Chapter 1

Gene Genealogies and Genetic Data

The goal of population genetics is to understand the forces that produce and maintain genetic variation within species. These forces include mutation, recombination, natural selection, population structure, and randomness in the birth and death of individuals within populations. Theoretical population genetics considers these factors in a mathematical setting, and has spawned subfields of applied mathematics, probability theory, and statistics. Now a century old — see Provine (1971) for a history — population genetics has acquired new relevance with the astonishing leaps of biotechnology over the past few years. For example, it is now straightforward to gather the most direct data concerning genetic variation, specifically DNA sequence data, to the extent that the entire 3×10^9 base pair human genome has been completed (International Human Genome Sequencing Consortium, 2001). Sequences showing variation within species are accumulating rapidly for many different organisms, and theoretical population genetics provides a wealth of pre-existing knowledge about how to interpret the patterns observed (Hartl, 2000; Hedrick, 2000; Gillespie, 1998; Hartl and Clark, 1997). This knowledge has been focussed and extended in coalescent theory, which describes the connection between demographic history and genetic data, and provides a framework for extracting information from samples of DNA sequences.

Even the barest understanding of biology forces us to grapple with history. All organisms use strings of nucleic acids to store the information required for life. Most organisms use DNA to encode this information, but some organisms use the closely-related molecule RNA. Some organisms, such as ourselves, are *diploid* and carry two complete copies of this information, while others are *haploid* and carry only a single copy of the genome. In DNA there are four bases: Adenine, Guanine, Cytosine, and Thymine, and we use the letters A, G, C, and T to represent these. RNA also uses four bases, but Thymine is replaced with Uracil, represented by the letter U. In the DNA double helix, A pairs with T and G pairs with C, so that genomic DNA sequences are double-stranded. Throughout this book, a *sequence* is taken to mean the ordered string of bases on one of these strands. For our purposes, it will often not matter which strand it is, because we will typically be considering a hypothetical piece of the genome, or genetic *locus*, which does not encode a function. Non-functional parts of the genome are very informative about history because mutations in such regions accumulate at a more constant rate over time than those within protein-coding or otherwise functional loci. When we observe some number of differences between a pair of sequences at the same locus, for example the two copies possessed by a single diploid organism or the sequences in two different haploid individuals, we know that the history back to their common ancestor must be long enough for the requisite mutations to have occurred with reasonable probability.

# SNPs	Poisson	Coalescent	Observed
0	$8,256 \pm 52$	$8,767 \pm 50$	$8,796 \pm 43$
1	$3,040 \pm 49$	$2,332 \pm 46$	$2,247 \pm 44$
2	617 ± 24	663 ± 26	668 ± 24
3	99 ± 9	200 ± 15	214 ± 14
4	16 ± 4	66 ± 9	102 ± 10

Table 1.1: Redrawn from Table 3 of The International SNP Map Working Group (2001)

The fact that there is a population-level process which influences patterns of genetic variation via these times to common ancestry, is surprisingly under-appreciated, even though it is obvious in nearly any DNA dataset. For a recent example, consider the initial population genetic analysis of the 1.42 million human single-nucleotide polymorphisms, or SNPs, discovered by the The International SNP Map Working Group (2001). One analysis (their Table 3) compiles the distribution of the number of differences, *i.e.* SNPs, per locus in randomly-sampled pairs of chromosomes. This distribution was compared to two possible theoretical predictions, the Poisson and the Coalescent, as shown in Table 1.1. The Poisson prediction would hold if there was no variation in times to common ancestry among loci in the genome. This is not true for humans and is extremely unlikely to be true in any species. The coalescent prediction explains the data significantly better than the Poisson, and the reason is that it captures the population-level processes that cause the times to common ancestry to differ among genetic loci. In this book, we will see that these differences in times to common ancestry are in fact a major source of variation in polymorphism levels among loci in a genome. We will also see that the simple coalescent model, while far better than the Poisson, is often too simple to explain all aspects of variation. For example, the Coalescent predictions underestimate the upper tail of the distribution of the number of SNPs per locus in Table 1.1, and this reflects the fact that human history is more complicated than the standard coalescent model.

Coalescent theory is at once rooted in the long history of population genetics and born out of more recent advances in biotechnology. Thus it is well-suited to the analysis of genetic data. The coalescent describes the genetic ancestry of a sample and uses this to make predictions about patterns of genetic variation. The genetic ancestry of a sample, the *gene genealogy*, is the set of ancestral relationships among the members of the sample, including times to common ancestry. The retrospective approach of coalescent theory stands in contrast to the prospective view taken by classical population genetics (Ewens, 1990) in which the main concerns have been to predict changes in the frequencies of *alleles* forward in time and to describe patterns of genetic variation in an entire population. An allele is simply a type at a genetic locus. It is possible to make predictions about samples using the classical approach, but this usually requires first characterizing the properties of the whole population then imagining taking a sample from it. The demonstration that a relatively simple ancestral process for a sample exists and obviates the need of explicitly modelling the entire population was a major advance in population genetics. It is important to note, however, that many well-known results can be derived either by the classical or the coalescent approach, and in some circumstances the classical approach my be the more fruitful (see Chapter 6). Still, the coalescent, with its close connection to the sample, supplies a more intuitive and efficient framework than classical theory for making inferences about the genetics and demography of populations.

