

Population Genetics

Pejman Mahboubi¹, John Seger², Allen Rogers³

¹Department of Mathematics, UCLA

²Department of Biology, University of Utah

³Department of Anthropology, University of Utah

Genetic Drift, Heterozygosity, and Fixation

- Genetic drift or allelic drift is the change in the relative frequency with which a gene variant (allele) occurs in a population due to random sampling and chance.
- Define the homozygosity G to be the probability that two alleles drawn at random from the population without replacement are identical. And define the heterozygosity H by $H=1-G$ Then

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t.$$

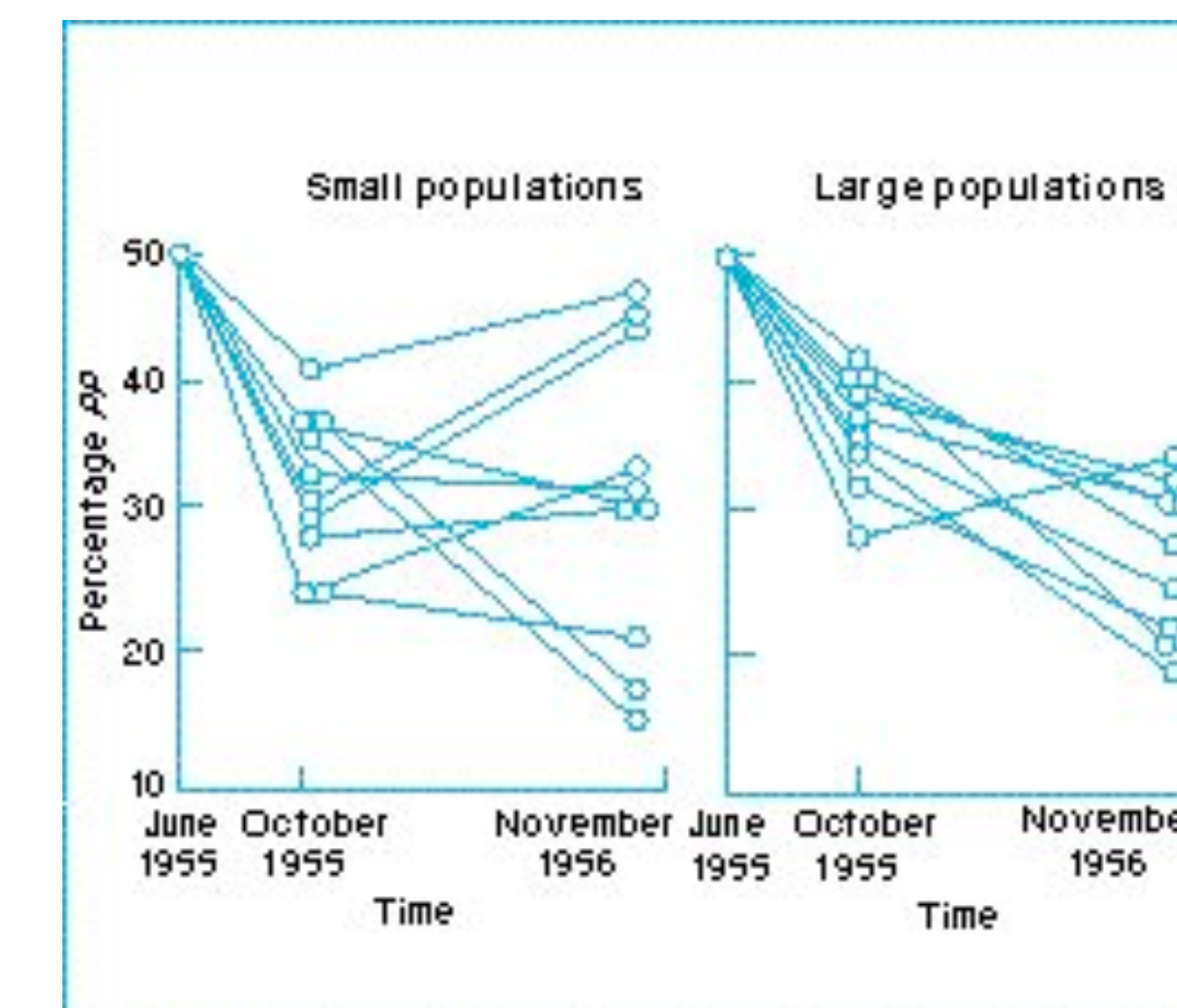
In the lack mutation H_t goes to zero as t goes to infinity, while in the presence of mutation it approaches an equilibrium H which is obtained from:

