Chapter 2
Mutation and Random Genetic Drift

2.1 The Wright–Fisher Model and the Kingman Coalescent

Evolution is a random process. Random events enter in many ways, from errors in copying genetic material to small and large scale environmental changes, but the most basic source of randomness that we must understand is due to reproduction in a finite population leading to *random genetic drift*. The simplest model of random genetic drift was developed independently by Sewall Wright and R.A. Fisher and is known as the Wright–Fisher model. We consider a population in which every individual is equally likely to mate with every other and in which all individuals experience the same conditions. Such a population is called *panmictic*. We also suppose that the population is *neutral* (everyone has an equal chance of reproductive success). Most species are either *haploid* meaning that they have a single copy of each chromosome (for example, most bacteria), or *diploid* meaning that they have two copies of each chromosome (for example, humans). We suppose that the population is haploid, so that each individual has exactly one parent. Although in a diploid population individuals have two parents, each *gene* can be traced to a single parental gene in the previous generation and so it is customary in this setting to model the genes in a diploid population of size N as a haploid population of size $2N$.[1] As we shall see in Sect. 5.6, this device fails once we are interested in tracing several genes at the same time.

Definition 2.1 (The neutral Wright–Fisher model). The neutral Wright–Fisher model for a panmictic, haploid population of constant size N is described as follows. The population of N genes evolves in discrete generations. Generation $(t+1)$ is formed from generation t by choosing N genes uniformly at random with replacement. That is, each gene in generation $(t+1)$ chooses its parent independently at random from those present in generation t.

[1] In fact we are assuming that the population is hermaphrodite here – so there are no separate sexes – and we are allowing a small chance of self-fertilisation. For a population subdivided into N_m males and N_f females we can still use the same model, but with an *effective* population size $4N_mN_f/(N_m+N_f)$ replacing N, see Example 2.9.

A. Etheridge, *Some Mathematical Models from Population Genetics*, Lecture Notes in Mathematics 2012, DOI 10.1007/978-3-642-16632-7_2,

From this definition it is an elementary matter to work out the genealogical trees that relate individuals in a sample from the population. Suppose first that we take a sample of size two. The probability that these two individuals share a common parent in the previous generation is $1/N$. If they do not, then the probability that their parents had a common parent is $1/N$, and so on. In other words, the time to the most recent common ancestor (MRCA) of the two individuals in the sample has a geometric distribution with success probability $1/N$. (The probability that their most recent common ancestor was T generations in the past is pq^{T-1} where $p = 1/N$ and $q = 1 - p$.) In particular, the expected number of generations back to their MRCA is N. Now typically we are interested in large populations, where our rather crude models have some hope of having something meaningful to say. Then it makes sense to measure time in units of size N and in those units the time to the MRCA of a sample of size two is approximately exponentially distributed with parameter one. More generally, consider a sample of size $k \geq 2$. The probability of three (or more) individuals from the sample sharing a common parent is $\mathscr{O}(1/N^2)$ and similarly the chance that two separate pairs of individuals are 'siblings' is $\mathscr{O}(1/N^2)$. This means that the time we must wait before we see such an event is $\mathscr{O}(N^2)$ generations. But before this happens (with probability tending to one as $N \to \infty$) all our ancestral lineages will have coalesced through pairwise coalescence events (each of which occurs within $\mathscr{O}(N)$ generations). Thus the time (in units of size N) before the present at which we first see a 'merger' of lineages ancestral to our sample is approximately exponentially distributed with rate $\binom{k}{2}$ and, when that merger takes place, it affects exactly two lineages chosen uniformly at random from the $\binom{k}{2}$ pairs available. After that we just trace the remaining $\binom{k-1}{2}$ pairs of lineages and the same picture holds.

Remark 2.2. Since we are dealing with a haploid population, each individual has only one parent and the genealogical trees get *smaller* as we go backwards in time, in contrast to our usual understanding of family trees (for a diploid population) which grow as we trace backwards in time. We'll return to this point in Sect. 2.7.

We shall loosely refer to the system of coalescing lineages that we have just described as Kingman's coalescent, but let us give a more formal definition. If we label individuals in our sample $\{1,2,\dots,k\}$, then our process of coalescing lineages defines a continuous time Markov process, $\{\pi_t\}_{t\geq 0}$, on the equivalence relations on $[k] = \{1,2,\dots,k\}$. Each equivalence class of π_t corresponds to an ancestor alive at time t before the present. It consists of the labels of all individuals in our sample descended from that ancestor.

Definition 2.3 (Kingman coalescent). A k-coalescent is a continuous time Markov chain on $\mathscr{E}_k$, the space of equivalence relations on $[k]$, with transition rates $q_{\xi,\eta}$ $(\xi, \eta \in \mathscr{E}_k)$ given by

$$q_{\xi,\eta} = \begin{cases} 1 \text{ if } \eta \text{ is obtained by coalescing two of the equivalence classes of } \xi, \\ 0 \text{ otherwise.} \end{cases}$$

The *Kingman coalescent* on $\mathbb{N}$ is a process of equivalence relations on $\mathbb{N}$ with the property that, for each k, its restriction to $[k]$ is a k-coalescent. By convention, we take the initial condition to be the trivial partition into singletons.

Remark 2.4 (Consistency). If we take a $(k+l)$-coalescent and restrict it to $[k]$, then we obtain a k-coalescent. In particular, if we take a sample of size $k+l$ from the population and restrict the genealogical trees relating the full sample to a randomly chosen subsample of size k, then we arrive at the same trees as if we had just taken a smaller sample in the first place. This sampling consistency is an essential part of the interpretation of the model.

Existence of the k-coalescent is clear (it is a finite state space Markov chain with bounded rates). The consistency allowed Kingman (1982) to pass to a projective limit.

Remark 2.5 (Terminology). In what follows we shall sometimes say that the genealogy of a sample (or population) of size k is determined by the Kingman coalescent. By this we mean that it is given by a k-coalescent.

To obtain the Kingman coalescent, we measured time in units of population size N and passed to an infinite population limit. Now let's examine what happens when we let $N \to \infty$ in our Wright–Fisher model. Suppose that the gene in question has two alleles which we label a and A say. For now we suppose that an offspring inherits the allelic type of its parent. We try to characterise the process, $\{p_t\}_{t\geq 0}$, which records the proportion of a-alleles in the population at each time $t \geq 0$. Notice that in the prelimiting model, $\{p_t\}_{t\geq 0}$ is a discrete time Markov chain on a finite state space with traps at 0 and 1.

Definition 2.6 (Fixation). If the proportion of one of the alleles in the population is one, then we say that the allele has *fixed*. The probability that a becomes fixed is its *fixation probability*.

To characterise the distribution of $\{p_t\}_{t\geq 0}$, we consider how $\mathbb{E}[u(p_t)]$ changes with time for sufficiently nice functions $u : [0,1] \to \mathbb{R}$. In the rescaling that we took to obtain the Kingman coalescent, the model evolves at time intervals of length $1/N$. Evidently, if a proportion p of the population is of type a in the current generation, then the expected *number* of type a individuals in the next generation is Np and the variance of that number is Npq (where $q = 1-p$). Thus the mean allele frequency remains the same and the variance is pq/N. Moreover $\mathbb{E}[\,(p_{1/N}-p)^k \big| \, p_0 = p] = \mathscr{O}(1/N^2)$ for all $k \geq 3$. Now the evolution of the process is homogeneous in time, so it is enough to consider what happens close to time zero. Using Taylor's Theorem, we obtain

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\left[u(p_t)\big|\, p_0 = p\right]\bigg|_{t=0} &\approx \frac{\left\{\mathbb{E}\left[u(p_{1/N})\big|\, p_0 = p\right] - u(p)\right\}}{1/N} \\
&= N\Bigg\{u'(p)\mathbb{E}[\,(p_{1/N}-p)\big|\, p_0 = p] \\
&\qquad + \frac{1}{2}u''(p)\mathbb{E}[\,(p_{1/N}-p)^2\big|\, p_0 = p] + \mathscr{O}\left(\frac{1}{N^2}\right)\Bigg\} \\
&= \frac{1}{2}p(1-p)u''(p) + \mathscr{O}\left(\frac{1}{N}\right).
\end{aligned}$$

Thus, in the limit as $N \to \infty$, if the process of allele frequencies converges to a well-defined stochastic process, then we expect that

$$\frac{d}{dt}\mathbb{E}\left[u(p_t)|\,p_0 = p\right]\Big|_{t=0} = \frac{1}{2}p(1-p)u''(p). \tag{2.1}$$

That is, we expect that in the limit, the distribution of the allele frequencies is governed by the solution to the Wright–Fisher stochastic differential equation:

$$dp_t = \sqrt{p_t(1-p_t)}dW_t, \tag{2.2}$$

where $\{W_t\}_{t\geq 0}$ is a standard Brownian motion.

