# Chapter 1.1-1.5

Siyi Wang

May 12, 2025

## Chapter 1.1- Chapter 1.2

### 0.1 Basic introductions

Linkage refers to the phenomenon where two loci physically close on the same chromosome are inherited together and are less likely to undergo recombination and separation during meiosis, thus serving as a positional tool for identifying unknown pathogenic genes. Gene mutations, such as deletions or insertions, can lead to various diseases by altering gene function.

High-throughput sequencing enables rapid and large-scale DNA/RNA sequencing, allowing for the analysis of inherited genetic traits, ancient DNA, and functional genomic studies.(Use for criminal, COVID-19, and pregnancy)

Genomes serve as biological data storage systems, encoding functional information in both coding and non-coding regions, with DNA structured as a double helix composed of four nucleotide bases (A, C, G, T) and a sugar-phosphate backbone, where bases form complementary pairs (A:T, C:G) and sequences are oriented 5 $\rightarrow$3 . Humans possess 23 chromosome pairs (22 autosomes and one pair of sex chromosomes, X/Y), with each cell containing approximately 6.6 billion base pairs (bp) compacted into chromatin through highly organized folding.

Here is a way to encode DNA to protein: Three nucleotides called a codon, 64 combinations in codons. In the Genetic Code, a start code is ATG, and end codes are TAA TAG TGA.

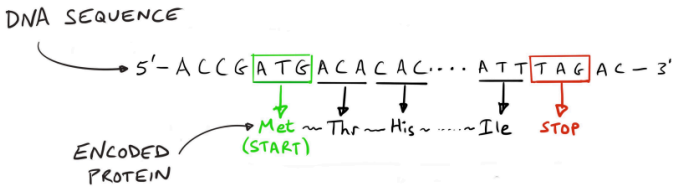


Figure 1: DNA Endoding

However, **DNA** needs to transform into **mRNA** (one strand of helix), then

change into **protein**. This process is called the Central Dogma, Transcription, which copies DNA to mRNA; here, RNA uses U to substitute for T in DNA, then translation, AUC to the 20 amino acid alphabet of protein.

A typical protein-coding gene consists of a 5  untranslated region (5 -UTR) followed by multiple exons interspersed with introns, a coding sequence, and a 3 -UTR. The gene is first transcribed in the nucleus into a pre-mRNA transcript, which undergoes splicing to remove introns and join exons, forming mature mRNA. This processed mRNA is then exported through nuclear pores to the cytoplasm, where it undergoes translation to synthesize the encoded protein.

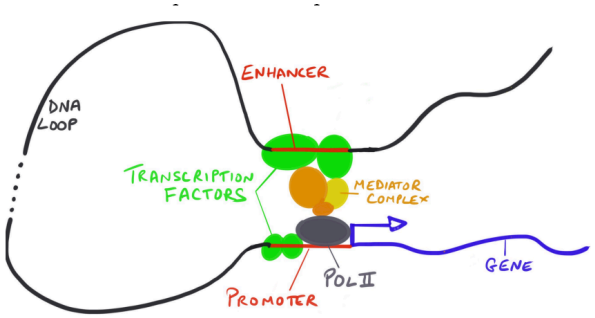


Figure 2: Gene regulation

The chromatin fibers (black curve) facilitate the formation of a DNA loop, bringing distal enhancer regions (top, orange) into close proximity with the target gene's promoter (bottom, red). The enhancer contains multiple cis-regulatory motifs recognized and bound by transcription factors (TFs, green spheres), whose activation domains recruit co-activators and histone-modifying enzymes to establish active chromatin marks. The Mediator complex (orange irregular shape) bridges the enhancer-bound TFs and the pre-initiation complex (PIC) at the promoter, relaying regulatory signals to transition RNA Polymerase II (Pol II, gray) from a paused to an active state. TF binding is sequence-specific, targeting preferred motifs, with binding sites determined by adjacent DNA elements; cooperative interactions among TFs form dynamic TF ensembles to fine-tune transcriptional output.

Mitosis and meiosis both involve DNA replication prior to division but differ fundamentally in division cycles, genetic outcomes, and biological functions. Mitosis consists of a single division, producing two genetically identical diploid daughter cells from one diploid parent cell, primarily supporting somatic growth, tissue repair, and asexual reproduction. In contrast, meiosis comprises two consecutive divisions (Meiosis I and II) following a single DNA replication event. Meiosis I reduces ploidy by pairing and segregating homologous chromosomes, introducing genetic recombination through crossing-over, and yielding two haploid cells. Meiosis II then separates sister chromatids, ultimately generating four genetically distinct haploid gametes, essential for sexual reproduction and the maintenance of genetic diversity within populations.

