interval.

the range of statistical error. For Europe, the two estimates are even closer. Thus, the STR estimates are in broad agreement with the mitochondrial estimates.

The confidence intervals are narrower with STR data than with mitochondrial data. The mitochondrial interval is six times as long as the STR interval for Africa and twice as long for Asia. For Europe, the mtDNA interval is infinite, but the STR interval is 68,000 years. Thus, STR estimates are appreciably more precise. This difference may reflect either the increased information available when one studies 60 loci rather than one, or possibly a lack of efficiency in the statistical method used for mtDNA.

Lecture 9

Alu Insertions

9.1 What is an *Alu* insertion?

Transposable elements are DNA sequences that can copy themselves and insert the copies into different parts of the genome.

- constitute large fractions of eukaryotic genomes
- "selfish DNA": they don't have any function that we know of

Short Interspersed Elements (SINEs) are short retroposons (typically 75–500bp) that lack the machinery needed to transpose themselves. They must be produced by active sites somewhere in the genome, but the details of this process are mysterious. They are present in huge numbers in eukaryotic genomes.

***Alu* insertions** are SINEs that contain roughly 300bp and are present in huge numbers within the human genome.

How many? *Alu*s constitute 10.6% of the genome, or just over 290 Mb [18].

Probably neutral (Substitution rate is comparable to that of other non-coding genomic regions.)

9.2 The "master gene" model

It is thought that only a few *Alu*s are capable of transposing. These give rise to all the others. Most of the *Alu* elements in our genomes are inert.

Families of *Alu*s are defined on the basis of mutations that (presumably) accumulated in the master copies. Many members of younger families are polymorphic in the human gene pool. This makes them especially interesting in intra-specific population genetics.

9.3 Properties that make *Alu*s interesting

1. Insertions are seldom if ever lost cleanly

Table 9.1: Sample sizes and allele frequencies

Population	Haploid sample size	Mean freq. of “+” allele
Africa	302	0.46251
Asia	154	0.55671
Europe	236	0.55919
India	728	0.54373

Data from Watkins et al. [35]

- They could be deleted, but this would ordinarily remove not only the *Alu* but also some largish region around it. Such a loss would not go undetected.
- They are degraded by point mutations, but this process is very slow. If the mutation rate is 10^{-9} per nucleotide per year, then half the the sites should still be unmutated after 700 million years.

2. We know the ancestral state: *Alu*-absent

3. At any given site in the genome, the probability per generation that an insertion will occur is very low.

4. We don’t have to worry about insertion occurring twice at the same place in the genome.

5. There are many *Alu*s, and we don’t have to estimate a mutation rate separately for each locus.

   (Maybe we should: *Alu*s seem slightly more likely to insert in AT-rich regions.)

6. Since there are so many of them, we can ignore the ones that are closely linked. Unlinked loci have nearly independent gene genealogies.

7. Consequences: Theory is simple, and it is easy to aggregate over loci. Genealogical trees are never plagued by shared-derived characters.

9.4 Average *Alu* frequencies

Table 9.1 shows the mean frequency of *Alu* insertions in several populations. These data raise two questions:

- Why is the mean *Alu* frequency roughly 0.5?
- Why is it lower in Africa than elsewhere?

Bulayeva et al. [2] attempt to answer these questions. Their analysis is summarized in the remainder of this section.

The first of these questions has a simple answer. Recall from the discussion of the site frequency spectrum that, at mutation-drift equilibrium, the frequency of sites that divide a sample into i mutants and $K - i$ non-mutants is proportional to $1/i$. But we ascertain *Alu* loci by looking at a single haploid genome. This is equivalent to saying that alleles with frequency p should occur with probability proportional to $1/p$. Meanwhile, we ascertain *Alu* loci by looking at a single haploid genome. If a locus has allele frequency p, we ascertain it with probability p. Thus, the probability that a locus in our sample will have allele frequency p is proportional to $1/p \times p = 1$. In other words, our sample should contain equal numbers of loci in every frequency category. The average of this uniform distribution is 1/2, in good agreement with the data.

9.4.1 Tests of two hypotheses

We are already in a position to test two hypotheses from the literature.

