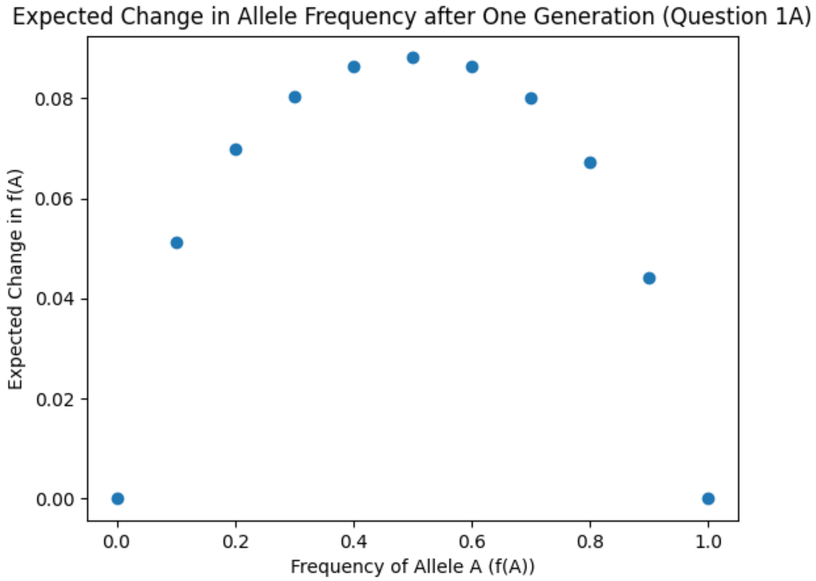
Note: Here is the link to my work: https://github.com/vhaghani26/GGG-201D

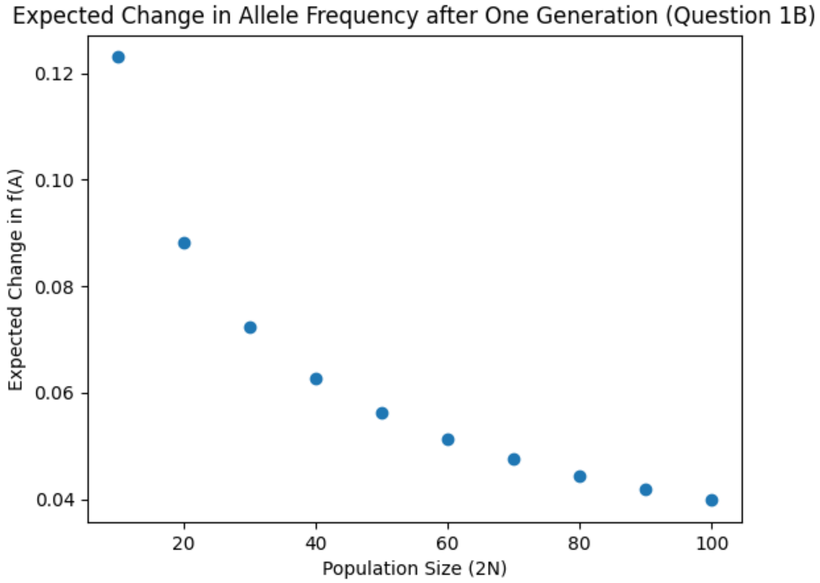
Question 1

A) Use a program such as R or Excel to generate a scatter plot that shows how expected allele frequency change from genetic drift depends on initial allele frequency. The x-axis should be initial allele frequency and range from 0 to 1. The y-axis should be expected change in allele frequency after one generation. Perform calculations in steps of 0.1 for a population size of 2N=20.

Note: I couldn't figure out how to write the probability matching for expected frequencies into my script, so I just used the data from the script in Excel to do the last part.



B) Use the same program to generate a scatter plot that shows how expected allele frequency change from genetic drift depends on population size. The x-axis should be population size and range from 2N=10 to 2N=100. The y-axis should be expected change in allele frequency after one generation. Perform calculations in steps of 10 with an allele frequency of 0.5



Question 2

You sequence a locus in three individuals from a population and obtain the below data:

I1: GCTACTTTACCATTCTCAGCGAGACGTAAGATCAGGCCAGATCCACCTCG
GTTCCTTTAACATTCTCAGCGAGACGTAAGAGCAGGCCAGATCCACCGCC

I2: GCTCCTTTACCATCCTCAGCGAGACGTAAGATCAGGACAGATCCACCTCG
GCTACTTTACCATCCTCAGCGACACGTAAGATCAGGCCAGATCCACCTCG

I3: GCTACTTTACCATCCTCAGCGAGACGTAAGAGCAGGCCAGATCCACCTCC
GCTACTTTACCATCCTCAGCGAGACGTAGGAGCAGGACAGATCCACCTCG

A) How many segregating sites (s) are present in these data?