What we have *shown* is that, at least for large populations evolving according to the neutral Wright–Fisher model, if we measure time in units of N generations, then the distribution of allele frequencies should be approximately governed by the partial differential equation (2.1), and the genealogy of a sample from the population should be well-approximated by the Kingman coalescent. Notice that it is the random genetic drift, that is the random change in allele frequencies caused by the random variation in individual reproduction, that causes coalescence of ancestral lineages as we trace backwards in time.

In reality, a variety of factors affect the rate of genetic drift and these are often summarised by using an *effective* population size.

Definition 2.7 (Effective population size). The effective population size N_e of a population is the size of the Wright–Fisher population that would give the same rate of random drift.

Remark 2.8. In fact this definition is incomplete as there are several ways to define the rate of genetic drift and they do not necessarily yield the same expression for the effective population size. For the Wright–Fisher model for a population of size N, we have the following three properties:

1. The maximum nonunit eigenvalue of the transition matrix is $1-1/N$.
2. The probability that two genes taken at random are descendants of the same parent is $1/N$.
3. Writing $p(t)$ for the proportion of a-alleles in generation t and $\mathtt{var}(p(t))$ for the corresponding variance, given $p(t)$, $\mathtt{var}(p(t+1)) = p(t)(1-p(t))/N$.

One can try to find an N_e corresponding to any of these properties, and this leads to *eigenvalue* effective population size, *inbreeding* effective population size and *variance* effective population size. Ewens (1982) discusses this in more detail. Nordborg and Krone (2002) define the *coalescent* effective size as the amount by which time must be rescaled in order to recover the Kingman coalescent as the genealogy in the limit as population size tends to infinity. Such an effective size may not exist, but there are strong arguments for not defining an effective population size in settings where one cannot (asymptotically) reduce to Kingman's coalescent. This is discussed further in Sjödin et al. (2005).

For a diploid population, modelled as a haploid population of size $2N$, the corresponding quantity will be $2N_e$.

Example 2.9 (Populations that are subdivided into males and females). Suppose that a diploid population is subdivided into N_m males and N_f females, then

$$N_e = \frac{4N_mN_f}{N_m + N_f}. \tag{2.3}$$

To see why, take a sample of two genes from the current generation. Each sits in a diploid individual and has probability $1/2$ of being inherited from the father of that individual and $1/2$ of being inherited from the mother.[2] If they are both inherited from fathers, which happens with probability $1/4$, then they have probability $1/2N_m$ of being descended from the same gene, and similarly, if both are inherited from a female they came from the same parental gene with probability $1/2N_f$. Thus the chance of coalescence in the previous generation is

$$\frac{1}{4}\frac{1}{2N_m} + \frac{1}{4}\frac{1}{2N_f} = \frac{N_m + N_f}{8N_mN_f} = \frac{1}{2N_e}$$

with N_e given by (2.3).

What we have derived here is the inbreeding effective population size, but the methods of Sect. 6.3 can be used to show that in this example this corresponds to the coalescent effective population size (see Nordborg and Krone (2002) for more details.) □

So how does (2.2) do as a model? Of course it is too simplistic to apply to most naturally occurring populations, but we can compare it to experimental data. Buri (1956) reports an experiment on populations of *Drosophila melanogaster*. Just over one hundred populations were propagated, each from eight males and eight females. The experiment measures the frequency of an allele of a gene that slightly alters the eye colour (without affecting fitness or reproductive success of the carrier). We'll denote it by a. He reports the change in the *variance* in allele frequency across the different populations with time. All populations are started with exactly half a and half A (which in this context just means 'not a') alleles. The variance starts at zero (all populations have the same frequencies) and then grows because of the random genetic drift until it reaches a maximum when each population consists either entirely of a-alleles or entirely of A-alleles.

We write v_t for the variance in allele frequency across populations at time t in our rescaled time units, $v_t = \mathbb{E}[p_t^2] - \mathbb{E}[p_t]^2$. Using (2.1) and the Markov property of $\{p(t)\}_{t\geq 0}$ we have that

$$\frac{d}{dt}\mathbb{E}[p_t] = 0, \quad \frac{d}{dt}\mathbb{E}[p_t^2] = \mathbb{E}[p_t(1-p_t)] \quad \text{and} \quad \frac{d}{dt}\mathbb{E}[p_t(1-p_t)] = -\mathbb{E}[p_t(1-p_t)].$$

[2] We are ignoring the possibility that we have sampled two distinct genes from the *same* individual. If this happens, then in the previous generation the ancestral lineages were necessarily in different individuals (the mother and father) and so correcting for this makes negligible difference to the inbreeding effective population size.

Combining these gives that $v_t \approx p_0(1-p_0)(1-\exp(-t))$. Writing V_t for the variance after t generations (in other words changing back to 'real' time units) this becomes

$$V_t \approx p_0(1-p_0)(1-\exp(-t/2N)).$$

The $2N$ is because Drosophila are diploid and in this case $N = 16$.[3] The theoretical prediction for the rate of increase in the variance turns out to be not very accurate, but it becomes good when instead of substituting the actual population size, one substitutes a smaller, effective, population size. Buri reports a best fit of $N_e = 11.5$. Buri's data and the theoretical predictions for $N_e = 16$ and $N_e = 11.5$ are plotted in the graph in Fig. 2.1.

Remark 2.10 (Large populations). The population size $N = 16$ does not perhaps seem particularly large. However, calculating directly with the Wright–Fisher model gives a variance after t generations of

$$p_0(1-p_0)\left(1-\frac{1}{N}\right)^t.$$

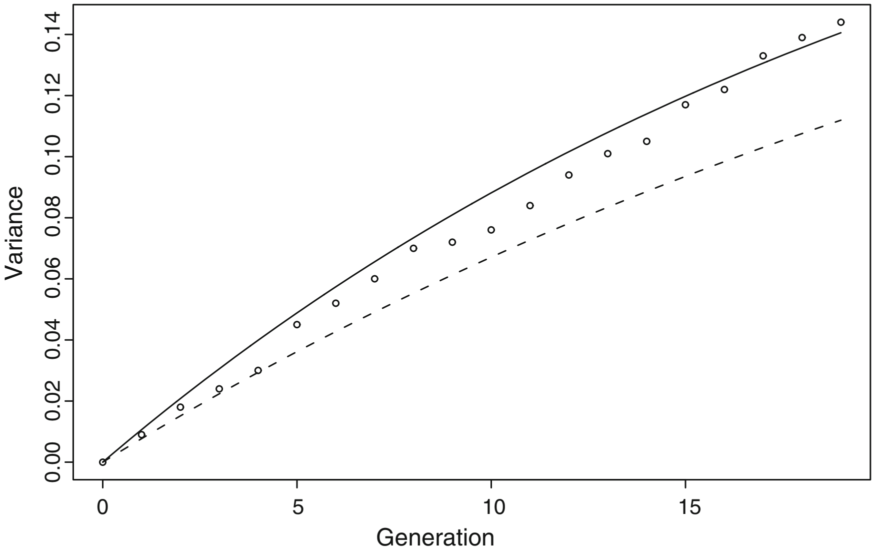

Fig. 2.1 Testing Wright's model of genetic drift. The graph shows experimental results of Buri (1956) based on just over 100 populations of *Drosophila melanogaster*, each propagated from 8 males and 8 females. Variance in allele frequency is plotted against time (in generations). Circles are data points, the dotted line is the theoretical prediction for $N_e = 16$ and the solid line is the theoretical prediction with $N_e = 11.5$

[3] Although the population was subdivided into males and females, the experiment maintained equal numbers of males and females so that the effective population size $N_e = 4N_m N_f/(N_m + N_f) = N$ (see Example 2.9).

Hence at the end of the experiment (after 19 generations) we are using $\exp(-19/32)$ as an approximation to $(1-1/32)^{19}$, giving a relative error of less that 1%.

But why did we need to use an effective population size here? At first sight Buri's populations appear to satisfy the assumptions of the Wright–Fisher model: they are panmictic and constant size, generation times are discrete and the allele under consideration does not affect fitness. In fact it is the Wright–Fisher reproduction mechanism itself that is at fault. It forces the variance of the offspring of a single individual to be $(1-1/N)$, but this does not reflect the true offspring distribution in the population. To see how offspring variance feeds into the effective population size we must consider a slightly more general model.

2.2 The Cannings Model

First a definition.

Definition 2.11 (Exchangeable random vector). A random vector $(\nu_1,\dots,\nu_N)$ is said to be *exchangeable* if its law is invariant under any permutation of the coordinates. That is,

$$(\nu_1,\dots,\nu_N) \overset{d}{=} (\nu_{\pi(1)},\dots,\nu_{\pi(N)})$$

for any permutation $\pi = (\pi(1),\dots,\pi(N))$ of $\{1,\dots,N\}$.

Definition 2.12 (Neutral Cannings Model). Consider a panmictic, haploid population of constant size N. Labelling the individuals in generation t by $\{1,\dots,N\}$, in a neutral *Cannings model*, generation $t+1$ is determined by an exchangeable random vector $(\nu_1(t),\dots,\nu_N(t))$ with $\sum_{k=1}^{N}\nu_k(t) = N$. Here, $\nu_k(t)$ denotes the number of children of the kth individual and the vectors $\{(\nu_1(t),\dots,\nu_N(t))\}_{t\in\mathbb{N}}$ are assumed to be independent and identically distributed.

