Population genetics: Coalescence theory I

Peter Beerli

November 6, 2005

1 Population models

To understand inferences based on sampling random relationships among a small sample of individuals of a contemporary population we need to know some basic models of population history. Several such models exist. Fisher (1929, and 1930) and Wright (1931) developed independently a simple population model. We call this model the *Wright-Fisher* population model. Alternatives are the Moran model, developed by Moran in 1958. A model that extends the Wright-Fisher model was found by Cannings in 1974. We will discuss the three models

1.1 Wright-Fisher population model

We assume that we have a population of N diploid individuals, so each individual has two gene copies at a specific locus. A population therefore has $2N$ gene copies. Every generation the individuals reproduce once and die. Each individual is releasing a very large number of gametes. The next generation is formed by picking random pairs from this pool of gametes. Most simple implementation of this model do not allow for change of population size through time, mutation, selection, immigration and other complicating population genetics forces. If we assume that we start such a population with two alleles A_1 and A_2, than we look at the number of one specific allele X, say the number of A_1 alleles. The future states of X can represent any of $0, 1, 2, ..., 2N$. Sampling from the gene pool can be replaced by a sampling with replacement from the population at time t. This means that $X(t+1)$ is a binomial random variable with index $2N$ and parameter $X(t)/(2N)$. We can express the transition probability from $X(t) = i$ to $X(t+1) = j$ as

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - (\frac{i}{2N}\right)^{2N-j}, \qquad i, j = 0, 1, 2, 3, ..., 2N$$