Ind	**Sequence**	**Gene Copy**
I1	GCTACTTTACCATTCTCAGCGAGACGTAAGATCAGGCCAGATCCACCTCG	A
	GTTCCTTTAACATTCTCAGCGAGACGTAAGAGCAGGCCAGATCCACCGCC	B
I2	GCTCCTTTACCATCCTCAGCGAGACGTAAGATCAGGACAGATCCACCTCG	C
	GCTACTTTACCATCCTCAGCGACACGTAAGATCAGGCCAGATCCACCTCG	D
I3	GCTACTTTACCATCCTCAGCGAGACGTAAGAGCAGGCCAGATCCACCTCC	E
	GCTACTTTACCATCCTCAGCGAGACGTAGGAGCAGGACAGATCCACCTCG	F

There are **10** segregating sites present.

B) What is pi (π) in these data?

AB	AC	AD	AE	AF	BC	BD	BE	BF	CD	CE	CF	DE	DF	EF	Sum
6	3	2	3	4	7	8	5	8	3	4	3	3	4	3	66

$$\pi = \frac{\text{sum of pairwise differences}}{\text{number pairwise comparisons}} = \frac{6}{15} = \mathbf{4.4}$$

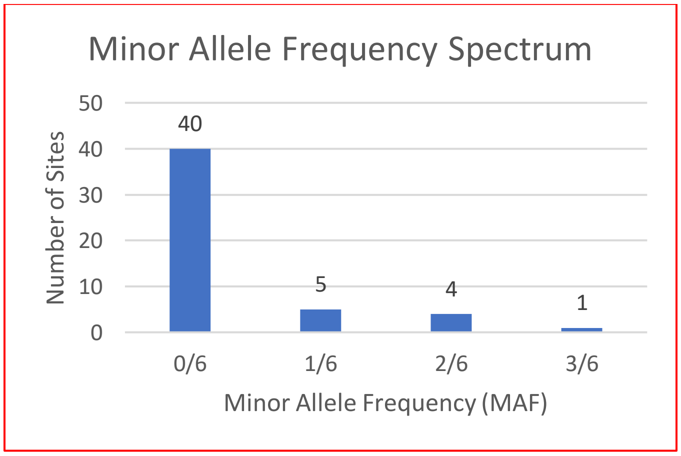
C) What are s and π expressed in per site values?

$$\pi_{\text{per site}} = \frac{\pi}{\text{number of sites}} = \frac{4.4}{50} = \mathbf{0.088}$$

$$s_{\text{per site}} = \frac{s}{\text{number of sites}} = \frac{10}{50} = \mathbf{0.2}$$

D) What is the minor allele frequency spectrum for these data?

s	1	2	3	4	5	6	7	8	9	10
MAF	1/6	2/6	1/6	2/6	1/6	1/6	3/6	2/6	1/6	2/6



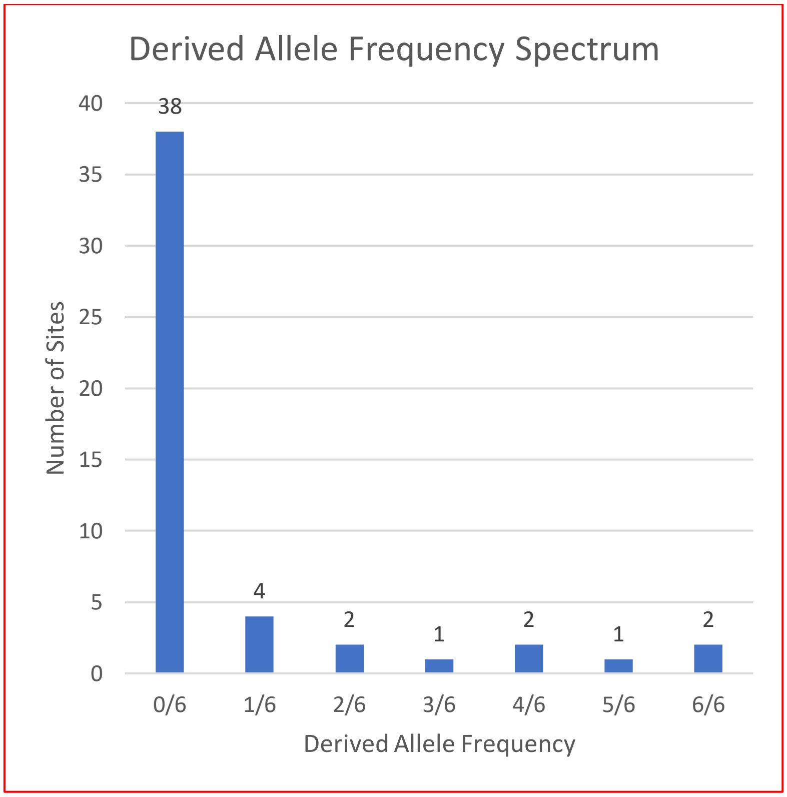
You next sequence the locus in a few closely related species and determine the ancestral sequence to be the following.

