Part I

Concepts and Methods in Bacterial Population Genetics

Chapter 1

The Coalescent of Bacterial Populations

Mikkel H. Schierup and Carsten Wiuf

1.1 BACKGROUND AND MOTIVATION

Recent years have seen an explosion in the number of available DNA sequences from many different species. Whereas small genomic regions routinely have been sequenced for more than 20 years and have improved our knowledge of genetic variation at the species and the population levels, new high-throughput techniques have made possible the sequencing of whole genomes and genomic regions for many individuals at an affordable price and in a realistic time frame. This offers unprecedented opportunities for studying genetic variation within and between species and the effects of variation on transcription, regulation, and expression. So, for example, population data sets for bacteria are now expected to consist of full genomes rather than single genes, and the limitations to evolutionary inference are more likely to be found in the analysis rather than in the generation of sequence data (see Chapter 7 of this book).

In the following, we will discuss a mathematical model—*the coalescent*—that describes the process of generating genetic data, with special reference to bacterial populations. For simplicity, we assume the data are in the form of DNA sequences; however, other forms of genetic markers can likewise be modeled. The sequences (or genes) are all homologous copies of the same genetic region in the genome of a species. The relevance of such a model becomes clear when we want to infer/learn details about the evolutionary processes that generated and shaped a sample of present-day sequences. This process may include inferring the mutation rate or demographic parameters, or assessing the age of mutations or common ancestors of sequences. The inferential analysis is retrospective; we seek to understand the evolutionary past of the sample (or population) through analysis of the present-day sequences.

Coalescent theory is the most widespread statistical framework for retrospective statistical analysis of genetic data. The term was coined by Kingman (1982a), who described the genealogy of a sample of n sequences and denoted the genealogical process the coalescent. In subsequent papers, Kingman (1982b,c) developed the theory further and within a few years, it was being studied widely. Kingman's (1980) work built on his own research

Bacterial Population Genetics in Infectious Disease, Edited by D. Ashley Robinson, Daniel Falush, and Edward J. Feil

as well as that of others, for example, Ewens (1972) and Watterson (1974). The coalescent was also independently discovered by Hudson (1983a,b) and Tajima (1983), and in unpublished notes by Bob Griffiths.

In this study, we will first show how simple models of reproduction can be formulated and will discuss their relationship to real bacterial populations. The simple models of reproduction underlie the basic (or standard) coalescent process, which is often used as a null model for statistical analysis. Subsequently, we will introduce some extensions of the basic model that allow for demography and recombination/gene conversion. The extensions predict measurable effects on a sample of sequence data, effects that in turn provide a means for interpreting the data. For further background on the coalescent, see the books by Wakeley (2008) and Hein et al. (2005).

1.2 POPULATION REPRODUCTION MODELS

A simple model of population reproduction was first suggested by Wright (1931) and Fisher (1930). This basic model provides the description of an idealized population and the transmission of genes from one generation to the next. In this study, we consider this model and two other similar models that might be useful for describing bacterial evolution. However, as our exposition is adapted to haploid populations, it may differ slightly from other examples in the literature.

A population of constant size N of haploid individuals forms the basis for our study. At time (generation) $t + 1$, N individuals are drawn from the population at time t—we then consider three different ways that each mimics reproduction in a true physical population (see Fig. 1.1). We use the terms "individuals," "sequences," and "genes" interchangeably in this section since for a haploid, nonrecombining organism, the history of any gene is the same as the history of the bacterial cells. The models we refer to in the study include the following:

Wright–Fisher (WF) model: N individuals are drawn randomly with replacement from the population at time t. The number of descendants of one individual in one time step is approximately Poisson distributed $P(k) = \exp(-1)/k!$.

Moran model: At time t, one individual is chosen randomly to reproduce and one individual is chosen to die. The same individual can be chosen to reproduce and then die. Thus, an individual has either zero, one, or two descendants. Zero and two with equal probability $p_0 = p_2 = (N-1)/N^2$, and one with probability $p_1 = 1 - 2p_2$.

Fission model: At time t, each individual has zero, one, or two descendants with probabilities p_0, p_1, and p_2, respectively. For the population to remain of constant size, we must have $p_0 = p_2 \leq 0.5$.

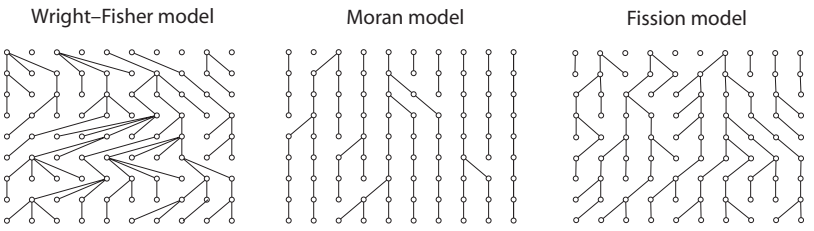


Figure 1.1 Eight generations of reproduction in the Wright–Fisher model, the Moran model, and the fission model, which have properties intermediate between the other two models (see text).

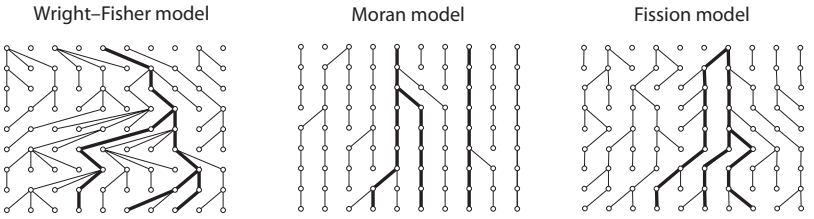


Figure 1.2 The genealogy of a sample of $n = 3$ genes in each of the three reproduction models. Note that coalescent events occur more rapidly in the Wright–Fisher model than in the Moran model. In this example, the three genes coalesce and find an MRCA five generations back, whereas in the Moran model, the three genes have not found an MRCA after eight generations, where there are still two ancestors to the sample. The fission model shows an intermediate pattern.