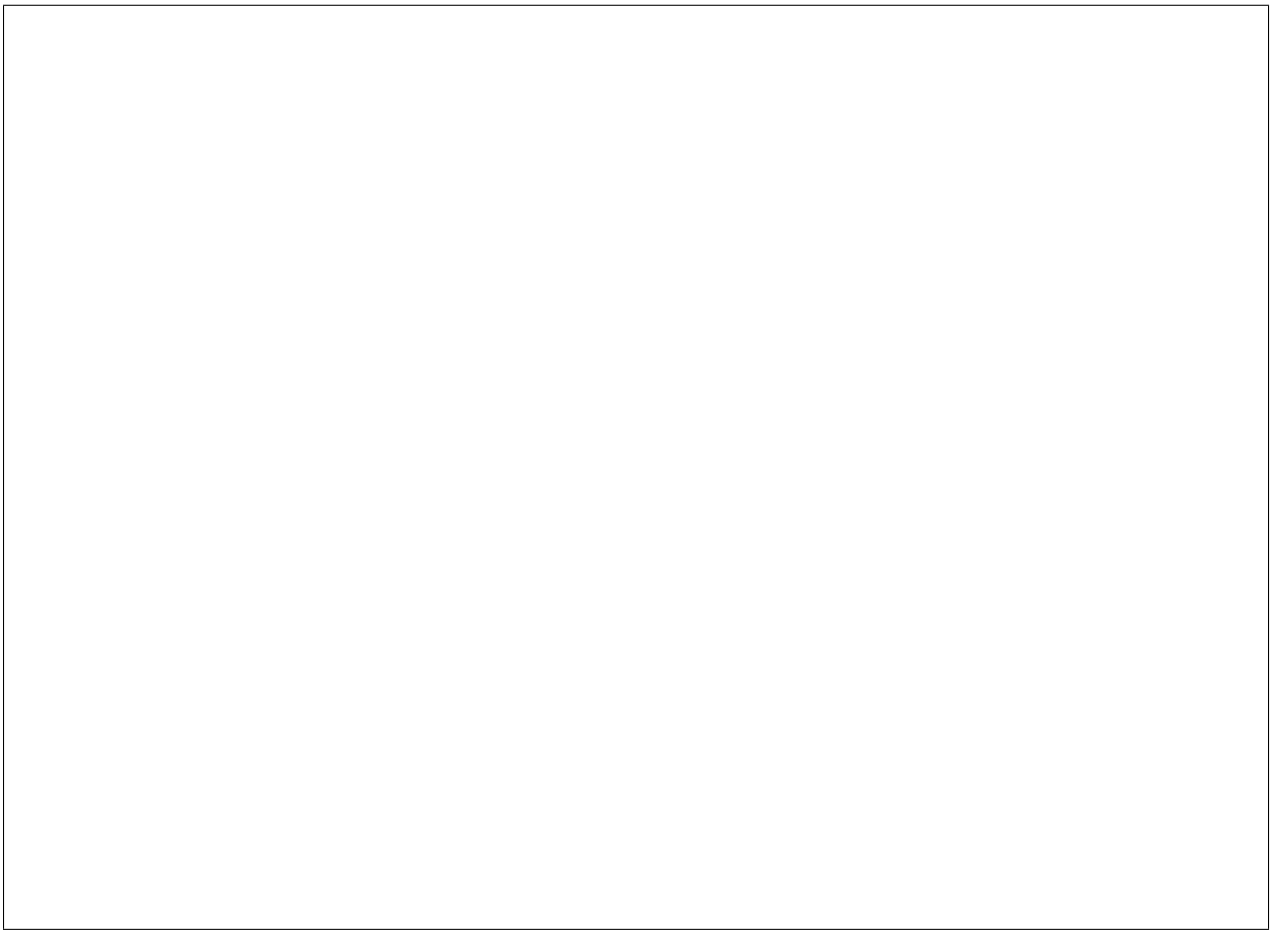

Figure 1: Example of a Wright-Fisher population model

We could now calculate (or approximate) average time to fixation of an allele or probability of fixation. The Wright-Fisher population model as explained here is prognostic as we look into the future and this makes it particularly difficult to derive quantities as average to fixation, Ewens (2004) gives and example of the time of fixation of an allele with frequency $i/(2N)$. For $i = 1$ one gets a mean absorption time

$$\bar{t}\left(\frac{1}{2N}\right) = 2 + \log 2N$$

We shall later see that we can derive similar estimates using *coalescence theory*.

Extending from two alleles to k alleles is easy as it uses simply the multinomial instead of the binomial distribution.

1.2 Moran's population model

Moran's model was derived for haploid populations, but many of its findings can be applied to diploids because if we assume neutrality then the alleles in a diploid population of size N behave like a haploid population of size $2N$. At time points $t_1, t_2, t_3, t_4, ...$, an "adult" individual is chosen at random to reproduce, after reproduction and "adult" individual is chosen randomly to die. Again with two alleles, A_1 and A_2, we can calculate transition probabilities. For comparison with the diploid Wright-Fisher model we still use $2N$ as the number of chromosomes in the population. If we have at time t i copies of allele A_1 then it at time $t+1$ there will be $i-1$ A_1 copies if an A_2 individual reproduced and an A_1 is chosen to die. This results in the probability

$$p(i, i-1) = \frac{i}{2N}\frac{2N-i}{2N}.$$

With probability $i/2N$ an A_1-individual dies and it gets replaced by an A_2-indvidual with probability $(2N-i)/2N$. For gaining an A_1 (equals loosing an A_2 first) we find

$$p(i, i+1) = \frac{i}{2N}\frac{2N-i}{2N},$$

and for no change

$$p(i, i) = \frac{i^2 + (2N-i)^2}{(2N)^2}.$$

The above probabilities form a transition probability matrix and, one can calculate the probability of fixation π of allele A_1 given that we have i individuals with allele A_1 now. Ewens (2004) calculates

$$\pi = \frac{i}{2N}.$$

Using this one can calculate the expected fixation time of a single allele A_1 as

$$\bar{t} = 2N(2N-1)$$

For a comparison with the Wright-Fisher population model we need to find a common measure of time as the events happening at times t_j in a Moran model are not equivalent to the events that happen at the times t'_j in a Wright-Fisher Model. We can express a common time g as the complete turnover of a population. If we set g=1 for the Wright-Fisher model then the Moran model needs on average around $2N$ time events to turn over, so we can express it generation time $g = 2N$. So we might express the waiting time to fixation in Wright-Fisher units as $\bar{t}_g(1/(2N)) = 2N - 1$.

