Chapter 2.1-2.7

Siyi Wang

May 12, 2025

Chapter 2.1

0.1 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**.

0.1.1 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.

0.1.2 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.

0.1.3 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.

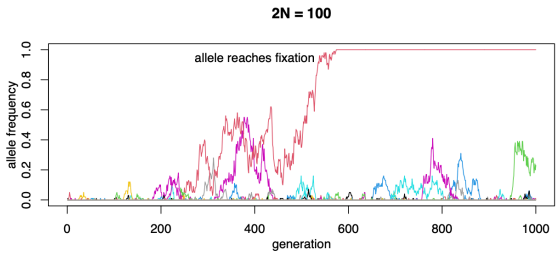

Figure 2.9: **Genetic drift of new mutations.** *Each line shows the simulated trajectory of a different mutation, starting at a random generation number, and drifting independently of the other mutations. This simulation included 200 mutations, most of which stayed rare and are hard to see on this plot.*

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.

0.2 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.

0.2.1 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.

0.2.2 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.

Chapter 2.2

0.2.3 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.

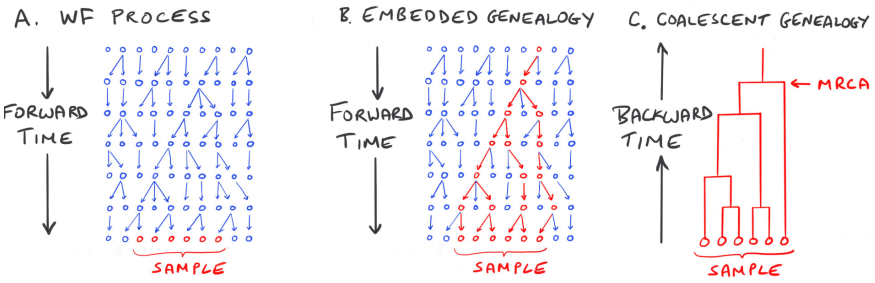

Figure 2: WF history

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.

0.2.4 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.

0.2.5 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.

0.2.6 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.

Chapter 2.3

0.2.7 Linkage Disequilibrium and the Role of Recombination

The genetic process of common ancestry results in correlations between genotypes at different SNP loci, a phenomenon known as Linkage Disequilibrium (LD). In experiments with *Drosophila melanogaster*, certain SNP combinations occur at higher frequencies than expected by chance, meaning that specific SNP combinations tend to be inherited together (for example, the haplotype "GTCTCC" appears simultaneously in four individuals).

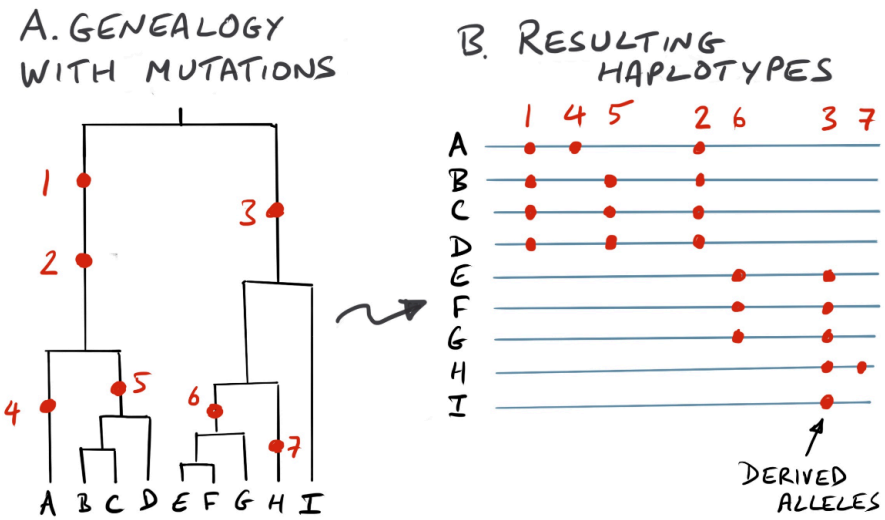

Figure 3: coalescent process with linkage

The figure 3 shows a coalescent tree without recombination. In the left panel, red dots represent mutations, while in the right panel the corresponding haplotype structure is displayed, with red dots indicating derived alleles. It can be observed that mutations occurring on the same branch are always inherited together (mutations 1 and 2 always co-occur), whereas mutations on adjacent branches may sometimes be inherited together (mutations 1 and 5 might appear together, but not consistently).