$$H = \frac{4Nu}{4Nu + 1} = \frac{\theta}{\theta + 1}.$$

- While genetic removes variation from the population, mutation restores the genetic variation. If the rate of mutation is u per generation, then in time L the expected number mutations S will be Lu

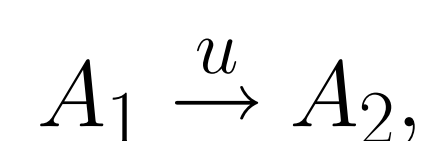
- If we ignore mutation, genetic drift removes the genetic variation from the population. When the relative frequency of one allele is 1, then that allele is fixed. The probability that an allele will be fixed is

$$p = \frac{i}{2N}.$$

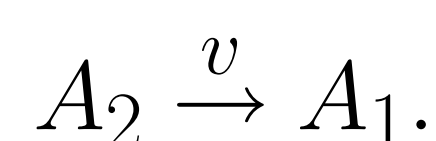


The Stationary Distribution

Under reversible mutation, the A_1 allele mutates to the A_2 allele with rate u



and A_2 mutates to the A_1 allele with rate v



When genetic drift is added to the model, it will cause p to jump around at random. The variance in the change is given by

$$v(p) = \frac{pq}{2N_e},$$

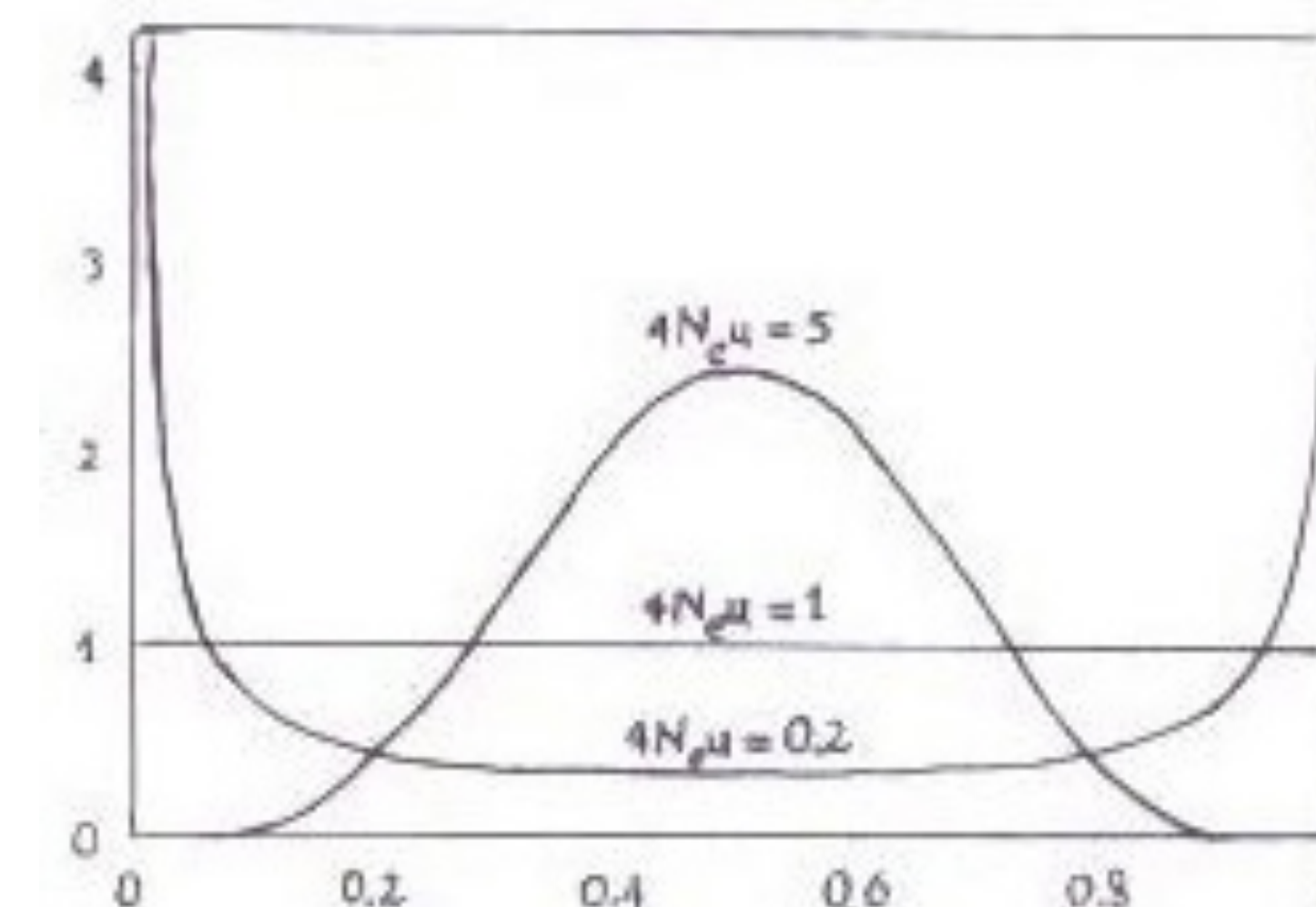
and the change in p in a single generation is