Homo erectus diaspora (HED) hypothesis is an otherwise-plausible hypothesis [11] that is falsified by these data:

- *Homo erectus* dispersed out of Africa 1.8 mya, or 72,000 generations ago.
- Consequently, many nuclear gene trees are roughly this deep.
- Neutral theory says that expected time to LCA is $4N_e$ generations.
- Setting $4N_e = 72,000$ gives $N_e = 18,000$.
- This is in fair agreement with N_e as estimated from a variety of nuclear loci.
- Simulations show that, under this hypothesis, the world mean *Alu* frequency would be smaller than 0.19. This did not happen.

Pleistocene population explosion

- Human mitochondrial mismatch distributions suggest that population expanded by several-hundred-fold between 30 kya and 130 kya.
- This would lead to an excess of low-frequency *Alu*s.
- Simulation shows that a history in which human numbers expanded from 3,000 to 300,000 at 5,000 generations ago would lead to a world mean *Alu* frequency between 0.30 and 0.40 in samples of 100 *Alu*s.
- This is also excluded by the data.

9.4.2 Why are *Alu* frequencies higher outside Africa?

To answer this question, we need a model.

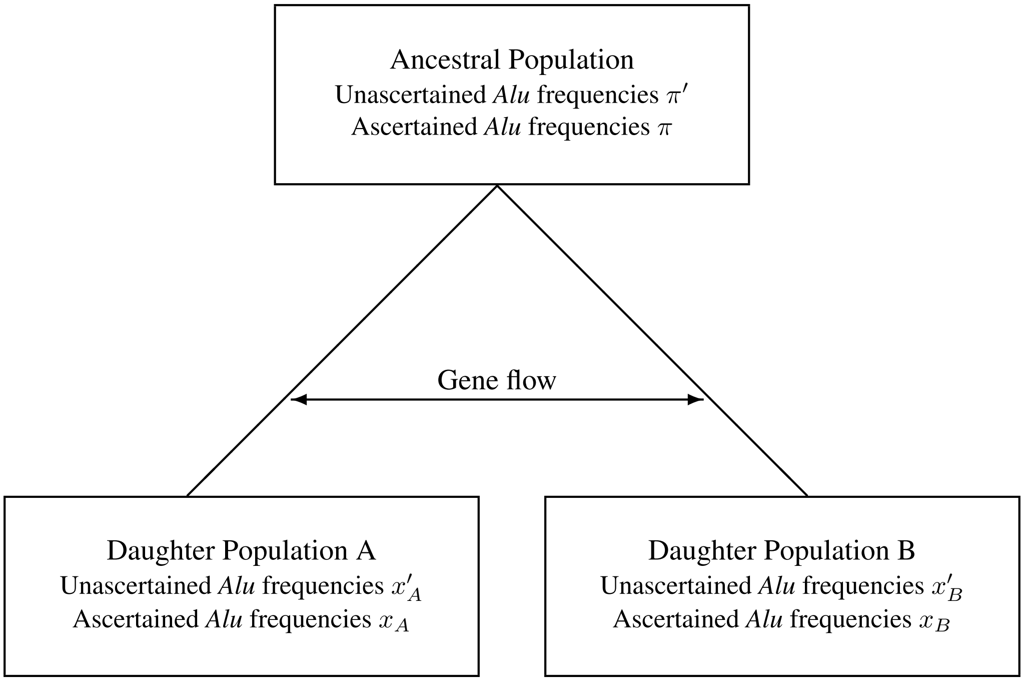
Ancestral Population
Unascertained *Alu* frequencies π'
Ascertained *Alu* frequencies π
Gene flow
Daughter Population A
Unascertained *Alu* frequencies x'_A
Ascertained *Alu* frequencies x_A
Daughter Population B
Unascertained *Alu* frequencies x'_B
Ascertained *Alu* frequencies x_B

Figure 9.1: Model of diaspora of contemporary populations from ancient source

Model

- An ancestral pop, in which *Alu* has frequency π'.
- Daughter pops A and B, in which *Alu* frequencies are x'_A and x'_B.
- Since diaspora, A and B have regularly exchanged migrants.
- Time is short: no *Alu*s have inserted since diaspora.
- x'_A and x'_B differ from π' because of genetic drift.
- We can only estimate π', x'_A, and x'_B, we can't observe them directly.
- Our estimates are biased upward.
- Why? Because we discover *Alu* loci by observing a single haploid genome drawn from population A. We are more likely to discover loci whose "+" allele is common in A.
- The biased versions of the allele frequencies are π, x_A, and x_B.

