Chapter 26

Coalescent trees

What happens to phylogenies when we consider individual copies of genes within populations? Trees do in fact exist, but are no longer trees of individuals. Copies of genes can be related by a tree, but different loci in the same individual are related by different trees, and the trees can change even within a gene. To make it clear how these trees form, let us consider a small population, mating at random in an idealized fashion. Figure 26.1 shows the genealogy of gene copies at a single locus in a random-mating population of 10 individuals.

The genealogy that is shown in Figure 26.1 differs from ordinary genealogies in that it shows connections between gene copies, rather than between individuals. Each line goes from a gene up to a gene that is descended from it. The mating system is that of an idealized Wright-Fisher model, commonly used in theoretical evolutionary genetics to investigate the effects of genetic drift. According to that model, each gene at a locus is from a randomly chosen parent, copied from one of its two genes at random. The population is thus in effect monoecious, and selfing occasionally occurs, when the two genes in an offspring happen to be descended from the same parent. This may seem biologically unrealistic, but in evolutionary genetics the effects of other mating systems are often taken into account by computing an effective population number N_e and putting it in place of the actual population number. This has been extensively investigated and is found to work surprisingly well.

The genealogy in Figure 26.1 is the result of a computer simulation of 11 generations of descent in a Wright-Fisher model with 10 individuals. It is almost impossible to comprehend. In an effort to make it easier to look at, we can erase the circles that indicate individuals (Figure 26.2). The result is still too tangled to convey much. If we abandon any attempt to put genes from the same individual near each other, we can swap gene copies left-to-right and untangle the genealogy. The result is shown in Figure 26.3. No lines cross. The figure resembles a branching river system, with small tributaries at the top feeding ever-larger rivers that flow