Kingman (1982a,b,c) proved the existence of the coalescent process. He showed that what he called the 'n-coalescent' holds for a wide range of populations with different breeding structures (see Chapter 3). Hudson (1983a) and Tajima (1983) explored many biologically relevant aspects of the coalescent process and presented more intuitive derivations starting with the most commonly-used population model: the Wright-Fisher model described in Section 3.1.1. The seeds of the coalescent were planted several decades before this, in the 1940's, by Gustave Malécot, who introduced the idea of following a pair of gene copies back to their common ancestor (Malécot, 1946; Malécot, 1948; Nagylaki, 1989; Slatkin and Veuille, 2002) and notion of identity by descent, a concept which is readily interpreted in terms of pairwise coalescence times (Hudson, 1990). Genealogical approaches to samples larger than two appeared later, in response to the first direct measurements of molecular variation (Harris, 1966; Lewontin and Hubby, 1966). These include Ewens (1972) who described the distribution of allele counts in a sample under the infinite-alleles model of selectively neutral mutation, and Watterson (1975) who gave an explicitly genealogical derivation of the number of *segregating sites*, or polymorhic sites, in a sample of sequences under the infinite-sites model of mutation without recombination. These mutation models are described in Section 1.2 below. In addition, Griffiths's (1980) theory of lines of descent under the infinite alleles model has the coalescent at its heart. Lines of descent are sets of descendents of mutations, and Tavaré (1984) shows how the structure of the coalescent is recovered from these models by setting the mutation rate to zero. Finally, Kingman (2000) draws some connections between the coalescent and earlier work on models of stepwise mutation (Ohta and Kimura, 1973; Moran, 1975).

Population genetics theory is mathematical, but in the case of coalescent theory the math is made tangible by constant reference to a simple graphical structure: the gene genealogy. Coalescent theory thus capitalizes on the long-standing familiarity of evolutionary biologists with tree structures. For example, the only figure in *The Origin of Species* (Darwin, 1859) is a hypothetical phylogeny, *i.e.* a tree representing patterns of descent among species. Readers used to "tree thinking," which is the subject of Section 1.1, will have little trouble seeing the close connection between genealogical trees and patterns in sampled data. However, there is a danger in carrying over too much from the field of phylogenetics. Gene genealogies are by their nature unobservable, and are treated as random variables due to the randomness of the process of reproduction within populations. While in phylogenetic studies the tree structure itself is significant, within species it is often the case that particular gene genealogies provide little information about population-level processes and events. These issues are manifest in current controversies surrounding the use of nested clade analysis (Templeton *et al.*, 1995; Templeton, 1998) or, more broadly, about differing sensibilities between the fields of intraspecific phylogeography (Avise *et al.*, 1987; Avise, 1989; Avise, 2000) and coalescent theory. See Knowles and Maddison (2002) and Hey and Machado (2002) for two perspectives on these issues. Under the coalescent we typically value genealogies not for their specific structures but rather as intermediaries in shaping genetic variation.

Another reason for the wide acceptance of coalescent theory, and the recognition of its utility, is that this single ancestral process holds for an extremely broad range of demographies, and even arises in wholly unexpected places (see Chapters 3 and 7). Next to Mendel's Laws, the coalescent may be the best justified and farthest reaching stochastic mathematical model in biology. Mendel's Laws and their rediscovery at the turn of last century — again see Provine (1971) — spurred the growth of theoretical population genetics, and this in turn formed the basis of the Neo-Darwinian Evolutionary Synthesis (Mayr, 1942; Dobzhansky, 1937) which frames current discussion about evolution within biology. The coalescent already defines most work in population genetics, and future promises are great. Due to the recent surge of research into the history, genomics, and medical genetics of humans, we can expect coalescent theory to become the source of practical bioinformatic methods and tools with which we can view the recent history of humans and other species.

1.1 Genealogies and Genealogical Thinking

Imagine that we have taken a sample of DNA sequences from some population of organisms. A collection of copies of the same genetic locus, for instance a particular protein-coding gene, constitutes a sample. For the moment, consider a locus without intragenic recombination; recombination is treated in Chapter 6. The history of the sequences back to their most recent common ancestor is the gene genealogy of the sample. When DNA is replicated, two copies descend from a single common ancestor. When this is viewed backward in time, and both descendent lineages are represented in the sample, it is called a coalescent event. Except at the origin of DNA-based life, all DNA sequences descend from other DNA sequences. Thus, the gene genealogy of the sample exists: the entire sample can be traced back to a single common ancestor through a series of coalescent events. This is of course true even for samples from different species, simply due to the facts of DNA replication, but the determinants of genealogical shape are different for inter-species and intra-species samples. The shapes of inter-species gene genealogies, often called gene trees (Pamilo and Nei, 1988; Maddison, 1997), for the most part coincide with the phylogenetic trees of the species from which the samples were taken. Gene genealogies for samples from a single species are more strongly influenced by the process of *random genetic drift*, which is the term applied to the effects of randomness in the birth and death of orgamisms in a population.

A note on terminology before continuing: from here on gene genealogies will be referred to simply as genealogies. It should be understood that this refers to the genetic ancestry of a sample at some locus in a genome and not to the usual definition of a genealogy being the family history of a set of individuals. Chang (1999) and Rohde *et al.* (2004) have recently studied some aspects of common ancestry in family histories in populations of biparental organisms, with a focus on humans, and use the word genealogy in the usual sense.

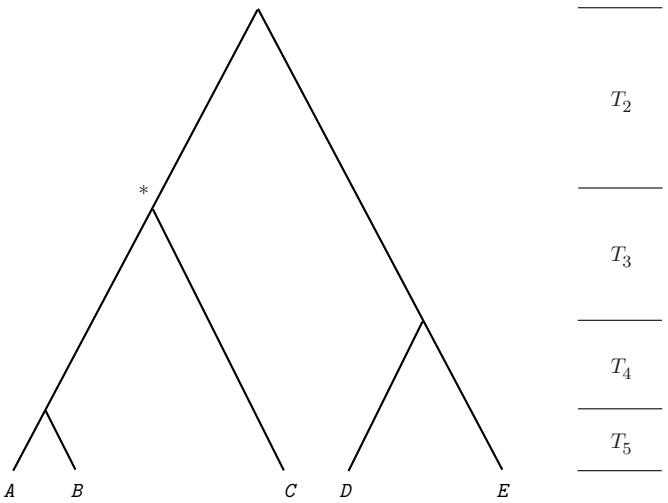


Figure 1.1: One possible genealogy of a sample of size five.