R-statistics Define

$$\begin{aligned} R_{AA} &= \frac{E[(x'_A - \pi')^2]}{\pi'(1-\pi')} \\ R_{AB} &= \frac{E[(x'_A - \pi')(x'_B - \pi')]}{\pi'(1-\pi')} \end{aligned}$$

R_{AA} is analogous to Wright's F_{ST} and measures the amount of drift that A has experienced since the diaspora. R_{AB} is a normalized covariance. It is positive if A and B exchange migrants and zero otherwise.

Results Harpending (unpublished) shows that

$$E[x_A] = E[\pi] + R_{AA}E[1-\pi] \tag{9.1}$$

$$E[x_B] = E[\pi] + R_{AB}E[1-\pi] \tag{9.2}$$

These results are interesting because they express R-statistics in terms of means. Ordinarily, one needs variances, which are harder to estimate than means. We could estimate the expectations by averaging over loci.

To use these formulas with data, Bulayeva et al. assume

- The ancestral population was African.
- The African population has remained large, so Africa has experienced little genetic drift, and π is approximately equal to the African allele frequency in the data.
- The *Alu* polymorphisms were ascertained in a European, so that x_A equals the European allele frequency in the data.

The last assumption is needed because *Alu* loci were discovered in the published sequence of the Human Genome Project, and the source of this sequence is not publically known.

Now, equations 9.1 and 9.2 indicate that *Alu* allele frequencies should be higher in non-African populations than in Africa. This is just what the data show. Furthermore, the difference provides an estimate of R_{AA}:

$$\begin{aligned} \hat{R}_{AA} &= \frac{E[x_A - \pi]}{E[1-\pi]} \\ &= \frac{0.56 - 0.46}{0.55} = 0.18 \end{aligned}$$

where the expectations have been approximated by averages over loci.

For an isolated population, R_{AA} follows

$$R_{AA}(t) = 1 - e^{-t/2N_e}$$

where t is time in generations and N_e is the effective size of the population [5]. If $R = 0.18$, then $t/2N$ = .22. If the diaspora occurred 40,000 years (or 1600 generations) ago, the European effective size would be $N_e = 4000$, and the census size perhaps 10,000.

Comments

- The estimate of N_e is in fair agreement with others, even though the method of analysis is very different.
- The method requires knowledge of ancestral state.
- Bias is ordinarily bad, but here it is crucial to the analysis.
- Thus, this analysis makes use of the unusual features of *Alu* elements.

9.5 The study of Sherry et al

This section will not be used in the 2004 version of the course.

Sherry et al. [26] report a series of *Alu* polymorphisms that were ascertained in HeLa cells. (This means that an element did not appear in the study unless it was present in HeLa.) They also ignored all elements that were present in chimpanzees. (I don't know how many chimpanzees they looked at.) They then screened a panel of 122 humans for each of the 57 elements that they had identified. Of the 57 elements identified, 13 turned out to be polymorphic in humans. The remaining 44 were fixed in humans and absent in chimps.

9.5.1 Consequences of insertions in different parts of the gene tree

1. A locus that is polymorphic in humans inserted inside the human gene tree. This is the portion labeled "a" in figure 9.2.
2. A locus that is fixed in humans and absent in chimps inserted along the branch leading from the chimp/human MRCA to the human MRCA. In the figure, this is the sum of parts b and c.

Part a of the tree

The expected length of part a of the tree (including all branches) is

$$L_a = 4N \sum_{i=1}^{K-1} \frac{1}{i}$$

as shown in section 5.1 and it's depth is

$$T_a = 4N(1 - 1/K)$$

as shown in section 4.4. Here, K is the number of haploid genomes in the sample and equals 246. Furthermore, these DNA polymorphisms are diploid, so N can be interpreted as the population size. Thus,

$$\begin{aligned} L_a &= 4N \times 6.08 \approx 24N \quad \text{generations} \\ T_a &= 4N(1 - 1/246) \approx 4N \quad \text{generations} \end{aligned}$$