$$m(p) = \Delta p = -up + (1-p)v,$$

In the 1930s Sewall Wright showed that the density $f(p)$ of the p is given by

$$\phi(p) = c \frac{e^{2 \int_0^p (m(x)/v(x)) dx}}{v(p)},$$

where c is just a constant that makes the whole probability equal to 1. The graph of the density is shown in the picture for different values of $4Nu$

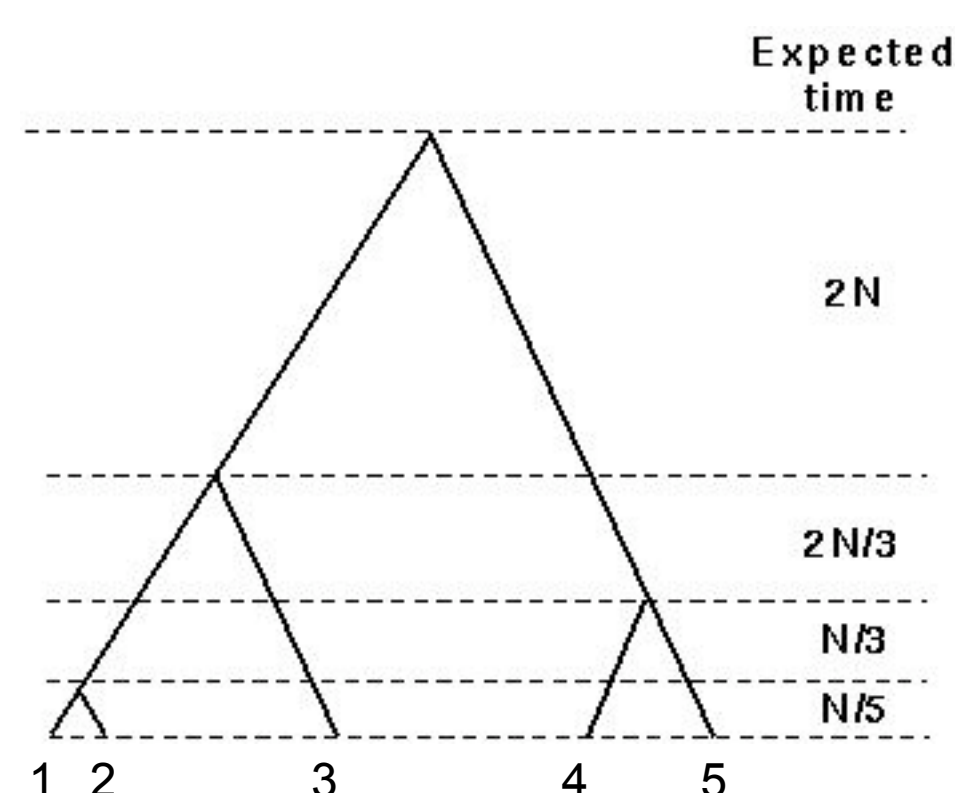


The Coalescent

- A coalescent is the lineage of alleles in a sample of size n traced backward in time to their common ancestor allele.
- The average total length T of the tree is

$$T = 2N,$$

where N is the total population.



- Let t_i be the time interval in which there are i lineages in the coalescence tree. Then

$$E[t_i] = \frac{N}{i(i-1)}$$

- Then the total time L that is spent in the tree is

$$E[L] = \sum_{i=2}^n i E[t_i] = N \sum_{i=1}^{n-1} \frac{1}{i} = Na, \quad \text{where } a = \sum_{i=1}^{n-1} \frac{1}{i}.$$

- Then the expected number of total mutations in the tree S is

$$E[S] = \theta a, \quad \text{where } \theta = Nu$$

The Site Frequency Spectrum

For a sample of size n write M_j for the number of sites at which exactly j individual carry the mutation, then the vector (M_1, \dots, M_n) is called the site frequency spectrum and