Figure 1.1 shows one possible genealogy of a sample of five gene copies. The sampled items appear at the bottom and are assumed to exist at the present time. Time is measured back from the present into the past, so we say that the sample was taken at time zero. The lines in the figure trace upwards, or backwards in time, from the present-day sample to the single common

ancestor shown at the top of the figure. These lines, or lineages, represent all of the ancestors of the sample up to the time of the most recent common ancestor. The angles at which these are drawn impart no information; the length of a lineage is just its vertical height. On their way up the lineages coalesce, in pairs, as expected from the facts of DNA replication. These coalescent events create junctures, called nodes, which occur at intervals shown to the right. Typically, T_i is used to designate the time during which there were i ancestral lineages. For example, the time back to the first coalescent event in figure 1.1 is labeled T_5 because during this time there were exactly five ancestral lineages. If n sequences are sampled, then i in T_i ranges from 2 to n.

There are, of course, many different ways to draw a genealogy besides the style in figure 1.1, and most of these appear in the literature. However they are drawn, coalescent genealogies without recombination are rooted bifurcating trees. Rooted refers to the fact that the deepest branch (uppermost in figure 1.1) is anchored by the common ancestor of the entire sample. Bifurcating refers to the fact that each node has just three lineages attached to it, one ancestral and two descendant. Again, this must be the case because this is how DNA is replicated. The root of the genealogy is the most recent common ancestor of the entire sample. In figure 1.1, four coalescent events take the sample back to its most recent common ancestor. Without recombination, a sample of n items requires exactly $n-1$ coalescent events. Excluding for a moment the root, each node defines a sub-tree that includes some members of the sample and excludes the rest. These can be likened to the clades of the phylogeneticist. For example, the asterisked node in figure 1.1 represents the common ancestor of three sequences on the left: A, B, and C. More generally, each branch in the genealogy splits the sample into two non-overlapping subsets. Every member of the sample is either on one side of a particular branch or on the other. This is important because the mutation events that produce the variation we observe in a sample must have happened on the lineages ancestral to the sample.

Every genealogy has exactly n external branches, the ones which connect each member of the sample to the rest of the genealogy. Then, as we follow the coalescent process back in time, each coalescent event terminates two branches by joining them and creates a new ancestral branch. This tracing of ancestral lineages stops at the $(n-1)$-st coalescent event, *e.g.* at the top of figure 1.1. This ultimate ancestral lineage does exist, and in some cases we may wish to follow it further back in time. Typically, we do not, and we count a total of either $2n-2$ branches if we wish to view the genealogy as an unrooted tree, or $2n-3$ branches if we wish to distinguish the two branches on either side of the root. In contrast to the n external branches, which separate single sequences from the rest, the $n-2$ internal branches partition the sample into subsets which both contain at least two members of the sample. The fact that each branch divides the sample into two distinct subsets may seem obvious, but it is also fundamental to an understanding of how genealogical structure patterns the data we observe. Consider a single mutation at some nucleotide site in the history of the sample. Each branch in the genealogy represents the opportunity for a mutation to create a particular pattern in the data, a particular kind of polymorphic site. When a mutation happens on the genealogy, the site at which it occurs will be polymorphic in the sample. This is why we do not typically care to follow the ancestral lineage of the entire sample beyond the time of the most recent common ancestor. If a mutation occurs on this branch, it will be shared by all the members of the sample and will not be polymorphic in the sample. One case in which we do become interested in these mutations is when samples come from multiple species (see Chapter 5).

Figure 1.2 illustrates how single mutations can create different patterns in the data. In figure 1.2(a), the ancestral nucleotide at some site was A. Along the first branch down to the right of the root, a mutation changed it to T. This branch leads to two members of the sample, so those two members inherit the mutant nucleotide. In contrast, in figure 1.2(b) the mutation happened on an external rather than an internal branch, so the mutant base is present in just one of the sequences. Because there are $2n-2$ branches in each genealogy, a single mutation placed on a genealogy will produce one of $2n-2$ possible patterns. However, if the ancestral

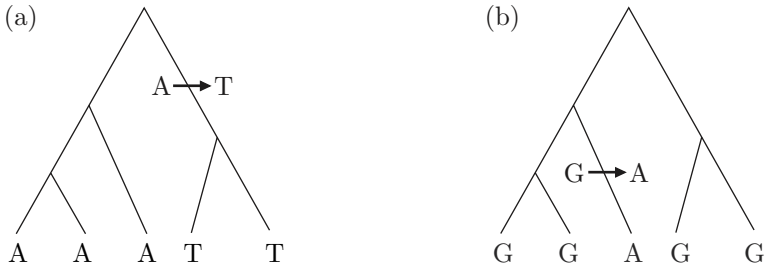
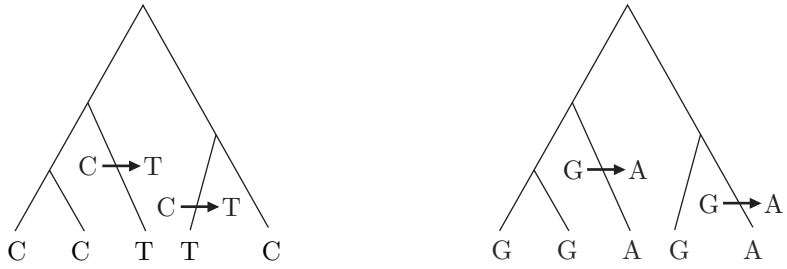


Figure 1.2: Two examples of a single mutation in the history of a sample.

state at the site where the mutation occurred is unknown, only $2n - 3$ patterns are possible. If we know that just one mutation has occurred at a site, then we know there is a branch in the genealogy of the sample that partitions the sample identically to the pattern at that site. If this is the case we say that the site pattern is *compatible* with the genealogy. The two site patterns in figure 1.2 are of course compatible with that genealogy. In fact the pattern in figure 1.2(b) is compatible with any genealogy because all genealogies have n external branches. The pattern in figure 1.2(a) would be incompatible with any genealogy that did not unite those two members on the right to the exclusion of the rest of the sample.



Sequence 1	. . . C . . . G . . .
Sequence 2	. . . C . . . G . . .
Sequence 3	. . . T . . . A . . .
Sequence 4	. . . T . . . G . . .
Sequence 5	. . . C . . . A . . .

Figure 1.3: Two site patterns which are incompatible with their genealogy and with each other.

