

Trees of Life (BIOL30004)  
Lecture 2  
Coalescent Theory I

Mark Beaumont

Spring 2017

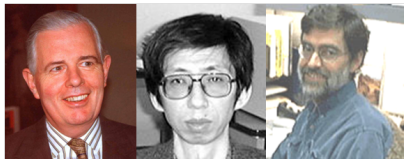
## Summary of Lecture 2

1. Introduce the Wright-Fisher model.
2. Genealogies in the Wright-Fisher model.
3. How to obtain the coalescent from the Wright-Fisher model.
4. Properties of coalescent models.
5. Simulating gene genealogies.

## The Coalescent

- ▶ The theory of the coalescent was given in two key papers by Sir John Kingman (formerly Vice-Chancellor of Bristol):
  - ▶ Kingman, J. F. C. (1982) The coalescent. *Stochastic Processes and their Applications* **13**: 235–248.
  - ▶ Kingman, J. F. C. (1982) On the genealogy of large populations. *Journal of Applied Probability* **19A**: 27–43.(These are difficult, technical, papers and I don't expect you to understand them, but you might want to have a look.)
- ▶ It was also developed independently by a couple of biologists:
  - ▶ Hudson, R.R., (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**: 203-217
  - ▶ Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437-460.

(These are much easier to read.)



## Genealogies and the Wright-Fisher Model

The properties of genealogies can be quite easily understood from an idealised, clonal, model of reproduction where each individual in a fixed population of size  $2N$  'chooses' its parent at random. This is called the Wright-Fisher model. (If you look up Wright-Fisher on the web you may get a bunch of formulae, but they are describing essentially the same thing . . . ).

Note that we say the population size is  $2N$ . This is because conventionally the population is diploid, so there are  $2N$  gene copies in the population, which we model as being from a haploid population.

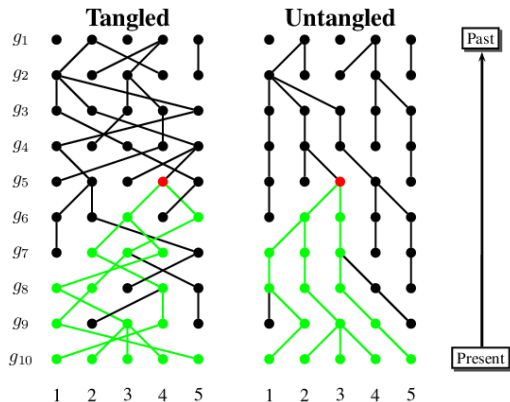
Sometimes it is just easier to think of  $N$  as haploid, and I will make it clear when I do.

# Wright-Fisher Examples

This website has a toy Wright-Fisher simulator which illustrates the genealogies you get from this model.

Note the population size  $N$  is haploid in this example.

[▶ Wright-Fisher](#)



## Coalescent basics I

In the Wright-Fisher model, the probability that two genes are copies of the same gene (*i.e.* they coalesce) in the previous generation is

$$\frac{1}{2N}.$$

Why?

Each gene 'chooses' its parent at random.

Take a pair of genes and let one choose its parent; then the probability the other chooses the same parent is  $1/(2N)$ .

Another way to look at it is to arrange all  $(2N)^2$  possible pairs of parents chosen by the two genes on a grid (matrix), of which  $2N$  (those on the diagonal) will be identical, giving

$$(2N)/(2N)^2 = 1/(2N)$$

## Coalescent basics II

What is the probability that two genes coalesce in the next generation after  $t$  generations?

We can work this through by thinking of the analogous situation:

What is the chance of getting a six on my first throw of a die [corresponding to  $t = 0$  in the way I have set it up]?. The answer is obviously

$$1/6$$

What is the chance on the second throw [ $t = 1$ ]?. The answer is

$$\left(1 - \frac{1}{6}\right) \frac{1}{6}.$$

*I.e* the probability of not getting it on the first throw, but getting it on the second.

