# Genomic Sequences

A **genome** is the complete set of DNA in an organism. It contains both **coding regions (genes)** and **non-coding regions**.

## What is Genomic Sequencing?

The **genome** of an organism is its complete set of **genetic material**, composed of a unique **DNA or RNA sequence**.

Each sequence consists of chemical building blocks called **nucleotide bases** — **adenine (A), thymine (T), cytosine (C), guanine (G)** for DNA, and **uracil (U)** replaces thymine in RNA. Determining the **exact order of these nucleotide bases** is known as **genomic sequencing** or simply **sequencing**.

### Genomes as Genetic Fingerprints

Every living organism, including **bacteria, viruses, and fungi**, has a distinct genomic sequence that acts as a **genetic fingerprint**.

By analyzing these sequences, scientists can:

- Identify disease-causing microbes.
- Track their origin and transmission.
- Study how they evolve over time.

### Whole-Genome Sequencing (WGS)

**Whole-Genome Sequencing (WGS)** is a laboratory technique used to determine the **order of nearly all nucleotide bases** in the genome of an organism.

In the context of **pathogens**, WGS helps:

I. Detect genetic variations or mutations.
II. Understand how microorganisms **spread within a population**.
III. Monitor **antibiotic resistance** and **outbreak sources**.

When applied at a **population or community level**, WGS provides valuable insights into **disease transmission patterns** and **microbial evolution**.

### Next-Generation Sequencing (NGS)

**Next-Generation Sequencing (NGS)** refers to a group of advanced sequencing technologies capable of processing **millions of DNA fragments simultaneously**.

Introduced in **2004**, NGS has largely replaced the earlier **Sanger sequencing method**.

## Key Features of NGS

- **High-throughput:** Enables sequencing of many samples in parallel.
- **Speed:** Generates large volumes of data in a short time.
- **Cost-effectiveness:** Dramatically reduces the cost per genome.
- **Accuracy:** Provides deep coverage for reliable analysis.

NGS has **revolutionized genomics** by making **whole-genome sequencing** faster, more affordable, and more accessible for research, clinical diagnostics, and public health surveillance.

**Some key points:**

- The genome of an organism is made up of a unique DNA or RNA sequence.
- Whole-genome sequencing (WGS) determines the order of all, or most, of the nucleotides in the genome.
- The information encoded in the genomes of bacteria, viruses, and fungi provide researchers with unique genetic "fingerprints."

# The genome sequencing process

## Step 1: Extraction

DNA or RNA is first **isolated (extracted)** from the biological source such as **bacteria, viruses, or other organisms**.

This step ensures that pure nucleic acids are obtained for further processing.

*Figure 1. DNA must be extracted from the cell.*

## Step 2: Library Preparation

The DNA or RNA to be sequenced must be specially prepared before it can be put into the sequencing machine. Although specific methods vary depending on the sequencing platform, the general steps include:

- **Conversion:** If the sample contains RNA or single-stranded DNA, it is converted into **double-stranded DNA**.
- **Fragmentation:** DNA strands are **broken into smaller fragments** of a desired length.
- **End Modification:** Special **adapters or tags** are added to the fragment ends so that the sequencing machine can recognize and process them.

Once these steps are completed, the prepared sample is called a **library**, ready to be loaded into the sequencer.

*Figure 2. DNA is cut before sequencing.*

## Step 3: Sequencing

In this step, the **sequencer reads the nucleotide bases (A, T, G, C)** of the DNA fragments.

Different sequencing technologies use different principles:

- **Fluorescent-based detection:** Each base emits a distinct color as it is incorporated, allowing identification.
- **Electrical signal detection:** Changes in electrical current are recorded as DNA strands pass through tiny pores (example: nanopore sequencing).


*Figure 3. Sequencers use*

The output is a large volume of **raw sequence data**, representing millions of short DNA reads.

## Step 4: Data Analysis

The raw sequence data are then **processed and analyzed** using bioinformatics tools.

Key steps include:

A. Assembly or Alignment: Short reads are either assembled into a full sequence or aligned against a reference genome.
B. Variant Identification: The new sequence is compared to known reference data to detect mutations, variations, or unique patterns.
C. Interpretation: These variations help scientists infer evolutionary or ancestral relationships, identify new species or pathogens, and study genetic traits or diseases.


*Figure 4. The sequence reads are aligned against a reference sequence.*

Example: The **human genome** consists of approximately **3.2 billion base pairs**.

# Types of Regions in a Genome

**1. Coding Regions (Genes)**

- These are parts of DNA that are **transcribed into RNA** and then **translated into proteins**.
- They contain the **instructions for building proteins** that perform vital cellular functions.
- Within a gene, the **coding sequence** is not continuous, it is divided into:
    - **Exons** : segments that **actually code** for protein.
    - **Introns** : **non-coding segments** within a gene that are **removed** during RNA splicing.