A pair of polymorphic sites are incompatible with each other if there is no single genealogy (here a bifurcating tree) upon which a single mutation at each site could have produced the observed patterns. This definition is not very practical because it implies that in order to use it we would need to check every possible genealogy. In fact, we can assess the compatibility of two polymorphic sites directly, by comparing the subsets each makes of the sample. Namely, if both subsets at one site overlap with both subsets at the other site, then they are incompatible. The "four-gamete" test provides a simple test of compatibility between a pair of polymorphic

sites (Hudson and Kaplan, 1985). The two sites are incompatible if all four possible gametes are present in the data. The two sites in figure 1.3 because the four possible two-site patterns — CG, CA, TA, and TG — are observed. We will return to this notion in Chapter 6. There are two possible explanations for incompatible sites: either they have incompatible genealogies, or they are the result of multiple mutations. Of course, these two explanations are not mutually exclusive. Figure 1.3 shows an example of the second possibility, and of the sequence data that would result if only these two sites received mutations.

1.2 Mutation and Mutation Models

Mutation is the bridge from genealogies to genetic data because the structure of the genealogy, in which each branch divides the sample into two groups, is revealed only if polymorphisms exist among the sampled sequences. Lewontin and Hubby (1966) and Harris (1966) made the first measurements of genetic variation within populations, of *Drosophila* and humans, respectively. These were indirect measurements, not of DNA sequences but of allelic variation at the protein level detectable by gel electrophoresis, yet they showed clearly that genetic variation is abundant. A series of better and better technologies since that time have refined our picture of genetic variation. Here, the focus is on DNA sequence data, and the first such dataset in population genetics was published by Kreitman (1983), coincident with the development of coalescent theory. Single-nucleotide polymorphisms, or SNPs, do appear to be the most common type of genetic variation, but insertions, deletions, and length variation in sequence repeats such as microsatellites are also fairly common. The degree to which we can observe or infer genealogical structure depends on the rate and pattern of mutation, which in turn depends on what kind of data we have gathered. For example, microsatellite loci have a relatively high rate of mutation, so that multiple mutations may be common, and this makes the correspondence between genealogies and data less clear. By focussing on SNPs we treat the bulk of genetic variation in a greatly simplfied setting.

For every type of genetic data there is a mutation model, and we can divide these roughly into two groups: allele-based models and nucleotide-sequence models. An important distinction between the two is that allele-based models typically do not encode any information about the historical relationships among alleles in a sample, while models for nucleotide sequences naturally generate such information in the patterns of polymorphism among sites in the sample. Here, we will focus on two particular models: the infinite alleles model (Malécot, 1946; Kimura and Crow, 1964) and the infinite-sites model (Kimura, 1969; Watterson, 1975). However, any mutation model can be accommodated under the coalescent. The important simplifying assumption in modelling mutations within the coalescent is that all variation is selectively neutral. In Chapters 5 and 6 we will see the ways in which non-neutral variation can be modelled. Under neutrality, variation by definition does not affect the reproductive success of organisms and has no influence on the structure of the genealogy of a sample. Thus, the genealogical process and the mutation process can be separated, and any mutation model can in principle be applied by considering changes along the branches of each genealogy. In practice, detailed analytical results can be found relatively easily under the infinite alleles and infinite-sites models, while other models are generally implemented using computer simulations (see Chapter 8).

The infinite alleles model assumes that every time a mutation occurs it introduces a new allele into the population. Work on this model has had a profound impact in probability theory and statistics as well as in population genetics, due mostly to the discovery of the Ewens sampling formula (Ewens, 1972; Karlin and McGregor, 1972). For example, a whole chapter is devoted to Ewens distributions in a recent text (Johnson *et al.*, 1997). Historically, work on the infinite alleles model was associated with investigations of identity by descent. The concept of identity by descent, which Malécot (1946) introduced, posits that two or more gene copies are descended

from a common ancestor without mutation. That is, their identity is a direct reflection of their common ancestry, as opposed to more general case of identity in state, which includes the possibility that gene copies are identical due to multiple, convergent mutations. The infinite alleles model can be formulated in many different ways, but here we will view it through its relationship with the infinite-sites model for sequence data described below. Specifically, the infinite alleles model is equivalent to the infinite-sites model with the caveat that all we can measure is whether the sequences at the locus, also called *haplotypes*, are the same or different. We will consider these issues in detail in Section 4.2.

The infinite-sites model assumes that each mutation occurs at a previously unmutated site. In addition, it is necessary to specify some assumptions about recombination, since this too can create new alleles. For the moment we will follow Watterson (1975) in assuming that recombination does not occur within the genetic locus under consideration, but we note that Kimura (1969) proposed an infinite-sites model with free recombination between all sites, and in Chapter 6 we will see the infinite-sites mutation model with arbitrary levels of recombination. When a mutation happens under the infinite-sites model without intra-locus recombination, it necessarily creates a new allele, so the infinite-sites model is also an infinite alleles model. The infinite-sites model imagines a very large number of possible positions, each with a very small mutation rate. This assumption is very often appropriate for DNA sequences, in which the rate of mutation per nucleotide site is typically low, on the order of 10^{-8} to 10^{-9} per generation in many "higher" organisms (Drake *et al.*, 1998).

Of course, if a great deal of time has passed between an ancestor and its descendent, it will be likely that multiple mutations will have occurred. In this case, we must use detailed models of DNA sequence change and take the possibility of convergence in state into account; see Li (1997) for a review of these models. Thus, the mutation rate per site, per generation is not the only thing that determines whether the infinite-sites model is appropriate. The time over which mutations might have occurred, or the length of the genealogy, is also important. As we will see in Chapter 3, the length of the genealogy is a function of the sample size and of the effective size of the population (see Section 3.2.3), and it is possible that these will make multiple mutations common. However, the levels of polymorphism within many organisms are low enough that multiple mutations seem unlikely. For example, in humans the genomic rate of polymorphism — that is, the fraction of sites that are polymorphic when a pair of sequences are compared — is on the order of 10^{-3} (Cargill *et al.*, 1999; Stephens *et al.*, 2001), and in *Drosophila* it is on the order of 10^{-2} (Wang *et al.*, 1997; Kliman *et al.*, 2000). If every site has the potential to mutate, the fact that human genomes are roughly 99.9% identical (99% in *Drosophila*) implies that two or more mutations at a single site is unlikely, so the infinite-sites assumption may be acceptable for nuclear DNA sequences as a first approximation.