Notice that, mathematically, neutrality is captured by exchangeability.

The Wright–Fisher model is the special case of the Cannings model in which $(\nu_1(t),\dots,\nu_N(t))$ has the multinomial distribution with N trials and equal weights.

Let's examine the genealogy of a sample from a large population evolving according to a more general Cannings model. Let c_N denote the probability that two individuals chosen at random from some generation have a common parent in the previous generation. Then (dropping the argument t)

$$c_N = \frac{\mathbb{E}[\nu_1(\nu_1-1)]}{N-1}.$$

To see this, condition on the vector $(\nu_1,\nu_2,\dots,\nu_N)$ that determines the division of offspring into families. The chance that two offspring (sampled at random and *without* replacement) both fall among the ν_1 individuals that make up the first family is just $\nu_1(\nu_1-1)/N(N-1)$. Now average over the distribution of the vector $(\nu_1,\nu_2,\dots,\nu_N)$. This gives the probability that both offspring are in the first family. Using exchangeability, the probability that they both belong to the same family (but any one of the N available) is just N times this probability, that is $\mathbb{E}[\nu_1(\nu_1-1)/(N-1)]$ as required. (For the Wright–Fisher model, $c_N = 1/N$.) The time until the MRCA of a random sample of size two from the population will be geometric with success probability c_N. This will determine the right time scaling to get convergence to a nontrivial limit as $N \to \infty$. We are going to assume that $c_N \to 0$ as $N \to \infty$. Now consider a sample of size three. The chance that they *all* have a common parent is

$$\frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{(N-1)(N-2)}.$$

Thus, if we measure time in units of $1/c_N$, provided that

$$\frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{N^2 c_N} \to 0 \qquad \text{as } N \to \infty, \tag{2.4}$$

in the limit as $N \to \infty$ we will only ever see pairwise mergers. In fact it turns out Möhle (2000) that the condition in (2.4) guarantees both that $c_N \to 0$ and that

$$\frac{\mathbb{E}\left[\nu_1(\nu-1)\nu_2(\nu_2-1)\right]}{N^2 c_N} \to 0 \quad \text{as } N \to \infty,$$

so that, measuring time in units of $1/c_N$, asymptotically we will not see *simultaneous* mergers of two different pairs of ancestral lineages. In the limit as $N \to \infty$ we recover Kingman's coalescent.

Lemma 2.13. *If we sample k individuals from a population evolving according to the neutral Cannings model of Definition 2.12 and if the condition (2.4) is satisfied, then for large N, when measured in time units of* $1/c_N$ *generations, the genealogy of the sample is approximately a k-coalescent.*

Similar calculations to those that we did for the Wright–Fisher model show that, again measuring time in units of $1/c_N$ generations and under assumption (2.4), the distribution of allele frequencies for a sufficiently large population evolving according to the Cannings model will be governed (approximately) by the partial differential equation (2.1). The only difference from the Wright–Fisher setting is that now when we wish to compare to data we must remember that c_N is (approximately) $\mathtt{var}(\nu_1)/N$, where `var` denotes variance. In our previous language, the *effective population size* is $N_e = N/\mathtt{var}(\nu_1)$. In particular, the greater the variance in offspring number, the smaller the effective population size and the faster the rate of random drift.

Remark 2.14 (Robustness of Kingman's coalescent). In passing to an infinite population limit, we aim to find an approximation that reflects the key features of our population (in this case that it is neutral, panmictic and of constant size), but which is insensitive to the fine details of the prelimiting model. As we can already see, the Kingman coalescent approximates a wide variety of local structures and it is this robustness that makes it such a powerful tool. Forwards in time we have taken a *diffusion approximation*, approximating the Wright–Fisher model by a Wright–Fisher diffusion. The importance of diffusion approximations in population genetics can be traced to the seminal work of Feller (1951).

2.3 Selfing

In footnote 1 we remarked that in considering a haploid population of size $2N$ in place of a diploid population of size N, since each individual samples its two parents independently *with* replacement, we are allowing a small probability of *self*-fertilisation. For that model, the probability of self-fertilisation is very small (for large populations), but for many plant populations a significant proportion of offspring are produced through self-fertilisation, or *selfing*. What effect does this have on the genealogy of a sample from such a population?

We consider a population of N diploid individuals. Let us write s for the expected fraction of offspring to be produced by selfing (in which case both genes in the offspring are sampled from the *same* parent) and $1-s$ for the expected fraction to be produced by random mating. To understand what is happening we trace the history of two ancestral lineages. At any time in the past they can be in one of three states:

1. Two lineages in distinct individuals;
2. two lineages in the same individual;
3. coalesced.

Suppose that the two lineages are in distinct individuals. They will remain in this state for a geometrically distributed number of generations with parameter $1/N$. At that time, with probability $1/2$ they are derived from the same parental chromosome and they coalesce and with probability $1/2$ they move to the second state, two lineages in the same individual. If this individual was produced by selfing, which happens with probability $s+\mathscr{O}(1/N)$,[4] then with probability $1/2$ the lineages are derived from the same parental chromosome, and so they coalesce, and with probability $1/2$ they are derived from different parental chromosomes and they remain in state 2. Thus two lineages in state 2 remain there for a geometrically distributed number of generations with parameter

$$\frac{s}{2}+1-s+\mathscr{O}\left(\frac{1}{N}\right)$$

at which time the lineages coalesce with probability

[4] The $\mathscr{O}(1/N)$ correction in these calculations is because random mating carries a small probability of selfing.

$$\frac{s/2}{s/2+(1-s)}+\mathscr{O}\left(\frac{1}{N}\right)\approx\frac{s}{2-s},$$

otherwise the system returns to the first state. In particular, we only stay in the second state for $\mathscr{O}(1)$ generations.

If we measure time in units of N generations and let $N \to \infty$, then the second state becomes instantaneous. If the process starts in this state, then it instantaneously coalesces (with probability $s/(2-s)$) or else moves to the first state. Similarly a proportion $s/(2-s)$ of transitions from the first state to the second state will be followed by an instantaneous coalescence while the rest will be followed by instantaneous return to the first state. The overall rate of transition from the first to the third state (in rescaled time) is then

$$\frac{1}{2}+\frac{1}{2}\frac{s}{2-s}=\frac{1}{2-s}.$$

To see this, the first term corresponds to the (rescaled) rate at which two lineages in distinct individuals sample the same parental chromosome, the second is the rate at which they sample different chromosomes within the same individual – that is move to the now instantaneous state 2 – multiplied by the probability that they exit state 2 through coalescence. Alternatively, by measuring time in units of $(2-s)N$, a pair of lineages waits an exponentially distributed amount of time with parameter one before coalescing.

This argument can be extended to arbitrary finite samples from the population. It is tedious because we must keep track of many possible states. The argument above is from Nordborg and Donnelly (1997). A rigorous mathematical proof (using the techniques of Sect. 6.3) can be found in Möhle (1998). We have the following result.

Lemma 2.15. *In a diploid population as above in which a portion s of offspring are produced by selfing and the remainder by random mating, as the population size N tends to infinity, the genealogy of a sample is determined by a Kingman coalescent in which each pair of lineages coalesces at rate $2N_e$ where the effective population size $N_e = \frac{2-s}{2}N$.*

2.4 Adding Mutations

A mutation is formally defined as a “heritable change in the genetic material (DNA or RNA) of an organism”. Mutations occur in many forms, but for simplicity we concentrate on *point mutations* which occur when there is a change from one base pair to another at a single position in the DNA sequence. Because of the redundancy in the genetic code some point mutations do not lead to a change in the sequence of amino acids. These are called *synonymous mutations*. Mutations are the ultimate source of all genetic variation; without them there would be no evolution. Although mutation rates are relatively slow, the mixing of mutations from different lineages that results from genetic recombination (see Sect. 5.6) rapidly leads to an enormous number of combinations on which natural selection can act. Mutation rates vary according to the type of mutation, the location on the genome and the organism involved, with the highest rates being in viruses.[5]

Typically in our models we assume a constant probability μ per individual per generation of a mutation at a given base or more generally at a given locus. If we follow a particular ancestral lineage in our population, then we must wait a geometrically distributed number of generations (with mean $1/\mu$) until we see a mutation. Assuming that $2N_e\mu$, that is the mutation rate multiplied by the effective population size, is of order one, this will, in rescaled time, be approximately exponential. Moreover, under this condition, the probability that we see both a coalescence and a mutation in our sample in a single generation is $\mathscr{O}(1/N_e^2)$. So just as in our derivation of the Kingman coalescent, we see that if there are currently k lineages ancestral to our sample, the time (in rescaled units) we must trace back until we see *some* event is (approximately) the minimum of k independent exponential random variables each with parameter $2N_e\mu$ and an independent exponential random variable with parameter $\binom{k}{2}$. Another way to say this is we can add mutations to Kingman's coalescent by simply superposing a Poisson process of mutations on the ancestral lineages. Notice that in order to ensure that the types in the sample are consistent with the pattern of mutations stemming from such a Poisson process, a type must first be assigned to the MRCA and then we work our way back through the coalescent tree assigning types to ancestral lineages. This is illustrated by example in Fig. 2.2. There are several important models of mutation. Perhaps the simplest is the parent-independent mutation model.