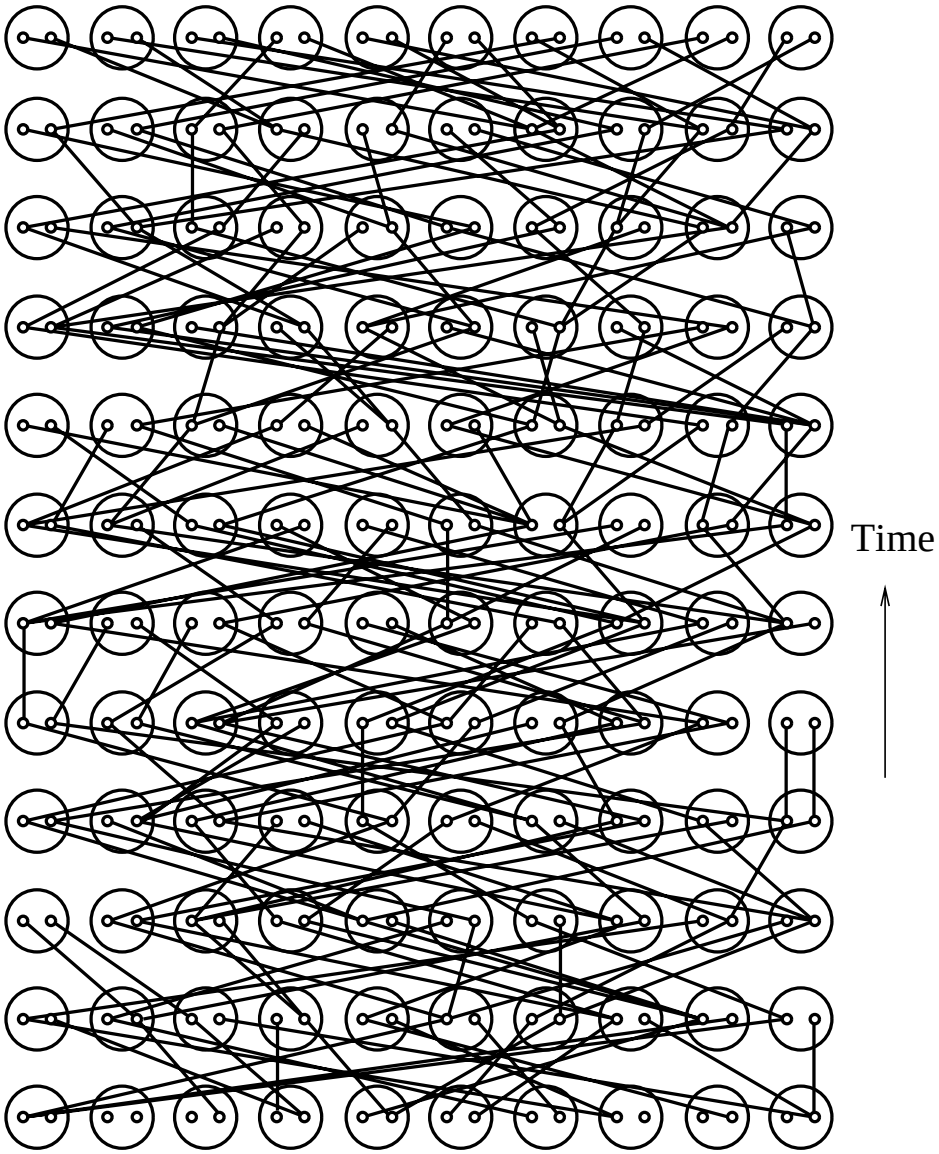

Figure 26.1: A genealogy of gene copies in a random-mating population of size 10, for 11 generations. Lines connect genes to their descendant copies in offspring. The model of reproduction is a Wright-Fisher model. Large circles are individuals, small ones are gene copies.