There are cases in which the infinite-sites model is obviously not applicable, such as in human mitochondrial DNA and in viral genomes where the per site mutation rate is much higher than the figures given above. Again, simulations offer an efficient way to incorporate any model of mutation. Besides their potential applicability to DNA data, the infinite-sites and infinite alleles models are attractive because of the relative ease with which analytical predictions can be generated. There is an inherent connection under these models between genealogy, mutation, and polymorphism, which fosters understanding about the forces that produce and maintain genetic variation in natural populations and the ways in which genealogical structure influences patterns of polymorphism. For example, under the infinite-sites model, when some number of mutations has occurred in the history of a particular sample, these will have happened independently at exactly that number of sites. We can think of them as being throw down randomly upon the genealogy of the sample. They will be distributed in some manner among the branches of the genealogy, and each mutation will produce a polymorphic site that partitions the sample just as the branch does along which it occurred.

1.3 Measures of DNA Sequence Polymorphism

A major aim of population genetics is simply to quantify levels of genetic variation. For a sample of DNA sequences, one possible measure of variation would be the raw data themselves, such as those pictured in figure 1.4 below. This is of course the most detailed information available, including the numbers of the different kinds of polymorphic sites and their distribution along the sequence. Computational methods for handling such data are considered in Chapter 8. These methods are computationally intensive, sometimes prohibitively so, and a lot of recent effort has gone into improving them. One avenue of improvement is to use data summaries that capture, as far as possible, the information in the data that bears on the question at hand. Historically, workers have focused on summary measures of genetic data, although it has not been primarily for this reason. Summary statistics can have a direct connection with some quantity of biological significance. For example, the average number of pairwise sequence differences is an extension to sequence data of the concept of *heterozygosity* of individuals in a diploid species (Tajima, 1993), which we will consider in Sections 3.1.1 and 3.1.2. In addition, theoretical predictions for summary measures can be generated relatively easily using the coalescent, which can advance our understanding of how evolutionary forces act to shape genetic variation. Predictions can be compared to observed data, often telling us something significant about biology or demonstrating that our collection of summary measures is insufficient to resolve an issue of interest.

There are of course a great number of ways to summarize the information in a sample of DNA sequences. Three historically important summary statistics are introduced here: the number of segregating sites in the sample, the average number of pairwise differences, and the *site frequencies*. These measures have motivated a lot of work in coalescent theory, and in the coming chapters we will uncover relationship between them and the various evolutionary forces that affect variation. Assume, as in Section 1.1 above, that we have taken a sample of DNA sequences from some genetic locus from a population or species. Figure 1.4 shows an example for the case of $n = 5$ sequences sampled. There are four polymorphic sites, and only these are shown in the figure. The intervening sites all show the same base in every sequence and are not displayed. Their presence is indicated by '. . . ' in the figure. In practice, we might have sequenced from tens to thousands of base pairs of DNA in order to find four polymorphic sites. Notice that the first two sites in figure 1.4 are the ones shown in figure 1.2. The other two sites are also each derived from a single mutation on a branch of the genealogy in figure 1.2. Thus, these data are consistent with the infinite-sites model of mutation.

Sequence	1	.	.	.	A	.	.	.	G	.	.	.	C	.	.	.	G	.	.	.
Sequence	2	.	.	.	A	.	.	.	G	.	.	.	T	.	.	.	G	.	.	.
Sequence	3	.	.	.	A	.	.	.	A	.	.	.	T	.	.	.	T	.	.	.
Sequence	4	.	.	.	T	.	.	.	G	.	.	.	T	.	.	.	T	.	.	.
Sequence	5	.	.	.	T	.	.	.	G	.	.	.	T	.	.	.	T	.	.	.

Figure 1.4: A dataset of DNA sequences.

The first and simplest measure of DNA sequence polymorphism is the number of segregating sites in the sample, which we will denote S. The word segregating just means polymorphic, and has its roots in the discovery of Mendelian inheritance. The data in figure 1.4 have $S = 4$ segregating sites. Note again that under the infinite-sites model, each mutation happens at a unique site, so every mutation that occurs along any lineage in the history of the sample will appear as a segregating site. In this case, the number of mutations and the number of

segregating sites are the same. Even before embarking on a theoretical treatment, we can make some robust predictions about the number of segregating sites in a sample. First, we expect that if we sequence a longer stretch of DNA we will find more segregating sites. If the mutation rate per site is constant along the sequence, the number of segregating sites should, on average, be proportional to the length of the sequences. Since the mutation rate for the whole sequence is equal to the per site mutation rate times the number of base pairs, another way of looking at this is to say that S should be proportional mutation rate. The relationship will be linear if the infinite-sites models holds. Next, we might guess that S should increase with the sample size, n. This is indeed the case, and we will be able to make this statement more specific after we developing the machinery of the coalescent in Chapters 3 and 4.

The second measure of DNA sequence polymorphism is the average number of differences between pairs of sequences in the sample, which historically has been called π (Tajima, 1983). It is calculated by comparing each sequence to every other one, counting the number of differences between them, and taking the average of these. Thus,

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij} \tag{1.1}$$

in which k_{ij} is the number of differences between sequence i and sequence j. For example, $k_{12} = 1$ and $k_{34} = 2$ for the data in figure 1.4. The symbol $\binom{n}{2}$ represents the number of pairwise comparisons made, *i.e.* the number of terms in the double sum in equation 1.1. We will encounter this notation again in Section 2.1. When all ten pairwise comparisons for the data in figure 1.4 are made we find that $\pi = 2$. Note that in general π will not be an integer, but S always is.