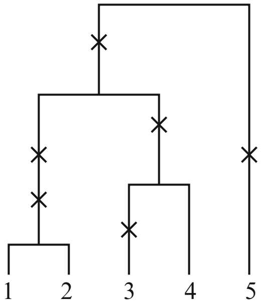

Fig. 2.2 Adding mutations to the Kingman coalescent. Mutations are added to the Kingman coalescent by throwing down an independent Poisson process of mutations on each branch. In order to ensure that the types in the sample are consistent with the pattern of mutations, one must first assign a type to the MRCA and then work back through the tree. In this example, we have used '×' to denote a mutation on a branch of the coalescent tree. Notice that here the individuals labelled 1 and 2 in the sample must have the same type

[5] See Fig. 12.23 in Barton et al. (2007). Rates shown there vary from $\mathscr{O}(10^{-4})$ per base pair per generation in RNA viruses like HIV to $\mathscr{O}(10^{-10})$ or $\mathscr{O}(10^{-11})$ in organisms like humans and mice. By contrast, there is a relatively uniform rate of mutation *per genome* per replication across diverse organisms.

Definition 2.16 (Parent independent mutation). In the *parent-independent* mutation model, a gene is assumed to occur in one of a finite number of types. Mutations occur at a constant rate per individual, independent of the current type of the individual. The type created by the mutation event is chosen according to a probability distribution which is also independent of the type of the parent.

More generally one can allow the probability of mutation to different types to depend on the current state of an individual, in which case the type of a lineage is governed by a Markov chain on the space of possible types.

Definition 2.17 (Infinitely many alleles model). In the *infinitely many alleles* model, every time a mutation occurs, it is to a new allele, never seen before in the population.

The infinitely many alleles model can be seen as the limit of the parent-independent mutation model when the number of alleles tends to infinity. It is useful in providing a link between the classical notion of *probability of identity* and the coalescent. In the infinitely many alleles model, two genes will be identical (that is they will have the same allelic state) if there has been no mutation since their MRCA. If their MRCA occurred T generations in the past, and the mutation rate per individual per generation is μ, then we see that this has probability $(1-\mu)^{2T} \approx e^{-2\tilde{\mu}\tau}$, where $\tau = T/N_e$ is the time to the MRCA in the coalescent timescale and $\tilde{\mu} = N_e\mu$ is the scaled mutation rate. Averaging out over the distribution of τ, the probability of identity is $\mathbb{E}[\exp(-2\tilde{\mu}\tau)]$, that is the Laplace transform of the distribution of the time, τ, to the MRCA.

Remark 2.18 (Mutation rates and nucleotide diversity). Since the expected number of generations since the MRCA of two genes sampled at random from a diploid population (under Kingman's coalescent) is $2N_e$, on average we expect them to differ by $4N_e\mu$ mutations per base pair. This can be counted directly if we are dealing with DNA sequences. The proportion of nucleotides that differ between two randomly chosen sequences is called the *nucleotide diversity* and is usually denoted by π. The crucial parameter $4N_e\mu$ is denoted by θ. Notice then that if we measure time in units of $2N_e$ generations (as is usual for the Kingman coalescent for a diploid population) then the rate at which we see mutations falling on each ancestral lineage is $\theta/2$. This explains the choice of scaling for the mutation rate in much of what follows. We shall use the same notation when μ is no longer the mutation rate per base pair, but rather the mutation rate for a locus or a whole gene.

If we sample a single nucleotide at random then with high probability all individuals in our sample will be identical. (A locus is usually defined to be polymorphic if the frequency of the most common type is less than 0.99. In humans, the chance of *heterozygosity* at a randomly chosen nucleotide is about 0.0008. In *Drosophila* it is an order of magnitude bigger, but still only about 1%, Lynch and Conery (2003), Fig. 1.) If the rate of mutation does not vary too greatly between bases then this justifies the so-called *infinitely many sites* model in which each time we see a polymorphic site in our sample we assume that it is due to a unique mutation.

Definition 2.19 (Infinitely many sites model). In the *infinitely many sites* model, every time a mutation occurs on a lineage it is at a new position on the DNA sequence.

It is sometimes convenient to model the genome as continuous, for example as $[0,1]$, in which case we suppose that each new mutation occurs at a position chosen according to an independent uniformly distributed random variable on $[0,1]$.

Notice that whereas in the infinitely many alleles model individuals only carry information about the most recent mutation on their ancestral lineage, in the infinitely many sites model they retain information about *all* mutations experienced by their ancestors.

2.5 Inferring Genealogies From Data

The genealogy of a sample from a population contains a great deal of information, but we cannot observe it directly. Instead we try to infer it from the pattern of mutations in the sample. We assume the infinitely many sites mutation model. Once a mutation occurs, it will be carried by all descendants of that individual and from this we can reconstruct at least partial information about the genealogical trees. If we suppose, for simplicity, that we know which is the ancestral type at each locus, then we can construct the so-called *gene tree*. The gene tree has mutations as its vertices. Figure 2.3 shows how this works in an example. Although a given pattern of mutations may be consistent with several different coalescent trees, if it is compatible with this model then it will be consistent with an essentially unique gene tree. The gene tree is unique up to permutations of labels along single lineages (for example 1 and 2 in the example in Fig. 2.3). However, there may be *many* different corresponding coalescent trees with mutation. For example, the gene tree in Fig. 2.3 is compatible with the coalescent tree of Fig. 2.4. It is also compatible with the coalescent in which b and c coalesce before a and b. More generally, if there are insufficient mutations then coalescent trees with many different topologies

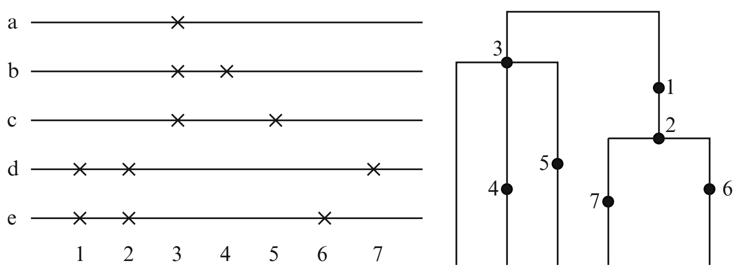

Fig. 2.3 Reconstructing a gene tree. The picture on the left represents a possible pattern of mutations in a sample of size 5. We suppose for simplicity that we know which is the ancestral type at each locus, so that an '×' in the picture indicates that an individual carries a mutation at that locus. On the right is a gene tree compatible with this pattern

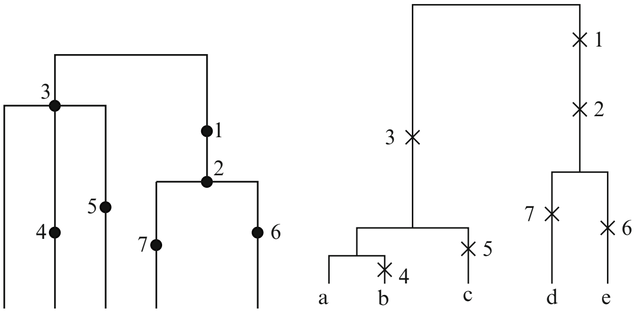

Fig. 2.4 Gene trees and coalescent trees. A given gene tree may be compatible with more than one coalescent tree. The coalescent tree on the right is compatible with the gene tree on the left. It would also be compatible with one in which *b* and *c* coalesce before *a* and *b*

may be compatible with a gene tree. (As an extreme example suppose that there were just one mutation shared by all but one individual in the sample.)

There are simple conditions to check that data is compatible with this model and efficient algorithms for reconstructing the gene trees. If the ancestral type is not known, then an unrooted tree is constructed. To recover a rooted tree one can compare to a more distantly related sequence (called an *outgroup*).

This procedure tells us something about the shape of the genealogical tree, but nothing about the lengths of the edges. However, since mutations are assumed to fall at an (approximately) exponential rate, some information about the time represented by an edge is available from the number of mutations occurring there. For much more on ancestral inference from gene trees we refer to Griffiths (2002). In practice, of course, things are not quite this simple. There are two principal problems. The first is *convergence*: if a site is evolving quickly, or if two sequences in our sample are very distantly related, then the same mutation may occur twice. The second is *recombination*, which we'll describe in more detail in Sect. 5.6. The result of recombination is that different stretches of our DNA sequence have different genealogies.

2.6 Some Properties of Kingman's Coalescent

We now return to Kingman's coalescent and record some of its elementary properties (and some of their consequences).