In the WF model, the entire population is replaced in each time step, whereas in the Moran model, it takes in the order of N time steps before the population is replaced by new individuals. The WF model is often referred to as a nonoverlapping generation model, while the Moran model is referred to as an overlapping generation model, because an individual that does not die continues to the next generation. Figure 1.1 shows eight time steps for each of the three models.

All these models rely on a number of essential, simplifying assumptions: (i) The population is selectively neutral; all alleles are equally fit; (ii) the population has no demographic structure; (iii) the genes are not recombining. We will later discuss how to incorporate recombination and demography, but not selection.

In the study under discussion, we could use these models to trace the genealogical relationship of a sample of n genes backward in time. In Fig. 1.2, this relationship is shown for a sample of size 3 for each of the three reproduction models. In the WF model, the first two genes find a common ancestor two generations back, whereas all three genes share a common ancestor five generations back. The first ancestor of the complete sample is called the *most recent common ancestor* (MRCA) to distinguish it from other ancestors of the sample further back in time. In the Moran model, the three genes have not yet found a common ancestor after eight time steps, but if we progressed far enough back in time, they would eventually find one, since in each time step there is a positive probability for this to happen. The fission model is intermediate between the WF and the Moran model in that coalescent events happen at a slower rate in the fission model than in the WF model, and at a faster rate in the fission model than in the Moran model. In Fig. 1.2, an MRCA is found after eight time steps for the fission model.

1.3 TIME AND THE EFFECTIVE POPULATION SIZE

As the above description suggests, the genealogical history depends on the reproductive model. However, for a large population (large N), all three models show remarkable similarities (Kingman, 1982a–c). To demonstrate this, we first describe the coalescent structure of a sample of size n, taken from the WF model. The probability that *none* of the n genes find a common ancestor in the previous generation is

$$\frac{N-1}{N}\frac{N-2}{N}\cdots\frac{N-n+1}{N}=\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)\approx 1-\frac{n(n-1)}{2N}. \quad (1.1)$$

The latter approximation holds for large N only. The first gene chooses a parent at random; the second can choose among the remaining $N-1$ genes, the third among $N-2$ genes, and so on. Consequently, the probability that none of the n genes have found common ancestors in the previous t time steps is

$$P\left(T_n^N > t\right) \approx \left(1 - \frac{n(n-1)}{2N}\right)^t, \tag{1.2}$$

where T_n^N denotes the waiting time until the first common ancestor event (superscript N refers to the dependency on population size N). The probability that more than two genes coalesce in the same generation becomes negligible for large N, and henceforth it is ignored in our exposition. The coalescing pair of genes is chosen randomly among all genes in the sample.

Equation 1.2 depends on the population size N. However, if time is scaled in units of N generations, Equation 1.2 takes the approximate form

$$P\left(T_n > v\right) \approx \exp\left(-\frac{n(n-1)}{2} v\right), \tag{1.3}$$

where now $T_n = T_n^N / N$. The argument that changes the product in Equation 1.2 into an exponential term in Equation 1.3 relies on N being large and n being relatively small. The right side can be recognized as an exponential variable with rate $n(n-1)/2$. Consequently, the genealogy of a sample is described by a series of waiting times $T_n, T_{n-1}, \ldots, T_2$ between successive coalescent events; each waiting time is an exponential variable with rate depending on the current number of ancestors. Equation 1.3 has the further important consequence that the genealogy of the sample depends on N only through a scaling of time. For the WF model, the scaling is linear in population size N.

Kingman (1982c) showed that for a variety of reproductive models, including the models discussed here, time can be scaled such that the time between coalescent events is approximately an exponential variable with rate $k(k-1)/2$, where k is the number of current ancestors (Fig. 1.3). At each coalescent event, two genes are chosen randomly to coalesce.

The scaling factor is known as the *effective population size*, N_e; see Ewens (2005) for discussion and formal definitions. The number N_e depends on N and on the reproductive mechanism in the following way:

$$N_e = \frac{N}{\sigma^2}, \tag{1.4}$$

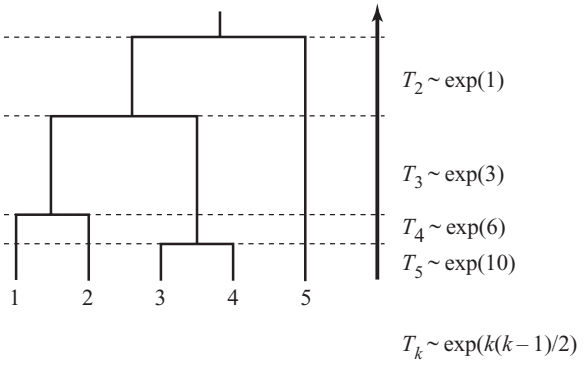


Figure 1.3 The genealogy is described by a series of coalescent events. The waiting times between coalescent events are exponentially distributed with intensities shown in the figure. The intensities depend on the squared number of sequences and therefore grow dramatically with an increasing number of sequences (see also Fig. 1.4). The coalescing pair is choosing randomly among all possible pairs of genes.

where σ^2 is the variance in offspring number (number of lineages subtending an individual in the next generation). For the three models discussed here, we have for large N $N_e^{(WF)} = N$ in the WF model (as already stated), $N_e^{(M)} = N^2/2$ in the Moran model, and $N_e^{(F)} = N/(2p_2)$ in the fission model. Note that if $p_2 = 1/N$, then the fission model is similar to the Moran model, and if $p_2 = 0.5$, then the fission model is similar to the WF model. Hence, in this sense, the fission model embraces both other models, though all of the models differ at the detailed level.

One interpretation of the effective population size is that it is the corresponding size of a similar WF model. For example, a Moran model with population size N corresponds to a WF model with population size $N^2/2$. (Sometimes, the effective population size is defined differently for overlapping generation models; see Ewens, 2005 and below.) Also, if a real physical population has effective population size N_e, then it is similar, with respect to time by generations, to a WF model also with size N_e.

