Population genetics: Coalescence theory II

Peter Beerli

October 31, 2011

1 The variance of the coalescence process

The coalescent is an accumulation of waiting times. We can think of it as standard queuing process where the times are exponentially distributed with rate $k(k-1)/(2 \times 2N)$ [for most elaboration in this chapter I use the Wright-Fisher model as a guide, for the Moran model the rate would be $k(k-1)/(2 \times (2N)^2)$]. The coalescent makes no assumptions about the interaction of the intervals, we will assume that the intervals with $k = n$ lineages is independent from the interval with $k = n-1$ lineages, and we further assume that the exponential distribution is a good approximation to the process (which it is), then we find that the variance of the time to the most recent common ancestor

$$\sigma^2(T_{\text{MRCA}}) = \sigma^2(u_n) + \sigma^2(u_{n-1}) + \sigma^2(u_{n-2}) + ... + \sigma^2(u_k) + ... + \sigma^2(u_2)$$
$$\sigma^2(T_{\text{MRCA}}) = \sum_{k=2}^{n} \left(\frac{k(k-1)}{4N}\right)^2$$

This expression looks so simple , but it expands into a mess

$$\sigma^2(T_{\text{MRCA}}) = \sum_{k=2}^{n} \left(\frac{k(k-1)}{4N}\right)^2 \sigma^2(T_{\text{MRCA}}) = \frac{1}{n^2(n+1)^2} F(1,n,n)$$

where F is the generalized hypergeometric function. Figure 2 gives an example of genealogies that express the variance in the depth of the tree, can vary widely. Any particular realization is not necessarily a good explanation of the process. The distribution in figure 2 was created from 10000 tree depth with 10 tips, the range is large as the above variance calculations suggests. a typical neutral coalescent genealogy will be near the mode of the distribution, but it is easy to see that even far off the mode there is still some probability mass.