# Chapter 1.3

## 0.2 DNA Structure and Base Pairing

DNA is a double-stranded structure composed of two complementary strands. Each strand follows strict base-pairing rules:

- **A (adenine)** always pairs with **T (thymine)**
- **C (cytosine)** always pairs with **G (guanine)**

However, when studying SNPs, the focus is on single-strand variations rather than the entire double-stranded structure.

## 0.3 SNP Annotation and Allele Nomenclature

When analyzing SNPs, we typically examine nucleotide variations on one **single strand** rather than both. For instance:

- Due to the double-stranded nature of DNA, an **A/G SNP** on one strand will appear as a **T/C SNP** on the complementary strand; however, only one strand's variation is recorded.

Key points include:

- Two copies of a gene (one from each parent) are called **alleles**.
- The **directionality** of SNPs (the **Strand Issue**) affects their naming.

SNPs are usually annotated based on the **reference genome** or the direction of protein coding. There are three common ways to name alleles:

- **Reference allele** (the allele present in the reference genome)
- **Alternate allele** (the variant allele)
- **Major/Minor allele** (indicating the more or less common allele, with **MAF** representing Minor Allele Frequency)

### 0.3.1 Genotype Notation

Individuals with the **AA** or **GG** genotype are **homozygotes**, whereas those with the **AG** genotype are **heterozygotes**.

## 0.4 The Hardy-Weinberg Principle

The Hardy-Weinberg model is based on several key assumptions:

- **Distinct generations:** The parental and offspring populations are separate.
- **Random mating:** Parents mate without regard to genotype.

- Allele frequencies in the parental population are defined as:
  - $p$: Frequency of the **A** allele
  - $q$: Frequency of the **G** allele, with $p + q = 1$

Under these assumptions, the genotype frequencies in the offspring are expected to follow the ratio: $p^2 : 2pq : q^2$. The model further assumes:

- Random mating
- No selection
- Non-overlapping generations

Deviation from Hardy-Weinberg proportions in a SNP often indicates potential genotyping errors.

### 0.4.1 Historical Background

The Hardy-Weinberg Principle was independently proposed in 1908 by the British mathematician **G. H. Hardy** and the German physician **Wilhelm Weinberg**.

> *"A little mathematics of the multiplication table is enough to prove this..."*

## 0.5 Population Genetics and Haplotype Structure

African populations exhibit the highest levels of heterozygosity, reflecting the African origin of modern humans and their dispersal over the past 100,000 years. Additional key points include:

- The **Human Genome** comprises approximately **3.2 billion base pairs**.
- Each individual carries about **1.5 to 3 million heterozygous SNPs**, with numbers varying by ancestry.
- The **1000 Genomes Project** identified **85 million SNPs**, including **8 million** common SNPs (about one every **400 base pairs** with frequency $> 5\%$).
- Nearly all possible SNP alleles are found globally due to a large population size and high mutation rate.
- The genomic difference between humans and chimpanzees is approximately **1.37%** (around **40 million SNP differences**), which is **15 times greater** than the difference between the two chromosomes within an individual.

A **haplotype** represents the arrangement of alleles on a single chromosome, whereas a **genotype** is the combination of paired alleles. Traditional sequencing methods reveal genotypes but cannot assign alleles to their respective chromosomes. For example, in heterozygous SNPs (e.g., genotype: AG), the possible haplotype combinations are:

- **A** on the maternal chromosome and **G** on the paternal chromosome, or
- **G** on the maternal chromosome and **A** on the paternal chromosome.

(The development of **long-read sequencing** has helped overcome this limitation.)

**Haplotype Phase:** Humans possess two homologous chromosomes, and each SNP can exist in different phase combinations. The **Genotype Matrix** represents allele counts:

- Each row corresponds to an individual.
- Each column corresponds to a SNP.
- **0** indicates major allele homozygosity.
- **1** indicates heterozygosity.
- **2** indicates minor allele homozygosity.

In practical applications, such as IGF1 studies, genotype variation is visualized using color schemes to display SNP differences across populations. Although most common variants are shared among human populations, allele frequencies vary, and many SNPs are rare. Additionally, variants at different loci often co-occur, resulting in **Linkage Disequilibrium** patterns.