The fission model most closely resembles an idealized bacterial population where individuals divide by fission. In each time step, a certain proportion of cells divide (<50%), a proportion does not divide, and a proportion dies (<50%) in order for the population size to remain constant (growing populations are treated below).

1.3.1 Algorithm 1

Based on the exposition above, an algorithm for simulating the genealogy of a sample of n genes is:

1. Start with $k = n$ genes.
2. Simulate an exponential variable with rate $k(k-1)/2$.
3. Choose two genes randomly among the k genes to coalesce.
4. Put k equal to $k - 1$.
5. If $k > 1$, go to 1; otherwise, stop.

To calculate time in terms of generations, multiply all coalescent times by N_e. This algorithm was used for an initial $n = 50$ genes in order to generate Fig. 1.4.

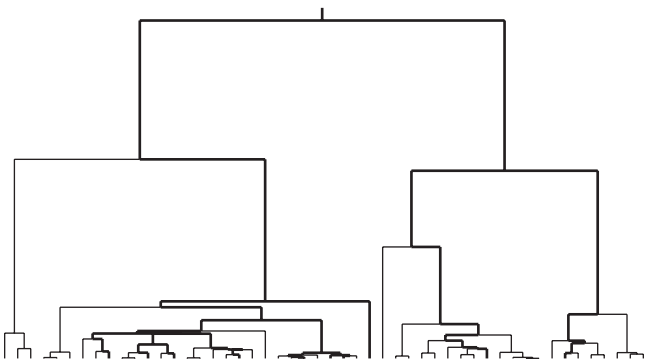


Figure 1.4 An example genealogy of 50 genes under the basic coalescent process. Thick lines track the genealogy of a subsample of size 10. Two features are noteworthy: (i) Coalescent events occur rapidly with many sequences; and (ii) the subsample shares most of the deep branches in the genealogy and the MRCA with the entire sample.

1.4 THE GENEALOGY OF A SAMPLE OF SIZE *n*

In this section, we draw some conclusions from the results of the previous section. The mean and the variance of the (scaled) waiting time while there are k ancestors, $k = 2, \ldots, n$, are, respectively,

$$E(T_k) = \frac{2}{k(k-1)} \tag{1.5}$$

$$Var(T_k) = \frac{4}{k^2(k-1)^2}. \tag{1.6}$$

Thus, more time is spent on average when there are few ancestors than when there are many ancestors (see also Fig. 1.3), and the variance in coalescence times is dominated by the variance when there are few ancestors. The time W_n until the MRCA is found is just the sum of the waiting times T_k; that is, $W_n = \sum_{k=2}^{n} T_k$, which has mean and variance given by

$$E(W_n) = 2\left(1 - \frac{1}{n}\right) \tag{1.7}$$

$$Var(W_n) = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} \approx 1.16. \tag{1.8}$$

The latter approximation holds for large sample sizes n. We note some immediate consequences of Equations 1.7 and 1.8: (i) The mean depth of the genealogy of any sample is bounded by 2; hence, an MRCA will *always* be reached, even for very large samples; (ii) even in a large sample, about half of the time is spent while the sample has two ancestors, since $E(T_2) = 1$; (iii) the time while there are two ancestors is much more variable than the remaining time, since $Var(T_2) = 1$, but also $Var(W_n) \approx 1.16$. Thus, unlinked genes might by chance have very different times until their MRCA.

Another quantity of interest is the total size of the genealogy L_n. It is given by $L_n = \sum_{k=2}^{n} kT_k$, because each of the k ancestors contributes T_k to the total size (see Fig. 1.3). It has mean and variance given by

$$E(L_n) = \sum_{k=2}^{n} \frac{2}{k-1} \approx 2\log(n) \tag{1.9}$$

$$Var(L_n) = \sum_{k=2}^{n} \frac{4}{(k-1)^2} \approx 6.58. \tag{1.10}$$

The approximations hold for large sample sizes n. In contrast to the mean depth of the genealogy, the mean of the total size grows without bounds for increasing sample size. However, it grows very slowly, and adding a few more genes only adds a little to the total branch length.

Figure 1.4 shows a sample of size 10 embedded in a larger sample of size 50. In a typical genealogy, the deep branches are shared between the two samples, and adding more genes mainly results in small twigs on the coalescent tree. Consequently, there is high probability that the MRCA of the large sample is also the MRCA of the embedded sample. With the sample sizes of Fig. 1.4, the probability that the embedded sample shares the MRCA with the large sample is 85%. If the larger sample is the entire population (or bacterial species), the probability of MRCA sharing is $(n - 1)/(n + 1)$, where n is the size

of the embedded sample (Hein et al., 2005). For $n = 20$, the probability is above 90%, and for $n = 100$, the probability becomes 98%. Thus, the genealogy of a few genes shares important features with the genealogy of the entire population.

1.5 FROM COALESCENT TIME TO REAL TIME

In the above exposition, time is measured in generations or in units of the effective population size N_e. However, it is often of interest to be able to infer the actual physical time in a genealogy. This is possible from sequence data if the mutation rate per time step is known (see below) or if there is an independent estimate of the effective population size. As an example, in *Escherichia coli*, the effective population size may be as large as 50 million (Charlesworth and Eyre-Walker, 2006; Charlesworth, 2009), and if we assume 200 generations per year in the wild, the expected coalescence time for two randomly picked bacteria (if clonal reproduction) would be $N_e = 50$ million generations or 250,000 years. This might be contrasted to humans where the generally agreed numbers are an effective diploid population size of 10,000 and a generation time of 20 years, implying an expected coalescent time of $2N_e = 20{,}000$ generations or 400,000 years, which is surprisingly close to the coalescent time in years in *E. coli.*

The corresponding WF model for the *E. coli* population has $N = 50$ million, whereas the corresponding Moran model has $N = \sqrt{2N_e} \approx 10{,}000$. For the fission model, N depends on the probability of leaving two descendants, p_2.