Ancestral: GCTCCTTTACCATCCTCAGGGACACGTAAGAGCAGGCCAGACCCACCTCC

E) What is the derived allele frequency spectrum for these data?

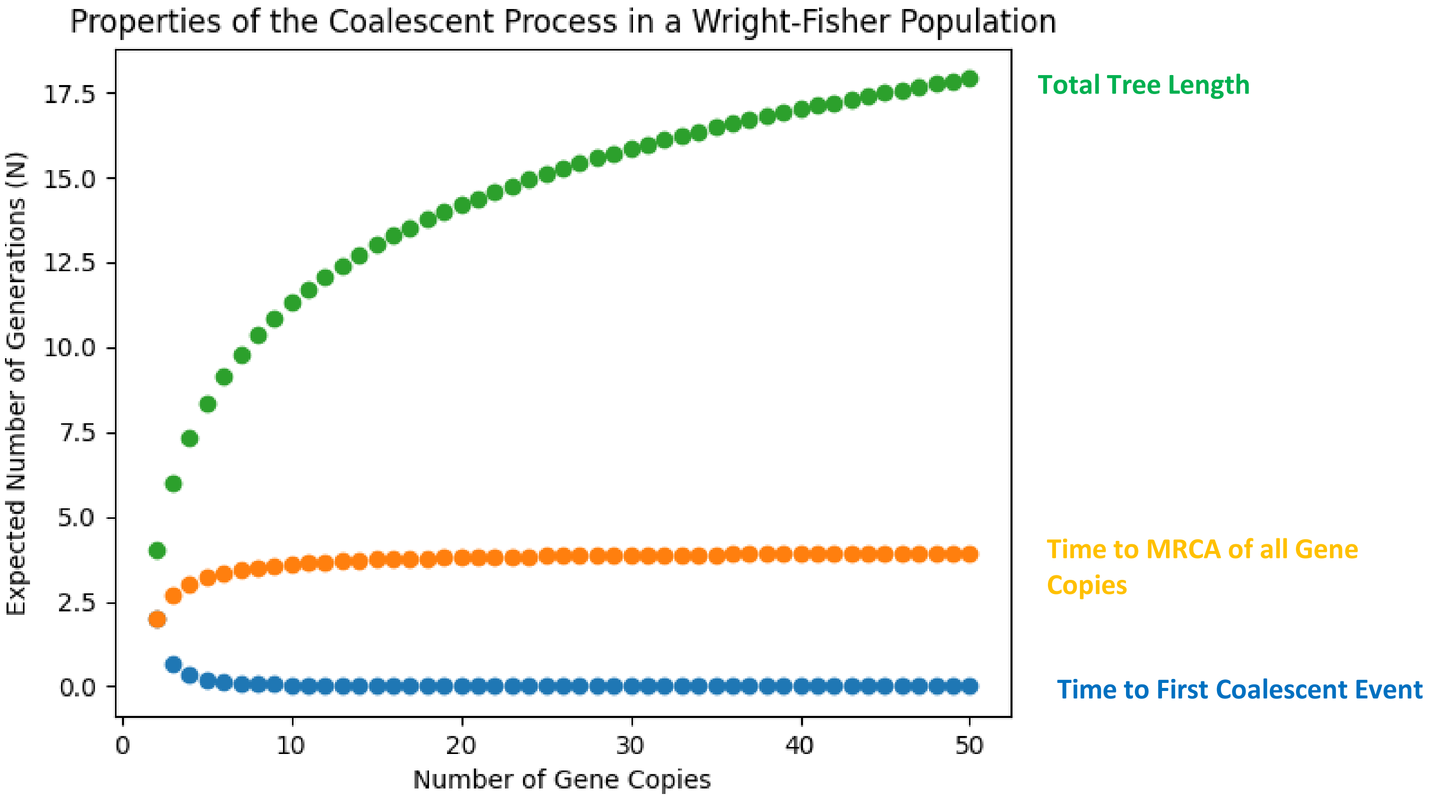
s	1	2	3	4	5	6	7	8	9	10
DAF	1/6	4/6	1/6	2/6	5/6	1/6	3/6	2/6	1/6	4/6

There are also two sites (first is between s4 and s5, second is between s8 and s9) where all of the derived sequences have the same allele at that site, but the sites are different from the ancestral, resulting in 6/6 at both these sites.



Question 3

Use a program such as R or Excel to generate a scatter plot that shows the properties of the coalescent process in a Wright-Fisher population. The x-axis should be number of gene copies and range from 2- 50. The y-axis should be expected number of generations in N units. Perform calculations in steps of one gene copy and plot the following three expectations: (1) time to the first coalescent event; (2) time to the most recent common ancestor of all gene copies; (3) total tree length


Properties of the Coalescent Process in a Wright-Fisher Population
Expected Number of Generations (N)
Number of Gene Copies
Total Tree Length
Time to MRCA of all Gene Copies
Time to First Coalescent Event