EMBO Course in Population Genomics, practicals on Coalescent Theory

Question 1. **Group 1**
Consider a diploid population of size N (with $2N$ chromosomes) that has two alleles, C and c, segregating in i and $2N - i$ copies respectively, at $t = 0$. The population is evolving following the standard Wright-Fisher model.

(a) Calculate the probability that allele c will be present in k copies in the first generation ($t = 1$).

(b) What is the probability of observing allele C in $2N$ copies in the first generation ($t = 1$).

(c) What is the probability that a sample of size $n = 2$ taken from generation t share the same parent at $t - 1$?

(d) With what probability k offspring at generation t share the same parent at $t - 1$?

Question 1. **Group 2**
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?
(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 $E[T_4^2]$.
(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.
(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?

Question 3. **Group 3**
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.

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

(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).

(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.

Question 4. **Group 4** 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?

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

The Figure 2 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).

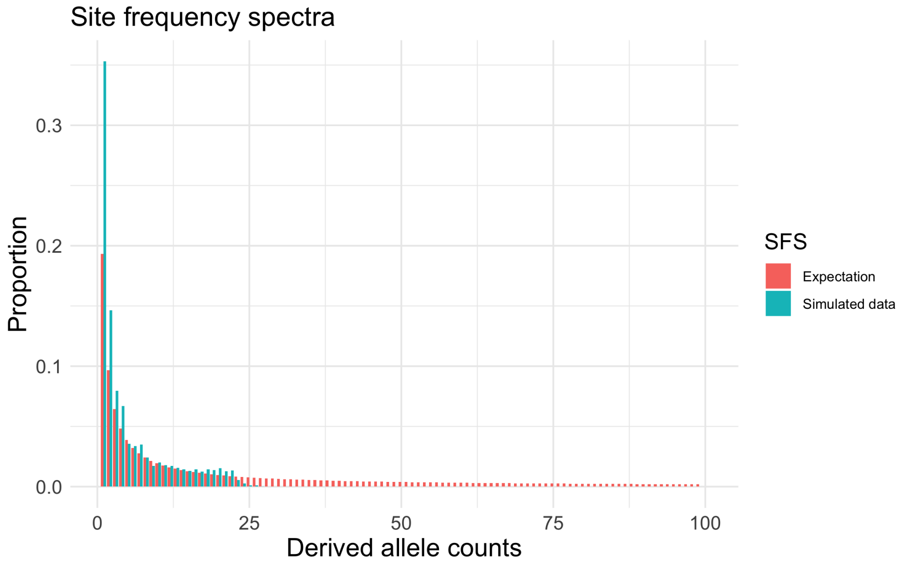

Figure 1: Site frequency spectra for two scenarios: a simulated dataset, and the expected SFS for a population with constant population size.

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

(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.

(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.

OPTIONAL. **Group 1+2+3+4** Run the small program to simulate allele frequencies through time under a Wright-Fisher model for the simplest case, e.g. neutral alleles, constant population size. TIf you still have time, try the assignments below as well.

```
# WFsim.r  simple simulations of the Wright-Fisher model.
# Code used by Andy Clark in 2018.

# Initialize popsize (N), number of samples (samp), number of
# generations (ngen) and starting frequency (startfreq)

# Note this is for a haploid population of size N.

N<-200
nsamp<-8
ngen<-200
startfreq<-.5
pcur<-matrix(1:N)

#Wright-Fisher is simply recurrent binomial sampling over generations

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)
}


# ASSIGNMENT 1: Conduct runs of the above simulation, and
# calculate the variance in allele frequencies across 20
# independent sample populations for each of generations 20
# through 200 in steps of 20.  Plot the results.
# Hint: the variance of allele freq at gen i is var(pcur[i,])




# ASSIGNMENT 2: Plot the expected time to fixation (when allele
# frequency is either 0 or 1) for 20 populations with population
# sizes 50, 100, 150 and 200, starting from p = 0.5.
```