We note that the above calculations rest entirely on the mathematical formalism set up in Section 1.3 and the desire to equate models with each other. The three models all have different features and capture different aspects of a biological reality. Hence, it is not reasonable per se to say that a certain number of time steps in the Moran model correspond to a number of times steps in the WF model.

1.6 MUTATIONS

Under neutrality, mutations do not affect the number of offspring produced by an individual, and we can impose mutations onto the genealogy after having generated the genealogy, rather than doing it at the same time as generating the genealogy. Figure 1.5 shows the occurrence of three mutations placed at random on the branches under the WF model of reproduction. Only two of these mutations make it to the present generation. Here we assume that mutations happen at a constant rate of u per gene per time steps,

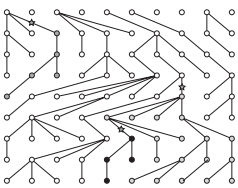


Figure 1.5 The basic coalescent with mutations (shown with stars) imposed. In this example, mutations occurred in generations 1, 4, and 6. The first mutation was lost from the population after three generations, while the third mutation is nested into the second mutation. Thus, there are three types of sequences at the present time (bottom generation)—the original plus two mutated sequences. In the example, they have population frequencies of 10%, 40%, and 50%, respectively.

irrespective of the underlying model. This corresponds to mutations arriving according to a Poisson process on individual lineages.

Since the three models are different, the rate u might be interpreted differently in the three models. In particular, in the Moran and fission models, genes mutate also outside reproduction (see Sniegowski, 2004), where this is suggested as a reasonable scenario for bacterial populations.

With the above definition, the length of the genealogy is directly proportional to the expected number of mutations in a sample; in each time step, there is probability u that the gene mutates; hence, the expected number of mutations is simply the total number of time steps (branch length) times the probability of a mutation. Consequently,

$$E(S_n) = \frac{\theta}{2} E(L_n) = \theta \sum_{k=2}^{n} \frac{1}{k-1} \approx \theta \log(n), \tag{1.11}$$

where S_n denotes the number of mutations in the history of a sample of size n and $\theta = 2N_e u$ is the scaled mutation rate. Thus, if the effective population size is doubled and the mutation rate is halved, then θ remains the same, and we are not able to estimate u and N_e separately from the sample. Equation 1.9 has the further consequence that adding further sequences from other individuals to the sample is not expected to add many more mutations to the data set because the logarithm is a slowly growing function. In contrast, the expected number of mutations increases linearly with sequence length. If mutations happen only during replication, then Equation 1.11 is true for the Moran model and the fission model with $\theta = Nu$, that is, taking the effective size to be $N/2$ in both cases.

These considerations have consequences for parameter inference. For example, for demographic inference, one should aim for longer sequences (potentially from different areas of the genome) rather than for large samples size. Doubling the sample size from 100 to 200 will only increase the expected number of mutations by 13%, whereas doubling the sequence length doubles the expected number of mutations.

A commonly reported estimator of the mutation rate is Watterson's (1975) estimator,

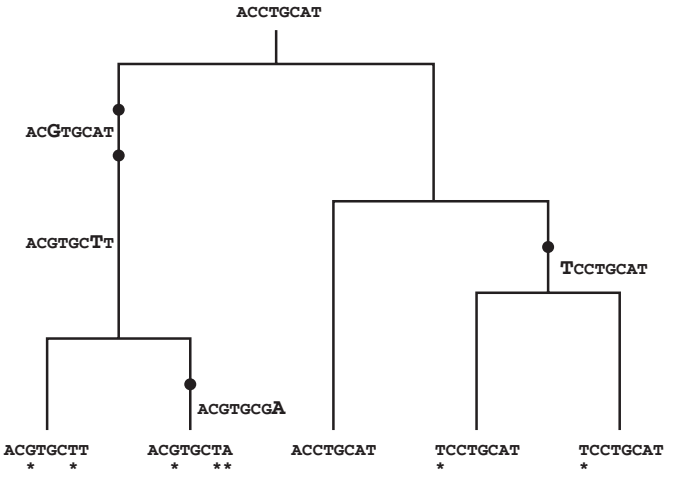
$$\hat{\theta}_W = S_n \Big/ \sum_{k=2}^{n} \frac{1}{k-1}, \tag{1.12}$$

which directly utilizes Equation 1.11 by replacing the expected number of mutations with the observed number. Another estimator, which also has found common support, is Tajima's (1989) estimator,

$$\hat{\theta}_T = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}, \tag{1.13}$$

where π_{ij} denotes the number of nucleotide differences between sequences i and j in the sample. This estimator exploits the fact that the number of mutations between a pair of sequences is expected to be θ (Eq. 1.11 with $n = 2$) and considers the average of differences among all possible pairs.

The estimators $\hat{\theta}_T$ and $\hat{\theta}_W$ put a different weight on the mutations in a genealogy. Figure 1.6 shows an example genealogy of five sequences where four mutations have occurred. Watterson's estimator puts equal weight to these mutations, whereas Tajima's estimator puts a larger weight on mutations further up in the genealogy. For instance, in the present example, a mutation carried by two sequences is counted in six comparisons, whereas a mutation carried by only one sequence is counted in four comparisons. Thus,




```
ACGTGCTT
ACGTGCTA      Sn = 4  => θW = 4/(1 + 1/2 + 1/3) = 2.3
ACCTGCAT
TCCTGCAT      θT = (1 + 2 + 3 + 3 + 3 + 4 + 4 + 1 + 1 + 0)/10 = 2.2
TCCTGCAT
```

Figure 1.6 An example data set. The effect of mutations is shown in the DNA sequences; at each mutation event (marked by a circle), a nucleotide changes. Sequences 4 and 5 are identical. Below the tree, the calculations leading to Watterson's and Tajima's estimators of the mutation rate are shown. Note that in this example, the two estimators are very similar, and we would not reject the basic coalescent model using Tajima's *D*. The asterisks indicate positions that have changed compared to the root sequence.

if a genealogy has longer inner branches than expected, Tajima's estimator will exceed Watterson's. This fact can be exploited to devise a statistical test for whether sequence data fit the basic coalescent. Tajima (1989) proposed the statistic

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{Std\left(\hat{\theta}_T - \hat{\theta}_W\right)}, \tag{1.14}$$

now commonly known as Tajima's *D*, which standardizes the difference of the two estimators (*std* denotes the standard deviation). The distribution of *D* is not known explicitly but can be evaluated by simulation. However, it is sufficiently close to a standard normal distribution, and a rule of thumb is that a Tajima's *D* value >2 or <–2 can be considered significant. This might be used to draw demographic inferences (see the next section).