## Coalescent Basics III

What is the chance on the third throw [ $t = 2$ ]? The answer is

$$\left(1 - \frac{1}{6}\right)^2 \frac{1}{6}.$$

*i.e* The probability of not getting it on the first and second throws, but getting it on the third.

In general we can see that the probability that our random variable  $T$  equals any particular value  $t$  is

$$p(T = t) = \left(1 - \frac{1}{6}\right)^t \frac{1}{6}.$$

This type of distribution is known as a geometric distribution, and in our case, for the coalescent, we have:

$$p(T = t) = \left(1 - \frac{1}{2N}\right)^t \frac{1}{2N}.$$



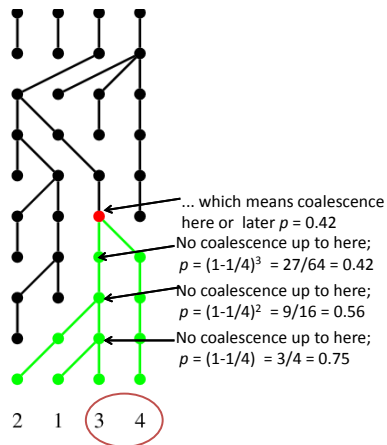
## Coalescent basics IV

Standard coalescent theory applies to continuous-time, rather than the discrete-time of the Wright-Fisher.

We can get to the continuous-time version by assuming that  $N$  is large.

Consider the probability that two genes coalesce after generation  $t$ , without specifying which generation in particular it happened in:

$$P(T > t) = \left(1 - \frac{1}{2N}\right)^t.$$



## Coalescent basics V

For  $N$  large this probability is approximated by

$$P(T > t) = \exp(-t/(2N)).$$

Even for  $N = 10$  this isn't too bad: e.g. for generation 5 we have  $\exp(-5/20) = 0.779$  whereas  $(1 - 1/20)^5 = 0.774$

So then

$$P(T \leq t) = 1 - \exp(-t/(2N)).$$

This is a cumulative distribution function and the corresponding density function (obtained by differentiating with respect to  $t$ ), is just an exponential distribution in which the rate of coalescence is  $1/(2N)$ :

$$\frac{1}{2N} \exp(-t/(2N)).$$

## Coalescent basics VI

Note that we can get this result less rigorously (but possibly more intuitively) by just taking the original geometric distribution:

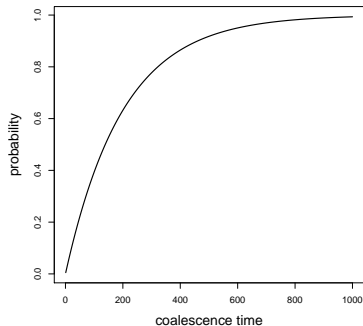
$$\rho(T = t) = \left(1 - \frac{1}{2N}\right)^t \frac{1}{2N}$$

and substituting  $\exp(-t/(2N))$  for  $\left(1 - \frac{1}{2N}\right)^t$ .

# Coalescent basics VII

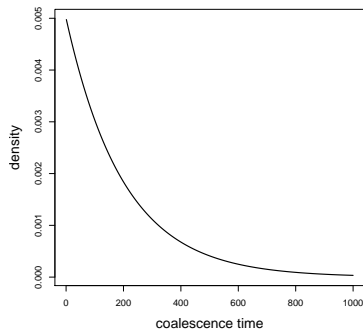
The c.d.f. for  $N=100$

$$P(T \leq t) = 1 - \exp(-t/(2N)).$$



The density function for  $N=100$

$$p(t) = \frac{1}{2N} \exp(-t/(2N)).$$



## Coalescent basics VIII

The mean of this distribution is  $2N$ .

(Think back to getting a six on the throws of the die: on average we expect to hit a six on the sixth throw.)