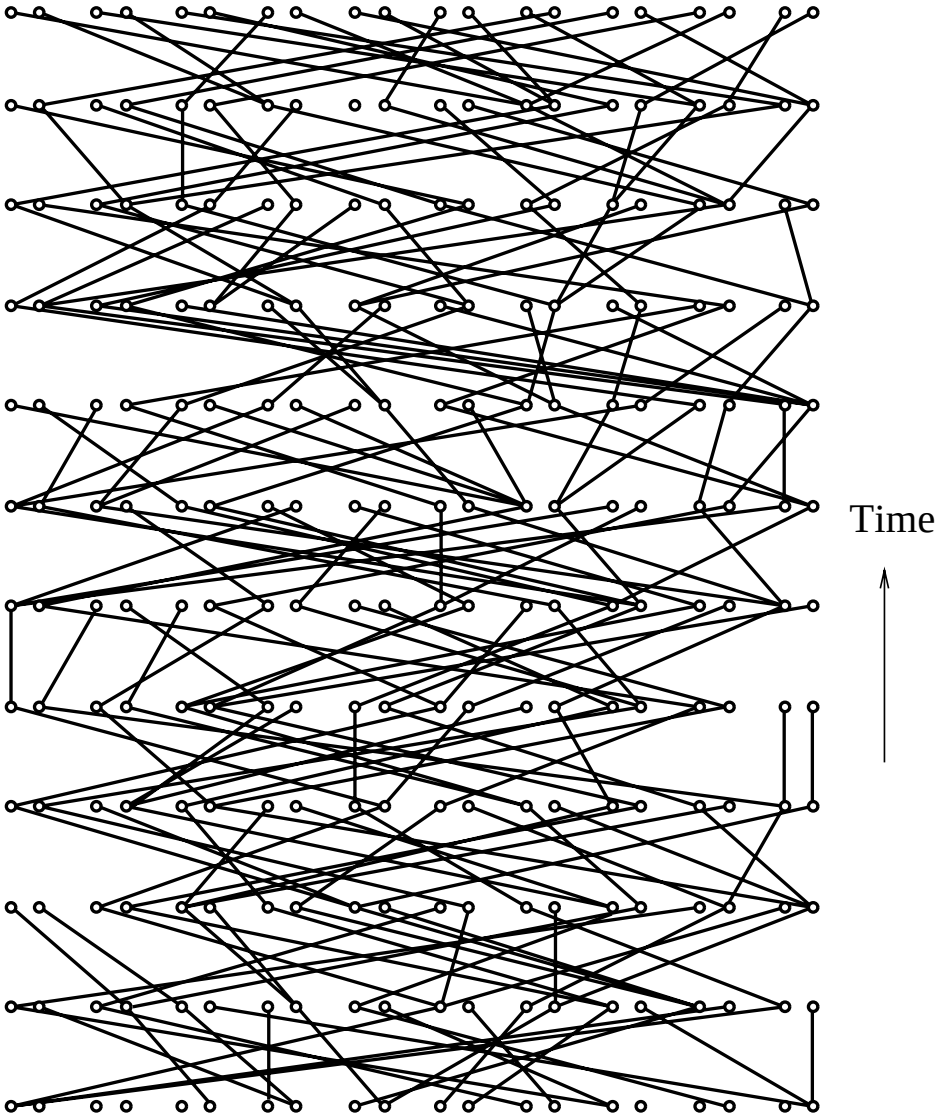

Figure 26.2: The same genealogy of genes as in Figure 26.1, with the individuals erased.

toward the bottom. In fact, of all the genes in the top generation, the first 6 are all descended from the leftmost gene in the bottom generation, and the next 14 are all descended from the gene number 10 in that generation. All other copies present in the bottom generation did not leave descendants by the time of the top generation.

Usually we are actually considering, not the entire genealogy of genes in a population, but the genealogy of a sample from the population. Figure 26.4 shows the genealogy of a particular sample of 3 copies from the current (top) generation in the genealogy of the previous figures. The members of the sample are related by a genealogical tree.

Kingman's coalescent

The structure of these trees of gene copies that form in random-mating populations was greatly illuminated by the probabilist J. F. C. Kingman (1982a, 1982b). Kingman's result is an approximation, but such a good one that few evolutionary geneticists have tried to investigate the exact structure of such trees (nor will I). Kingman's results are generalizations of a result for two copies that was obtained by the famous evolutionary geneticist Sewall Wright (1931). Wright noted that in a finite population of size N, which is monoecious and has selfing allowed, the probability that two gene copies come from the same copy in the preceding generation is $1/(2N)$. In each generation there is the same probability. The distribution of the number of generations until the two copies finally have a common ancestor is thus exactly the same as the distribution of the number of times one must toss a coin until "heads" is obtained, where the probability of "heads" is $1/(2N)$ on each toss.

That distribution is called a geometric distribution. It has mean $2N$. It is very well approximated, as Wright noted, by an exponential distribution which also has mean $2N$. Kingman's result is the extension of this result to a population with k copies of the gene. Going back in time, there will be a number of generations until two or more of these k copies have a common ancestor. Rather than following Kingman's algebra in detail, we can use a result from my own paper (1971) on genetic drift with multiple alleles. We compute the probability that none of the k alleles in the current generation came from the same copy in the preceding generation, i.e., that all of them came from distinct copies.

The first copy came from some copy in the preceding generation. The second has probability $1 - 1/(2N)$ of coming from a different one. Given that (so that two copies in the preceding generation are now represented), the chance that the third copy came from a copy different from both of these is $1 - 2/(2N)$. Given that, the fourth copy has probability $1 - 3/(2N)$ of coming from a different copy from all of these. Continuing in this fashion, the probability that all of them came from different copies is

$$G_{kk} = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right)\left(1 - \frac{3}{2N}\right)\cdots\left(1 - \frac{k-1}{2N}\right) \qquad (26.1)$$

