EMBO Course in Population Genomics, practicals on Coalescent Theory

Question 1. **Group 1**
The coalescence times T_i in the standard coalescent process are independent and exponentially distributed as

$$f_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}\ t}$$

when time is measured appropriately.

(a) What is the distribution of T_2, T_3, and T_4?
Answer:

$$f_{T_2}(t) = \binom{2}{2} e^{-\ \binom{2}{2}\ t} = e^{-t},$$

$$f_{T_3}(t) = \binom{3}{2} e^{-\ \binom{3}{2}\ t} = 3\ e^{-3\ t},$$

$$f_{T_4}(t) = \binom{4}{2} e^{-\ \binom{4}{2}\ t} = 6\ e^{-6\ t}.$$

(b) Write down the average and variance of waiting times T_2, T_3, and T_4. Also, write down the value of $E[T_2^2]$, $E[T_3^2]$, and $E[T_4^2]$. Answer:

$$E[T_2] = 1,\ E[T_3] = \frac{1}{3},\ E[T_4] = \frac{1}{6}.$$

$$Var[T_2] = 1,\ Var[T_3] = \left(\frac{1}{3}\right)^2 = \frac{1}{9},\ Var[T_4] = \left(\frac{1}{6}\right)^2 = \frac{1}{36}.$$

By definition,
$E[i^2] = Var[i] + (E[i])^2$, therefore,
$E[T_2^2] = 2$, $E[T_3^2] = \frac{2}{9}$, and $E[T_4^2] = \frac{2}{36}$.

(c) If one infers that $T_3 > T_4 > T_2$ upon looking at a gene genealogy for a sample of size 4, would this be consistent with the standard coalescent process? Justify your answer. **Answer:** Yes, it would be consistent with the standard coalescent process. T_2, T_3, and

T_4 are random variables. So, for some genealogies, $T_3 > T_4 > T_2$ holds. Though $E[T_2] > E[T_3] > E[T_4]$ will always be true.

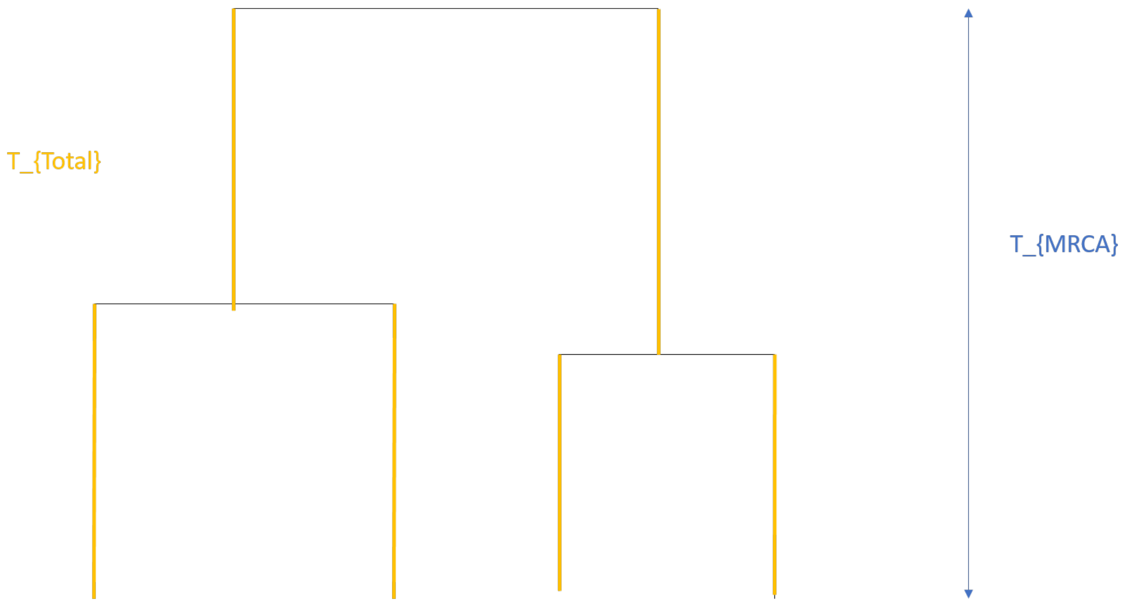
(d) Does $\binom{i}{2}$ in the expression given above have some meaning in standard coalescent process (Kingman's coalescent)? If yes, what is it? **Answer:** This quantity represents the

rate of coalescence events. Since only two lineages merge at the same in the standard coalesence, it is equal to $\binom{i}{2}$ the number of pairs among i lineages.

Question 2. **Group 2**

T_{MRCA} is the time to the most recent common ancestor of the entire sample and T_{total} is the total length of all the branches in the genealogy.

(a) Draw a coalescent tree with sample size $n = 4$ and indicate on the tree what T_{MRCA} and T_{total} correspond to.