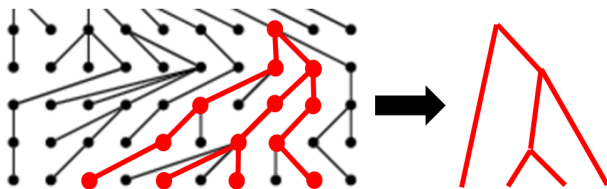
This is the expected time we'd have to wait for coalescence of two lineages.

So we expect the  $T_{\text{MRCA}}$  for two genes chosen at random in the current population to occur  $2N$  generations ago.

## Coalescent basics IX

What is the time until the first coalescence for a sample of size  $k$ ?

As before, we ask what is the probability that any pair of genes has a common ancestor in the previous generation. We assume that  $k$  is much smaller than  $N$ , so only one pair can coalesce at a time.



## Coalescent basics X

So then there are  $k(k - 1)/2$  possible pairs. *I.e.* this is the number of ways of choosing any 2 objects from  $k$  objects, written as  $\binom{k}{2}$ .

If  $k \ll N$  so that we can ignore multiple coalescence then there are  $\binom{k}{2}$  chances of coalescence, each of which occurs with a probability  $1/(2N)$ . So the probability that any pair of lineages coalescence in the previous generation is

$$\binom{k}{2}/(2N).$$

## Coalescent basics XI

Following exactly the same reasoning as for a single pair, for a sample of size  $k \ll N$  and large  $N$  the distribution of waiting times until first coalescence (between any pair of the  $k$  lineages) is approximately exponential with rate:

$$\binom{k}{2}/(2N) = \frac{k(k-1)}{4N}.$$

*I.e.* writing it out in full:

$$p(t) = \frac{k(k-1)}{4N} \exp(-[k(k-1)/(4N)]t).$$



## Coalescent basics XII

### Scaling time in the coalescent

In more formal treatments time is scaled so that one unit of scaled time corresponds to  $2N$  generations, and then we can let  $N \rightarrow \infty$ , and in this limit the (scaled) waiting times are exponentially distributed with rate  $k(k-1)/2$ . (But often easier to keep demographic parameters explicit when we want to use genetic data to infer them.)

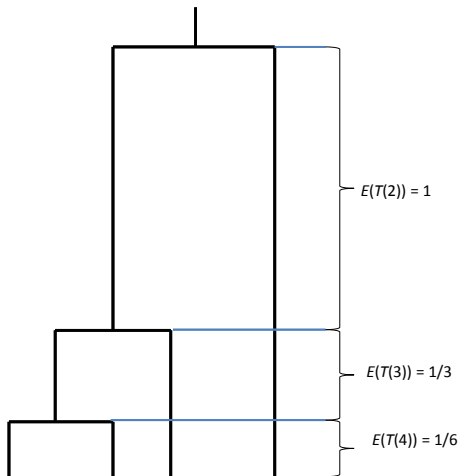
## Modelling a genealogy I

The genealogy consists of  $n - 1$  events with time intervals between events,  $T(k)$ , that depend on the number of lineages  $k = n, \dots, 2$ , where  $n$  is the actual sample size. At each event a randomly chosen pair of lineages coalesce until we reach the most recent common ancestor (MRCA). So we can simulate a genealogy as follows:

1. To begin with we have  $n$  lineages, so set  $k = n$  and  $T_{\text{current}} = 0$
2. We then simulate an exponential random variable  $T(k)$  with rate  $k(k - 1)/2$  (let's assume we are working with scaled time. . .).
3. Choose 2 lineages at random and coalesce them.
4. Set  $T_{\text{current}} = T_{\text{current}} + T(k)$
5. Set  $k = k - 1$
6. Repeat steps 2–5 until  $k = 1$ .
7. You have then created a coalescent genealogy, and the  $T_{\text{MRCA}}$  is given by the value of  $T_{\text{current}}$