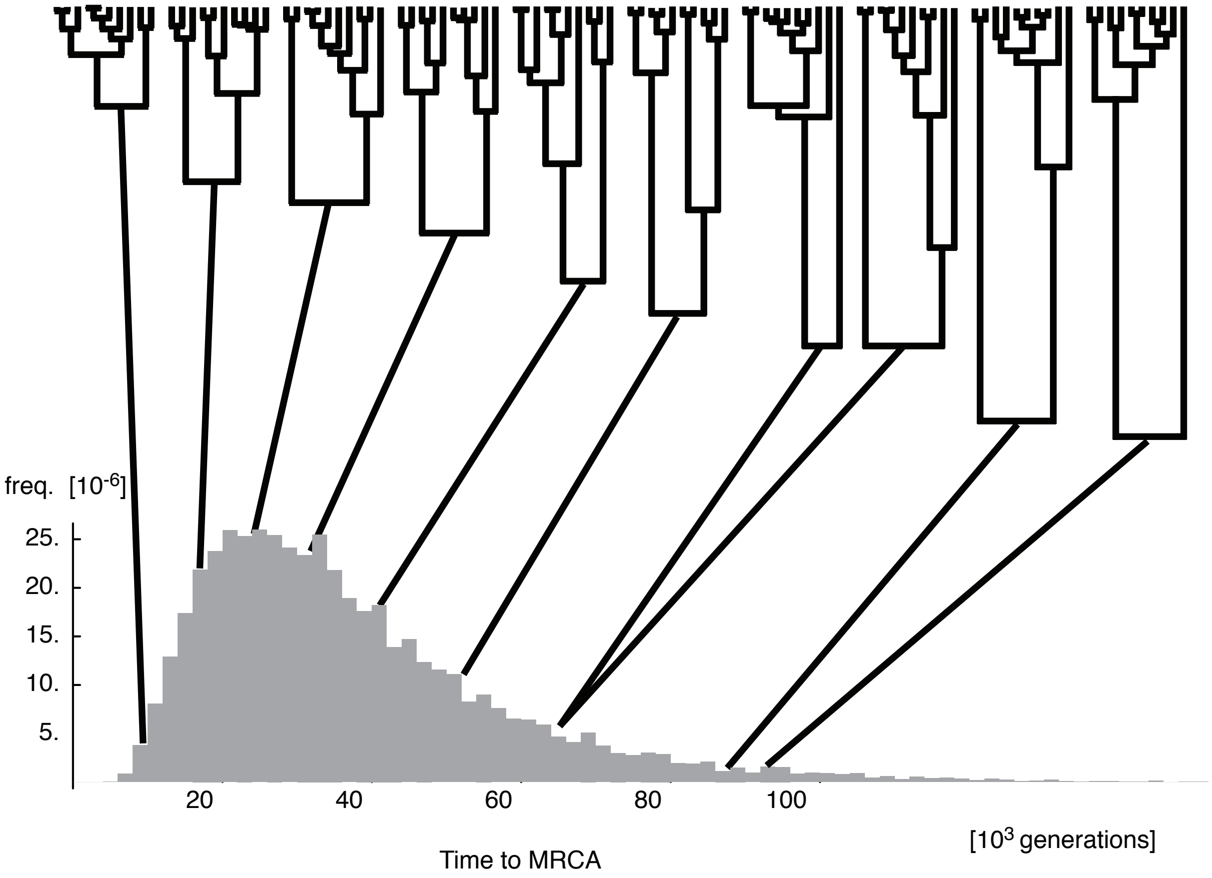

Figure 1: Example of ten random coalescent trees generated with the same population size of 10,000. The distribution of T_{MRCA} was generated using 10000 simulated trees generated with $N = 10,000$

2 Sampling issues

Often biologists ask whether they should sample more individuals or indpendent loci. The coalescent can assist in answering such questions. If the populations or species of interest are haploids with no recombination then the only way to improve the answer is to increase the number of sampled genetic material (more sequence per individual) and the number of individuals. But because of the structure of a coalescent tree after about 10 individuals most of them will have highly correlated histories and so additional individual do not improve our knowledge at the root of the genealogy. Turn back to the variance calculations and compare the contribution of an additional individual to the variance when we add an individual to a set of n. The variance increase strongly on the first few individuals.

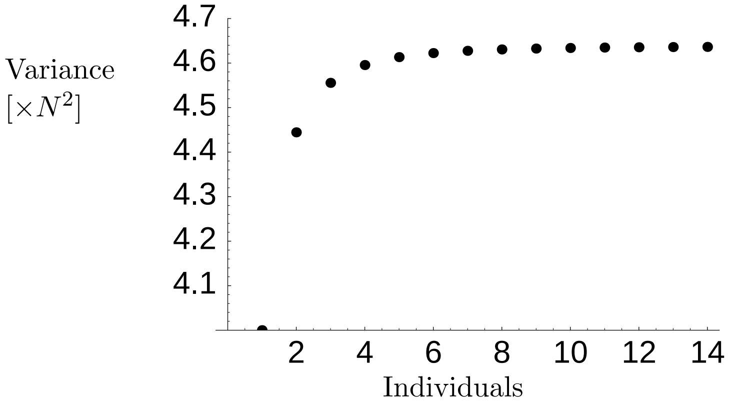

Figure 2: Variance contribution per individual sampled.

3 Extensions if the simple coalescent

3.1 Population growth

Population growth can be modeled in several ways and several authors have worked on this: sudden expansion, exponential growth and logistic growth was modeled using the coalescent. The case of exponential growth or shrinking will be explained in more detail, but in principle we can treat all growth cases can be treated the same way.

Take for example the exponential growth case. Here we add a growth rate g to the existing parameter, the population size N. using the population size today N_0 and looking backward in time we can construct the relationship

$$N(t) = N(0)e^{-tg}$$

where N_t is the size t generation in the past, g is the exponential growth rate, and t is the time in generations. Hudson, Kingman and others recognized that the standard coalescent can be extended by manipulating the the time scale. In the standard coalescent the time scale is constant, but in a growing population the time scale is proportional to the $N(0)$ and $N(t)$, we could think about changing the time scale in such a way that we integrate this proportionality, and we calculate the change of time scale as

$$\begin{aligned} d\tau =& \frac{N(0)}{N(t)} \\ \tau =& \int d\tau = \int \frac{N(0)}{N(t)} d\tau \\ =& \int \frac{N(0)}{N(0)e^{-g\tau}} d\tau \\ =& \frac{1}{g}(e^{gt} - 1) \end{aligned}$$