Figure 9.2: Figure 1 from Sherry et al

Part b of the tree

If the human/chimp speciation event occurred 4.5 myr (perhaps 225,000 generations) ago, then the length of part b is

$$T_b = 225,000 - 4N \quad \text{generations}$$

Part c of the tree

This part of the tree represents the coalescence time of a pair of genes within the species ancestral to humans and chimps. On average, it should equal

$$T_c = 2N_c \quad \text{generations}$$

where N_c is the size of the ancestral species. Takahata et al. [28] estimate that $N_c \approx 100,000$. If this is right, then

$$T_c = 200,000 \quad \text{generations}$$

9.5.2 Data analysis

The total branch length is

$$L_a + T_b + T_c = 20N + 425000$$

The sample includes insertions that occur anywhere within these branches. Only a fraction of these mutations are polymorphic, however: those occuring in region a, which has length $24N$. The fraction of polymorphic insertions ought to equal (on average) the fraction of the total branch length that holds polymorphic insertions. In other words,

$$\frac{13}{57} = \frac{L_a}{L_a + T_b + T_c} = \frac{24N}{20N + 425000}$$

This gives

$$N = 4986$$

as an estimate of the effective human population size. This is very close to the estimate obtained from mtDNA.

9.6 Ascertainment bias

Unfortunately, the estimate just obtained is biased. The problem is that we only included *Alu*s that were present in HeLa cells.

9.7 Exercises

⋆ EXERCISE **9–1** How would the estimate of N have changed if 26 of the 57 *Alu*s has been polymorphic?

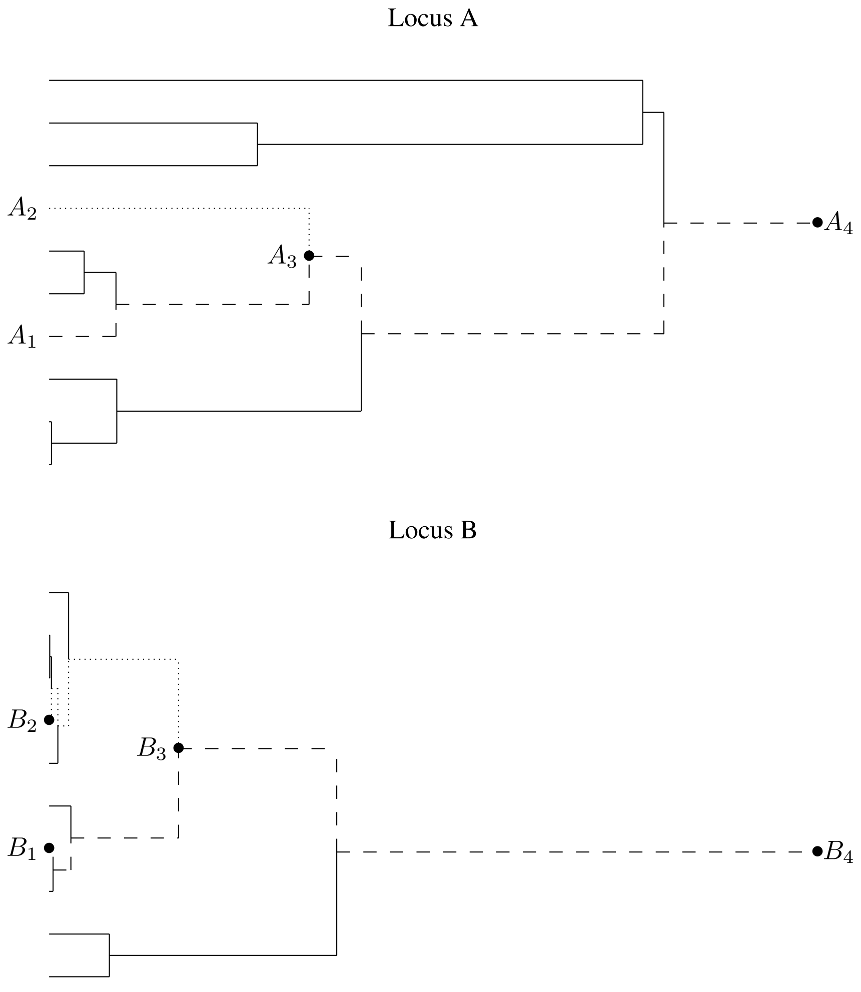

Figure 9.3: Haploid (and diploid) ascertainment on a large and a small genealogy

Appendix A

Mean, Variance and Covariance

This is not really a lab exercise. It is a summary and description of concepts of mean, variance, and covariance. It is meant to familiarize you with notation and concepts that are used elsewhere.

