Chapter 2.1-2.2

Siyi Wang

February 19, 2025

1 Chapter 2.1

1.1 Genetic Drift and Allele Dynamics

Genetic drift occurs randomly and affects the frequency of alleles in a population. New mutations generally vanish within a few generations; only in very rare cases do mutations, by chance, increase in frequency within the population. Most SNPs in an individual are **ancestral alleles**mutations that originated hundreds of thousands of years ago in ancestors living in sub-Saharan Africa. At the site of a mutation, the new allele is referred to as the **derived allele**.

1.1.1 Initial Frequency of Derived Alleles

Let N denote the number of individuals in a population. The initial frequency p of a derived allele is given by:

$$p = \frac{1}{2N},$$

where the factor of 2 accounts for the fact that chromosomes occur in pairs.

Mathematically, genetic drift will eventually drive the allele frequency p to either 0 (loss) or 1 (fixation). In other words, new alleles are typically lost within a few generations, while fixation requires thousands of generations.

1.1.2 Wright-Fisher Model

The Wright-Fisher model first assumes a population with a fixed number of individuals, N, meaning that there are $2N$ copies of alleles at each locus. This model is based on two key assumptions:

1. The population has discrete generations and the size of each generation is constant.

2. Individuals mate randomly, and the alleles in the next generation are obtained through random sampling.

These assumptions ensure fair and random allele transmission.

To simulate this random transmission, binomial sampling is used to generate the genotypes of the next generation. In this process, one allele is sampled (with replacement) at a time, and this is repeated $2N$ times to form a new generation. The variance of the allele frequency in the next generation, p_1, is given by:

$$\text{Var}(p_1) = \frac{p(1-p)}{2N},$$

indicating that the variation in allele frequency is inversely proportional to the population size. The standard deviation, $\text{SD}(p_1)$, provides a measure of the change in allele frequency, such that 95% of the time, p_1 will be within two standard deviations of p.

1.1.3 Wright-Fisher Model as a Markov Chain

The Wright-Fisher model extends over multiple consecutive generations. In this model, the outcome of the binomial sampling in one generation becomes the starting point for the sampling in the next generation, thereby forming a Markov chainoften described as a random walk. In this random walk, the allele frequency fluctuates randomly between 0 (loss) and 1 (fixation) until it eventually reaches one of these two boundaries, which act as absorbing states.

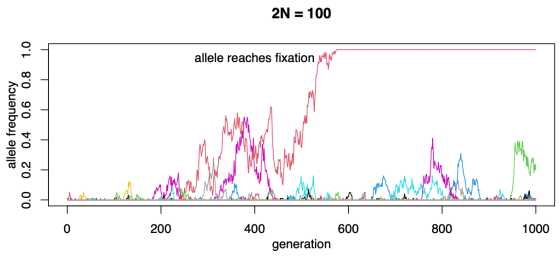

Figure 2.9: **Genetic drift of new mutations.** *Each line shows the simulated trajectory of a different mutation, starting at a random generation number, and drifting independently of the other mutations. This simulation included 200 mutations, most of which stayed rare and are hard to see on this plot.*

Figure 1: Drifting Process

As illustrated in Figure 1, the simulation shows how each of 200 mutations drifts independently. Most mutations remain rare, while a few may reach fixation. Moreover, the probability that a derived allele with a current frequency p eventually fixes is also p.

1.2 Measurement of Genetic Diversity: Expected Heterozygosity

The measure of genetic diversity is expected heterozygosity, which reflects the balance between mutation (which increases genetic variation) and drift (which decreases genetic variation). This balance determines the overall level of genetic diversity in a population. Expected heterozygosity, representing the extent of

neutral genetic variation within a population, is calculated using the population genetic parameter:

$$\theta = 4N\mu,$$

where N denotes the population size and μ is the mutation rate. This formula indicates that larger populations, by virtue of their greater number of alleles, can counteract the allele loss due to genetic drift, while a higher mutation rate introduces more new variants, thereby increasing heterozygosity.

1.2.1 Effective Population Size (N_e)

Effective population size, N_e, is used to adjust the population size in an idealized model in order to better reflect the realistic effects of genetic drift and genetic variation. When population sizes fluctuate over time, N_e is strongly influenced by the smallest population sizes, because genetic drift occurs more rapidly during those generations.

1.2.2 Application of the Wright-Fisher Model to Haplotypes

In this application, we simulate the evolution of entire haplotype sequences over a genomic region of length L by modeling both mutation and drift. In each generation, mutations occur randomly at any position in the sequence at a rate μ. For the sampling process, each haploid sequence in the next generation is randomly drawn from the previous generation, analogous to placing all $2N$ haplotypes in a bag, drawing one haplotype at a time to form the next generation, recording the new haplotype, and then replacing the drawn haplotype back into the bag.

The simulation proceeds as follows:

1. **Initialization:** Create an initial genotype matrix G with $2N$ rows (representing $2N$ haplotypes) and L columns (representing L nucleotide sites). Each entry in G is an integer (0, 1, 2, or 3) corresponding to one of the four possible nucleotides. Initially, all entries in G are set to 0.

2. **Mutation:** For each nucleotide position in G, introduce mutations with probability μ, altering the nucleotide value.