1.7 DEMOGRAPHY

It is in fact very rare for a population of any species to be of constant size and to mate randomly, as is assumed in the coalescent model. Bacterial populations, for example, have the capacity to very rapidly change population size from a few cells to billions. They can go through dramatic population bottlenecks due to, for example, drugs or during shifts from one host to the next for pathogenic or commensal species. Some bacterial species confined to specific hosts are mainly transmitted from mother to offspring, and they will therefore display a type of population subdivision. Prominent examples of the latter

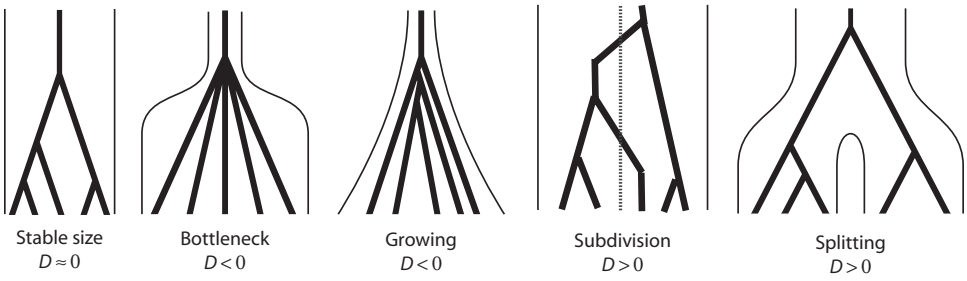


Figure 1.7 Four common demographic scenarios that create deviations from the basic reproduction model with a stable population size. Bottleneck: The rate of coalescence increases dramatically at the time of the bottleneck. Growing: The coalescent rate increases gradually back in time. Subdivision: Initial coalescent events occur preferentially within subpopulations, whereas the last coalescent events need to wait for migration between demes to occur. Splitting: Coalescence can only occur within populations until the time where the two populations merge (viewed backward in time). Subdivision and splitting lead to decreasing coalescent rates back in time, resulting in positive values of Tajima's D. In contrast, growth and bottleneck result in negative values of Tajima's D.

include studies of human migration patterns inferred from the population structure of bacterial species (Falush et al., 2003; Moodley et al., 2009). Even free-living bacterial species are not necessarily very mobile, and one would expect that bacterial cells close to each other are related by fewer cell divisions than bacterial cells far apart. This leads to many different types of population subdivision that are not reflected in the basic coalescent model (see Chapter 6 of this book). Figure 1.7 shows a cartoon of four different demographic population stratifications that deviate from the basic coalescent model.

Growing population: In a growing population, the rate of coalescence increases back in time because the chance of finding a common ancestor is larger in a small population than in a large one. Indeed, the coalescence rate is proportional to the population size. Thus, if the population size has been growing exponentially, then the coalescent rate measured in the present population size will be exponentially increasing back in time. This implies that the last coalescent events (those farthest away from the present) occur relatively faster than the first coalescent events compared with the basic coalescent. Consequently, the internal branches of the coalescent tree are comparatively shorter, which in turn implies that Tajima's D (Eq. (1.14) should be negative. Large negative values of Tajima's D have indeed been interpreted as evidence for population growth in many studies (see, e.g., Venkatesan et al., 2007).

Population bottleneck: A population bottleneck viewed back in time is a fast and dramatic decrease in population size. During the bottleneck, the coalescent rate is therefore much higher than outside the bottleneck. Therefore, the effect on Tajima's D will often resemble that of population growth. If the bottleneck lasts for a very short while, it is possible that not all ancestral lineages coalesce during the bottleneck. In that case, the coalescent genealogy would have a time interval where many coalescent events occurred at almost the same time.

Population subdivision: When bacteria occupy separated habitats, for example, distinct hosts, they can have a stable pattern of subdivision with cell division occurring within each subpopulation and occasional migration between subpopulations. This situation is modeled by equilibrium models, among these the popular n-island model. Population subdivision implies that lineages can only coalesce within

demes, so lineages from different demes will need to migrate to the same subpopulation before coalescence can occur. The number of migration events (viewed back in time) is proportional to the number of lineages, whereas the coalescent rate is proportional to the square of the number of lineages (Eq. 1.3). The consequence is that for a sample of individuals, the first coalescent events will be relatively fast because they occur between pairs of lineages in the same subpopulation. The last coalescent events often need to wait for migration events to bring together lineages in the same subpopulation, so if the migration rate is low, we expect that the last coalescent events take a comparatively long time. This implies that the resulting coalescent tree has longer internal branches than the basic coalescent tree and that Tajima's D is expected to be positive.

Population splitting (and merging): It is not possible in general to predict the effect of nonequilibrium population subdivision on the coalescent tree; hence, more subtle ways than Tajima's D are required to detect this scenario. Much progress has been made in predicting population splitting for human populations in the past using single nucleotide polymorphism (SNP) data (Li et al., 2008).

1.8 RECOMBINATION AND GENE CONVERSION

Many bacterial species are very amenable to coalescent-based analysis because they reproduce clonally. However, exchange of genetic material between cells of the same species is also prevalent in many species. This can occur in different ways (see Chapter 4 of this book), but the main effect is the same, namely, that there will no longer be a single coalescent tree describing the fate of the complete genome (or genomic region). This complicates analysis, but it is also the basis for association mapping of a phenotype of interest to a particular loci.