1.3 Canning's (Exchangeable) population model

The Canning's model can be viewed as an intermediate between the Wright-Fisher model and the Moran model. Because depending on the reproduction function it can mimic the other two models.

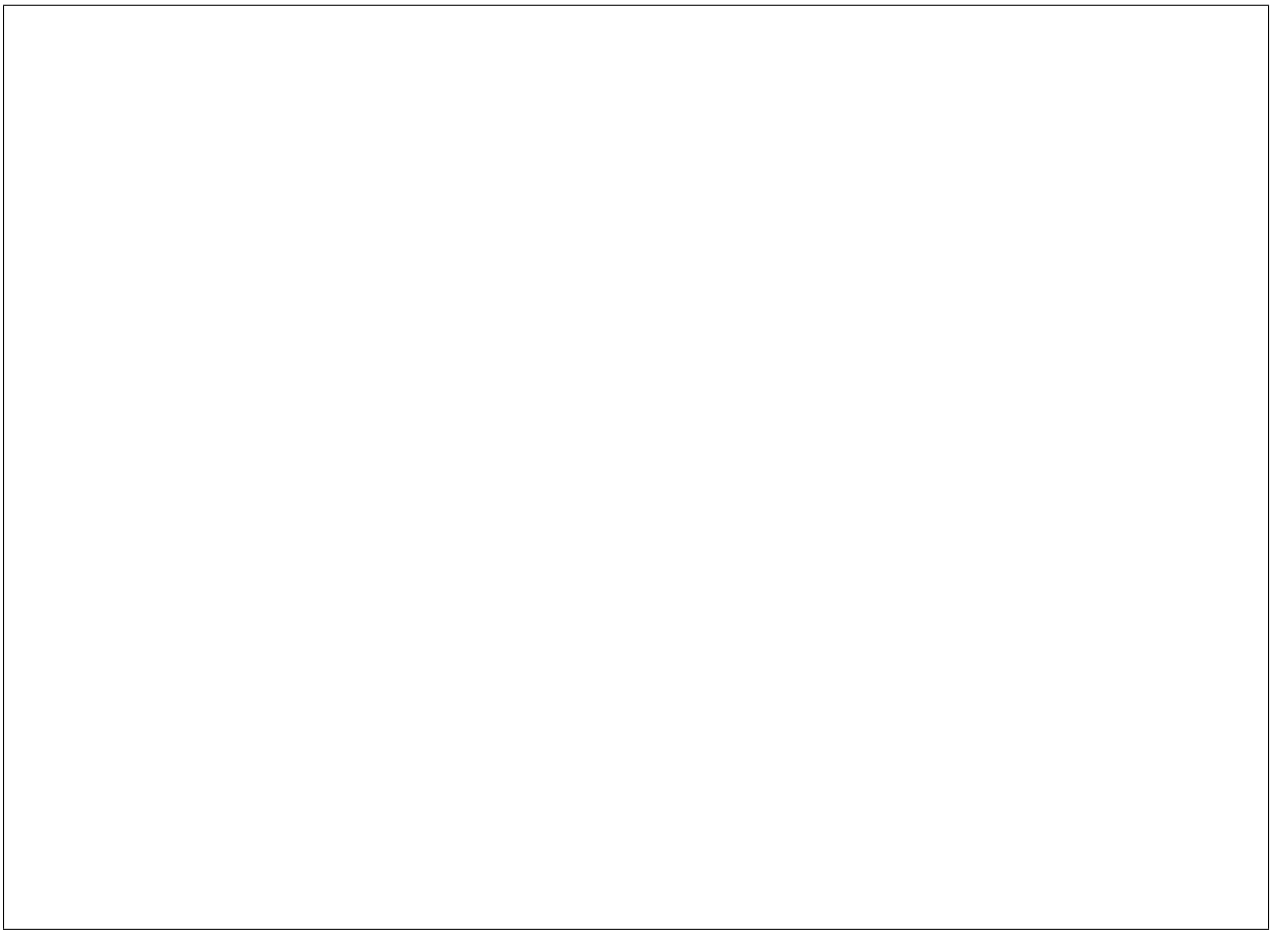

Figure 2: Example of Moran's population model

In its most basic version it looks like the Wright-Fisher model . Consider a a "population" of genes of size $2N$ reproducing at time points $t_1, t_2, ...$ The transition between the old generation and the new generation can be very general (not like the Wright-Fisher model) as long as the model guarantees that all alleles at time t have the same distribution of descendants at time $t + 1$: they need to have the same offspring probability distribution. This distribution has a mean of 1 offspring with variance σ^2. We can construct offspring distributions that fit this general description that are far from the multinomial distribution needed for the Wright-Fisher model.

2 The coalescent – coalescence theory

2.1 Historical note

Up to 1982 most development in population genetics was prospective and developed expectations based on situations of today. Most work did provide expectations about the future. With the easy availability of genetic data retrospective analyses did catch up only in phylogenetics (starting in the sixties). Only Malécot who pioneered "looking backwards in time" in 1948 to develop results in population genetics. Kingman expressed this "looking backwards in time" approach as the coalescence of sampled lineages. He was not the only one working on such problem at the the time as with many great solutions it was in the air, see Hudson (1983) and Tajima (1983).

2.2 The coalescent