Linkage causes SNPs to be co-inherited, thereby forming LD; in the absence of recombination, genetic variations (SNPs) occur on specific branches of the evolutionary tree, meaning that all descendants will inherit that variation. Recombination disrupts LD by shuffling specific segments on the chromosome, thereby generating new haplotypes.

0.2.8 Linkage Disequilibrium between Two SNPs

Assume there are two SNPs: for the first SNP, the alleles are A and a; for the second SNP, the alleles are B and b. The four possible haplotypes and their frequencies are AB, Ab, aB, and ab, denoted by p_{AB}, p_{Ab}, p_{aB}, and p_{ab} respectively. Under the assumption of independence, we have $p_{AB} = p_A p_B$. Therefore, the LD value is defined as

$$D = p_{AB} - p_A p_B.$$

If $D = 0$, the SNPs are independent and in linkage equilibrium; otherwise, LD exists. We often use a standardized LD measure, D', whose absolute value is less than 1, representing the effect of recombination between the SNPs:

$$D' = \frac{D}{\max(D)}.$$

Another important metric is r^2, which quantifies the strength of LD. When $r^2 = 0$, the two SNPs are completely independent; when $r^2 = 1$, they are completely linked (i.e., only two haplotypes exist, such as AB/ab or Ab/aB). Specifically,

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}.$$

The strength of LD decreases with increasing genetic distance. For example, if the recombination probability per generation is c, then after t generations the LD decays according to:

$$D_t = (1 - c)^t D_0.$$

Thus, if two SNPs are 20 Mb apart (with $c \approx 0.2$), after 10 generations LD is reduced to 10% of its original value ($0.1D_0$). For unlinked SNPs ($c = 0.5$), which are completely independent, LD essentially disappears after 10 generations. When the recombination rate exceeds 0.1 cM, recombination significantly weakens LD, thereby breaking up haplotype structures.

0.2.9 Non-uniform Decay of LD and Variability in Recombination Rates

LD does not decay uniformly across the genome. In our previous assumption of no recombination, all individuals would trace back to a single ancestor, resulting in complete linkage of SNPs and no decay of LD. However, the Ancestral Recombination Graph (ARG) shows that, when tracing backward in time,

recombination events break haplotypes into segments that may trace back to different common ancestors. For instance, closely located SNPs may share the same common ancestor (resulting in strong LD), whereas SNPs that are farther apart may trace back to different ancestors (resulting in weaker LD). Recombination breakpoints divide the haplotype into independent regions, leading to a decay of LD with increasing genomic distance.

The LD matrix visually displays the strength of LD using a color-coded heatmap, with red indicating high LD and white indicating low LD. Therefore, adjacent SNPs in the genome typically exhibit strong LD (red areas at the bottom), while LD gradually decreases with increasing distance (the color changes from red to white).

0.2.10 Recombination Hotspots and PRDM9

Recombination is concentrated in certain small regions, known as recombination hotspots. Typically, 8090% of recombination occurs in less than 10% of the human genome. In these hotspots, LD drops sharply, while in adjacent regions it remains high. Moreover, LD heatmaps differ among populations, suggesting that they may be under genetic control.

A study in 2010 demonstrated that the *PRDM9* gene controls the locations of human recombination hotspots by recognizing specific DNA sequences and promoting recombination in those regions. Paradoxically, due to gene conversionwhere certain mutated alleles can be replaced (i.e., hotspot sequences are replaced by non-hotspot sequences from the homologous chromosome)one would expect hotspot DNA sequences to gradually disappear. However, in reality, hotspots persist. This may be because the zinc finger structure of *PRDM9* can evolve rapidly; old hotspots gradually lose functionality while new hotspots are continuously and rapidly created, thus maintaining recombination. Interestingly, different populations possess different *PRDM9* alleles, leading to distinct recombination hotspot patterns, and recombination hotspots in chimpanzees do not overlap with those in humans.