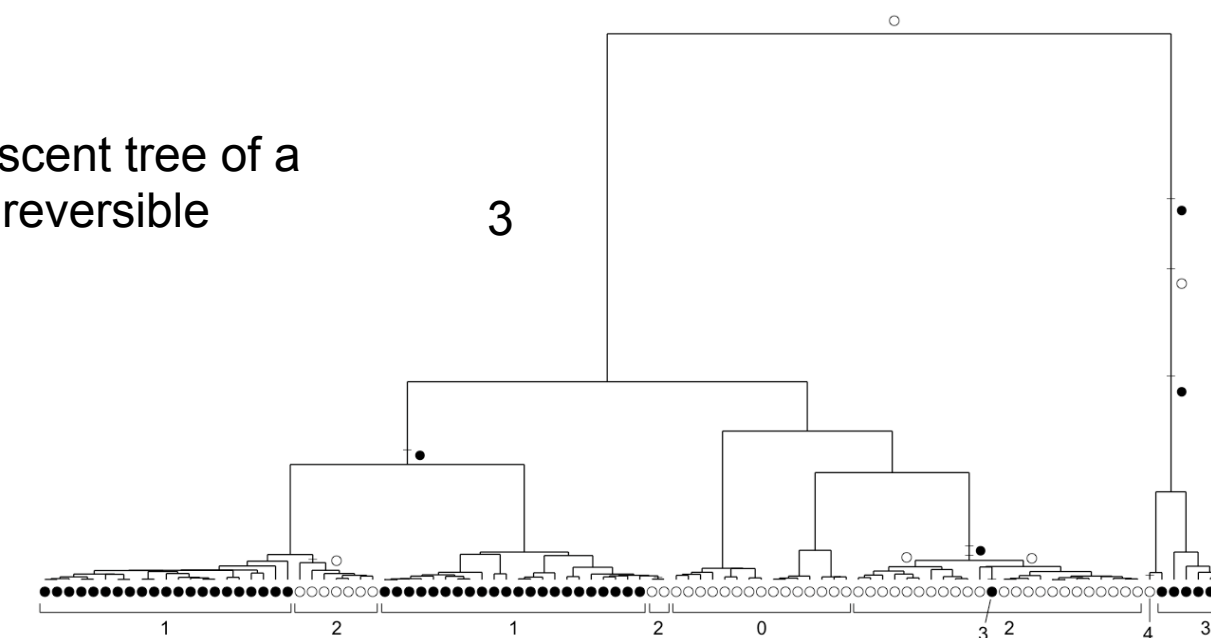
$$EM_j = \frac{\theta}{j}.$$

- Since the number of mutations S is a Poisson random variable with parameter ES , then since for a usual sample size the finite harmonic sum is of order 5

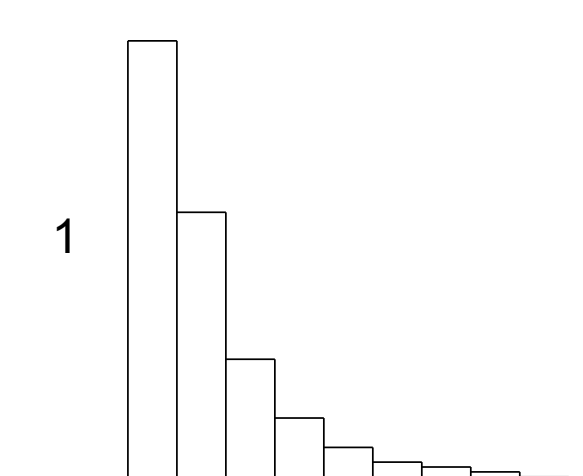
$$P(S=0) = e^{-\theta a} \sim 1 \quad \text{for } \theta \ll 1,$$

$$P(S=1) = e^{-\theta a} \theta a \ll 1 \quad \text{for } \theta \ll 1.$$

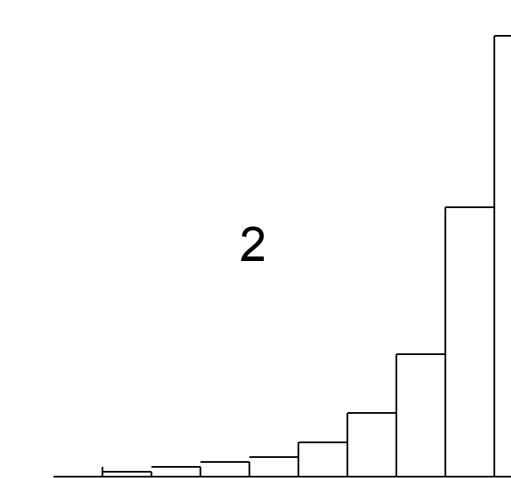
Picture 3 shows a coalescent tree of a sample of size 100 with reversible mutations.



If $S=0$, given the original state is A_1 , then $p=1$. When $S=1$, then using the site frequency spectrum (SFS) we compute the probability distribution of having $j=1, \dots, n$ of A_1 allele. Indeed, when $S=1$, the probability of a mutation being a singleton is same as the probability that only one A_2 allele exists. This gives us the graph 1.



- If we repeat the same computation, given the original state is A_2 , we will get graph 2.
- Averaging this two graphs, gives us an approximation of the Sewall Wright's density curve for the case that θ is small.



Generalization, and future Plans

- when θ is larger, this method breaks down, because in the conditioning on S requires including more terms, i.e. terms corresponding to $S=2$ larger, in our computation. Unfortunately, the current SFS won't be longer useful, as it because then it won't be the frequency of either A_1 or A_2 .

- One way to overcome this obstacle is to try to find a generalization of the SFS to more than one mutation.

- Figure 4 shows the linear relation between the number of the tips of the sub-tree and the length of the subtree. This is an interesting relation which we worked on it, and hope it provides a solution to the general approximation of Sewall Wright's Formula. We didn't prove this phenomenon yet.