(b) Find $E[T_{MRCA}]$ and $E[T_{total}]$ for a sample of size n. Write down all the steps involved.

Answer:

$$T_{MRCA} = \sum_{i=2}^{n} T_i$$

where n is the sample size, since T_i are independent random variables, therefore,

$$E[T_{MRCA}] = \sum_{i=2}^{n} E[T_i],$$

$$E[T_{MRCA}] = \sum_{i=2}^{n} \frac{2}{i(i-1)} = 2\sum_{i=2}^{n} \Big(\frac{1}{i-1} - \frac{1}{i}\Big),$$

$$E[T_{MRCA}] = 2\Big(\frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} + ... - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n}\Big) = 2\Big(1 - \frac{1}{n}\Big).$$

$$T_{total} = \sum_{i=2}^{n} i\ T_i,$$

using $E[cX] = c\ E[X]$,

$$E[T_{total}] = \sum_{i=2}^{n} i\ E[T_i],$$

$$E[T_{total}] = \sum_{i=2}^{n} \frac{2\ i}{i(i-1)} = \sum_{i=2}^{n} \frac{2}{(i-1)}.$$

(c) Calculate the value of $E[T_{MRCA}]$ as n goes to infinity.

Answer:

$$\lim_{n\to\infty} E[T_{MRCA}] = 2$$

(d) Write down the $E[T_{MRCA}]$ and $E[T_{total}]$ when sample size is 4. Also, compute $\frac{E[T_{MRCA}]}{E[T_2]}$ (definition of T_2 is given in the previous question).

Answer:

$$E[T_{MRCA}] = 2\Big(1 - \frac{1}{n}\Big) = 2\Big(1 - \frac{1}{4}\Big) = \frac{3}{2} = 1.5,$$

$$E[T_{total}] = \sum_{i=2}^{4} \frac{2}{(i-1)} = \frac{11}{3} = 3.67.$$

$$\frac{E[T_{MRCA}]}{E[T_2]} = \frac{1.5}{1} = 1.5.$$

(e) Can $\frac{E[T_{MRCA}]}{E[T_{total}]} > 1$ be true ? Justify your answer.

No, we have that $E[T_{total}] < E[T_{MRCA}]$ cannot be true. The branches that are summed to get to T_{MRCA} are included in the sum of the T_{total} calculation.

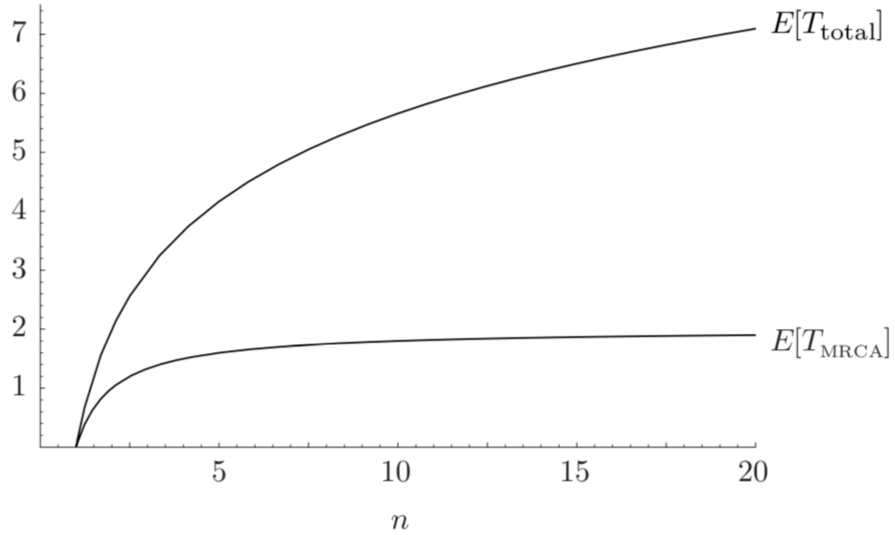


Figure 1: Reference: An Introduction to Coalescence Theory by John Wakeley.

Conclusion: $E[T_{total}] > E[T_{MRCA}]$ **will always be true.**

Plot $E[T_{MRCA}]$ and $E[T_{total}]$ as a function of n.

Question 3. **Group 3** Consider the data below, where n=10 haploid sequences were simulated. Each line represents an individual and each column a segregating site S. The ancestral allele is denoted by "0" and the derived allele is denoted by "1." We observe a total of S=20 segregating sites.

```
00001000010000000000
00000000100100010011
00000000100100010011
00000000000101110001
00000000000101110001
10000000000101010001
01010111001010011001
10000000000101010001
00100000100100010111
00000000000101110001
```

(a) What is the frequency of the **derived** allele in each segregating site?
Answer: 2 1 1 1 1 1 1 1 3 1 1 8 1 5 3 9 1 1 3 9

(b) Compute the SFS for the data above, using the allele frequencies you just computed.