## Modelling a genealogy II

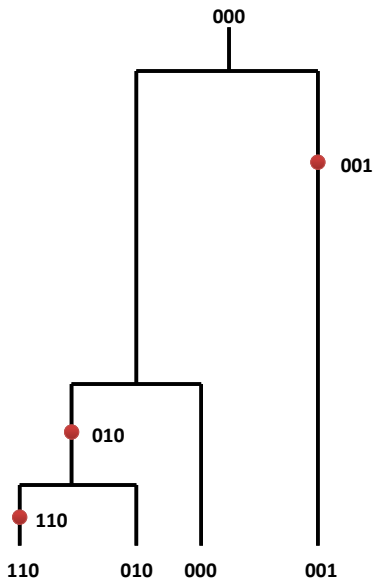
The expected waiting time for coalescence is the reciprocal of the rate. *i.e.* it's the mean of the exponential distribution with that particular rate. (I use  $E()$  to mean 'expectation of'.) So in this example, using scaled time, with 4 lineages the expected coalescence times are respectively  $1/6$ ,  $1/3$  and  $1$ .



## Modelling a genealogy III

Mutations occur with a probability  $\mu$  when the DNA is copied from generation to generation. Because mutation is rare and we are considering many generations, the number of mutations along any lineage is Poisson with mean equal  $\mu T_l$ , where  $T_l$  is the length of the lineage.

In this example there are 3 mutations on the genealogy. We focus only on the variable nucleotides. The ancestral version of the nucleotide is conventionally labelled '0'. When a nucleotide at a particular position mutates the ancestral type '0' changes to a different nucleotide denoted by '1'.



## The time of the most recent common ancestor

We saw previously that for a sample of size two the expected  $T_{\text{MRCA}} = 2N$ ; what is the expected time for a sample more generally?

This can be easily calculated because the expectation of a sum of independent random variables is just the sum of the expectations, *i.e.*:

$$2N \times (1 + 1/3 + 1/6 \dots).$$

If we work it out, then the expected number of generations until the MRCA is:

$$E\left(\sum_{k=2}^n T(k)\right) = 4N(1 - 1/n).$$

This is quite an interesting result because it says that as the sample size gets larger the expected  $T_{\text{MRCA}} = 4N$  — just twice as long as for a random pair of lineages.

## Measuring it in scaled time:

The expected scaled time for any two genes to coalesce is 1.

The expected scaled time until the MRCA is:

$$E\left(\sum_{k=2}^n T(k)\right) = 2(1 - 1/n)$$

Mutations occur along each branch at a rate  $\theta/2$ , where  $\theta = 4N\mu$  and can be superimposed on the (scaled) genealogy.

So the expected number of mutations in the genealogy of a sample of size two is  $\theta$ .

## The number of mutations in a genealogy

- ▶ The expected number of mutations in a genealogy will be equal to the mutation rate,  $\mu$ , multiplied by the total sum of branch lengths,  $L$ .
- ▶ The sum of the branch lengths,  $L$ , has expectation

$$E(L) = 2N \times \left( 2 + 3 \times \frac{2}{3 \times 2} + 4 \times \frac{2}{4 \times 3} + \dots \right) = 4N \sum_{i=1}^{n-1} \frac{1}{i}$$

- ▶ So the expected number of mutations is

$$\theta \sum_{i=1}^{n-1} \frac{1}{i}$$

## The infinite sites mutation model I

- ▶ The mutation rate per nucleotide site is typically so low compared to the population size that to a very good approximation we can assume that no mutation hits the same site twice in a genealogy.
- ▶ So,  $S$ , the number of single nucleotide polymorphisms (SNPs) in a section of sequence, gives an estimate of  $\theta$ :

$$\hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}.$$

This is known as Watterson's estimator.



## The infinite sites mutation model II

- ▶ Another estimator for  $\theta$  can be obtained by taking all pairs of sequences from the sample, and counting the number of pairwise differences in the two sequences. The average of this over all  $n(n - 1)/2$  pairs is usually called  $\pi$ , the average pairwise difference. As we have seen, the expected number of mutations between any pair of sequences is  $\theta$ , giving another estimator  $\hat{\theta}_\pi = \pi$ .
- ▶ So, in principle, if we know the mutation rate, we can obtain an estimate of  $N$  from the nucleotide data, using either of these two estimators.