## 0.6 Forms of Genomic Variation

## 0.7 Small-scale Variation (1bp – ∼100bp)

- **SNP (Single Nucleotide Polymorphism):** A change in a single base pair, e.g., ACGTCAGTGT…vs ACGTCAATGT…
- **Indel (Insertion/Deletion):** The insertion or deletion of base pairs, e.g., ACGTCAGTGT…vs ACGTC---GTGT…Indels in **exons** can lead to **frameshift mutations** affecting protein translation.
- **STR (Short Tandem Repeat):** The amplification of short sequence repeats, e.g., A C G C A C A C A C A G…vs A C G C A C A C A G…

## 0.8 Intermediate to Large-scale Variation

- **VNTR (Variable Number Tandem Repeats)**
- **CNV (Copy Number Variation)**

## 0.9 Large-scale Variation (100bp – 1Mb)

- **Deletion:** Loss of a DNA segment.
- **Duplication:** Increase in the number of copies of a DNA segment.
- **Inversion:** Reversal of the orientation of a DNA segment.
- **Complex Structural Variation:** Variations involving multiple mutational events.

## 0.10 Main Types of Protein-Coding Sequence Variants

SNPs affect gene function primarily through alterations in **protein-coding sequences** and **gene regulation**. Most SNP changes are neutral, with less than **10%** having significant biological effects.

1. **Synonymous:** Changes in the DNA sequence (e.g., `AGA → AGG`) that do not affect the protein.
2. **Missense:** Amino acid changes (e.g., `AGA → GGA`) that may affect protein function, particularly in key functional domains.
3. **Nonsense:** Mutations introducing a premature stop codon (e.g., `TAA, TAG, TGA`), often resulting in a truncated, nonfunctional protein.
4. **Frameshift:** Insertions or deletions that shift the reading frame and typically disrupt protein function severely.
5. **Splice Site Disruption:** Mutations at exon-intron boundaries (introns usually begin with `GT` and end with `AG`) can affect RNA splicing and alter gene expression.

SNPs that affect gene expression often function through subtle, "fine-tuning" mechanisms involving interactions between DNA, proteins, and transcription factors. Although regulatory SNPs rarely cause single-gene disorders on their own, they are major drivers of phenotypic variation and evolutionary change. In many cases, hundreds or thousands of regulatory SNPs work together to shape phenotypic diversity.

## 0.11 Case Studies and Historical Examples

## 0.12 Hemophilia in the Royal Families – A Historical Case of an SNP Mutation

Hemophilia is caused by a mutation on the X chromosome, making males more susceptible while females generally act as carriers. Queen Victoria's lineage extensively propagated the hemophilia mutation, ultimately affecting several European royal families. In 2009, genetic analysis revealed that the mutation disrupted an RNA splicing site, resulting in a frameshift. This aberration led

to the insertion of extraneous amino acids and the premature termination of translation at a TAA stop codon, producing a nonfunctional protein.

## 0.13 Structural Variations and Evolutionary Adaptations

### 0.13.1 AMY1 – An Evolutionary Adaptation to a High-Starch Diet

Structural variations can affect gene function by altering protein sequences or modulating gene expression. Copy number variations (CNVs) may induce genetic syndromes in heterozygotes that compromise fitness, leading to reduced reproductive capacity—especially in copy-number sensitive genes—and may even cause disease. While amplification of the copy number in certain genes (e.g., *AMY1*) may confer an evolutionary advantage, most large-scale CNVs result in severe genetic defects due to haploinsufficiency. Fluorescence in situ hybridization (FISH), which uses fluorescent labeling of specific loci, is often employed to detect gene copy number variations.

### 0.13.2 Chromosome Segregation Errors and Aneuploidy

Aneuploidy is a phenomenon characterized by an abnormal number of chromosomes and is primarily transmitted by the mother. Unlike point mutations, chromosome segregation errors result from numerical abnormalities rather than changes in nucleotide sequences. These chromosomal errors are especially common in oocytes of older mothers and can lead to conditions such as Trisomy 21 (Down syndrome). The high incidence of chromosomal abnormalities in older oocytes is a significant factor contributing to infertility. Female oocytes exhibit a higher error rate during meiosis, mainly due to:

- Prolonged arrest in meiosis I for several decades, which leads to deterioration of the chromosome segregation mechanisms.
- Kinetochore drift and instability of the microtubule spindle, increasing the likelihood of aneuploidy.

# Chapter 1.4

DNA sequencing is a transformative tool that has rendered genome sequencing both affordable and efficient. DNA sequencing technology has revolutionized research in biology and medicine. It is not limited to genome determination but is widely employed in various fields, including cancer research (to detect genetic changes associated with uncontrolled cell growth), infectious disease tracking (by identifying pathogens in air, water, or human samples, as demonstrated in the surveillance of SARS-CoV-2), microbial ecology (through the analysis of microbial communities in the human gut, agricultural environments, or soil), and gene function analysis.