A sample of n gene copies is taken at the present time and we are interested in the ancestral relationship of these gene copies. We express time τ increasing the further back in real time we go: $\tau_1 < \tau_2$ means that τ_2 is further in the past than τ_1. Kingman (1982) and Ewens (2004) describe this backwards in time process with equivalence classes. Two copies are in the same equivalence class at time τ when they have a common ancestor at that time. At time $\tau = 0$ each individual

Figure 3: Example of the coalescence process

gene can be considered in its own equivalence class and we could express this for a sample of $n = 8$ as

$$\phi_0 = \{(a), (b), (c), (d), (e), (f), (g), (h)\}$$

Kingman's n-coalescent describes the moves from ϕ_0 to a single equivalence class

$$\phi_n = \{(a, b, c, d, e, f, g, h)\}.$$

All individuals are in some equivalence relation ξ and we can find a new equivalence relation η by joining two of the equivalence classes in ξ. This joining process is called a *coalescence*, and a series of such joinings is called the *coalescent* or *coalescence process*. Figure 4 gives an example of the relationship of a sample and the equivalence classes describing the process. It is assumed that the probability of a coalescence depends on time waiting time $\delta\tau$

$$\text{Prob(process in } \eta \text{ at time } tau + \delta\tau \mid \text{process in } \xi \text{ at time } \tau) = \delta\tau$$

(ignoring higher order terms), and if k is the number of equivalence classes in ξ then

$$\text{Prob(process in } \xi \text{ at time } tau + \delta\tau \mid \text{process in } \xi \text{ at time } \tau) = 1 - \frac{k(k-1)}{2}\delta\tau = 1 - \binom{k}{2}\delta\tau$$

These result in the rates of the coalescent event of τ and the rate for the waiting that that event happens $\binom{k}{2}\tau$. We will see that if we apply the right time scale to τ then we will end up in the more familiar terms that are common in the applied population genetics literature.