0.2.11 Simplified Haplotype Copying Models and Their Applications

Calculating the Ancestral Recombination Graph (ARG) is extremely complex, making it impractical for large-scale datasets. Consequently, in 2003 scientists proposed simplified models known as haplotype copying models. Due to linkage disequilibrium (LD), adjacent SNPs are more likely to originate from the same haplotype, while longer segments that span recombination hotspots may be copied from a different haplotype.

Hidden Markov Models (HMMs) can be employed to decompose a diploid genotype into two complete haplotypes, effectively determining whether each SNP belongs to the paternal or maternal chromosome (haplotype phasing). Moreover, the LD structure enables the inference of unobserved SNP genotypes via a reference panel, thereby facilitating genotype imputation.

Chapter 2.4

0.2.12 Population Structure and Genetic Drift

Population structure arises from geographic isolation, cultural and social factors, and historical demographic events, leading individuals to preferentially breed within limited areas. This causes allele frequencies to differ among populations and influences the pattern of genetic variation in the human genome. Most human genetic variations are shared, consistent with the Out-of-Africa model around 80,000 years ago. African populations possess the highest genetic diversity, while non-African populations experienced bottlenecks that reduced allele diversity. Consequently, some ancestral alleles are common to all populations, whereas certain SNPs are unique to specific populations as mutation alleles that arose after population separation. In the absence of migration, allele frequencies evolve independently.

In the Wright-Fisher drift model, the changes in allele frequency in the ancestral population prior to drift and the independent drift of frequencies in the two populations after splitting are distinct processes. The Nicholson-Donnelly approximation can be used to calculate the frequency of a particular SNP in the current population after drift:

$$p_T \sim \text{Normal}\left(p_A, \frac{T}{2N} p_A(1-p_A)\right),$$

where the variance is given by $\frac{T}{2N} p_A(1-p_A)$. The larger T is or the smaller N is, the stronger the effect of drift. For example, in the case of Tibetans and Han Chinese, due to the unique environment, the SNP in the *EPAS1* gene is particularly high in Tibetans, which is a signal of natural selection.

0.2.13 Migration and Gene Flow

If migration occurs, alleles flow between populations, which mitigates the effects of genetic drift. This generally means that the gene pool of one population is influenced by that of another. Under the Wright-Fisher model, the migration rate, m, is defined as the proportion of alleles in each generation that are introduced from other populations. A higher m leads to more similar allele frequencies among populations. For instance, when $m = 0$, the alleles in the next generation are entirely drawn from the previous generation of the same population, resulting in stronger drift effects. Conversely, migration increases gene flow and reduces the random fluctuations in allele frequencies caused by drift. Consequently, migration can rapidly diminish the differences in allele frequencies over a short period; if the migration rate is sufficiently high, two populations will eventually converge to the same allele frequencies in the long term. It is noteworthy that if samples are taken from the same population, the coalescence time is approximately $2N$ generations, whereas if the samples come from different populations, the coalescence time is $2N + T$ generations (where T is the divergence time).

0.2.14 Population Differentiation and Incomplete Lineage Sorting

F_{ST} is an index for measuring population differentiation, defined as the variance in allele frequencies among populations divided by the maximum possible variance. A value of 0 indicates that the allele frequencies of two populations are identical, while a value of 1 indicates complete differentiation with no shared alleles. Thus, if the allele frequency differences between two populations are large, F_{ST} will be high, and vice versa. Data show that the F_{ST} between Han and European populations is only 0.106, suggesting that the genetic differences between East Asians and Europeans are relatively small. This supports the Out-of-Africa theory, indicating that most genetic variation among modern humans is shared globally.

The coalescent model is employed to analyze the evolution of different gene segments among humans and other primates. In theory, all genes should indicate that humans and chimpanzees are most closely related, since their most recent common ancestor is estimated to have lived about 67 million years ago. However, in reality, approximately 30% of the genome supports a closer relationship between either humans and gorillas or chimpanzees and gorillas. This discrepancy is due to incomplete lineage sorting, where the coalescence time was too short for some genes to fully coalesce in the common ancestor of chimpanzees and humans, leading to mismatches between gene trees and the species tree.