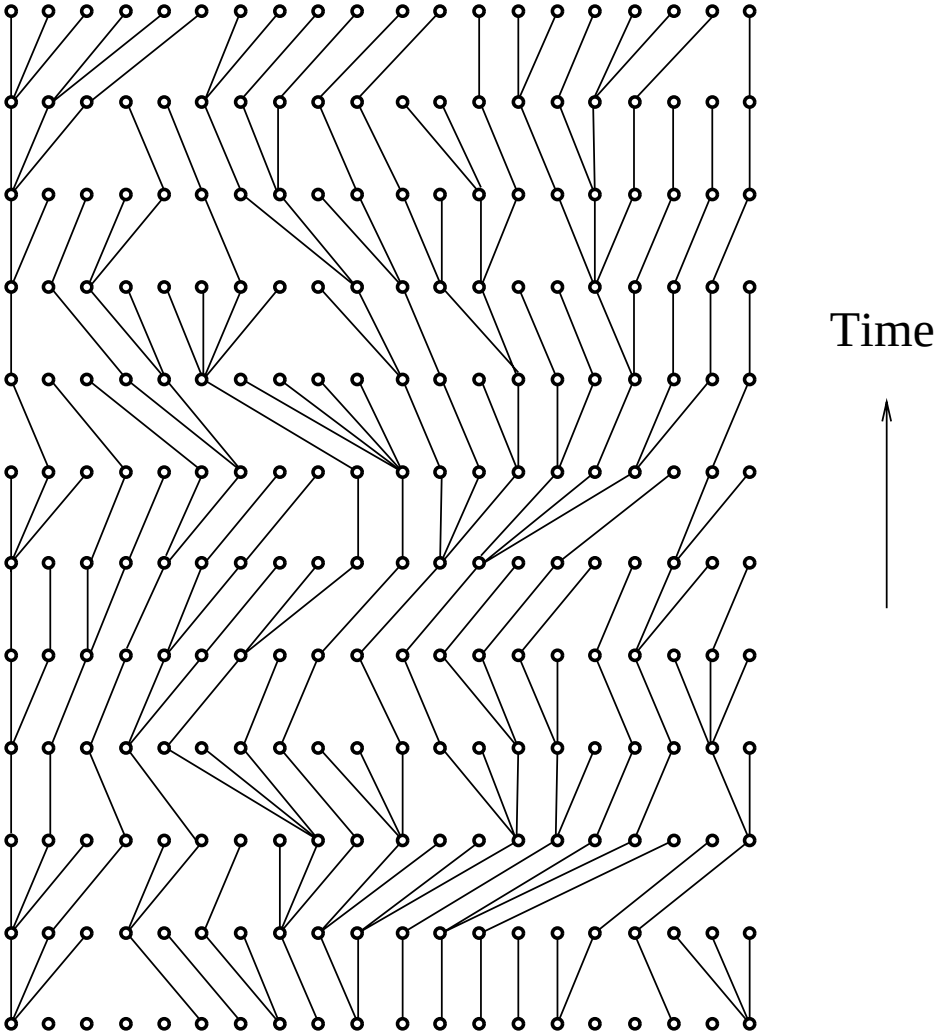

Figure 26.3: The same genealogy of genes as in Figure 26.2, with lines swapped left-to-right to untangle it, removing all crossed lines.