Lemma 2.20. *Let W_k denote the time to the most recent common ancestor of a sample of k genes whose genealogy is determined by Kingman's coalescent. Then*

$$\mathbb{E}[W_k] = 2\left(1 - \frac{1}{k}\right).$$

Proof. Since $W_k = T_k + T_{k-1} + \cdots + T_2$ where T_i is exponentially distributed with rate $\binom{i}{2}$ we have

$$\begin{aligned}\mathbb{E}[W_k] &= \sum_{i=2}^{k} \frac{2}{i(i-1)} \\ &= 2\sum_{i=2}^{k}\left[\frac{1}{i-1} - \frac{1}{i}\right] \\ &= 2\left(1 - \frac{1}{k}\right).\end{aligned}$$

□

Thus the mean time to the MRCA of the whole population (k infinite) is only twice that for a sample of size two. The picture is that for a large sample, as we trace backwards in time, we see a burst of quick coalescence followed by a long period with just a few ancestors. As a result, adding more and more individuals to our sample adds surprisingly little information. Moreover, since, in 'real' time, the standard deviation of the time when there are exactly two ancestral lineages is N_e generations (or twice that for a diploid population), the tree is always highly variable irrespective of the sample size. Figure 2.5 is a simulation of the Kingman coalescent for a sample of size 1,000.

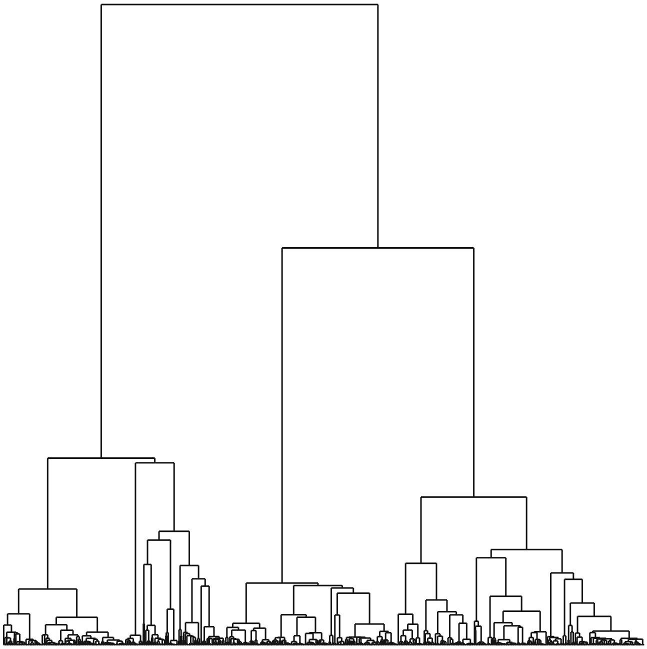

Fig. 2.5 Simulation of the Kingman coalescent. The picture is a single realisation in a simulation (courtesy of Bob Griffiths) of a Kingman coalescent for a sample of size 1,000. Notice the initial period of very rapid coalescence. For a large proportion of the time back to the MRCA, only two or three ancestral lineages remain

Lemma 2.21. *Let $L^{(k)}$ denote the total length of the genealogical tree relating a sample of size k. Under the Kingman coalescent, $L^{(k)}/2$ is distributed as the maximum of $(k-1)$ independent exponential random variables. In particular,*

$$\frac{1}{2}L^{(k)} - \log k \xrightarrow{d} X \qquad \textit{as } k \to \infty,$$

where X has a Gumbel distribution with density $\exp(-x - e^{-x})$.

Proof. The length of the tree here is measured until the time of the MRCA of the sample. Notice that if E is an exponentially distributed random variable with parameter one, then for $\gamma > 0$, writing $X_\gamma = E/\gamma$, we have $\mathbb{P}[X_\gamma > t] = \mathbb{P}[E > \gamma t] = \exp(-\gamma t)$, so that X_γ is exponentially distributed with parameter γ.

Now, in this notation, for each $2 \le j \le k$, the portion of $L^{(k)}$ corresponding to the time when there are exactly j ancestral lineages is $jX_{\binom{j}{2}}$ and the random variables $X_{\binom{j}{2}}$ are independent for different j. Thus

$$\begin{aligned} L^{(k)} = \sum_{j=2}^{k} jX_{\binom{j}{2}} &= \sum_{j=2}^{k} \frac{j}{\binom{j}{2}} E_j \\ &= \sum_{j=2}^{k} \frac{2}{j-1} E_j, \end{aligned}$$

where the E_j are independent exponentially distributed random variables with parameter one. From this

$$\frac{1}{2}L^{(k)} = \sum_{i=1}^{k-1} \frac{1}{i} E_{i+1} = \sum_{i=1}^{k-1} X_i = \sum_{j=1}^{k-2} X_{k-j-1}, \tag{2.5}$$

where the random variables X_i are independent exponential random variables with parameter i.

Now suppose that we have $k-1$ independent exponential random variables, each with parameter one, and arrange them in increasing order, $E^{(1)} < E^{(2)} < \cdots < E^{(k-1)}$. Then $E^{(1)}$ has an exponential distribution with parameter $(k-1)$ and, as a result of the lack of memory property of the exponential distribution, for $1 \le j \le k-2$, $E^{(j+1)} - E^{(j)}$ has an exponential distribution with parameter $k-j-1$. Thus the right hand side of (2.5) is distributed exactly as the maximum of $k-1$ independent exponentially distributed random variables, each with parameter one.

In particular,

$$\mathbb{P}\left[\frac{1}{2}L^{(k)} < x\right] = \left(\mathbb{P}[E_1 < x]\right)^{k-1} = \left(1 - e^{-x}\right)^{k-1},$$

and so

$$\begin{aligned}\mathbb{P}\left[\frac{1}{2}L^{(k)} - \log k < x\right] &= \left(1 - e^{-(x+\log k)}\right)^{k-1} \qquad \text{for } x > -\log k \\ &= \left(1 - \frac{1}{k}e^{-x}\right)^{k-1} \\ &\to \exp(-e^{-x}) \qquad \text{as } k \to \infty.\end{aligned}$$

□

Remark 2.22. Although $L^{(k)}$ has mean $2\log k$, the variance, $\mathtt{var}(L^{(k)})$, is bounded as $k \to \infty$.

Conditional on $L^{(k)}$, under the infinitely many sites model, the number of mutations that we see in our sample is Poisson with parameter $\theta L^{(k)}/2$ (recall Remark 2.18). Each site at which we see a mutation is called a *segregating site* or *SNP* (single nucleotide polymorphism). Writing $S^{(k)}$ for the number of segregating sites, we see that

$$\frac{2S^{(k)} - \theta L^{(k)}}{\sqrt{2\theta L^{(k)}}}$$

is asymptotically normally distributed with mean zero and variance one. Thus if we know the asymptotic distribution of $L^{(k)}$ we can deduce the asymptotic distribution of $S^{(k)}$.

Definition 2.23 (Watterson's estimator). Watterson proposed the following estimator for the mutation rate:

$$\hat{\theta} = \frac{2S^{(k)}}{\mathbb{E}[L^{(k)}]} = \frac{S^{(k)}}{\sum_{i=1}^{k-1}\frac{1}{i}}.$$

As a result of Lemma 2.21 we see that Watterson's estimator is asymptotically normal. However, since $L^{(k)}$ grows like $\log k$, in practice the convergence is extremely slow.

2.7 Genealogies and Pedigrees

We have seen that under our neutral population models, in finite time everyone in our population traces back to a single common ancestor. It follows immediately (by symmetry) that if an allele starts with frequency p_0 in the population, and there is no mutation, then the probability that it is eventually fixed (that is, carried by everyone) is just p_0. As a special case, the probability that a particular gene present in a single individual now will leave descendants in the indefinite future is $1/N$.

On the other hand, if we trace back family trees in a diploid population, then each individual has two parents, four grandparents and so on and, in a finite population, we quickly exhaust the population. Of course, in practice the ancestors are not all unique, but nonetheless we expect a significant proportion of the population to be included somewhere in our family tree. We shall refer to this family tree as the *pedigree* of the individual.

The following lemma illustrates the fact that if we trace far enough back in time, *most* individuals in the ancestral population will be in the pedigree of a given individual now.

Lemma 2.24. *Suppose that in a large diploid (but for simplicity hermaphrodite) population of size N, evolving in discrete generations, each individual chooses* two *parents uniformly at random from the previous generation. Then the probability that a randomly chosen individual from the population t generations in the past is in the pedigree of a given individual in the current population converges to about* 0.8 *as* $t \to \infty$.