Chapter 2.5

0.2.15 Natural Selection and the Modern Synthesis

Natural selection originates from Darwin's "On the Origin of Species," where the two most important concepts are "survival of the fittest" and "descent with modification." Therefore, evolution requires three fundamental conditionsvariation, inheritance, and competitionto enhance survival. Historically, the then-popular theory of blending inheritance suggested that offspring traits were simply the average of their parents' traits, which would lead to the gradual disappearance of genetic variation. In 1866, Mendel proposed that alleles determine hereditary information, thereby allowing genetic variation to be preserved across generations. The modern synthesis integrates the mechanisms of genetic transmission, changes in allele frequencies, natural selection, and genetic drift.

0.2.16 Fitness and Selection Coefficients

Fitness refers to the ability of a genotype to produce offspring in the next generation. Genotypes with lower fitness are more likely to be predated upon and eventually disappear. Mathematically, if we set the fitness of the ancestral genotype AA to 1, then the fitness of the heterozygote Aa is $1 + hs$, and the fitness of the mutant genotype aa is $1 + s$.

If $s > 0$, the mutation is beneficial, representing positive selection; the frequency of allele a increases over generations, possibly reaching fixation. If $s = 0$, allele a is neutral and is not affected by selection. Conversely, if $s < 0$, allele a is deleterious, and negative selection works to remove harmful mutations, causing the frequency of a to decrease and potentially be lost over time.

The parameter h is the dominance coefficient, which measures the fitness contribution of the heterozygote Aa. When $h = 0$, the mutation is recessive, meaning that only aa is subject to selection while Aa remains unaffected; such mutations may remain hidden for a long time until their frequency increases sufficiently to be acted upon by selection. When $h = 1$, the mutation is dominant, leading to rapid spread or elimination.

0.2.17 Combined Effects of Genetic Drift and Selection in Finite Populations

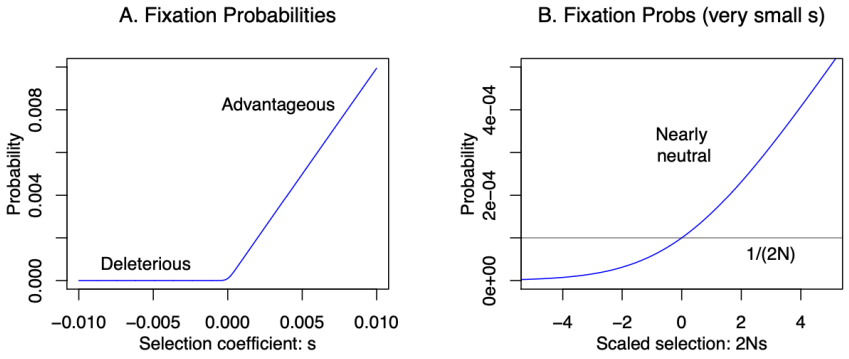

Figure 4: Fixation Probability

In a finite population, genetic drift and natural selection act together, so we must consider three scenarios. When $2Ns \approx 1$, the effects of drift and selection are roughly equivalent. When $2Ns < 1$, drift dominates and selection is essentially ineffective. Conversely, when $2Ns > 1$, selection is strong enough to overcome drift. Therefore, under strong selection, the fixation trajectory of a mutation approximates deterministic growth, while weak selection causes allele frequencies to fluctuate widely under drift, making fixation difficult.

When $s > 0$, the mutation is beneficial, but it can still be lost due to drift:

$$P_{\text{fix}} \approx \frac{1 - e^{-s}}{1 - e^{-2Ns}}.$$

This formula illustratesas shown that if s is very small, even a beneficial mutation may be lost. When $s = 0$, the fixation probability is only $1/(2N)$, com-

pletely determined by genetic drift. For $s > 0$ (advantageous mutations), the mutation is more likely to be fixed (the fixation probability is sensitive to selection pressure). Figure 4 part B on the right of the figure represents the relationship between fixation probability and scaled selection; the gray area indicates that near $s \approx 0$, the fixation probability remains nearly constant, being influenced only by drift. When s is very small (but nonzero), selection has only a slight effect on fixation probability. When $2Ns \gg 1$ (right side), the fixation probability increases significantly, indicating that strong selection effectively drives the spread of mutations. Conversely, when $2Ns \ll -1$ (left side), the fixation probability declines rapidly, showing that strong selection effectively suppresses mutation fixation. Thus, under strong selection pressure, the fixation probability becomes independent of population size, whereas under weak selection, allele drift remains the predominant factor.