### 0.13.3 History

Sanger sequencing revolutionized DNA research, but it was expensive and slow, making it suitable only for small-scale studies. In contrast, next-generation sequencing (NGS), particularly Illumina sequencing, dramatically increased speed and efficiency, rendering it well-suited for large-scale genomic analyses. NGS enables whole-genome sequencing at a fraction of the cost compared to earlier methods and can detect complex variants and DNA modifications, although its error rate is relatively high.

Three major applications of three types of DNA sequencing in human genomics are as follows:

**Genome Resequencing:** This approach is used to identify genomic differences among individuals. It requires a reference genome, and mutations or variations are detected through alignment-based comparisons.

**De Novo Genome Sequencing:** This method is employed for sequencing the genome of an unknown species or for analyzing complex genomic regions. In the absence of a reference genome, the entire genome is assembled from scratch.

**Sequencing as a Molecular Counting Tool:** DNA sequencing can also serve as a quantitative tool. It is widely applied in gene expression and regulatory analysis, cancer research, and CRISPR screening.

Short-read sequencing is inexpensive and widely applied; however, due to the short length of its reads, it is challenging to precisely determine the genomic location of each read when reconstructing complex genomic regions. Shotgun sequencing overcomes the limitations of short-read sequencing by fragmenting the DNA, sequencing the fragments, and then reassembling them. Nonetheless, it still faces challenges related to sequencing errors (approximately one error per 1,000 nucleotides).

### 0.13.4 Standard Pipeline for Human Genome Sequencing

DNA is first extracted and the genome is fragmented to enable shotgun sequencing. The resulting reads are then aligned to a reference genome (Read Mapping), and genotype differences are subsequently inferred. This process allows us to determine which gene mutations may be associated with disease. However, challenges in read mapping arise due to repeated sequences; certain sequences occur multiple times in the genome, making it difficult to accurately determine the source of each fragment.

### 0.13.5 Genome Coverage

Genome coverage is defined as the average number of times a given genomic position is sequenced. A 30X coverage is the standard for high-quality genome sequencing, ensuring higher reliability for the detection of SNPs and structural variants.

### 0.13.6 SNP Calling

SNP calling distinguishes among sequencing errors, homozygous mutations (e.g., AA), and heterozygous mutations (e.g., AT). Moreover, the inability of short-read sequencing to differentiate alleles derived from the paternal and maternal chromosomes necessitates haplotype phasing, which can be achieved using long-read sequencing for comprehensive haplotype analysis.

### 0.13.7 Structural Variant Detection

Short reads can also be used to detect larger structural variants. For example, a heterozygous deletion may reduce the sequencing coverage to approximately 50% of the average genomic coverage. Structural variants can further be identified by examining abnormal distances between read pairs; if the distance between paired reads on the chromosome exceeds the expected DNA fragment size, it may indicate the presence of a deletion in the intervening region.

### 0.13.8 Some other Sequencing

Exome sequencing employs a laboratory technique to preselect all the DNA fragments that span gene exons, thereby reducing costs; however, its detection range is considerably limited. Genotyping determines a person's genotype at a specific set of preselected SNP positions. It can assay between 500,000 and 2,000,000 SNPs, enabling large-scale screening at a cost of less than $100 per sample, and is therefore widely used in academic research.

## Chatper 1.5

Exome sequencing employs a laboratory technique to preselect all the DNA fragments that span gene exons, thereby reducing costs; however, its detection range is considerably limited. Genotyping determines a person's genotype at a specific set of preselected SNP positions. It can assay between 500,000 and 2,000,000 SNPs, enabling large-scale screening at a cost of less than $100 per sample, and is therefore widely used in academic research.

Germline mutations are heritable, transmitted to offspring, and influence both evolution and genetic diseases. They can be identified via family trio sequencing, wherein the genomes of the parents and the child are sequenced. If the child carries a mutation that is absent in the parental genomes, this indicates that the mutation is a de novo event occurring in the germline.

In contrast, somatic mutations are not inheritable, although they may lead to cancer or other age-related diseases. On average, each child carries approximately 70 mutations; most of these mutations are benign, but some can result in disease.

The mutation rate of the human genome can be calculated as follows:

$$\text{Mutation rate} = \frac{70 \text{ mutations}}{2 \cdot (2.68 \times 10^9) \text{ bp}} \approx 1.3 \times 10^{-8} \text{ mutations per bp.}$$

Equivalently, this corresponds to a mutation rate of about $4.0 \times 10^{-10}$ per base pair per year of the parent's age.