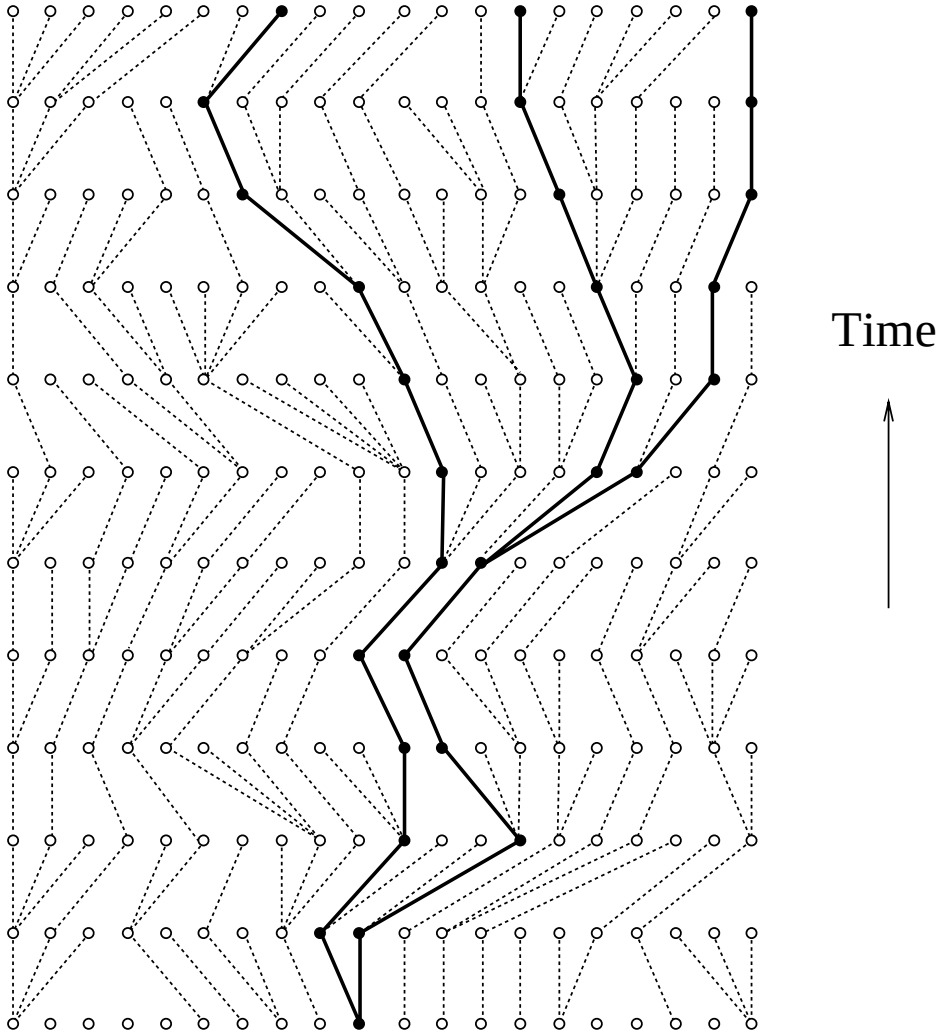

Figure 26.4: A genealogy of three gene copies sampled from the final generation of the preceding figures.

The right-hand side can be multiplied out, and yields

$$G_{kk} = 1 - (1 + 2 + 3 + \ldots + (k-1))/(2N) + \text{terms in } \frac{1}{N^2} \tag{26.2}$$

The sum of the integers from 1 to $k-1$ is well-known to be $k(k-1)/2$. Kingman's results amount to showing that ignoring the terms in $1/N^2$ is a good approximation. This will be true as long as the quantity $k(k-1)$ is much smaller than the population size N, which is usually the case. The events that are envisaged in the second term on the right-hand side of equation 26.2 are those in which precisely two of the genes are copies of the same parent gene. So Kingman's approximation in effect says that events in which three or more lineages collide are rare compared to ones in which two lineages collide.

We can then say that, to good approximation, in each generation a coin is tossed which has probability

$$1 - G_{kk} \approx \frac{k(k-1)}{4N} \tag{26.3}$$

of "heads". The number of tosses (generations) that are needed to get a "heads" is geometrically distributed, with mean being the reciprocal of the "heads" probability. Calling this time u_k we have its expectation as

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)} \tag{26.4}$$

The time is also well-approximated by an exponential distribution with the same expectation.

It should also be obvious from the process which lineages are the ones that collide – a random pair. Thus Kingman's recipe for constructing a genealogical tree of k gene copies is simply:

1. Go back a number of generations drawn from an exponential distribution with expectation $4N/(k(k-1))$.

2. Combine two randomly chosen lineages.

3. Decrease k by 1.

4. If $k = 1$, stop. Otherwise go to step 1.

The resulting stochastic process was called by Kingman the *n-coalescent*. The name has stuck (though without the n): genealogical trees of ancestry of multiple gene copies are widely known as coalescents. We should keep in mind that Kingman's coalescent is an approximation, in which it is impossible for three lineages to collide simultaneously. But as long as $k(k-1) \ll N$ it is a very good approximation.