Kingman focused on the Canning model, and since the Wright-Fisher model is a special case of the results carry over easily. The coalescent is an approximation to these models because it was developed on a continuous time scale whereas the Canning and Wright-Fisher population models have discrete time. Any findings using this coalescent machinery needs to be rescaled to the time scale of these discrete time models. In the coalescent framework one has only a single coalescent per infinitesimal time period. This forces us to restrict the use of the coalescent to discrete time models were we can guarantee that there is not more than one coalescent event occurring per time period. For example, for a Wright-Fisher population we can allow only one coalescent event per generation. this sound rather restrictive but as long as the sample n is much smaller than the population N this situation rarely occurs. Fu (year?) calculates that a sample needs to be less than the square-root of N

$$n < \sqrt{N},$$

when we use the coalescent for a model such as the Wright-Fisher or some version of the Canning model.

2.3 The coalescent and the Wright-Fisher population model

The coalescent process is in effect a sequence of $n-1$ Poisson processes[1], with rates

$$r_k = \frac{k(k-1)}{2}, k = n, n-1, n-2, ..., 2$$

describing the Poisson process at which two of the equivalence classes merge when there are k equivalence classes.

Since these events are coming from a Poisson distribution we can calculate the expectation for each interval, which is is 1/rate, here $2/(k(k-1))$. All mergers of the equivalence classes are independent of each other so the expectation of the whole coalescence process is

$$\mathbb{E}(T_{\text{MRCA}}) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \frac{1}{k(k-1)}$$

Comparison of the content of the sum with the coalescence rate makes clear that the Wright-Fisher population model is $2\times$ the standard coalescent units and we need to multiply by $2N$ to arrive at the more familiar generation time scale.

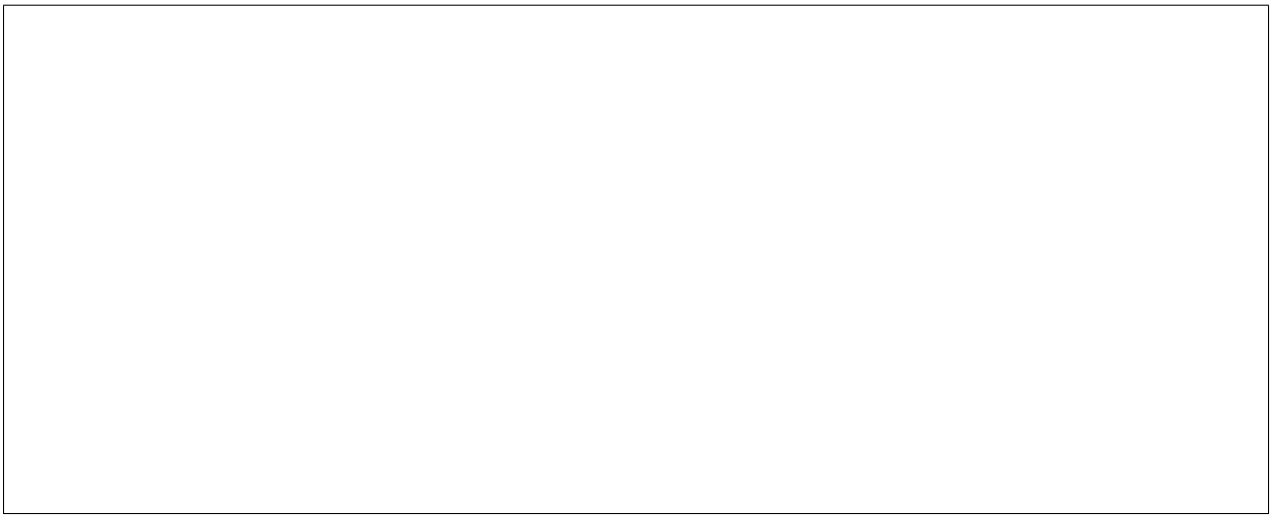

Figure 4: Expected times in the Wright-Fisher coalescent process

[1]From Mathworld: A Poisson process is a process satisfying the following properties:

1. The numbers of changes in non-overlapping intervals are independent for all intervals.
2. The probability of exactly one change in a sufficiently small interval is , where is the probability of one change and is the number of trials.
3. The probability of two or more changes in a sufficiently small interval is essentially 0.
   In the limit of the number of trials becoming large, the resulting distribution is called a Poisson distribution.