# The generality of the coalescent model I

We have derived the coalescent from consideration of the Wright-Fisher model, which may seem rather contrived.

In fact a lot of (very mathematical) work has been carried out to examine the limiting behaviour of many life-history models (overlapping generations, inbreeding, non-random mating, variance in reproductive success), as you let  $N$  get large, and they generally tend to the coalescent (with some exceptions. . .).

Intuitively you can see why this might be so: if the sample size is very small compared to the population size, and the population size is very large, then there are many generations between coalescent events, and the nitty-gritty details of what happens in these generations gets averaged out.

## The generality of the coalescent model II

However although the genealogies follow the coalescent predictions very closely, the scaling of time may vary between life-history models. For example, when there is a variance  $\sigma^2$  in the number of offspring among individuals, we need to work with  $2N/\sigma^2$ . But since we normally don't know what  $\sigma^2$  is, it just gets parcelled up in  $N$ . So  $N$  becomes an “effective population size” and may have very little relation to any census value of population size that would be measured by an ecologist.

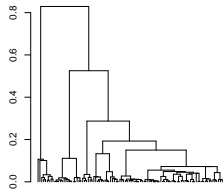
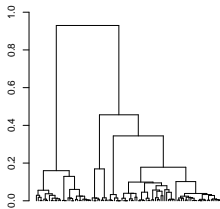
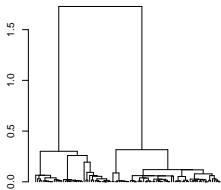
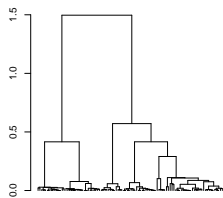
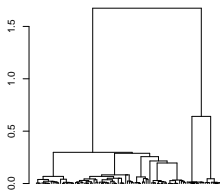
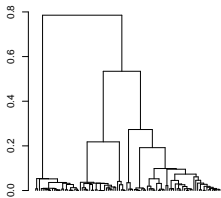
Hammering home the point: if you have a very good estimate of the mutation rate, and good sequence data, you can get a good estimate of  $N$ , but only if the population behaved like the idealised Wright-Fisher model would the value of  $N$  you estimate have any bearing on the actual number of individuals in the population.

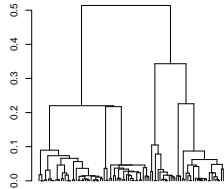
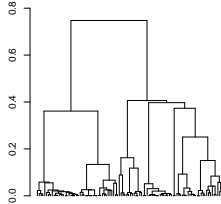
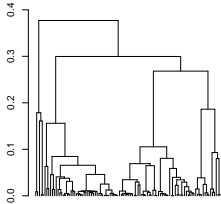
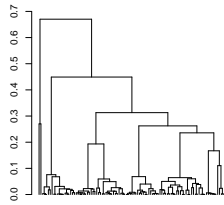
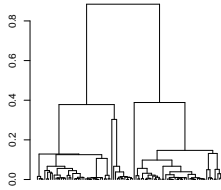
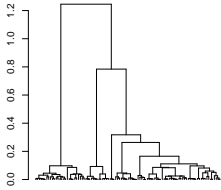
## The variability of coalescent genealogies