A.1 The mean

Given a set of numbers such as

$$x_1 = 10, \quad x_2 = 12, \quad x_3 = 10, \quad x_4 = 8$$

the mean is

$$\bar{x} = (10 + 12 + 10 + 8)/4 = 10$$

which can also be written in several other ways, including

$$\bar{x} = (x_1 + x_2 + x_3 + x_4)/4 = \left(\sum_{i=1}^{4} x_i\right)/4$$

In this example there are four elements in the sum. Often the number of elements is represented by a symbol such as N. Then

$$\bar{x} = N^{-1} \sum_{i=1}^{N} x_i$$

The limits of summation are sometimes omitted, as in

$$\bar{x} = N^{-1} \sum_{i} x_i$$

which means that the sum is over all values of i, which here runs from 1 through 4.

It is often more convenient to express the mean in terms of the frequencies of elements with particular values. For example, let n_x denote the number of elements with value x. In our example, $n_1 = n_2 = \cdots = n_7 = 0$, $n_8 = 1$, $n_9 = 0$, $n_{10} = 2$, $n_{11} = 0$, and $n_{12} = 1$. With this notation,

$$\bar{x} = N^{-1} \sum_{x} x n_x$$

and

$$N = \sum_x n_x$$

Yet another formulation defines the relative frequency of elements with value x by $p_x = n_x/N$. Then

$$p_8 = p_{12} = 1/4, \quad \text{and} \quad p_{10} = 1/2$$

and the mean can be written as

$$\bar{x} = \sum_x x p_x$$

In this formulation, p_x is the relative frequency of value x in the data.

The same formula also turns up in probability theory. Suppose that some variable, X, takes values that cannot be predicted exactly, but which occur with specified probabilities, $p_1, p_2, \ldots$. Then X is called a "random variable" and its expectation, $E[x]$, is defined as

$$E[X] = \sum_x x p_x$$

Note that this is the same as the preceding formulation of the mean.

Finally, suppose that X is a continuous variable, such as stature, that takes values over the range from a to b. Then you cannot enumerate the possible values that it may take. Its expectation, or average value, is written as

$$E[X] = \int_a^b x f(x) dx$$

where $f(x)dx$ can be thought of as the probability that X takes a value within the (very small) interval from x to $x + dx$, and is thus analagous to p_x in the discrete formulation.

A.2 Variance

The variance of a series of numbers is the average squared difference from the mean. For example, the mean of the numbers listed above is 10, so the variance is

$$V = ((10-10)^2 + (12-10)^2 + (10-10)^2 + (8-10)^2)/4 = 2$$

As with the mean, there is a variety of ways to represent the variance, including

$$\begin{aligned} V &= N^{-1} \sum_i (x_i - \bar{x})^2 && \text{(A.1)} \\ &= \sum_x (x - \bar{x}^2) p_x \\ &= \sum_x x^2 p_x - \bar{x}^2 && \text{(A.2)} \\ &= \overline{x^2} - \bar{x}^2 && \text{(A.3)} \end{aligned}$$

The last line says that the variance is the mean square of x minus the squared mean.

Equation A.1 provides the most straightforward method of calculating the variance, but it requires two passes through the data. The first pass calculates the mean, and the second sums the squared deviations from the mean. Equation A.3 is often more convenient because it requires only one pass through the data.

⋆ EXERCISE **A–1** Verify that these formulas are equivalent.

In probability theory, the (theoretical) variance is defined as

$$V[X] = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$$

These expressions hold irrespective of whether the random variable is continuous or discrete, but they have slightly different interpretations in the two cases. For discrete random variables,

$$E[X^2] = \sum_x x^2 p_x$$

but for continuous random variables

$$E[X^2] = \int x^2 f(x) dx$$

A.3 Covariances

The covariance of two sets of numbers, $x_1, \ldots, x_N$, and $y_1, \ldots, y_N$, is

$$Cov(x, y) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x}$ and $\bar{y}$ are the means of the two sets of numbers. Alternate expressions for the covariance include

$$\begin{aligned} Cov(x, y) &= \sum_x \sum_y (x - \bar{x})(y - \bar{y}) p_{x,y} \\ &= \sum_{xy} (x - \bar{x})(y - \bar{y}) p_{x,y} \\ &= \sum_{xy} xy p_{x,y} - \bar{x}\bar{y} \qquad \text{(A.4)} \end{aligned}$$

Here $p_{x,y}$ is the relative frequency of pairs of numbers with values (x, y) and the sum is taken over all possible of pairs of numbers. The notation $\sum_{xy}$ means the same thing as $\sum_x \sum_y$.