Moreover, DNA storage and replication are remarkably precise. DNA in human germ cells can be stored for 20–40 years or longer and replicated hundreds of times, yet the overall error rate is only about one point mutation per 100 million base pairs.

### 0.13.9 Mutation Rates and DNA Damage

Mutation rates not only increase with parental age, but paternal contributions are greater. Germline mutations occur more frequently in males than in females, whereas errors in chromosomal segregation are primarily attributable to the mother. It was previously believed that the higher mutation rate in males was due to the increased number of germ cell divisions; however, recent research suggests that DNA damage may be the predominant cause.

Due to the low frequency of somatic mutations in the genome, higher-precision sequencing methods are required (e.g., duplex sequencing, in which a mutation is only confirmed if it is detected on both strands of DNA). Similar to germline mutations, somatic mutation rates increase with age; however, different tissues exhibit varying mutation rates that are not necessarily correlated with cell division rates but rather with DNA damage. Approximately 70,000 DNA lesions occur daily, yet only about one in a million of these lesions ultimately becomes a mutation. Thus, while DNA damage is common, most lesions do not result in mutations because cells possess robust DNA repair mechanisms.

Cancer is a class of diseases characterized by the uncontrolled replication of somatic cells, typically involving mutations in multiple genes. These mutator genotypes can either inhibit or enhance cell division. Mutations that promote rapid cell division or metastasis are selectively retained in cancer cells. Consequently, cancers often evolve higher mutation rates, usually as a result of mutations in DNA repair or proofreading genes.

### 0.13.10 Types and Mechanisms of Mutations

Single nucleotide mutations are the most common type, accounting for the vast majority of genomic mutations. Short Tandem Repeats (STRs) exhibit much higher mutation rates than single nucleotide polymorphisms (SNPs) due to replication slippage, and they play a key role in certain genetic diseases (e.g., Huntington's disease). Structural variants are less common but have widespread effects; approximately 60% of these variants are large deletions, while 30% are duplications, which may lead to gene loss or dosage effects. Aneuploidy, which occurs relatively frequently during fertilization, refers to abnormal chromosome numbers. However, due to the low survival rate of aneuploid cells, such abnormalities can lead to conditions such as Down syndrome.

### 0.13.11 Mutation Patterns: Transitions, Transversions, and CpG Effects

Transitions (A↔G and C↔T) are more common than transversions (A↔C, A↔T, G↔C, and G↔T) because their chemical structures are more similar, making them more likely to occur during DNA replication. Approximately two-thirds of point mutations are transitions.

The mutation rate at CpG sites is extremely high because methylated cytosine (5-methylcytosine, 5-mC) can spontaneously deaminate to thymine (T), and DNA repair mechanisms often fail to correct this error.

Furthermore, the mutation rate of mitochondrial DNA is up to 50 times higher than that of the nuclear genome, primarily due to its reduced DNA repair capacity, making it an important target for genetic research.

### 0.13.12 Short Tandem Repeats and Structural Variants

Short Tandem Repeats (STRs) are repetitive sequences (e.g., CACACACA…) that are highly prone to mutations. During DNA replication, replication slippage can occur when one DNA strand forms a loop structure, leading to insertions or deletions. This mechanism underlies the high mutation rate observed in STRs; for example, Huntington's disease is caused by an expansion of the CAG repeat in the *Huntingtin* gene.

Structural variants affect large segments of DNA and can arise through multiple mechanisms:

- **Recombination Errors:** Misalignment between repetitive sequences can cause large-scale structural rearrangements.
- **DNA Replication Errors:** Tandem repeats are especially susceptible to errors during replication, potentially resulting in extra copies of the sequence.
- **Double-Strand Break Repair Errors:** Inaccurate repair of double-strand breaks may lead to significant insertions or deletions.

These structural changes can have profound effects on human health, contributing to a variety of genetic disorders.

### 0.13.13 Aneuploidy and Chromosome Segregation Errors

Aneuploidy is a phenomenon characterized by an abnormal number of chromosomes, primarily transmitted by the mother. Unlike point mutations—which involve changes in the nucleotide sequence—chromosome segregation errors result from abnormal chromosome numbers. These chromosomal abnormalities are mainly inherited from mothers, especially older mothers, and can lead to conditions such as Trisomy 21. The high rate of chromosomal abnormalities observed in the oocytes of older mothers is a significant factor contributing to infertility. Consequently, compared to males, female oocytes exhibit a higher error rate during meiosis, primarily due to:

- Prolonged arrest in meiosis I for several decades, which leads to a deterioration of the chromosome segregation mechanisms.
- Kinetochore drift and the instability of the microtubule spindle, which increase the incidence of chromosome aneuploidy.