Both S and π represent fairly drastic summaries of any DNA data. They reduce all of the information contained in the data, *e.g.* those in figure 1.4, to a single number. For some purposes this may be sufficient, and for simplicity even desirable, but for others we may want to capture more information about how genetic variation is structured. Given the correspondence between genealogical structure and observable DNA data under the infinite-sites mutation model, an obvious choice would be to count the number of each possible kind of polymorphic site, distinguishing between sites that partition the sample in different ways. Using this criterion, each of the four polymorphic sites in the data of figure 1.4 is unique, and we would count one each for four different kinds sites (and zero for all other possible patterns). There are a very large number of such patterns when the sample size is not small, and instead of distinguishing all of these, most work has focussed on site frequencies, which lump together all the patterns that have the same counts of the two segregating bases. For example, the second and third sites in figure 1.4 both have one sequence different from the other four, so they would be counted together. Site frequencies provide an intermediate measure, between the total data and the extreme summaries S and π.

Assume for the moment that we know which is the ancestral nucleotide and which is the mutant at each site in our data. For example, the ancestral base in figure 1.2(a), which produced the first polymorphic site in the data in figure 1.4, is A. We might know the ancestral base at each site by sequencing the same locus in a closely related species, so that the infinite-sites model holds between species as well as within. In this case we are able to count the number of sites ξ_i at which the mutant base is present in i copies and the ancestral base is in $n - i$ copies in the sample. If the infinite-sites model holds, then there are never more than two bases segregating at a polymorphic site, and we know that each polymorphism is the result of a single mutation. The i in ξ_i can range from 1 to $n - 1$, otherwise the site is not polymorphic, and the counts ξ_i over i are called the "unfolded" site-frequency spectrum. The situation would not be so simple if multiple mutations occurred at single sites, such that three or four bases might be present at a polymorphic site, but we will ignore this possibility for the moment.

Often we do not know which is the mutant and which is the ancestral base at a polymorphic site, and we cannot distinguish a pattern where the mutant base is in i copies from one in which the mutant is in $n-i$ copies. Then we are limited to counting the number of sites η_i at which one base is present in i copies and the other is present in $n-i$ copies:

$$\eta_i = \frac{\xi_i + \xi_{n-i}}{1+\delta_{i,n-i}} \qquad 1 \le i \le [n/2] \tag{1.2}$$

in which $[n/2]$ is the largest integer less than or equal to $n/2$. In this case the largest site frequency we can count is $n/2$ when n is even, and $(n-1)/2$ when n is odd. For example, for the data in figure 1.4, we count $\eta_1 = 2$ (sites two and three) and $\eta_2 = 2$ (sites one and four). The symbol $\delta_{i,j}$, Kronnecker's δ, is equal to zero when $i \neq j$ and equal to one when $i = j$. The reason for its presence in equation 1.2 is that, when n is even, $\xi_{n/2}$ and $\xi_{n-n/2}$ count the exact same sites. The smallest site frequency we can count is always η_1 and this is often referred to as the number of singletons in the sample. The distribution of η_i over i is called the "folded" site-frequency spectrum.

All three measures, S, π and η_i (or ξ_i) summarize the information contained in sequence data, but in different ways. More specifically, they bear a certain relationship to one another: both S and π are simple functions of the η_i (or ξ_i). For S the relationship is obvious:

$$S = \sum_{i=1}^{[n/2]} \eta_i \tag{1.3}$$

since both simply count the polymorphic sites in the sample. For π we have

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{[n/2]} i(n-i)\eta_i. \tag{1.4}$$

When a polymorphic site divides the sample into i and $n-i$ sequences, it will show a pairwise difference in exactly $i(n-i)$ of the $\binom{n}{2}$ possible pairwise comparisons. Instead of forming all pairs and counting total numbers of differences along the sequence, we can consider each polymorphic site in turn and calculate its contribution to π. To obtain the relationship of S and π to ξ_i, we replace η_i with ξ_i in both equations, 1.3 and 1.4, and take the sums over $1 \le i \le n-1$.

These three measures capture different aspects of the information in the data about the genealogical history of the sample. Under the infinite-sites model of mutation, the number of segregating sites, S, is simply the total number of mutations in the history of the sample. The counts of sites frequencies, η_i, are the numbers of mutations that occurred on lineages which left either i or $n-i$ descendents in the sample. The average number of pairwise differences, π, traces all the lineages that connect each pair of tips of the genealogy and counts the numbers of mutations on them. Thus, S and η_i count each mutation exactly once whereas π weighs sites depending on how the branch where the mutation occurred, or the site itself, divides the sample. The term $i(n-i)$ in equation 1.4 is largest when $i = [n/2]$ and smallest when $i = 1$, so middle-frequency sites contribute disproportionately to π.

It is important to consider what information has been sacrificed when these data summaries are adopted. Because both S and π are further summaries of the site frequencies, we can consider η_i (or ξ_i). An important aspect of the data that is missed entirely in the site frequencies is the distribution of the polymorphism among chromosomes or haplotypes in the sample. For example, the data in figure 1.5 give the same values of η_i as the data in figure 1.4. These two data sets have the same number of polymorphisms, segregating in the same frequencies, but differ in how the mutant and ancestral bases at each are distributed among the sequences. The data in figure 1.5

```
Sequence 1   . . . A . . . G . . . T . . . G . . .
Sequence 2   . . . T . . . G . . . T . . . G . . .
Sequence 3   . . . T . . . G . . . T . . . T . . .
Sequence 4   . . . A . . . G . . . T . . . T . . .
Sequence 5   . . . A . . . A . . . C . . . T . . .
```

Figure 1.5: Another dataset of DNA sequences.

fail the four gamete test (compare the first and last sites), so if we are confident that only a single mutation has occurred at each site, some recomination must have occurred in the history (see Chapter 6). The data in figure 1.4 pass the four-gamete test. This difference between the two datasets cannot be inferred from the site-frequency counts η_i (or ξ_i) so we have definitely lost some information. In addition, none of these summary measures depend on the distribution of the sites along the sequence or chromosome. Whether or not this is a problem will of course depend on what we are trying to infer from the data.