We interested in t [in generations] and not the fictional time τ, but the time scale in τ can use the standard coalescence so all wee need to do is to assemble the bits:

$$p(u = t_e - t_s | N(0), g) = e^{-\frac{\left(e^{g\text{te}} - e^{g\text{ts}}\right)(k-1)k}{4gN(0)}}$$

where t_s and t_e are the absolute start and end time of the interval. The probability of a genealogy of a growing or shrinking population is therefore

$$\text{Prob}(G|N(0), g) = \prod p(u = t_e - t_s|N(0), g)\frac{2}{4N(0)e^{-gt_e}} = \prod e^{-\frac{\left(e^{g\text{te}} - e^{g\text{ts}}\right)(k-1)k}{4gN(0)}} \frac{2}{4N(0)e^{-gt_e}}$$

To understand growing populations it helps to realize that when the population is small then the rate of coalescence is large ($k(k-1)/(4N)$) and therefore the time intervals to coalescences are short, whereas when the population is large the rate of coalescence is small. This produces on average genealogies for an exponentially growning population with longer branches at the tips and shorter branches at the root than the standard coalescent, but often this is not easy to see at all (Figure 3).

3.2 Recombination

Without recombination every site on a chromosome has the same coalescent as its neighbors. Recombination is breaking up this relationship and so it can happen that sites 1-100 in a sample

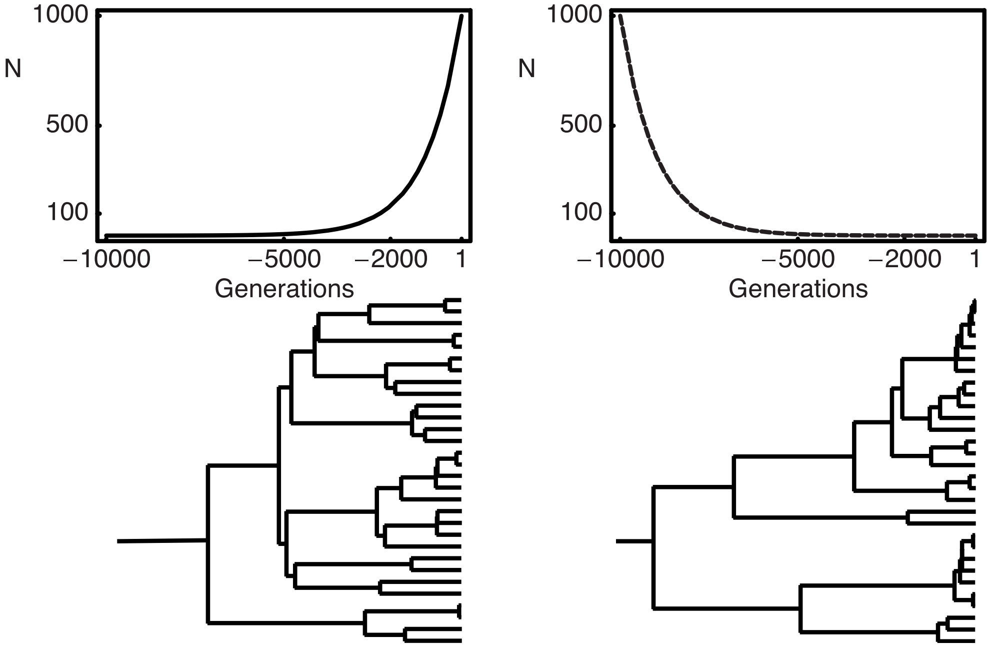

Figure 3: Growing and shrinking population

has a different genealogy than sites 101-500. Figure 4 shows a possible example. We can express an recombination event as a branching downwards process and we can incorporate this into the probability calculations, where the waiting times are now not only dependent on the rate of coalescence but also on the rate of recombination. The rates of recombination depend the magnitude of the recombination parameter ρ and the number of possible recombination sites: with sequences for example of length 10 we have only 9 possible sites for recombination and once each site has completely coalesced there is no further information available

Migration

Instead of simply having samples from a single population we could have samples form multiple populations and could investigate what effect this subdivision has on the coalescent. Again we could think that at any time we consider rates to coalescence and now rates for migration events (events where one lineages moves to the other population). Migration models can have many parameters, for example a simple two population model can have 4 parameters (Fig 5)

A typical coalescent tree with migration can be depicted in two ways. We could move the lineages between the populations (that produces a tangled mess with many migration or then we can express the migration events on the tree (see Figure 6).