Idea of Proof. First note that since N is large, the random number of descendants left by a single individual is approximately Poisson with parameter two (being, if we ignore the possibility of an individual choosing the same parent twice, Binomial with $2N$ trials and success probability $1/N$). Let $P(t)$ be the probability that an individual alive t generations ago does *not* belong to the pedigree of our chosen individual. Then, since none of that individual's descendants can be in the pedigree, we have $P(t+1) \approx \exp(-2+2P(t))$.[6]

The equation $p = \exp(-2+2p)$ can be solved (at least numerically). To see this, we first rearrange to obtain $(-2p)\exp(-2p) = -2\exp(-2)$. Now the equation $z = W(z)\exp(W(z))$ defines the *Lambert W function*, also known as the *product log* function. In general it is multivalued, but for $z \in (-1/e, 0)$ there are just two branches and choosing the one with $W(z) \geq -1$ gives a unique solution. This yields $p = -W(-2e^{-2})/2$ which is close to 0.2. □

The same calculation tells us that the 80% of individuals that are in the pedigree of our chosen individual are actually in the pedigree of *everyone* in the current population. The conclusion is that although most of us will have descendants alive into the indefinite future, a particular gene is highly unlikely to be transmitted.

In fact much finer results than these are known. Chang (1999) shows that if we go back $\sim \log_2 N$ generations[7] then we can expect to see an individual in the population who is ancestral to *every* present-day individual. Tracing back $\sim 1.77 \log_2 N$ generations *all* those individuals who are ancestors will be ancestors of every present-day individual.

[6] Here we are supposing that the probability of being in the pedigree is independent for each of the Poisson number of individuals. Although not *quite* true, the idea is that this probability is determined while the family trees of descendants of the different individuals are still small, before the dependence becomes important. We refer to Chang (1999) for a rigorous proof.

[7] We are using the notation $f(N) \sim g(N)$ to mean $f(N)/g(N) \to 1$ as $N \to \infty$.

Remark 2.25. In Baird et al. (2003) a branching process model is considered which traces the pedigree descendants of an individual *forwards* in time in a diploid population and asks whether that individual contributes *any* genetic material to the population t generations into the future. The genome is represented by the interval $[0,1]$. As a result of *recombination* (see Sect. 5.6), each offspring inherits, with equal probability, either the block $[0,U]$ or the block $[U,1]$ of genome from the 'pedigree parent', with the complement coming from the other parent (assumed unrelated). The random variable U is uniformly distributed on $[0,1]$ and is independent for each offspring. Whereas the probability of transmission of a particular gene in such a branching process model is $\mathscr{O}(1/t)$ (corresponding to the probability that a critical branching process survives until time t) if one asks whether *some* material from a block of genome has been transmitted, the rate of decay of survival probability is much slower (of $\mathscr{O}(1/\log t)$). This effect is akin to the birthday problem, since we are just asking that *some* block be transmitted, we are not specifying a particular block.

2.8 The Moran Model

We now return to the main theme of this chapter, random genetic drift, and introduce a second important model, the *Moran model* (due to Moran (1958)). Although less popular with biologists than the Wright–Fisher model, mathematically it is often more convenient. For example, in a population divided into two allelic types (as in Sect. 2.1), the frequency of the a-allele is governed by a birth and death process which greatly simplifies its analysis. Moreover, as we shall see, the genealogy of a sample from a population evolving according to a Moran model is *exactly* determined by Kingman's coalescent.

There are two essential differences between the Wright–Fisher model and the Moran model:

1. Whereas the Wright–Fisher model evolves in discrete generations, in the Moran model generations overlap.
2. In the Wright–Fisher model an individual can have up to N offspring, but in the Moran model an individual always has zero or two offspring.

Definition 2.26 (The neutral Moran model). A population of N genes evolves according to the *Moran model* if at exponential rate $\binom{N}{2}$ a pair of genes is sampled uniformly at random from the population, one dies and the other splits in two.

Remark 2.27. There is no agreement in the literature as to how to choose the rate at which pairs of individuals are chosen, this choice is convenient as it means that the genealogy of the population is determined by Kingman's coalescent, with no need for a further time change. With this choice of parameters, therefore, we can compare the predictions of the Moran model to those of the Wright–Fisher or Cannings models in the *coalescent* timescale. However, some care is needed in interpreting the model in 'real' time units.

Remark 2.28. The embedded discrete time Markov jump chain is a Cannings model in which the vector $(\nu_1(t),\dots,\nu_N(t))$ is uniformly distributed on all the permutations of $(2,0,1,1,\dots,1)$.

A more formal way to describe the model is as follows. We suppose that individuals in our population at time zero are labelled $1,\dots,N$. Associated to each pair of labels (i,j) is an independent rate one Poisson process that we denote by $\pi_{(i,j)}$. Since there are only a finite number of these, the points of distinct $\pi_{(i,j)}$'s are distinct. At a point of the Poisson process $\pi_{(i,j)}$, the individuals (genes) currently labelled (i,j) are involved in a reproduction event in which one dies and the other reproduces (with equal probabilities). The two offspring adopt the labels i and j. This is represented graphically in Fig. 2.6.

We can recover the ancestry of a sample by tracing backwards in time. If an ancestral line is at the tip of an arrow, then it *coalesces* with that at the root. If it is at the root it will be unaffected. For the population of Fig. 2.6 this is illustrated in Fig. 2.7. It is not hard to convince oneself that the genealogical trees relating individuals in a random sample are then precisely those generated by Kingman's coalescent. For example, follow a sample of size two backwards in time. The labels of the two individuals will change with time, let's call them $(i(t),j(t))$ say, but because of the lack of memory property of the exponential distribution, the time until we see an arrow joining the pair $(i(t),j(t))$ is still going to be exponential parameter one; if a label changes before coalescence, we simply piece together the random time before the label change with the remaining random time after the label change until we see coalescence. In particular then we see that, for large populations, from the point of view of the genealogy of a sample it makes little difference whether we consider a Wright–Fisher model or a Moran model.

Remark 2.29 (Adding mutations). We should like to add mutations to the Moran model in such a way that we can readily make comparisons with the Wright–Fisher model. For this reason, we separate the processes of mutation and reproduction so

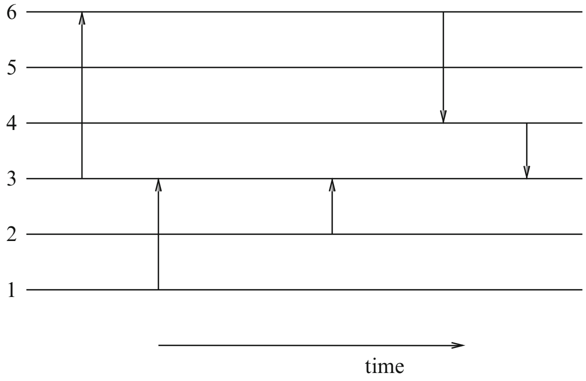

Fig. 2.6 Graphical representation of the Moran model. We draw an arrow between the lines labelled (i,j) at each point of $\pi_{(i,j)}$. The arrow $i \to j$ indicates that i reproduced and j died, $i \leftarrow j$ indicates that j reproduced and i died

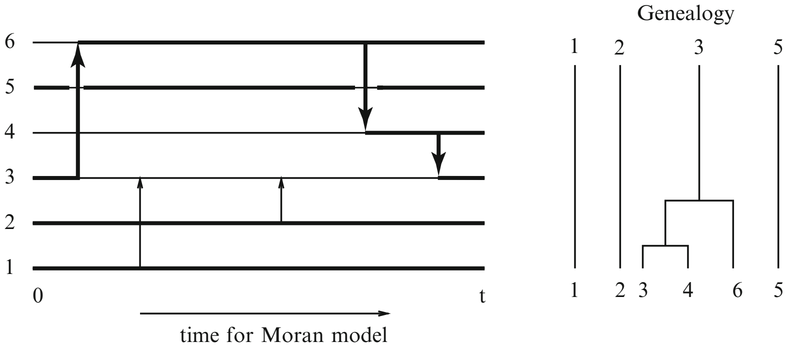

Fig. 2.7 Genealogies under the Moran model. The picture on the right shows the genealogical trees relating individuals in the population on the left, tracing back from time t to time 0

that mutations fall on the genealogical tree relating individuals in the sample according to a Poisson process, just as in Sect. 2.4. Since we are already in the timescale of the Kingman coalescent (c.f. Remark 2.27), it is natural to suppose that each individual accumulates mutations at a constant rate (irrespective of population size). In order to incorporate a range of different mutation models, we model this by supposing that in between reproduction events, the type of each individual, independently, evolves according to a mutation process (typically, but not necessarily, a finite state space Markov chain).

2.9 The Site Frequency Spectrum

In this section we exploit the relationship with the Moran model to continue our investigation of the Kingman coalescent.

The simplest statistic for a sample under the infinitely many sites mutation model is the number of segregating sites, whose distribution we discussed in Sect. 2.6, but one can also ask for more detailed information.

Definition 2.30 (Site frequency spectrum). For a sample of size k under the infinitely many sites mutation model, write $M_j(k)$ for the number of sites at which exactly j individuals carry a mutation. The vector $(M_1(k), M_2(k), \ldots, M_k(k))$ is called the *site frequency spectrum* of the sample.

This is illustrated in Fig. 2.8.