Despite the fact that a great deal of information is lost in reducing a rich object like a sample of DNA sequences into these three very simple and very drastic data summaries, these measures S, π, and η_i will remain in focus in the coming chapters. There are two main reasons for this. First, it is fairly easy to generate predictions about them using the coalescent, and a lot of theoretical and empirical work has focussed on them. Thus, these measures underlie current understanding of how genealogical processes structure observed patterns of genetic variation. Although they are fairly simple summaries, there is value in doing this because, as we will see in Chapter 4, S, π, and η_i provide information about the relative lengths of different types of branches in the genealogy of the sample, *i.e.* branches that have i descedents in the sample. As a consequence, a battery of statistical tests of the standard coalescent model have been devised using combinations of S, π, and η_i. Deviations from the predictions of the standard coalescent model indicate that one or more assumptions of the model are wrong, that the structure of the genealogy of the sample has been molded by natural selection, changes in population size, geographic subdivision, or other factors. Later, when we consider extensions of the standard coalescent model, these measures will of course become insufficient and we will need to consider more informative statistics. The following section provides our first taste of this.

1.4 Variation at the PDHA1 Locus in Humans

Actual DNA sequence data can be considerably more complex than the hypothetical sequences discussed above. For example samples are very often taken over the range of a species, in an attempt to assess the degree of population structure and to understand the historical processes and events that gave rise to current patterns of polymorphism. Harris and Hey (1999) reported sequence data from a 4.2 kilo-base region of the PDHA1 (pyruvate dehydrogenase *E1* α subunit) locus in a geographically diverse sample of 35 humans. These data are shown in figure 1.6. The figure shows the eleven unique haplotypes that were uncovered in sequencing the 35 samples. Also shown is the single haplotype found when two chimpanzee samples were sequenced for the same region. The chimpanzee sequence can be used to root the human polymorphisms, *i.e.* to determine the ancestral state at each polymorphic site. There are $S = 25$ segregating sites, and $\pi = 6.07$ among all the human samples. The position numbers in Harris and Hey's (1999) alignment of these data are shown above the sequences. Since eleven haplotypes were

found when 35 humans were sequenced, many humans must be identical at the PDHA1 locus. This was certainly true in the sample. Finally, the samples came from eight different places on the globe, four in Africa and four outside Africa. Within Africa these were coded: B, South African Bantu speakers; S, Senegalese; K, Khoison from the Angola/Namibia border; P, Pygmy from the Central African Republic. Outside Africa these were coded: C, China; V, Vietnam; F, France; M, Mongolia. Because homonids originated in Africa there is a great deal of interest in understanding patterns of polymorphism within and between African and non-African samples. Figure 1.6 shows the counts of each haplotype in the samples from each locality.

	Base Position								
	11111112222222333344								
	14567023345911124491346 12	Population Counts							
	59479030537323661878008 00								
Haplotypes	74421520693615067959678 89	African				Non-African			
Chimpanzee	CCGGTTATGCCGAGAATACGGCGCC	B	S	K	P	F	C	V	M
A	--ACCC--TGT--AC-CC-----T-	-	-	-	-	-	2	1	1
B	--ACCC--TGT--AC-C------T-	-	-	-	-	5	5	4	-
B1	--ACCC--TGT--AC-C---A--T-	-	-	-	-	1	-	-	-
C	---CCC--TGT--AC-C------T-	1	2	-	2	-	-	-	-
D	-A-----C--*-T-----T--T---	-	1	-	1	-	-	-	-
E	TA-----C----------T--T---	1	-	-	-	-	-	-	-
F	-A----CC----------TA-----	-	1	-	-	-	-	-	-
G	-A-----C-------G--T---C-T	1	-	-	-	-	-	-	-
H	-A----CC--*----G--T---C--	-	2	-	-	-	-	-	-
I	-A-----C--*A------T-A-C--	1	-	2	-	-	-	-	-
J	-A-----C--*-------T------	-	-	1	-	-	-	-	-

Figure 1.6: Polymorphic base positions within humans at the PDHA1 locus.

A quick visual inspection of the data in figure 1.6 reveals an apparently deep split between two groups of haplotypes: A, B, B1, and C, versus D through J. Twelve sites in the data distinguish these groups perfectly, with A through C showing one base and D through J showing a different base. The first group contains all 19 sequences from outside Africa and five sequences from three different places within Africa. It may seem remarkable that this split is not between the African and non-African samples. However, it has been known for decades that these broad geographic labels, and the patterns of race and ethnicity that tend to go with them, account for a very small fraction of variation at the genetic level (Lewontin, 1972). With the multitudes of genomic sequence data now available, this conclusion continues to hold (Rosenberg *et al.*, 2002), but it is also possible to make more refined statements about human variation (Excoffier, 2002). For example, it appears that Africa houses more genetic variation than other large geographic regions, *e.g.* Asia or Europe. The data in figure 1.6 show this pattern as well. With this background, what was surprising about the PDHA1 data of Harris and Hey (1999) is that they actually do contain a site that distinguishes the African from the non-African samples (site 544). It is very unlikely to find a correspondence such as this, a fixed difference between samples from different localities, unless subdivision between the regions is strong (see Chapter 5). Later however, by sequencing more copies of the locus, it was shown that this site does not distinguish Africans from non-Africans (Yu and Li, 2000), it only distinguishes the African and non-African *samples* studied by Harris and Hey (1999).

A number of different summary measures have been proposed for the study of geographically-distributed samples, such as those in figure 1.6. These same measures can be used to analyze

data from different species. Nei and Li (1979) proposed the net number of nucleotide differences

$$\pi_{ij}^{(\text{net})} = \pi_{ij} - \frac{\pi_i + \pi_j}{2} \tag{1.5}$$

between two sub-populations i and j as a measure of divergence. For the PDHA1 data in figure 1.6, we might define just two sub-populations, Africa and non-Africa, or we might prefer to recognize the eight, more geographically restricted sub-populations. We might even define several hierarchical levels of population structure, *e.g.*, as in Slatkin and Voelm (1991). Let us assume a sample from just two sub-populations, where the first n_1 sequences are from one sub-population and the next n_2 sequences are from the other. The terms on the right in equation 1.5 would be the average number of pairwise differences between sub-populations 1 and 2,

$$\pi_{12} = \frac{1}{n_1 n_2} \sum_{r=1}^{n_1} \sum_{s=n_1+1}^{n_1+n_2} k_{rs} \tag{1.6}$$

and the average number of pairwise differences within each sub-population, *e.g.*

$$\pi_1 = \frac{1}{\binom{n_1}{2}} \sum_{r=1}^{n_1-1} \sum_{s=r+1}^{n_1} k_{rs},$$

$$\pi_2 = \frac{1}{\binom{n_2}{2}} \sum_{r=n_1+1}^{n_1+n_2-1} \sum_{s=r+1}^{n_1+n_2} k_{rs}. \tag{1.7}$$

It is straighforward to extend this to more than two sub-populations.