As the number of copies grows smaller, the expectation of the time for them to coalesce grows longer. The expected total time for k copies to coalesce is readily computed. Note that $1/(k(k-1)) = 1/(k-1) - 1/k$ so that

$$\begin{aligned} &\tfrac{4N}{k(k-1)} + \tfrac{4N}{(k-1)(k-2)} + \tfrac{4N}{(k-2)(k-3)} + \ldots + \tfrac{4N}{2} \\ &= 4N\left(\tfrac{1}{k-1} - \tfrac{1}{k} + \tfrac{1}{k-2} - \tfrac{1}{k-1} + \tfrac{1}{k-3} - \tfrac{1}{k-2} + \ldots + \tfrac{1}{1} - \tfrac{1}{2}\right) \qquad (26.5) \\ &= 4N\left(1 - \tfrac{1}{k}\right). \end{aligned}$$

The results are a bit surprising. When there are many copies it takes on average about $4N$ generations for all of their ancestral lineages to coalesce! But when there are two copies, it takes on average $2N$ generations. That implies that a bit more than half of the depth of a coalescent tree is spent waiting for the last two copies to coalesce. $(1-1/n)/(1-1/k)$ of the time is spent waiting for the last n copies to coalesce. So with 100 copies in all, 0.9 / 0.99 or 0.90909 of the time is spent waiting for the last 10 copies to coalesce. Only 9% of the time is spent on the first 90 coalescent events! (Of course, we mean "first" in the sense of going backwards in time). One gets the picture that lineages coalesce rather rapidly at first and then the process gradually slows down.

These figures are based on expectations, and as expectations of ratios are not quite the same things as ratios of expectations, they may be a bit off, but are a reliable guide to what coalescent trees look like.

One might also ask how unbalanced these random trees of lineages are. Farris (1976) and Slowinski and Guyer (1989) considered that the basal split of a random tree with k tips could have any number lineages from 1 through $k-1$ on the left-hand side. They have shown that all $k-1$ of these values are in fact equiprobable. This also gives us useful information about the effect of adding one lineage to a tree. If we have 100 lineages and add one lineage, what is the probability that it will connect to this tree below the pre-existing root? Their result shows that the probability is only 2 / 100 that the root of the 101-lineage tree separates one lineage from the rest. And even if it does, the chance is only 1 / 101 that this single lineage is the new one that we added. Thus the chance that the new lineage establishes a new root below the pre-existing one is only $2/(101 \times 100) = 0.000198$.

Figure 26.5 shows 9 realizations of a coalescent with 20 gene copies, all drawn to the same scale. This will show both the pattern of increasing lengths of time for coalescence to occur as the number of lineages decreases, and the enormous variability around that implied by the exponential distributions involved. It also shows a reasonable agreement with the Farris-Slowinski-Guyer uniform distribution of numbers of lineages on each side of the bottom split.

Figure 26.5 shows the tendency for the first few lineages to have in their ancestry the long lines at the bottom of the tree. It shows a sample of 50 gene copies. The ancestry of a random subsample of 10 of them is indicated by making the lines