The coalescent process results in a tree of a sample of n individuals. We call this often a genealogy as the the individuals are typically from the same species or population (hence population size).

Often the probability of the coalescent process for the Wright-Fisher population model is expressed as

$$p(G|N) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{4N}} \frac{2}{4N},$$

where u is expressed in generations. The expected T_{MRCA} is the same as above. Often we will not use a time scale in generations but generations $\times$ mutation rate and then we would express the above formula as

$$p(G|\Theta) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{\Theta}} \frac{2}{\Theta},$$

where Θ is $4 \times N_e \times \mu$, with μ as the mutation rate per generation and site (when using sequence data), and N_e as the effective population size. Under a strict Wright-Fisher population model $N = N_e$, but under more biological scenarios one needs to know more about the life history of the species to translate the N_e into real numbers.

2.4 The coalescent and the Moran population model

The coalescent is an *exact* representation of the Moran model because the problems with the multiple coalescent events in one generation do not occur. The Moran model allows only one lineage to change at a given time. Therefore the limitation to small sample size as we have seen for the Wright-Fisher model is not needed.

Using our findings of the discussion of the Moran model earlier, but instead of thinking forward in time, think backward in time. Looking backwards we see that the Moran process is similar structured like the coalescence process. We have n individuals that are reduced in their ancestry to $n-1, n-2, ...$ and eventually to one gene, the most common recent ancestor of the n sampled individuals. Assume that we are at a time where we have a sample of k individuals, these are descendants of $k-1$ parents of one of these parents was chosen to reproduce and the offspring is in ancestry of the sample of n genes. The probability of this event is

$$\frac{k(k-1)}{(2N)^2},$$

and with probability

$$1 - \frac{k(k-1)}{(2N)^2},$$

the ancestors remain at j. Tracing back this ancestry the number of death and birth events between the times when there are j and $j-1$ ancestors follows a geometric distribution with parameters $k(k-1)/(2N)^2$ and thus has a mean of

$$\mathbb{E}(u_j) = \frac{(2N)^2}{k(k-1)}$$

now we can assemble the expectation for the time to the most recent common ancestor

$$\begin{aligned}\mathbb{E}(T_{\text{MRCA}}) &= \sum_{k=2}^{n} \frac{(2N)^2}{k(k-1)} = (2N)^2(1-\frac{1}{n}) \\ &= (2N)^2 \sum_{k=2}^{n} \frac{1}{k(k-1)} \qquad (1)\end{aligned}$$

If we assume that we sampled the whole population, where $n = 2N$ than we derive the same result as with standard (forward) theory.

$$\mathbb{E}(T_{\text{MRCA}}) = (2N)^2(1-\frac{1}{n}) = (2N)^2(1-\frac{1}{2N}) = 2N(2N-1)$$

We also can make the same observation as we made with the Wright-Fisher population model. The coalescence time scale is by a factor $(2N)^2$ different from the Moran model time scale (formula 1).

We can express the probability of the genealogy under the Moran model using the fact that the exponential distribution is a good approximation to the geometric distribution

$$p(G|N) = \prod_{k=2}^{n} e^{-u_k \frac{k(k-1)}{2(2N)^2}} \frac{1}{(2N)^2},$$

where u is expressed in generations. The expected T_{MRCA} is the same as above. Often we will not use a time scale in generations but generations $\times$ mutation rate and then we would express the above formula as

$$p(G|\Theta) = \prod_{k=2}^{n} e^{-u_k \frac{2k(k-1)}{\Theta^2}} \frac{4}{\Theta^2},$$

where Θ is $4 \times N_e \times \mu$, with μ as the mutation rate per generation and site (when using sequence data), and N_e as the effective population size. Under a strict Wright-Fisher population model $N = N_e$, but under more biological scenarios one needs to know more about the life history of the species to translate the N_e into real numbers.