Bibliography

[1] A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L. Luca Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368:455–457, March 1994.

[2] Kazima Bulayeva, Lynn B. Jorde, Christopher Ostler, Scott Watkins, Oleg Bulayev, and Henry Harpending. Genetics and population history of Caucasus populations. *Human Biology*, 75(6):837–853, 2003.

[3] Rebecca L. Cann, Mark Stoneking, and Allan C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(1):31–36, January 1987.

[4] Andrew G. Clark, Kennety M. Weiss, Deborah A. Nickerson, Scott L. Taylor, Anne Buchanan, Jari Stengard, Veikko Salomaa, Erkki Vartiainen, Markus Perola, Eric Boerwinkle, and Charles F. Sing. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human Lipoprotein Lipase. *American Journal of Human Genetics*, 63:595–612, 1998.

[5] James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, New York, 1970.

[6] Yunxin Fu. Statistical properties of segregating sites. *Theoretical Population Biology*, 48(2):172–197, October 1995.

[7] Marc K. Halushka, Jian-Bing Fan, Kimberly Bently, Linda Hsie, Naiing Shen, Alan Weder, Richard Cooper, Robert Lipshutz, and Aravinda Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, 22:239–247, 1999.

[8] Rosalind M. Harding, S. M. Fullerton, R. C. Griffiths, Jacquelyn Bond, Martin J. Cox, Julie A. Schneider, Danielle S. Moulin, and J. B. Clegg. Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics*, 60:722–789, 1997.

[9] Henry C. Harpending, Mark A. Batzer, Michael Gurven, Lynn B. Jorde, Alan R. Rogers, and Stephen T. Sherry. Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences, USA*, 95:1961–1967, 1998.

[10] Eugene E. Harris and Jody Hey. X chromosome evidence for ancient human histories. *Proceedings of the National Academy of Sciences, USA*, 96:3320–3324, 1999. URL `http://www.pnas.org/cgi/content/full/96/6/3320`.

[11] John K. Hawks, K. Hunley, S. H. Lee, et al. Population bottlenecks and Pleistocene human evolution. *Molecular Biology and Evolution*, 17(1):2–22, 2000.

[12] Richard R. Hudson. Gene genealogies and the coalescent process. In Douglas Futuyma and Janis Antonovics, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press, Oxford, 1990.

[13] Lynn B. Jorde, M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley, P. A. Krakowiak, K. D. Carpenter, H. Soodyall, T. Jenkins, and Alan R. Rogers. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *American Journal of Human Genetics*, 57: 523–538, Sept 1995.

[14] Lynn B. Jorde, Alan R. Rogers, Michael Bamshad, W. Scott Watkins, Patrycia Krakowiak, Sandy Sung, Juha Kere, and Henry C. Harpending. Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences, USA*, 94:3100–3103, April 1997.

[15] Henrik Kaessmann, Florian Heißig, Arndt von Haeseler, and Svante Pääbo. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genetics*, 22:78–81, 1999.

[16] Marek Kimmel, Ranajit Chakraborty, J. Patrick King, Michael Bamshad, W. Scott Watkins, and Lynn B. Jorde. Signatures of population expansion in microsatellite repeat data. *Genetics*, 148:1921–1930, 1998. http://www.genetics.org/cgi/content/full/148/4/1921.

[17] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13:235–248, 1982.

[18] Eric S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[19] Wen-Hsiung Li. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, 85:331–337, 1977.

[20] Wen-Hsiung Li. *Molecular Evolution*. Sinauer, Sunderland, Mass., 1997.

[21] Michael W. Nachman, Vanessa L. Bauer, Susan L. Crowell, and Charles F. Aquadro. DNA variability and recombination rates at x-linked loci in humans. *Genetics*, 150:1133–1141, 1998.

[22] Masatoshi Nei, Gregory Livshits, and Tatsuya Ota. Genetic variation and evolution of human populations. In C. F. Sing and C. L. Hanis, editors, *Genetics of Cellular, Individual, Family, and Population Variability*, pages 239–252. Oxford University Press, Oxford, 1993.

[23] Alan R. Rogers. Genetic evidence for a Pleistocene population explosion. *Evolution*, 49(4):608–615, August 1995.

[24] Alan R. Rogers. Population structure and modern human origins. In Peter J. Donnelly and Simon Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 55–79. Springer-Verlag, New York, 1997.

[25] Alan R. Rogers and Henry C. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9:552–569, 1992.