Lemma 2.31. *Suppose that the genealogy of a sample is determined by the Kingman coalescent and that mutations occur at rate $\theta/2$ along each ancestral lineage. Under the infinitely many sites mutation model we have*

$$\mathbb{E}[M_j(k)] = \frac{\theta}{j}. \tag{2.6}$$

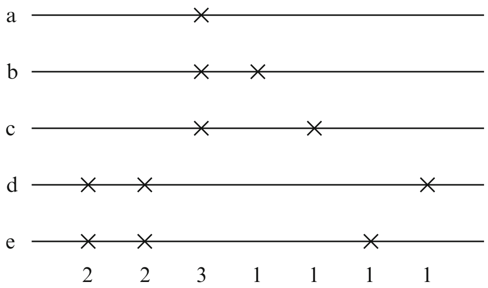

Fig. 2.8 The site frequency spectrum. In the example depicted here there are seven mutations, four of which are singletons, one occurs three times and two appear twice. (We are assuming that we know the ancestral type at each locus.) The site frequency spectrum is $(4,2,1,0,0)$

Proof of Lemma 2.31. We use the relationship between the Kingman coalescent and the Moran model. We emphasise that the 'population' in the Moran model below is not that from which we have sampled. It will have size k, the number of individuals in the *sample*. Suppose that a mutation arose at time $-t$ (that is t before the present) and denote individuals in our sample carrying that mutation as type a. For the corresponding Moran model (with population size k), we think of the mutation as arising at time zero and of the sample as the *whole population* at time t.

From the point of view of the Moran model, the probability that we see j type a individuals in the sample is the probability that a mutation arising on a single individual at time zero is carried by j individuals at time t later. We write X_t for the number of type a individuals at time t and $p(t,i,j) = \mathbb{P}[X_t = j | X_0 = i]$. In this notation, the probability that there are exactly j type a individuals in the sample is $p(t,1,j)$.

Since, under the infinitely many sites model, each mutation occurs at a different point on the genome and mutations occur at rate $\theta/2$ per individual (and the population size is k), the expected *total* number of sites at which we see a mutation carried by exactly j individuals is just

$$\mathbb{E}[M_j(k)] = \int_0^\infty k\frac{\theta}{2}p(t,1,j)dt. \tag{2.7}$$

Now $G(i,j) \equiv \int_0^\infty p(t,i,j)dt$ is just the expected total time that the process $\{X_t\}_{t\geq 0}$ spends in site j if it started from i and our next task is to calculate this.

Note that if $X_s = i$, then it moves to a new value at rate $i(k-i)$ (which is just the number of the $\binom{k}{2}$ ways of sampling a pair from the population in which the two individuals sampled are of different types) and when it does move, it is equally likely to move to $i-1$ or $i+1$. Let

$$T_i = \inf\{t > 0 : X_t = i\}$$

denote the first hitting time of site i. Then since 0 is a trap for the process we have

$$G(1,j) = \mathbb{P}[T_j < T_0 \big| X_0 = 1] \cdot G(j,j).$$

Now, because it is just a timechange of a simple random walk, for $0 \leq i \leq j$,

$$\mathbb{P}[T_0 < T_j \big| X_0 = i] = \frac{j-i}{j},$$

and similarly, for $j \leq l \leq k$,

$$\mathbb{P}[T_k < T_j \big| X_0 = l] = \frac{l-j}{k-j}.$$

Thus, partitioning on whether the first jump out of j is to $j-1$ or to $j+1$, we find that if it is currently at j, the probability that this is the *last* visit that X_t makes to j is

$$\rho = \frac{1}{2}\frac{1}{j} + \frac{1}{2}\frac{1}{k-j} = \frac{1}{2}\frac{k}{j(k-j)}.$$

In other words, if we start from j, the number of visits to j (including the current one) before either the allele is fixed in the population or it is lost is geometric with parameter ρ. Each visit lasts an exponentially distributed time with mean $\frac{1}{j(k-j)}$. Thus

$$G(1,j) = \frac{1}{j}G(j,j) = \frac{1}{j}\frac{1}{\rho}\frac{1}{j(k-j)} = \frac{2}{kj}.$$

Substituting into (2.7) completes the proof. □

Remark 2.32. The remarkable fact about this result is that the site frequency spectrum is almost independent of k. Increasing k only changes the allowed classes. The sceptical reader can work directly and, by conditioning on the first event as one traces backwards in time in the Kingman coalescent, check that the expected number of singletons is independent of k. This approach rapidly becomes tedious when checking the corresponding result for the other terms in the spectrum.

2.10 The Lookdown Process

The consistency of the k-coalescents for different values of $k \in \mathbb{N}$ allows us to recover all of them as projections of a single stochastic process, Kingman's coalescent. Since genealogical trees for the Moran model are precisely governed by the Kingman coalescent, it is reasonable to hope that we can also construct Moran models corresponding to different population sizes as projections of a single stochastic process. This is at the heart of the powerful Donnelly and Kurtz *lookdown process*.

To see how it works, we exploit the connection with the Kingman coalescent. Suppose that the population at the present time is labelled $\{1,2,\ldots,N\}$. Recall that the full description of the Kingman coalescent (or rather the N-coalescent) is as a process taking values among the set of equivalence relations on $\{1,2,\ldots,N\}$, with each ancestral lineage corresponding to a single equivalence class. Now suppose that we label each equivalence class by its smallest element. If blocks with labels $i < j$ coalesce, then after the coalescence the new block is necessarily labelled i. In our graphical representation of the Moran model, this just dictates the direction of the arrow corresponding to that coalescence event; it will always be the individual with the smaller label that gave birth. Backwards in time, our process is equivalent to one in which, as before, at the points of a rate one Poisson process $\pi_{(i,j)}$ arrows are drawn joining the labels i and j, but now the arrows are always in the same direction (upwards with our convention). The genealogies are still determined by the Kingman coalescent, we have simply chosen a convenient labelling, and so in particular they are precisely those of the Moran model. But what about forwards in time? What we saw backwards in time was that choosing the direction of the arrows corresponded to choosing a particular labelling of the population. If the distribution of the population is *exchangeable*, that is it doesn't depend on the labelling, then forwards in time too we should not have changed the distribution in our population. Our next task will be to check this, but first we need a formal definition.

Definition 2.33 (The N-particle lookdown process). The N-particle lookdown process will be denoted by the vector $(\zeta_1(t),\dots,\zeta_N(t))$. Each index is thought of as representing a 'level', with $\zeta_i(t)$ denoting the allelic type of the individual at level i at time t. The evolution of the process is described as follows. The individual at level k is equipped with an exponential clock with rate $(k-1)$, independent of all other individuals. At the times determined by the corresponding Poisson process it selects a level uniformly at random from $\{1,2,\dots,k-1\}$ and adopts the current type of the individual at that level. The levels of the individuals involved in the event do not change. In between lookdown events the type at each level evolves, independently, according to the mutation process.

Remark 2.34. Because of our convention over the interpretation of arrows, it is not at all clear from the above why one should call this the *lookdown* process. The explanation is that at rate $(k-1)$ the kth individual 'looks down' to a level chosen uniformly at random from those below and adopts the type of the individual at that level.

To see that the lookdown process and the Moran model produce the same distribution of types in the population, provided we start from an exchangeable initial condition, we examine their infinitesimal generators. Recall the definition of the generator of a continuous time Markov process.

Definition 2.35 (Generator of a continuous time Markov process). Let $\{X_t\}_{t\geq 0}$ be a real-valued continuous time Markov process. For simplicity suppose that it is time homogeneous. For a function $f:\mathbb{R}\to\mathbb{R}$ define

$$\mathscr{L}f(x)=\lim_{\delta t\downarrow 0}\frac{\mathbb{E}[f(X_{\delta t})-f(x)|X_0=x]}{\delta t}$$

if the limit exists. We'll call the set $\mathscr{D}(\mathscr{L})$ of functions for which the limit exists the *domain* of $\mathscr{L}$, and the operator $\mathscr{L}$, acting on $\mathscr{D}(\mathscr{L})$, the *infinitesimal generator* of $\{X_t\}_{t\geq 0}$.

If we know $\mathscr{L}$, then we can write down a differential equation for the way that $\mathbb{E}[f(X_t)]$ evolves with time. If $\mathscr{L}f$ is defined for sufficiently many different functions then this completely characterises the finite dimensional distributions of $\{X_t\}_{t\geq 0}$.

Let us write E for the space of possible allelic types for individuals in our population. The Moran model for a population of size N is then simply a continuous time Markov chain on E^N and its infinitesimal generator, K_N, evaluated on a function $f : E^N \to \mathbb{R}$, is given by

$$\begin{aligned} K_N f(x_1,x_2,\dots,x_N) = & \sum_{i=1}^{N} A_i f(x_1,x_2,\dots,x_N) \\ & + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\Phi_{ij} f(x_1,\dots,x_N) - f(x_1,\dots,x_N)\right], \end{aligned} \tag{2.8}$$

where $\Phi_{ij} f(x_1,\dots,x_N)$ is the function obtained from f by replacing x_j by x_i. The operator A_i is the generator of the mutation process, A, acting on the ith coordinate. (Recall that in the Moran model mutation was superposed as a Markov process along lineages.)