0.2.18 Purifying Selection, Codon Bias, Transposable Elements, and Genetic Load

Each human generation produces approximately 70 new mutations, most of which have no effect on the organism (neutral mutations); among those with functional effects, the majority are deleterious. Purifying selection is a form of natural selection that removes mutations under negative selection, preventing their accumulation in the population and thereby protecting the functional integrity of the genome. Typically, purifying selection mainly acts on deleterious mutations with fitness $s < 0$; as a result, harmful mutations usually remain at low frequencies because they are suppressed by purifying selection. The maximum frequency of a mutation is approximately $\frac{1}{2Nhs}$, indicating that strong purifying selection can quickly eliminate mutations, whereas weak purifying selection may allow them to persist longer. (Below are three evolutionary consequences of purifying selection.)

Nearly-neutral mutations have very weak selective effects, so genetic drift may dominate over selection. The same amino acid can be encoded by multiple codons (for example, glycine: GGA, GGC, GGG, GGT). Some codons are translated more efficiently, and as a result, some organisms exhibit a preference for certain codonsa phenomenon known as codon bias. In organisms with large populations (such as fruit flies), purifying selection can maintain strong codon bias; in species with smaller populations (like humans), nearly neutral drift tends to weaken codon bias.

Transposable elements (TEs) are DNA elements capable of self-replication and insertion into different regions of the genome. Because such insertions, particularly into coding regions, can disrupt gene function, most TEs are potentially harmful to the host genome. Approximately two-thirds of the human genome consists of TEs, with Alu elements accounting for about 10%. Although the fitness cost of an individual TE insertion is very small, their cumulative effect is significant; consequently, the genome has evolved epigenetic mechanisms to suppress TE activity.

The reduction in fitness in a population due to the presence of deleterious

mutations is known as genetic load. There are two forms of genetic load: the individual-level load, which refers to the unique deleterious mutations carried by each person that are gradually removed by purifying selection through drift, and the population-level load, which refers to deleterious mutations that are nearly neutral and cannot be efficiently removed by purifying selection, thereby accumulating over time. In theory, genetic load should continuously accumulate in a population and eventually lead to extinction; however, weak deleterious mutations can be offset by compensatory mutations, allowing organisms to maintain stable evolution.

Chapter 2.6

0.2.19 Polygenic Adaptation, Selective Sweeps, and Genetic Hitchhiking

Over the past 70,000 years, humans have successfully adapted to various environments (extreme cold, high temperatures, tropical rainforests, high altitudes, etc.). These environmental pressures have driven genetic and positive selection. Polygenic adaptation refers to traits such as height, weight, skin color, and metabolic characteristics that are usually controlled by many genes, with adaptation potentially involving subtle frequency changes at tens of thousands of gene loci. In this process, alleles with strong advantages rapidly increase in frequency within the population, forming selective sweeps. These sweeps carry along linked neutral variants via genetic hitchhiking, resulting in regions of reduced genetic diversity. In such regions, haplotypes become very similar and heterozygosity declinesthat is, an individual's two alleles in that region are more likely to be identical.

When selective sweeps occur, because the mutation is recent and rapidly increases in frequency, a beneficial mutation will carry a large segment of the surrounding genomic sequence along with it before recombination has time to break the linkage, forming a long haplotype extension. For example, the mutations associated with lactase persistence in East Africa are located within a long haplotype region. After a selective sweep, as new beneficial mutations quickly replace the old alleles, common alleles nearly disappear and rare alleles become enriched, since the new mutations have not yet accumulated to high frequencies.

The SLC24A5 gene, which controls human skin pigmentation, illustrates this further. In low-latitude regions (such as Africa), high melanin content helps prevent UV damage and folate degradation, whereas in high-latitude regions (such as Europe) reduced melanin enhances vitamin D synthesis. Consequently, a mutation (Ala111Thr) in this gene causes lighter skin and is nearly fixed in Europe (>90%) but is almost absent in Africa (<5%). In African populations, a mutation in the Duffy antigen gene results in its non-expression on red blood cells, conferring resistance to *P. vivax* malaria. The Duffy mutation is nearly fixed in Africa yet nearly absent in other regions, indicating that it underwent a strong selective sweep in Africa; however, because this mutation is found on

multiple different haplotypes, it is indicative of a soft sweep.