[26] Stephen T. Sherry, Henry C. Harpending, Mark A. Batzer, and Mark Stoneking. *Alu* evolution in human populations: Using the coalescent to estimate effective population size. *Genetics*, 147:1977–1982, 1997.

[27] Mark D. Shriver, Li Jin, Robert E. Ferrell, and Ranjan Deka. Microsatellite data support an early population expansion in Africa. *Genome Research*, 7:586–591, 1997.

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

[29] Simon Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26:119–164, 1984.

[30] Peter A. Underhill, Li Jin, Alice A. Lin, S. Quasim Mehdi, Trefor Jenkins, Douglas Vollrath, Ronald W. Davis, L. Luca Cavalli-Sforza, and Peter J. Oefner. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Research*, 7:996–1005, 1997.

[31] Linda Vigilant, Mark Stoneking, Henry Harpending, Kristen Hawkes, and Allan C. Wilson. African populations and the evolution of human mitochondrial DNA. *Science*, 253:1503–1507, 1991.

[32] John Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, Colorado, 2009.

[33] R. H. Ward, Barbara L. Frazier, Kerry Dew-Jager, and Svante Pääbo. Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences, USA*, 88:8720–8724, 1991.

[34] W. S. Watkins, M. J. Bamshad, A. E. Fraley, and L. B. Jorde. Population genetics of trinucleotide repeat polymorphisms. *Human Molecular Genetics*, 4:1485–1491, 1995.

[35] W. S. Watkins, A. R. Rogers, C. T. Ostler, S. Wooding, M. J. Bamshad, A-M. E. Brassington, M. L. Carroll, S. V. Nguyen, J. A. Walker, B. V. Prasad, P. G. Reddy, P. K. Das, M. A. Batzer, and L. B. Jorde. Genetic variation in world populations: Inferences from 100 *Alu* insertion polymorphisms. *Genome Research*, 13(7):1607–1618, 2003.

[36] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276, 1975.

[37] J.L. Weber and C. Wong. Mutation of human short tandem repeats. *Human Molecular Genetics*, 2(8): 1123–1128, Aug 1993.

[38] Stephen Wooding and Alan R. Rogers. The matrix coalescent and an application to human SNPs. *Genetics*, 161:1641–1650, 2002.

[39] Ewa Zietkiewicz, Vania Votova, Michal Jarnik, Maria Koran-Laskowska, Kenneth K. Kidd, David Modiano, Rosaria Scozzari, Mark Stoneking, Sarah Tishkoff, Mark Batzer, and Damian Labuda. Genetic structure of the ancestral population of modern humans. *Journal of Molecular Evolution*, 47: 146–155, 1998.

Appendix B

Answers to Exercises

⋆EXERCISE **1–1**

```
% Cmd line: seqstat eu10.dat -c
% 1st line of input specifies 10 subjects and 10 sites
% Results will include the reference sequence (line 1).

% Population 0
    sequences       : 10
    sites           : 10
    mismatch = 11 25 9 ;
    meanPairwiseDiff
        per sequence: 0.955556
        per site    : 0.0955556
   segregating sites: 4
   theta estimated from segregating sites
        per sequence: 1.41394
        per site    : 0.141394
    spectrum =  3 1 ;
    % Count of minor allele at each polymorphic site:
    %psite  site  count
         1     3      2
         2     6      1
         3     7      1
         4    10      1
```

⋆EXERCISE **2–1** The parameter values that maximize this expression are

$$\hat{P}_1 = \frac{N_1}{N_1 + N_2 + N_3} \tag{B.1}$$

$$\hat{P}_2 \quad = \quad \frac{N_2}{N_1 + N_2 + N_3} \tag{B.2}$$

⋆ EXERCISE **3–1** Expected heterozygosity equals 0.16 when $\theta \approx 0.19$.

⋆ EXERCISE **3–2** If $\theta = 0.19$ then $N = 47,500$.