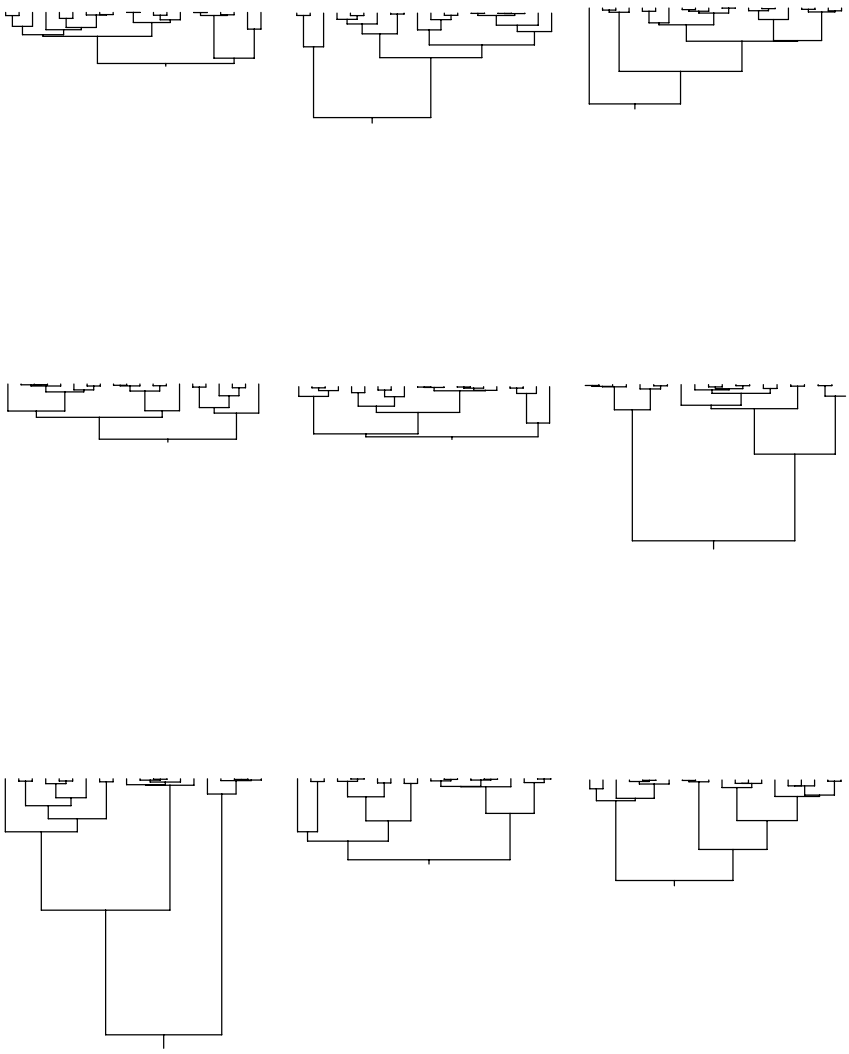

Figure 26.5: Nine outcomes of the coalescent process with 20 gene copies, drawn to the same scale.

bolder. Adding the 40 to the 10 actually adds no new lines to the bottom part of the tree: they tend to connect to the existing lines by short branches and rarely add much length to the tree.

Bugs in a box – an analogy

We can make a physical analogy (if a somewhat fanciful one) by considering a box full of hyperactive, indiscriminate, voracious and insatiable bugs. We put k bugs into the box. They run about without paying any attention to where they are going. Occasionally two bugs collide. When they do, one instantly eats the other. Being insatiable, it then resumes running as quickly as before. It is obvious what will happen. The number of bugs in the box gradually falls from k to $k-1$, to $k-2$, as the bugs coalesce, until finally only one bug is left.

The analogy is actually fairly precise. The number of pairs of bugs that can collide is $k(k-1)/2$. If there are $2N$ "places" in the box that can be occupied, the probability of a collision will be proportional to $k(k-1)/4N$. The size of the population corresponds to the size of the box. A box with twice as many "places" will slow the coalescence process down by a factor of two. So a simpleminded physical analysis of the bugs-in-a-box process will have the Kingman coalescent distribution as the probability distribution of its outcomes.

Effect of varying population size

We have been assuming that effective population size does not change through time. In reality it will, and we will also want to make inferences about its changes. Working backwards in time, when we get (back) to the point where the effective population size is $N(t)$, we will find there that the instantaneous rate of coalescence of k lineages is $k(k-1)/4N(t)$. If population size $N(t)$ is half the value that it has now, these coalescences will happen twice as fast as they do now. The effect is to make it appear that time is passing twice as fast. This suggests a simple time transformation that allows us to find the distribution of coalescence times in the case where the effective population size is $N(t)$ at time t ago.