0.2.20 Balancing Selection and Heterozygote Advantage

Some mutations provide an adaptive advantage in the heterozygous state but cause disease in the homozygous state. These mutations do not disappear completely; rather, they are maintained at a certain frequency, resulting in balancing selection. For example, in sickle cell disease involving the HBB gene, the mutation confers malaria resistance in the heterozygous (AS) state, but leads to sickle cell anemia in the homozygous (SS) state, maintaining a frequency of about 15% in Africa.

0.2.21 Polygenic Adaptation in Complex Traits

It has been previously mentioned that most complex traits (such as height, body shape, intelligence, and immune function) are controlled by multiple genes, each contributing a small effect. The hallmark of polygenic adaptation is that selection does not act on individual genes but accumulates through minor frequency changes, and this type of adaptation occurs more rapidly than traditional selective sweeps. For example, in the Illinois Maize Experiment, which began in 1896, maize varieties were selected for either high or low protein content. After 100 generations, the change in protein content far exceeded what could be achieved by single-gene mutations, demonstrating that polygenic adaptation can drive rapid evolution.

Chapter 2.7

In the process of evolution, proteins tend to accumulate amino acid variations, with each protein exhibiting a unique rate of change. In 1983, Motoo Kimura proposed the Neutral Theory, suggesting that most new mutations are functionally neutral or nearly neutral. This theory provides a null hypothesis for studying genetic variation: in the absence of natural selection, molecular evolution would be primarily driven by random genetic drift.

The Molecular Clock supports this neutral theory, demonstrating that differences in protein or DNA sequences between species approximately accumulate linearly along evolutionary timelines. The dN/dS ratio serves as a powerful tool for detecting natural selection's effects.

0.3 The dN/dS Ratio and Selection

The dN/dS **ratio** is defined as follows:

- dN: The average number of nonsynonymous substitutions per site (causing amino acid changes).
- dS: The average number of synonymous substitutions per site (without amino acid alterations).

Under the assumption of no selective pressure, the probability of fixing nonsynonymous and synonymous substitutions should be approximately equal, with:

$$\frac{dN}{dS} \approx 1$$

However, in reality, $dN/dS < 1$ is the most common scenario, reflecting purifying selection, which eliminates harmful mutations.

For instance, in human genes, the estimated average dN/dS is around 0.14, meaning that only about 14% of amino acid substitutions can be considered neutral or subject to weak selection.

Conversely, when $dN/dS > 1$, this indicates **positive selection**, suggesting that a significant proportion of beneficial mutations become fixed.

0.4 Application: Major Histocompatibility Complex (MHC)

A notable application of this principle emerges in the Major Histocompatibility Complex (MHC). The MHC genes, responsible for immune recognition, benefit from genetic diversity, as this enhances pathogen detection. Consequently, in the antigen-binding regions (ABR), researchers observe:

$$dN > dS \quad \Rightarrow \quad \frac{dN}{dS} > 1$$

This pattern demonstrates how balancing selection and diversifying selective pressures maintain amino acid variation.

However, the MHC represents a special case. Most genes accumulate far fewer adaptive changes, making the detection of positive selection ($dN/dS > 1$) a conservative approach. Meaningful results emerge only when selective pressures are sufficiently strong and sustained.

0.4.1 McDonald-Kreitman test

The McDonald-Kreitman test is a method for detecting natural selection by comparing polymorphism within populations and fixed differences between the species. As shown in the figure, the MK test categorizes variations by their nature: in data from two species, each mutation site can be classified as either fixed (different fixed alleles between species A and B) or polymorphic (variations existing within a species), which are further divided into non-synonymous and synonymous mutations.

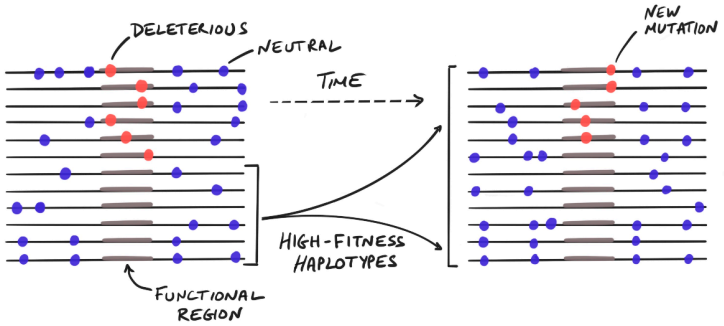

Figure 5: Background selection

In a neutral+deleterious baseline model, as the Figure 5 above, the neutral portion of non-synonymous mutations will drift randomly like synonymous mutations, so the proportion of non-synonymous variations remains the same in both polymorphic and fixed differences (the blue and pink proportions in the top and bottom sections are equal).