⋆ EXERCISE **3–3** In generations 0 through 20 the heterozygosity is:

Generation	Heterozygosity
0	0
1	.00001052631579
2	.00002105249972
3	.00003157855180
4	.00004210447203
5	.00005263026041
6	.00006315591694
7	.00006315591694
7	.00007368144162
8	.00008420683445
9	.00009473209544
10	.0001052572246
11	.0001157822219
12	.0001263070874
13	.0001368318210
14	.0001473564228
15	.0001578808928
16	.0001684052309
17	.0001789294372
18	.0001894535117
19	.0001999774543
20	.0002105012651

⋆ EXERCISE **3–4** A few representative values are:

p	$p(1-p)/(2N)$
0	0
.25	$0.1973684211 \times 10^{-5}$
.50	$0.2631578948 \times 10^{-5}$
.75	$0.1973684211 \times 10^{-5}$
1.00	0

⋆ EXERCISE **3–5** At $p = 1/2$, $\sqrt{(}p(1-p)/(2N)) \approx 0.0016$

⋆ EXERCISE **3–6** In the figure, $u = 0.005$ and $N = 2500$, so $\theta = 50$, the equilibrium value of heterozygosity is 0.98, and the equilibrium value of homozygosity is 0.02. This is just what the figure shows.

⋆ EXERCISE **4–1** The equation is $T = h + (1-h)(1+T)$. Re-arranging this expression gives $T = 1/h$.

⋆ EXERCISE **4–2** Beginning with the right side of the expression, we end up with the left side:

$$\frac{1}{i-1} - \frac{1}{i} = \frac{i}{i(i-1)} - \frac{i-1}{i(i-1)} = \frac{1}{i(i-1)}$$

⋆EXERCISE **4–3** The intervals are:

i	$4N/(i(i-1))$
10	111.1111111
9	138.8888889
8	178.5714286
7	238.0952381
6	333.3333333
5	500.0000000
4	833.3333333
3	1666.666667
2	5000.000000

⋆EXERCISE **4–4** Since the sample is large, the mean age of the LCA is close to $4N$ generations. Half of this period is accounted for by the interval during which the tree contained only two lineages.

⋆EXERCISE **5–1** Let x_{ij} denote the difference between the ith and jth sequences, and let M denote the number of pairs of sequences in the sample. Then $\pi = \sum x_{ij}/M$, where the sum runs over all i and j such that $i < j$. The expected value of π is

$$\begin{aligned}
E[\pi] &= E\left[M^{-1}\sum x_{ij}\right] && \text{(definition of } \pi\text{)} \\
&= M^{-1}E\left[\sum x_{ij}\right] && \text{(JEPr eqn. 7)} \\
&= M^{-1}\sum [Ex_{ij}] && \text{(JEPr eqn. 8)} \\
&= M^{-1}M\theta && \text{(eqn. 5.4 with } K=2\text{)} \\
&= \theta
\end{aligned}$$

Here, JEPr eqn. 7 tells us that $E[aX] = aE[X]$ for constant a and variable X, and JEPr eqn. 8 tells us that the expectation of a sum equals the corresponding sum of expectations. Eqn. 5.4 (with $K = 2$) tells us that the expected number of segregating sites in a sample of size 2 is equal to θ. For such samples, the number of segregating sites is equal to the number of differences between the two sequences.

⋆EXERCISE **6–1** That example was of a sample of $K = 10$ DNA sequences, which had 15 segregating sites. Thus, we can estimate θ as

$$\hat{\theta}_S = \frac{15}{\sum_{i=1}^{9}\frac{1}{i}} = 5.3$$

I'll use the symbols v_u and v_f to represent the unfolded and folded spectra respectively. The unfolded theoretical spectrum (assuming selective neutrality and constant population size) is

$$v_u = [\theta, \theta/2, \theta/3, \ldots, \theta/9]$$

The first entry in this vector is the expected number of singleton sites, the second is the expected number of doubleton sites, and so on. Substituting the estimated value of θ turns this into

$$v_u = [5.30, 2.65, 1.77, 1.33, 1.06, 0.88, 0.76, 0.66, 0.59]$$

The folded spectrum is constructed as follows:

$$\begin{aligned}
v_f &= [5.30 + 0.59, 2.65 + 0.66, 1.77 + 0.76, 1.33 + 0.88, 1.06] \\
&= [5.89, 3.31, 2.53, 2.21, 1.06]
\end{aligned}$$

Thus, we expect 5.89 sites at which the minor allele is present in 1 copy, 3.31 sites at which it is present in 2 copies, and so on. In the real data (see section 1.3, page 7), we had

$$\hat{v}_f = [6, 2, 2, 5, 0]$$

where the "hat" indicates that these values refer to data rather than from theory. The theoretical and observed spectra are similar, but certainly not identical. No inference can be drawn from this difference, because our sample is very small.