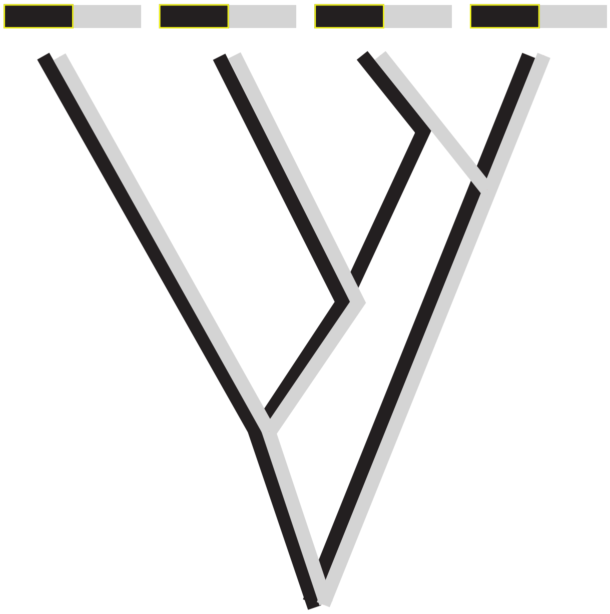

Figure 4: Recombination event on a genealogy

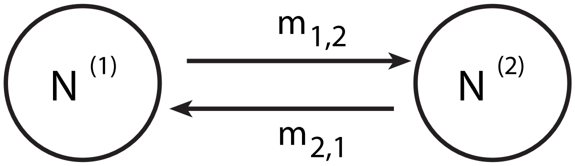

Figure 5: Migration model

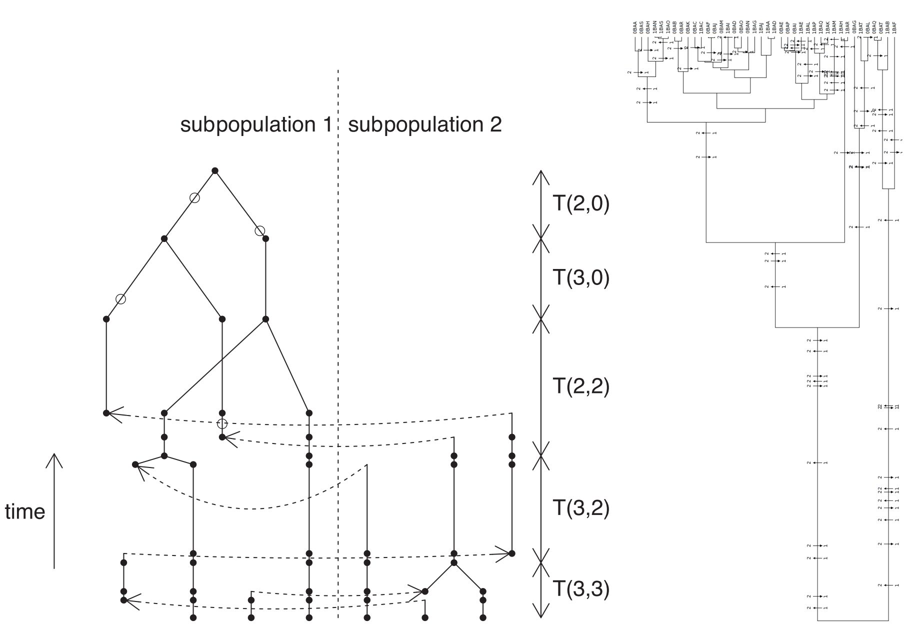

Figure 6: Two example representation of migration events on a tree, the two trees are not identical

For two populations we need to considered coalescences in population 1 and 2 and migration events that move lineages from 1 to 2 or 2 to 1. The probability of a genealogy with migration again is not all that difficult to calculate because all the events are independent of each other: we have exponential waiting time for each of these events

$$\text{Prob}(G|N_1, N_2, m_{12}, m_{21}) = \prod_j \exp(-u_j \sum_i^m \left(\frac{k_i(k_i-1)}{4N_i} + \sum_j k_i m_{ji}\right) \left\{ \begin{array}{l} \frac{2}{4N_1} \\ \frac{2}{4N_2} \\ m_{21} \\ m_{12} \end{array} \right.$$

4 Study questions

1. Why is not useful, in most cases, to sample more than 20 individuals for any population study? Can you a construct a case where it would be useful? What about migration? What about selection?

2. Taking into account data instead of simply the coalescent, what are your thoughts about the sampling discussion?

3. Give the rate of coalescence for a Wright-Fisher model.

4. why is the growth parameter treated differently than recombination or migration?

5. What happens when the migration rate in a two population problem approaches zero?