The generator of the N-particle lookdown process, L_N is given by

$$\begin{aligned} L_N f(x_1,x_2,\dots,x_N) = & \sum_{i=1}^{N} A_i f(x_1,x_2,\dots,x_N) \\ & + \sum_{1\leq i<j\leq N}\left[\Phi_{ij} f(x_1,x_2\dots,x_N) - f(x_1,x_2\dots,x_N)\right]. \end{aligned} \tag{2.9}$$

Assuming that we start both processes from the same exchangeable initial condition, we should like to show that the types $(\zeta_1(t),\zeta_2(t),\dots,\zeta_N(t))$ under the lookdown model and the types $(Z_1(t),Z_2(t),\dots,Z_N(t))$, say, under the original Moran process have the same distribution for each fixed $t>0$, even though the processes are manifestly different. Following Dawson (1993), we must check that the generators of the two processes agree on symmetric functions. Observe first that any symmetric function, f, satisfies

$$f(x_1,x_2,\dots,x_N) = \frac{1}{N!}\sum_{\pi} f\left(x_{\pi(1)},x_{\pi(2)},\dots,x_{\pi(N)}\right),$$

where the sum is over all permutations of $\{1,2,\dots,N\}$. Substituting this expression for f into (2.9), we recover (2.8). In other words, the generators of $(\zeta_1,\zeta_2,\dots,\zeta_N)$ and $(Z_1,Z_2,\dots,Z_N)$ agree on symmetric functions as required. (We are implicitly assuming uniqueness of the distribution on symmetric functions corresponding to this generator. It follows from duality with the N-coalescent, but we don't allow that to detain us here.)

The key observation now is that our Nth lookdown process is simply the first N levels of the $(N+k)$th lookdown process for any $k \geq 1$. The *infinite* lookdown process can then be constructed as a projective limit.

Theorem 2.36 (Donnelly and Kurtz 1996). *There is an infinite exchangeable particle system $\{W_i, i \in \mathbb{N}\}$ such that for each N,*

$$(W_1, W_2, \ldots, W_N) \overset{\mathscr{D}}{=} (\zeta_1, \zeta_2, \ldots, \zeta_N),$$

where $\zeta_1, \zeta_2, \ldots, \zeta_N$ is the N-particle lookdown process.

Remark 2.37. In fact more is true. It is known that the sequence of empirical measures $\frac{1}{N}\sum_{i=1}^{N} \delta_{Z_i(t)}$ converges to a *Fleming–Viot superprocess* as $N \to \infty$. Donnelly and Kurtz also show that

$$Y = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{W_i},$$

is a Fleming–Viot superprocess. A rapid introduction to Fleming–Viot superprocesses and further references can be found, for example, in Etheridge (2000). Rather than introduce the general Fleming–Viot superprocess, which takes its values among probability measures on the type space E, in Sect. 2.11 we shall consider what this limit looks like in the special case when E is a two-point set representing two alleles a and A, in which case it is enough to specify the evolution of the proportion of type a individuals in the population.

Since the genealogy of a sample of size k from the Moran model is a k-coalescent, and since we've seen that the genealogy of the first k levels in the lookdown process is also a k-coalescent, with this labelling we have a nice consistent way of sampling from a Moran model of arbitrary size. The genealogy of the sample is that of the first k levels in the lookdown process. The evolution of those levels does not depend on the population size – because we only ever look 'down' we don't see the population size N at all.

2.11 A More Simplistic Limit

Instead of discussing general Fleming–Viot superprocesses (which would allow us to consider essentially arbitrary type spaces) we now turn to identifying the limiting model for allele frequencies when our population is subdivided into just two types which, as usual, we label a and A. Just as in our discussion of the rescaled Wright–Fisher model, we consider the proportion, p_t, of individuals of type a at time t.

The only possible mutations are between the two types. We suppose that each type a individual mutates to type A at rate ν_1 and each type A individual mutates to type a at rate ν_2. Recall that for the Moran model we are already in the timescale of the Kingman coalescent and so we should think of $\nu_i = N\mu_i$ where μ_1 and μ_2 are the true mutation rates.

Remark 2.38. The idea that we can mutate backwards and forwards between types may seem at odds with our discussion of mutations in Sect. 2.4. Models of this type were introduced long before biologists knew about, and had access to, DNA sequences. Classically one might imagine a small number of alleles defined through phenotype, for example colour. In modern terms one can justify the model by pooling sequences into classes according to the corresponding phenotype.

The generator for the birth-death process of allele frequencies under the Moran model for a population of size N is then

$$\begin{aligned}\mathscr{L}_N f(p) = &\binom{N}{2} p(1-p)\left(f\left(p+\frac{1}{N}\right)-f(p)\right)\\ &+\binom{N}{2} p(1-p)\left(f\left(p-\frac{1}{N}\right)-f(p)\right)\\ &+N\nu_1 p\left(f\left(p-\frac{1}{N}\right)-f(p)\right)+N\nu_2(1-p)\left(f\left(p+\frac{1}{N}\right)-f(p)\right).\end{aligned} \tag{2.10}$$

To see this, note that the reproduction events in the Moran model take place at the points of a Poisson process with rate $\binom{N}{2}$ and at the time of such a transition there will only be a change in allele frequencies if the two individuals chosen to be involved in the reproduction event are of different allelic types. Thus, if the current proportion of a alleles in the population is p, then

$$\begin{aligned} p &\mapsto p+\frac{1}{N} \quad \text{with probability } p(1-p),\\ p &\mapsto p-\frac{1}{N} \quad \text{with probability } p(1-p)\end{aligned}$$

and there is no change with probability $1-2p(1-p)$. The chance that we see a reproduction event in a time interval of length δt is

$$\binom{N}{2}\delta t + \mathscr{O}((\delta t)^2)$$

and the probability of seeing more than one transition is $\mathscr{O}((\delta t)^2)$. For mutation events, at total rate $Np\nu_1$, one of the Np type a individuals will mutate to type A, resulting in a reduction of p by $1/N$ and at total rate $N(1-p)\nu_2$ one of the $N(1-p)$ type A individuals will mutate to type a. Putting all this together gives that for $f : [0,1] \to \mathbb{R}$ and $p = i/N$ for some $i \in \{0,1,\dots,N\}$, $\mathscr{L}_N f(p)$ is given by (2.10).

To see what our population process will look like for large N we take f to be three times continuously differentiable, and use Taylor's Theorem to find an approximation for $\mathscr{L}_N f$. Thus

$$\begin{aligned}\mathscr{L}_N f(p) = &\binom{N}{2} p(1-p)\left(f(p)+\frac{1}{N}f'(p)+\frac{1}{2N^2}f''(p)+\mathscr{O}\left(\frac{1}{N^3}\right)-f(p)\right)\\ &+\binom{N}{2} p(1-p)\left(f(p)-\frac{1}{N}f'(p)+\frac{1}{2N^2}f''(p)+\mathscr{O}\left(\frac{1}{N^3}\right)-f(p)\right)\\ &+Np\nu_1\left(f(p)-\frac{1}{N}f'(p)+\mathscr{O}\left(\frac{1}{N^2}\right)-f(p)\right)\\ &+N(1-p)\nu_2\left(f(p)+\frac{1}{N}f'(p)+\mathscr{O}\left(\frac{1}{N^2}\right)-f(p)\right)\\ =&\frac{1}{2}p(1-p)f''(p)+((1-p)\nu_2-p\nu_1)f'(p)+\mathscr{O}\left(\frac{1}{N}\right).\end{aligned}$$

We have proved the following.

Lemma 2.39. *As $N \to \infty$, the generator $\mathscr{L}_N$ of the process of allele frequencies under the neutral Moran model with mutation converges to $\mathscr{L}$, the generator of the Wright–Fisher diffusion with mutation, which is given by*

$$\mathscr{L}f(p) = \frac{d}{dt}\mathbb{E}[f(p_t)|p_0 = p]\bigg|_{t=0} = \frac{1}{2}p(1-p)f''(p)+(\nu_2-(\nu_1+\nu_2)p)f'(p). \tag{2.11}$$

Remark 2.40. Notice, in particular, that if we set $\nu_1 = \nu_2 = 0$ we obtain

$$\mathscr{L}f(p) = \frac{1}{2}p(1-p)f''(p),$$

which is exactly the generator that we obtained in the large population limit from our Wright–Fisher model. It is not hard to extend the work that we did there to include mutations and recover the full generator (2.11).

What we have written down is the generator of a one-dimensional diffusion. We should like to be able to use the convergence of generators that we have verified to justify using the corresponding one-dimensional diffusion as an approximation to the process of allele frequencies under the Moran, the Wright–Fisher and the Cannings models (on suitable timescales). We defer the statement of a theorem that provides that justification until Sect. 3.2. Evidently we also need to know that there is a unique Markov process with generator (2.11) and that we can actually calculate quantities of interest for it. Happily both are true.