The effect of recombination is described in Fig. 1.8. Forwards in time, a genetic element from one individual is exchanged with the homologous element in a recipient individual. Backward in time, this has the consequence of splitting the genetic material of one individual onto two ancestral individuals, and the genealogical histories of two positions sitting close to each other but on different sides of one of the black bars in Fig. 1.8 will differ. The backward process depends on the rate of recombination and on the length of the exchanged segments (Hudson, 1983a, 1994; Wiuf, 2001; Hein et al., 2005).

Assuming an individual undergoes recombination with probability r, we find that the number of time steps until a lineage has experienced recombination has a probability distribution,

$$P\left(T_{\mathrm{Rec}}^{N_e} > t\right) = (1-t)^t. \tag{1.15}$$

Assuming as in the previous sections that time is scaled in the effective population size, then

$$P(T_{\mathrm{Rec}} > v) = \left(1 - \frac{N_e r}{N_e}\right)^{N_e v} \approx \exp(-v\rho/2), \tag{1.16}$$

where $\rho = 2N_e r$ is the scaled recombination rate (similar to the scaled mutation rate) and $T_{\mathrm{Rec}} = T_{\mathrm{Rec}}^{N_e}/N_e$. Thus, lineages wait for recombination and coalescence to occur, and the ancestral sample is modified according to whatever happens first. The total rate of recombination is $n\rho/2$, while that of coalescence is $n(n-1)/2$. This gives the following algorithm for simulating a sample history (see also Fig. 1.9).

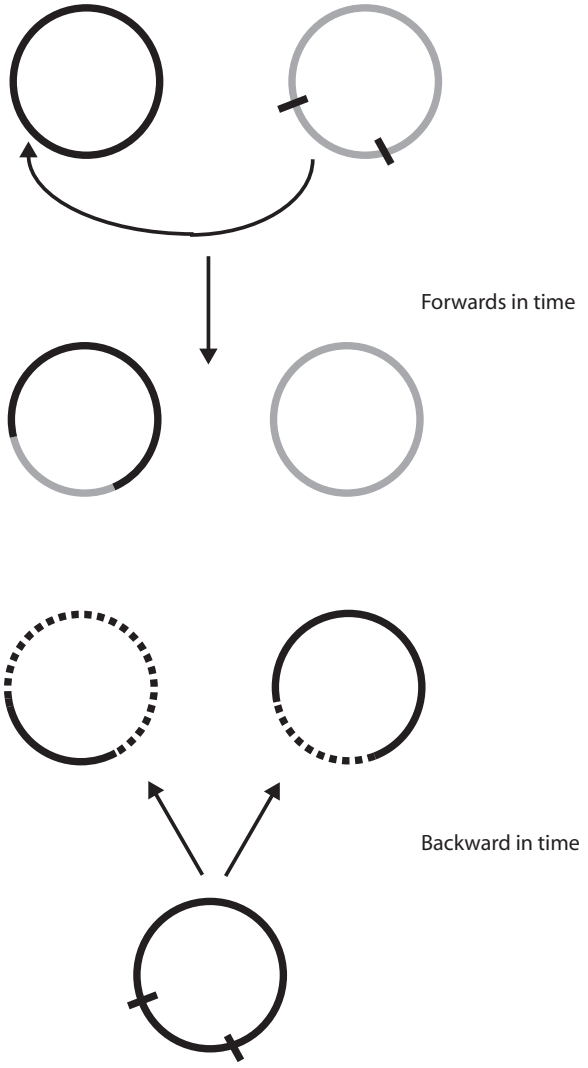


Figure 1.8 Schematic representations of bacterial recombination forwards and backward in time.

1.8.1 Algorithm 2

The algorithm is a modification of Algorithm 1. Do the following:

1. Start with $k = n$ genes.
2. Simulate an exponential variable with rate $k\rho/2 + k(k-1)/2$ (the sum of the rates for coalescence and recombination).
3. With probability $\dfrac{k\rho/2}{k(k-1)/2 + k\rho/2} = \dfrac{\rho}{k-1+\rho}$, perform a recombination event; otherwise, with probability $\dfrac{k-1}{k-1+\rho}$, perform a coalescent event
   a. If the result is a recombination event, choose a sequence at random and split it into two. This can be accomplished in different ways; for example, one could

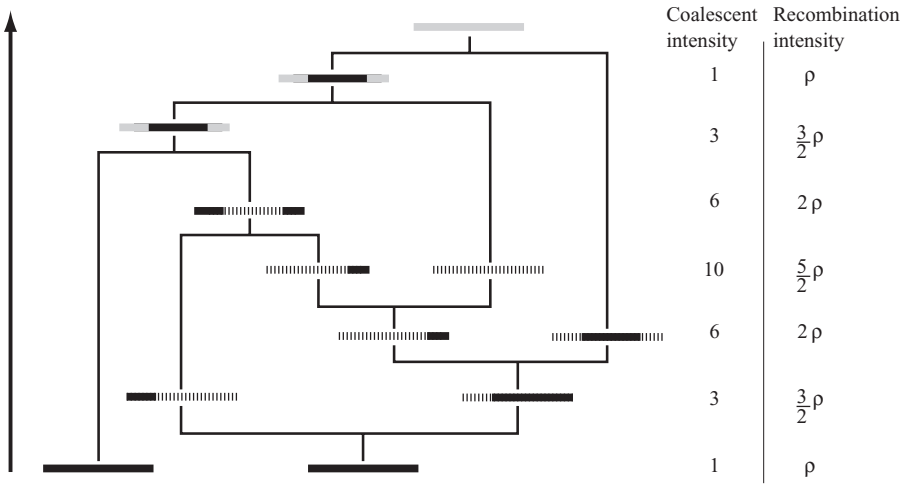


Figure 1.9 The coalescent process with recombination for a sample of two sequences. The intensities of coalescence and recombination at each time point are shown to the right, assuming that each sequence has a recombination rate of ρ/2. The third recombination event (counted from the present time) creates a sequence that is not ancestral to the present-day sample (the dotted black sequences). The solid black lines represent material ancestral to the present-day sample. Finally, the solid gray lines represent common ancestral material. Figure adapted from Hein et al. (2005).

choose two points at random, or one could choose one point at random and the other in a fixed distance from it.

b. If the result is a coalescent event, choose two sequences randomly among the k genes to coalesce.

