Similar calculations to those we did for the Wright-Fisher model show that again measuring time in units of $1/c_N$ generations and under assumption (3), 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 (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) $Var(\nu_1)/N$, where Var denotes variance. In our previous language, the *effective population size* is $N/Var(\nu_1)$. In particular, the greater the variance in offspring number, the smaller the effective population size and the faster the rate of random drift.

Remark 1.10 (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).*

1.3 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 §2.5) 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.

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 of order $1/N_e^2$. 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.

PICTURE

There are several important models of mutation. Perhaps the simplest is the parent independent mutation model.

Definition 1.11 (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 transitions of types of ancestral lineages are governed by a Markov chain on the space of possible types.

Definition 1.12 (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 if there has been no mutation since their MRCA. If their MRCA occurred at a time τ in the past, and the mutation rate per individual per generation is μ, then we see that this has probability $\exp(-2\mu\tau)$ and now averaging out over the distribution of τ, the probability of identity is $\mathbb{E}[\exp(-2\mu\tau)]$, that is the Laplace transform of the distribution of the time, τ, to the MRCA.

Remark 1.13 (Mutation rates and nucleotide diversity) *Since the expected time 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 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 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 &

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 1.14 (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.

1.4 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. For example, suppose that in a sample of size 5 we see the pattern of mutations in the figure below. We suppose for simplicity that we know which is the ancestral type at each locus, so that an 'x' in the picture indicates that an individual carries a mutation at that locus.

PICTURE

Then we can construct the so-called *gene tree*. The gene tree has mutations as its vertices. 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 a unique gene tree.

PICTURE

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. 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 §2.5. The result of recombination is that different stretches of our DNA sequence have different genealogies.

1.5 Some properties of Kingman's coalescent

Let us record some elementary properties of Kingman's coalescent (and some of their consequences).

Lemma 1.15 *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.

Lemma 1.16 *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,*

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

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 $2 \leq j \leq k$, 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

$$\frac{1}{2} L^{(k)} = \sum_{i=1}^{k-1} \frac{1}{i} E_{i+1} = \sum_{i=1}^{k-1} X_i, \tag{4}$$

where again the random variables X_i are independent.

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 $1 \leq j \leq k-2$, $E^{(j+1)} - E^{(j)}$ has an exponential distribution with parameter $k-j-1$. Thus the right hand side of equation (4) is distributed exactly as the maximum of $k-1$ independent exponentially distributed random variables, each with parameter one.

In particular,

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

and so

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

□

Conditional on $L^{(k)}$, under the infinitely many sites model, the number of mutations that we see in our sample is Poisson with parameter $\frac{1}{2}\theta L^{(k)}$ (recall Remark 1.13). 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 1.17 (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 1.16 we see that Watterson's estimator is asymptotically normal. However, since $L^{(k)}$ grows like $\log k$, in practice the convergence is extremely slow.

1.6 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 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 1.18 *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 $t \to \infty$.*

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 individuals descendants can be in the pedigree, we have $P(t+1) \approx \exp(-2 + 2P(t))$.

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 $z = W(z)\exp(W(z))$ 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 $p = -\frac{1}{2}W(-2e^{-2})$ which is close to 0.2. □

In fact 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.

Remark 1.19 *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 genetic material is represented at time zero by the interval* $[0, 1]$. *As a result of* recombination *(see §2.5), each offspring inherits, with equal probability, either a block* $[0, U]$ *or the block* $[U, 1]$ *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 of order* $1/t$ *(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 (order* $1/\log t$*). 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.*

1.7 The Moran model

We now turn to a second important model for random genetic drift, the *Moran model*. 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 §1.1), 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 1.20 (The neutral Moran model) *A population of N genes (labelled* $1, \ldots, N$*) 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 1.21 *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 1.22 *The embedded discrete time Markov chain is a Cannings model in which the vector* $(\nu_1(t), \ldots, \nu_N(t))$ *is uniformly distributed on all the permutations of* $(2, 0, 1, 1, \ldots, 1)$.

A more formal way to describe our model is as follows. We suppose that individuals in our population at time zero are labelled $1, \ldots, N$. 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.

Graphically:

[PICTURE]

where we have drawn 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.

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.

[PICTURE]

It is not hard to convince oneself that the genealogical trees from a 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.

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 that mutations fall on the genealogical tree relating individuals in the sample according to a Poisson process, just as in §1.3. Since we are already in the timescale of the Kingman coalescent (c.f. Remark 1.21), 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).