Answer: We consider the frequencies of the derived allele (1) and count the number of derived alleles at each site to get the following:

Derived allele count	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
# of sites	12	1	3	0	1	0	0	1	2

Table 1: Unfolded SFS

The Figure below shows two Site Frequency Spectra (SFS). The red distribution is the expected SFS for a given value of θ for a population of constant size, not subdivided (geographically or otherwise), and where mutations have no effect in fitness. Mutation rate per base pair is constant and loci are not linked. The blue distribution is the observed SFS for a population simulated under neutrality using *ms* (Hudson 1990).

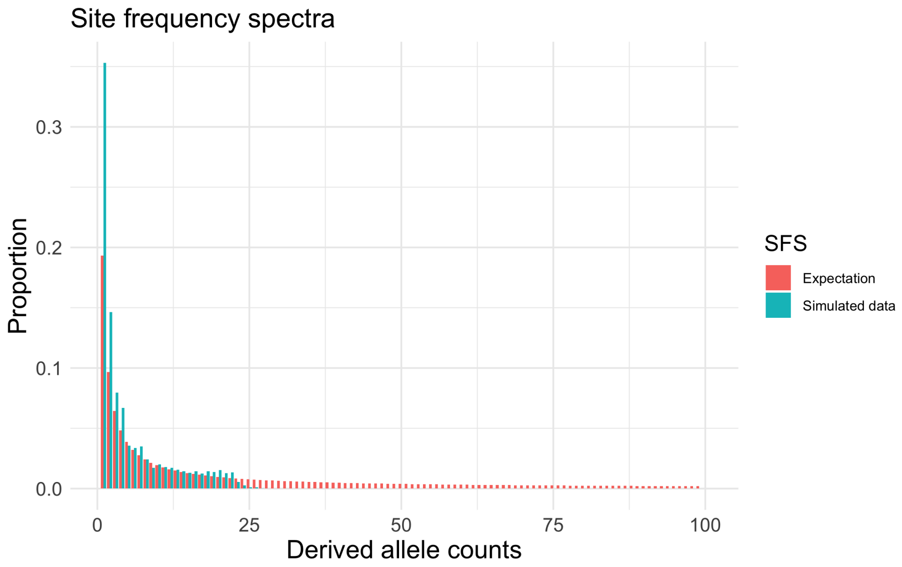


Figure 2: Site frequency spectra for two scenarios: a simulated dataset, and the expected SFS for a population with constant population size.

(c) For the red curve (expectation with constant N_e), we see 800 singletons. What is the expected number of doubletons?

Answer: $\mathrm{E}[S_1] = \theta = 800$. $\mathrm{E}[S_2] = \theta/2 = 400$. We expect to see 400 doubletons.

(d) We can see that the SFS for the simulations (blue) departs from the expectation (red). Besides some statistical noise, what is the most striking difference between these two distributions? Please use maximum two sentences to answer this question.

Answer: The most striking difference is that the blue curve has more than 50% singletons than the red curve. In other words, there seems to be an excess of singletons in the simulations relative to the expectation.

(e) Given that the same value of θ was used to generate both distributions, can you think of an explanation for the observed differences? Please use maximum three sentences to answer this question.

Answer: The explanation is that the simulated population went through some sort of demographic change. Here, the population was simulated with a bottleneck and exponential growth.

Question 4. **Group 4** Write a small program to simulate allele frequencies through time under a Wright-Fisher model for the simplest case, e.g. neutral alleles, constant (haploid) population size. Plot the trajectories. Try out several parameters and make observations. For instance: what happens when the population size increases or decreases?

```
# WFsim.r  simple simulations of the Wright-Fisher model.
# Code by Andy Clark (2018, EMBO course).

# Hints:
# Initialize popsize (N), number of samples (nsamp), number of
# generations (ngen) and starting frequency (startfreq) and
# keep track of the current allele frequency (pcur)
# Remember that Wright-Fisher is simply recurrent binomial sampling o
# over generations
```

Answer:

```
# WFsim.r  simple simulations of the Wright-Fisher model.
# Code by Andy Clark (2018, EMBO course).

N<-200
nsamp<-100
ngen<-400
startfreq<-.5

x<-rbinom(nsamp,N,startfreq)
p<-x/N
pcur<-p

for (i in 1:ngen){
x<-rbinom(nsamp,N,p)
p<-x/N
pcur<-rbind(pcur,p)
}

#Now plot these trajectories
gen<-seq(1,ngen+1)
for (k in 1:nsamp){
plot(gen,pcur[,k],type="l",ylim=c(0,1))
par(new=TRUE)
}
```