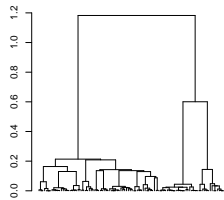
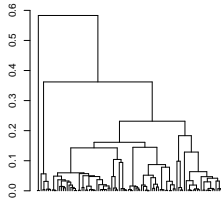
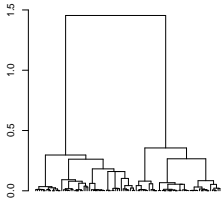
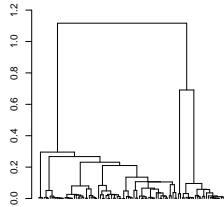
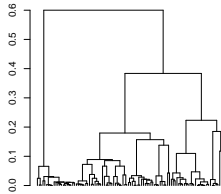
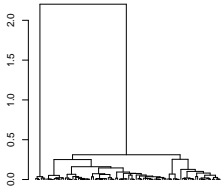
It is very easy to simulate genealogies. I have made available on Blackboard an R script and some instructions on how to simulate and visualise genealogies in the R statistical programming environment using a very famous program (among population geneticists . . .) called *ms* ('make samples'), written by Dick Hudson, one of the founders of coalescent theory. (Don't worry this is for your interest only; you won't be examined on it).

The next few slides will give examples of genealogies for a sample of size 100 using this program.

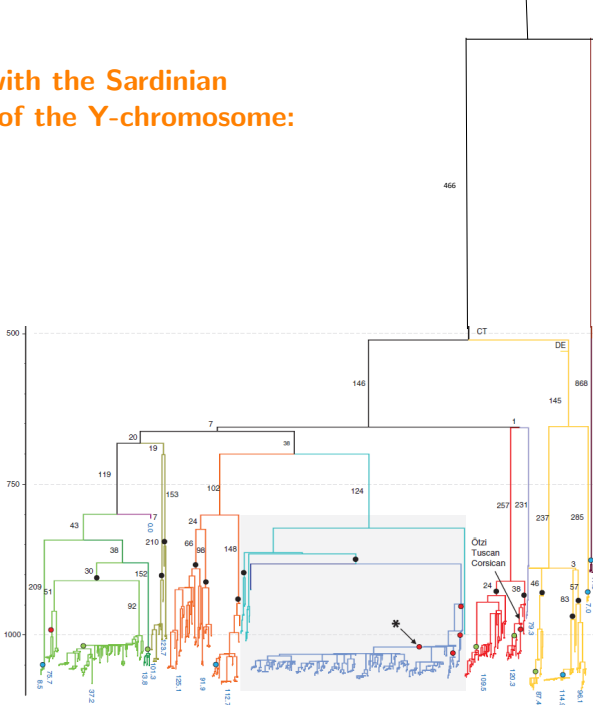
In the following slides note that *ms* scales time by  $4N$  rather than  $2N$ , so the expected height of the tree is 1 rather than 2 for a large sample.







# Compare with the Sardinian genealogy of the Y-chromosome:









## The next lecture





We will revisit the Y-chromosome papers of the previous lecture, and try and interpret some of the results in terms of the theory developed here.

We will then look at the effect of variable population size on genealogies.

## Further Reading I

-  Hein, J., Schierup, M., and Wiuf, C.  
Gene Genealogies, Variation and Evolution: A primer in coalescent theory.  
Oxford University Press.
-  Donnelly, P., and Tavaré, S. (1995)  
Coalescents and genealogical structure under neutrality.  
*Annual review of genetics*, 29(1), 401–421.
-  Hudson, R. R. (1990).  
Gene genealogies and the coalescent process.  
*Oxford surveys in evolutionary biology*, 7, 1–44.
-  Hudson, R.R. (1983)  
Testing the constant-rate neutral allele model with protein sequence data.  
*Evolution*, 37, 203-217

## Further Reading II

-  Kingman, J. F. C. (1982)  
The coalescent.  
*Stochastic Processes and their Applications* 13, 235–248.
-  Kingman, J. F. C. (1982)  
On the genealogy of large populations.  
*Journal of Applied Probability*, 19A, 27–43.
-  Rosenberg, N. A., and Nordborg, M. (2002)  
Genealogical trees, coalescent theory and the analysis of genetic polymorphisms.  
*Nature Reviews Genetics* 3(5), 380–390.
-  Tajima, F. (1983)  
Evolutionary relationship of DNA sequences in finite populations.  
*Genetics*, 105, 437-460.