3. **Wright-Fisher Sampling (Genetic Drift):** To simulate genetic drift, generate a new temporary genotype matrix G'. For each of the $2N$ rows in G', randomly select an integer u between 1 and $2N$ and copy the u-th row from G into G'. Once all $2N$ rows of G' have been filled, copy G' back to G to begin the next generation.

4. **Iteration:** Repeat the mutation and sampling steps for multiple generations until genetic variation reaches equilibrium.

2 Chapter 2.2

2.0.1 coalescent

Alleles that are identical by descent (IBD) represent the same chromosomal segments inherited among relatives. A common ancestor refers to an ancient shared ancestor within a population. For the Wright-Fisher (WF) model, we typically compute generations forward from an arbitrary starting point, as was done in the previous chapter. However, in the common ancestry model, we define the present as $t = 0$ and count generations backward from now, generating each generation by random sampling replacement from the previous generation.

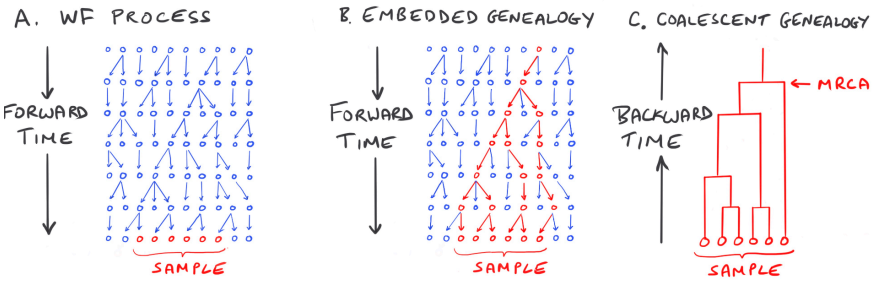

Figure 2: WF history

In the figure 2 above, each row represents a haploid genome. Through random sampling and inheritance across generations, the transmission of genotypes over time is depicted. The middle figure is a sample genetic pathway extracted and traced from the overall WF process. The samples marked by red circles represent selected genotypes, illustrating how they are passed from one generation to the next. The final part traces backward in time (upward), revealing how these samples gradually converge to a most recent common ancestor (MRCA). We can observe the coalescence points among samples in population genetics, as well as how they are connected via different lineages from the past to the present.

Each copy of this locus has a random parent chosen from among the $2N$ possible chromosomes in the previous generation. Therefore, the probability that both copies descend from the same parent is $\frac{1}{2N}$. Conversely, the probability that they do not share a common ancestor in the previous generation is $1 - \frac{1}{2N}$. By independence, the probability that no common ancestor is found after t generations is

$$\left(1 - \frac{1}{2N}\right)^t,$$

which is less than 1. As we multiply this probability over many generations, the value will gradually approach zero. This means that if we trace back far

enough into the past, we can guarantee that any pair of copies at a locus will eventually share a common ancestor.

2.0.2 Genetic Variation Patterns in Modern Samples

Modern patterns of genetic variation in samples reflect the interplay of common ancestry and mutation. Each branch of the gene tree represents the transmission history of a segment of the genome, and the length of a branch determines the potential number of mutations that can occur along it. The mutation rate (μ, defined as the per-generation, per-base pair mutation rate) and the branch length together determine the expected number of mutations, which follow a Poisson distribution. Moreover, any mutation on a branch is inherited by all samples descending from that branch. This gene tree illustrates how genetic diversity accumulates within a population.

2.0.3 Equivalence of Forward and Backward Approaches

Both forward and backward methods yield the same result. The coalescence time, T_2, for two samples is exponentially distributed with an average of $2N$ generations, and the average number of mutations per branch is μL. Therefore, the expected number of differences is

$$2 \cdot E(T_2) \cdot \mu L,$$

which is also expressed as $H = 4N\mu$. This model indicates that the average coalescence time for any random pair of homologous copies in the genome is $2N$ generations (approximately 1 million years). The average variation observed in the genome is due to a mutation that occurred roughly 500,000 years ago, with many variations originating even further back in time.

2.0.4 Population Bottlenecks and Population Growth

Population bottlenecks refer to reductions in population size, which are typically, though not always, followed by a recovery in population size, and can significantly increase the rate of genetic drift. In the Wright-Fisher model, bottleneck effects can be viewed as amplifying changes in allele frequencies: some alleles increase dramatically in frequency while others decrease. The coalescence rate is given by $\frac{k(k-1)}{4N}$ per generation; thus, when N decreases, the coalescence rate increases inversely.

Conversely, population growth has the opposite effect to population bottlenecks. Extremely large population sizes in recent times lower the coalescence rate, resulting in a substantial increase in the number of very rare variants. Population growth can be described by an exponential growth model. Although the model of unlimited growth is unrealistic, it reveals that when the population size expands dramatically, low-frequency variants increase significantly. Empirical genomic data also show that, compared to theoretical models, modern

populations have a higher proportion of rare variants, indicating that recent population growth has had a very significant impact on genetic diversity.

These analyses help us gain a deeper understanding of how historical population dynamics have shaped our genetic structure. Historical population size models are capable of fitting the complete site frequency spectrum (SFS) data.

3 Questions

1. What is the difference between effective population size (Ne) and actual population size (N)? Why is effective population size more important than actual population size in population genetics models?
2. In models, how do genetic drift and mutation rate jointly determine the genetic diversity of a population? How are these parameters estimated in practice?