4. If the result is a recombination event, put k equal to $k + 1$; if a coalescent event, put k equal to $k - 1$.

5. If $k > 1$, go to 1; otherwise, stop.

To get time in generations, multiply all times by N_e. During bacterial conjugation, the F factor is transferred, which can only happen once per replication cycle. In that case, it is reasonable to scale the recombination rate by $N/2$ rather than by N_e (i.e., similar to the discussion of Eq. 1.11).

This algorithm is illustrated in Fig. 1.9. A sample size of two waits for recombination and coalescence to occur. Here we only look at a small (linear) segment of the entire (circular) genome. The first two events are both recombination events and spread the ancestral material of the right sequence onto three ancestors. The third event is also a recombination event, but it creates an "empty" sequence in the sense that the recombination break point is in the part of the sequence that does not carry material ancestral to the present-day sequence. Hence, this sequence might be ignored. After the three recombination events, the first coalescent event happens, which brings together two pieces of ancestral material (see Fig. 1.9). The next event is also a coalescent event and at this event, some positions in the sample find an MRCA (shown in gray in the figure).

It is worth noticing that different positions might have different genealogies and MRCAs. Also, positions far apart might share some history; in Fig. 1.9, the leftmost and the rightmost positions share MRCA, but they do not share their entire genealogical history. Also, and in contrast to recombination in linear genomes, each recombination

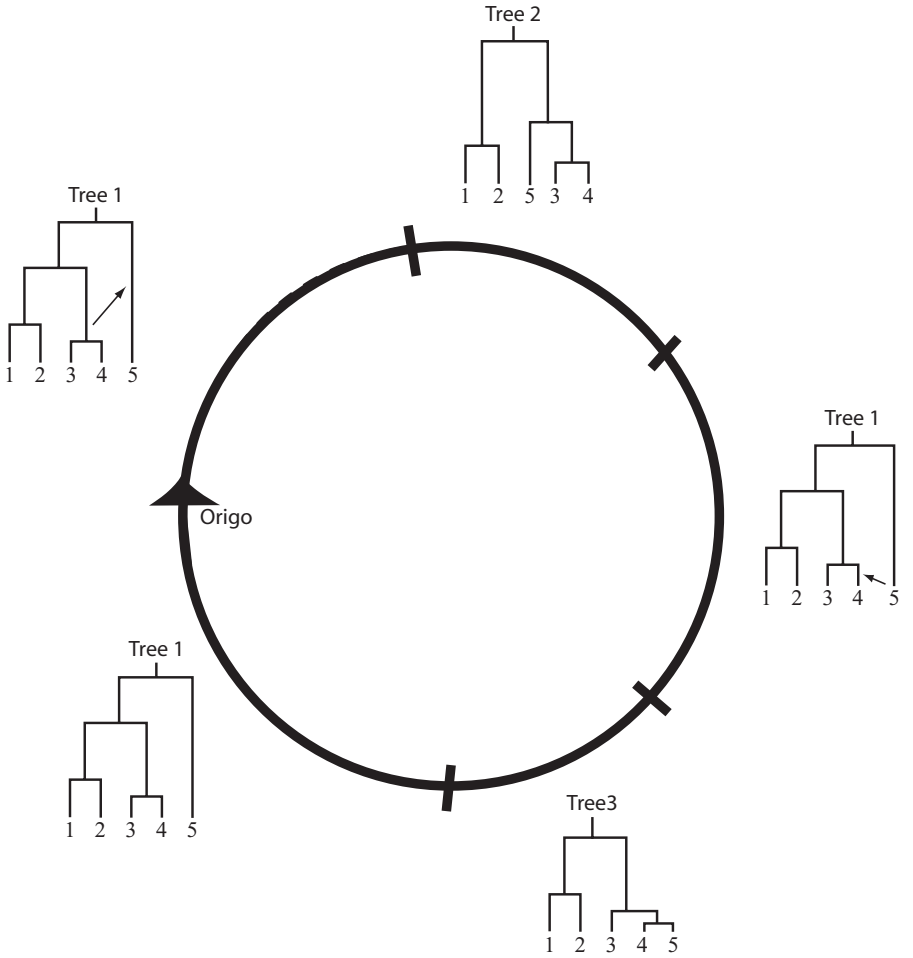


Figure 1.10 The consequences of the recombination process for a bacterial genome. Since the genome is circular, all recombination events resemble gene conversion events. Starting from the origin of replication (origo) moving in the direction of the arrow, a sample of five genes is related through coalescent tree 1. A recombination/gene conversion break point at the top results in a subtree transfer (indicated by an arrow) of the tree carrying sequences 3 and 4 to a different branch leading to coalescent tree 2. At the next break point (right break point of the gene conversion event), we return to tree 1. The next break point results in a subtree transfer of sequence 5 to the branch leading to sequence 4. This also leads to a different time of the MRCA in coalescent tree 3. At the final break point, we again return to coalescent tree 1.

event requires two break points (a beginning and an end of the segment being exchanged), and hence the recombination in circular genomes resembles gene conversion in linear genomes (Wiuf and Hein, 2000; Wiuf, 2001). Computationally, it is important to note that the coalescent with recombination is much more difficult to handle because the number of ancestral sequences might go up or go down, whereas the number of ancestral sequences in the pure coalescent process always goes down by one at each event.