In Chapter 5, we will see how predictions about these measures in a subdivided population can be made using the coalescent. For the PDHA1 data, we have $\pi_1 = 6.93$, $\pi_2 = 0.57$, and $\pi_{12} = 8.82$, in which sub-population 1 is Africa and sub-population 2 is non-Africa. Net divergence is $\pi_{ij}^{(\text{net})} = 5.07$ because many African/non-African pairs differ every one the twelve sites that distinguish the two major haplotype groups. Thus, these 'sub-populations' are diverged to some extent, although we will not be in a position to judge the significance of this until Chapter 5. Clearly, the fact that $\pi_1 > \pi_2$, which is consistent with the statement made above that African sub-populations harbor a greater proportion of the total variation among humans than do non-African sub-populations, explains some portion of the net divergence between these African and non-African samples. Equations 1.5 through 1.7 represent just one of many possible ways to summarize DNA sequence polymorphisms within and between sub-populations, and we will consider several others in Chapter 5.

With the exceptions of sites 1232 and 3306, the PDHA1 data of Harris and Hey (1999) are consistent with the infinite-sites mutation model without intra-locus recombination. That is, with the exception of 1232 and 3306, all pairs of sites in figure 1.6 pass the four-gamete test (Hudson and Kaplan, 1985). Harris and Hey (1999) followed common practice and ignored these two sites in their subsequent analyses which assumed the infinite-sites model of mutation. Of course, tossing out even a small number of data points is never the ideal course of action because some information is lost and, depending on the method of analysis, the results could be biased. However, when the proportion of sites ignored is small the error is probably not great. In addition, the majority of available methods and software to make inferences from DNA sequence data assume the infinite-sites model, so having infinite-sites data greatly increases the number of tools for analysis. In a similar vein, note that Harris and Hey (1999) coded the deletion among some members of their sample at site 1573 as the ancestral state (C) at that site. With

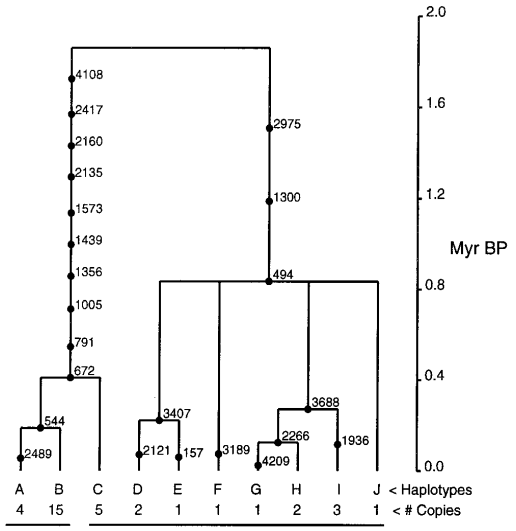


Figure 1.7: The gene tree for the PDHA1 data of Harris and Hey (1999).

these minor omissions, the PDHA1 data of Harris and Hey (1999) are fully consistent with the infinite-sites mutation model.

Due to the one-to-one correspondence between polymorphic sites and branches in the genealogy of the sample under the infinite-sites mutation model, a single gene tree can always be reconstructed from the infinite-sites data. Gusfield (1991) describes an efficient algorithm for inferring the structure of the tree from infinite-site data. The gene tree estimated from the PDHA1 data of Harris and Hey (1999) is shown in figure 1.7. It shows the thirteen mutations/sites that separate the two major haplotype groups in the data as well as the mutation/site 544 at which the mutant base is present in all the samples from outside Africa but none of the samples from within Africa. The timescale at the right of figure 1.7 was obtained by assuming a mutation rate of $\sim 10^{-9}$ per site per generation, a generation time of 20 years, and by assuming the humans and chimpanzees shared a common ancestor 5 million years ago (Harris and Hey, 1999). If the data are sparse relative to the sample size, then there must be some lineage(s) in the genealogy of the sample that did not experience any mutations. Thus a gene tree reconstructed from data usually leaves some branchings unresolved. For example, with the exception of haplotypes E, F, G, and J, all the tips of the gene tree in figure 1.7 actually comprise several coalescent events. There is no information in the data about their order. The same is true of the three coalescent events that occurred between the mutation/site 494 in the figure and mutation/site 3688. We must consider all possibilities to be equally likely. Chapter 8, describes how genealogies like the one in figure 1.7 can be obtained using Griffiths and Tavaré's (1994) Monte Carlo method to compute the likelihood of the data under the coalescent with infinite-sites mutation. For a recent review of this method, see Griffiths (2002).

1.5 Exercises

1. What is the total number of possible arrangements of the four bases, A, G, C, and T, at a single site in a sample of n sequences labelled $1, 2, \ldots, n$?

2. Draw a genealogy of a sample of six sequences, with mutations on the branches, such that $\xi_1 = 1$, $\xi_2 = 2$, and $\xi_4 = 3$.

3. Do the data below pass or fail the four-gamete test? Explain your answer.

Seq 1	...A...T...G...A...
Seq 2	...A...T...G...C...
Seq 3	...G...T...G...A...
Seq 4	...G...C...T...A...
Seq 5	...G...C...G...C...

4. Compute the site-frequency counts ξ_i, $i = 1, 2, 3, 4$, for the data above by assuming you have an outgroup sequence that is identical to sequence 5. What are folded site-frequency counts η_i for these data?

5. Show two different methods of computing the average number of pairwise sequence differences, π, for the data in excercise 3.

6. Draw a possible genealogy for the data in excerise 3, and show how the data might have been derived via mutations along the branches of the tree.