However, in Figure b, there are positively selected non-synonymous mutations that quickly become fixed and rarely exist in a polymorphic state. Consequently, the proportion of non-synonymous variations in fixed differences will be significantly higher than in polymorphic variations.

0.5 Implications

By comparing the proportion of non-synonymous mutations in fixed differences and polymorphic variations to determine if they deviate from neutral expectations (with synonymous mutation proportions remaining unchanged), the conclusion is drawn: there is strong support for adaptive evolution in the gene, with a considerable portion of amino acid substitutions being fixed by positive selection.

In humans, only 0-10% of non-synonymous fixed differences are adaptively fixed. Therefore, in primate genomes, the vast majority of amino acid differences can be explained by neutral drift, with truly positively selected fixations comprising only a small fraction. This finding is consistent with the neutral theory's expectations and emphasizes the human genome's evolution being primarily driven by negative selection and genetic drift.

0.6 Linked Selection

Linked Selection refers to the influence of linkage disequilibrium on the distribution of genetic variation in adjacent genomic regions.

0.7 Selective Sweep

Selective Sweep describes the phenomenon where a new beneficial mutation rapidly increases in frequency and becomes fixed in a population. As this occurs, neutral variations linked to the advantageous allele also rise in frequency and may even become fixed, leading to a significant reduction in genetic diversity in the surrounding region. This process is also known as the hitchhiking effect.

Quantitative models suggest that the extent of this effect can be expressed as:

$$r \times \frac{2\log(2N)}{s}$$

where r represents the recombination rate. The lower the recombination rate, the larger the affected segment, making it harder for neutral sites to escape the hitchhiking effect of the beneficial allele. Consequently, the loss of genetic diversity extends over a broader region.

Once the sweep is complete, most individuals in the adjacent genomic region will carry identical haplotype fragments, nearly eliminating polymorphism in the region. Experimental studies on *Drosophila* genomes have demonstrated the *hitchhiking effect*, revealing that genetic diversity in low-recombination regions is significantly lower than in high-recombination regions. This phenomenon has also been observed in the human genome.

0.8 Background Selection

Background Selection is another form of linked selection, focusing on the effect of negative selection on nearby neutral variations. The presence of deleterious mutations reduces the effective population size in the surrounding region: only chromosome copies free from harmful mutations have a higher probability of being passed on.

Mathematically, in the absence of recombination, the total fraction of chromosomes carrying deleterious mutations, denoted as f, is approximately:

$$f \approx L \times \mu/(hs)$$

where L is the number of base pairs susceptible to mutation, μ is the mutation rate per base pair, and hs represents the selective disadvantage for heterozygotes.

In the absence of recombination, if a proportion f of chromosomes each generation carries mutations that cause them to be eliminated, the effective population size in that region is approximately reduced by a factor of $(1 - f)$. Equivalently, the expected neutral diversity $E[\pi]$ is also reduced in proportion to $(1 - f)$.

In the presence of recombination, neutral sites have a chance to escape linkage with deleterious alleles and "survive." The extent to which background selection reduces genetic diversity depends on both the deleterious mutation rate and the recombination rate.

Charlesworth and colleagues proposed that the lower polymorphism observed in low-recombination regions of Drosophila may not primarily result from more frequent selective sweeps but rather from the enrichment of functional elements in these regions. The constant removal of deleterious mutations in such regions reduces neutral site variation through linkage effects.

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?
3.Is the decay of LD uniform? Does recombination occur at the same rate in all regions of the genome?
4. Are there other factors that affect incomplete lineage sorting?
5.Which genes are subject to the strongest purifying selection? Does purifying selection change with environmental variation?
6. What can i do of the high polymorphism of MHC genes makes it challenging to determine which variants are functional and which are neutral variations in population genetic analyses?