Suppose that we imagine a time scale where time passes at a rate proportional to $N(0)/N(t)$, where $N(0)$ is the effective population size now. Let us call this fictional time scale τ, where

$$d\tau = \frac{N(0)}{N(t)}\, dt. \tag{26.6}$$

The total amount of this fictional time that elapses going back from the present to time t ago will then be the integral

$$\tau = \int d\tau = \int \frac{N(0)}{N(t)}\, dt. \tag{26.7}$$

Whatever the course of population size change, as long as its inverse can be integrated, we can use equation 26.7 to derive the formula for the fictional time.

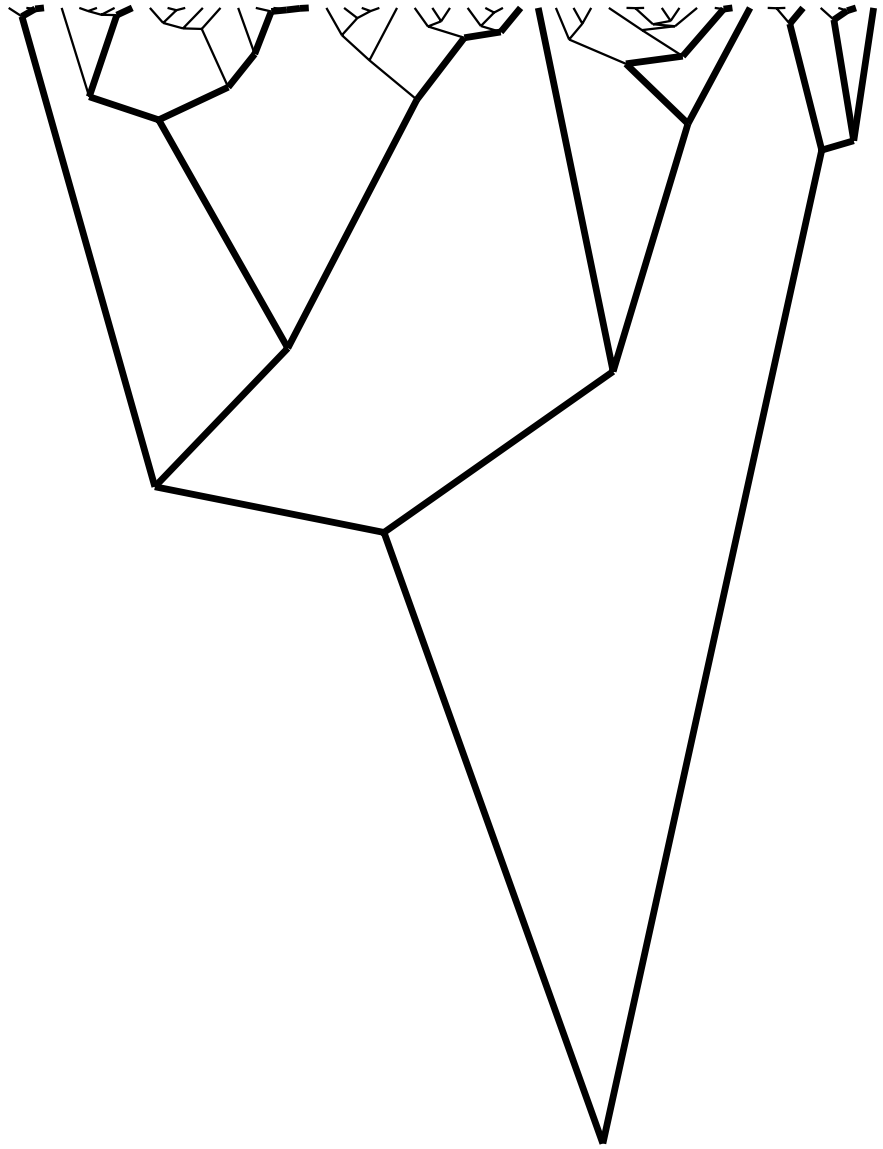

Figure 26.6: A sample genealogy of 50 gene copies, with the ancestry of a random 10 of them indicated by bold lines. Note that adding 40 more gene copies to the sample discloses no new lines in the bottom part of the diagram.