Figure 1.10 illustrates further some of the consequences of a circular genome. At the origin of replication, the sample is related through a single coalescent tree. Moving away from the origin, a recombination break point is encountered in a branch. The effect of the recombination event is to move the subtree subtending the branch to a different location

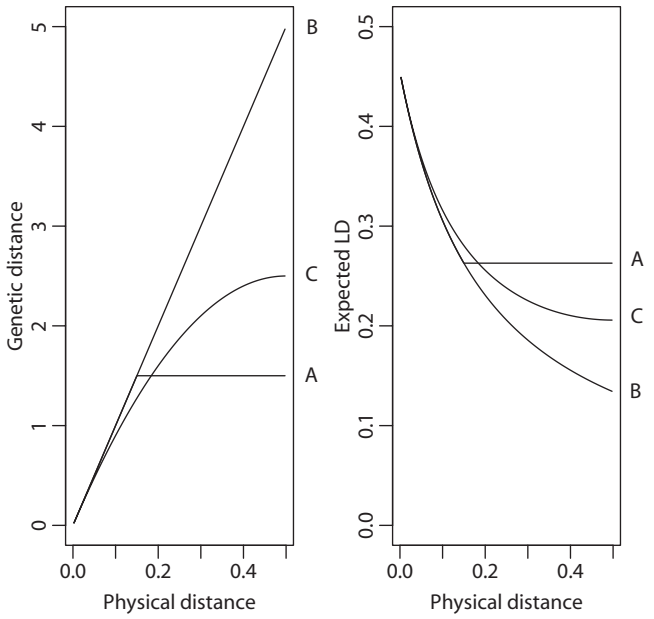


Figure 1.11 The left part of the figure shows the relationship between the physical and the genetic distances. Here the entire circular genome has length 1; hence, the maximal distance between two positions is 0.5. The genetic distance here is scaled to 10 (corresponding to $\rho/2 = 10$). Break points are chosen in the following way: (A) The first position is chosen randomly, the second at distance $L = 0.15$ away; (B) the first position is chosen randomly, the second at distance $L = 0.5$ away; (C) both positions are chosen randomly. The right figure shows the expected LD for each of the three models.

in the original tree. In Fig. 1.9, the sample size is only two and each recombination event potentially moves the common ancestor up or down (the subtree is just a single lineage). In Fig. 1.10, the subtree consisting of sequences 3 and 4 is moved to a different location, thereby creating a new tree. Moving further away from the origin, another recombination break point is encountered and we end up with the first tree again. This can likewise happen in linear genomes but is less frequent since the recombination process does not have the same similarity to gene conversion as in circular genomes.

For linear genomes, the linkage disequilibrium (LD) is expected to decay to zero over long distances because the genetic distance is roughly proportional to the physical distance. This is not the case for circular genomes. Figure 1.11 shows, for two different models, the relationship between the genetic and the physical distance, and the expected LD (measured by the quantity r^2) in a large sample,

$$E(r^2) \approx \frac{10+g}{22+13g+g^2}, \tag{17}$$

where g is the genetic distance between two positions in the genome. It is noteworthy that LD decays faster in a linear genome than in a circular genome.

1.9 SUMMARY

We have presented a basic powerful framework for modeling population variation data. In this framework, it appears that many properties are shared by apparently different reproductive models

and that the approximating coalescent process is a very robust approximation. Genetic processes, such as mutation and recombination, as well as demographic effects can easily be incorporated into the coalescent; the consequences of these additions/changes can be studied by simulation and can be compared to real data. In this chapter, we have focused on describing a variety of different models and processes and have ignored the statistical analysis of real data. For further background on statistical inference in population genetics, we refer to Balding et al. (2007).

REFERENCES

Balding, D. J., Bishop, M., and Cannings, C., eds. (2007) *Handbook of Statistical Genetics*, 3rd ed. Wiley, New York.

Charlesworth, B. (2009) Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. Advanced online publication.

Charlesworth, J. and Eyre-Walker, A. (2006) The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol* **23,** 1348–1356.

Ewens, W. (2005) *Mathematical Population Genetics*, 2nd ed. Springer, New York.

Ewens, W. J. (1972) Sampling Theory Of Selectively Neutral Alleles. *Theor Popul Biol* **3,** 87–112.

Falush, D., Wirth, T., Linz, B. et al. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* **299,** 1582–1585.

Fisher, R. (1930) *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Hein, J., Schierup, M. H., and Wiuf, C. (2005) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Hudson, R. R. (1983a) Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* **23,** 183–201.

Hudson, R. R. (1983b) Testing the constant-rate neutral allele model with protein-sequence data. *Evolution* **37,** 203–217.

Hudson, R. R. (1994) Analytical results concerning linkage disequilibrium in models with genetic-transformation and conjugation. *J Evol Biol* **7,** 535–548.

Kingman, J. F. C. (1980) *Mathematics of Genetic Diversity*. SIAM, Philadelphia, PA.

Kingman, J. F. C. (1982a) The coalescent. *Stoch Process Appl* **13,** 235–248.

Kingman, J. F. C. (1982b) *Exchangeability and the Evolution of Large Populations. Exchangeability in Probability and Statistics*, pp. 97–112. North-Holland, Amsterdam.

Kingman, J. F. C. (1982c) On the genealogy of large populations. *J Appl Probab* **19A,** 27–43.

Li, J. Z., Absher, D. M., Tang, H. et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104.

Moodley, Y., Linz, B., Yamaoka, Y. et al. (2009) The peopling of the Pacific from a bacterial perspective. *Science* **323,** 527–530.

Sniegowski, P. (2004) Evolution: Bacterial mutation in stationary phase. *Curr Biol* **14,** R245–R246.

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

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.

Venkatesan, M., Westbrook, C. J., Hauer, M. C., and Rasgon, J. L. (2007) Evidence for a population expansion in the West Nile virus vector *Culex tarsalis*. *Mol Biol Evol* **24,** 1208–1218.

Wakeley, J. (2008) *Coalescent Theory: An Introduction*. Roberts & Co., Greenwood Village.

Watterson, G. (1974) The sampling theory of selectively neutral alleles. *Adv Appl Probab* **6,** 463–488.

Watterson, G. A. (1975) Number of segregating sites in genetic models without recombination. *Theor Popul Biol* **7,** 256–276.

Wiuf, C. (2001) Recombination in human mitochondrial DNA? *Genetics* **159,** 749–756.

Wiuf, C. and Hein, J. (2000) The coalescent with gene conversion. *Genetics* **155,** 451–462.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16,** 97–159.