**Example:**

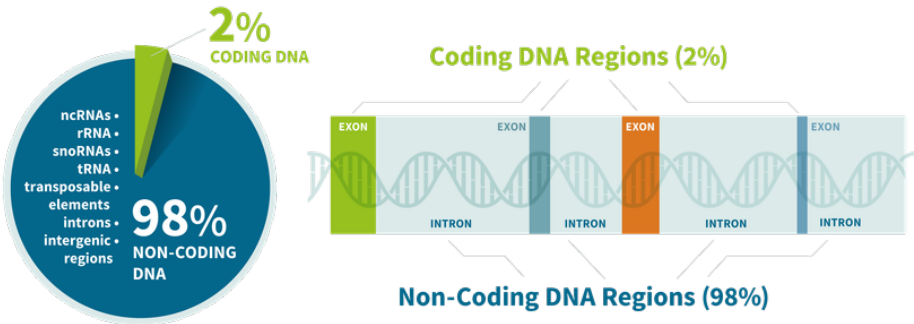
The **BRCA1** gene contains multiple **exons** and **introns**. Its exons code for a protein involved in **DNA repair**.

**2. Non-Coding Regions**

- These are DNA sequences that **do not code for proteins**, but may still have **important biological roles**.
- Once called *"junk DNA"*, many of these regions are now known to **regulate gene expression** or maintain **chromosome structure**.

**Examples:**

- **Promoters** and **Enhancers** : control when and how much a gene is expressed.
- **Introns** : non-coding parts **within genes**, removed during mRNA processing.
- **Repetitive Sequences** : contribute to genome **stability** and **evolution**.




Think of a genome like a **library**.

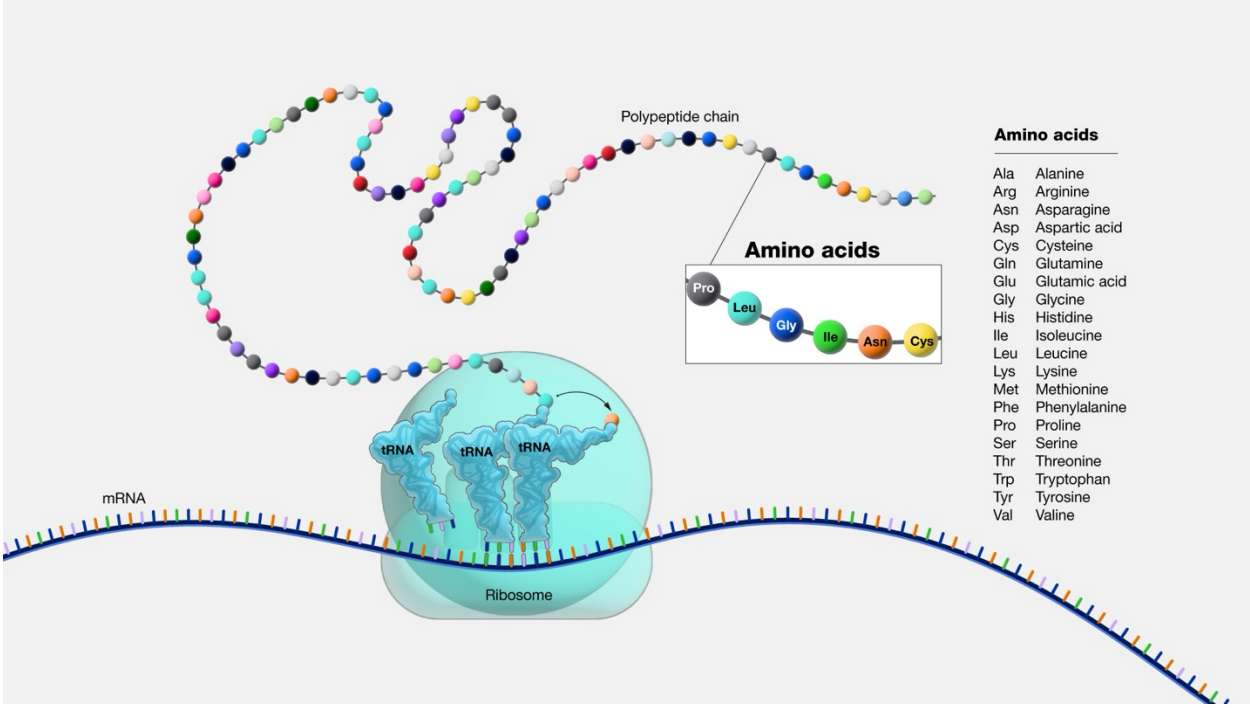
- Each **book** as a chromosome.

- Each **chapter** as a gene.
- Some pages are **blank notes** (non-coding DNA).

**Sequence Representation**

1. **Nucleotide Sequences**
    - DNA alphabet: A, T, G, C
    - RNA alphabet: A, U, G, C
2. **Protein Sequences**
    - 20 amino acid letters (A = Alanine, F = Phenylalanine, etc.)

An amino acid is the fundamental molecule that serves as the building block for proteins. There are 20 different amino acids. A protein consists of one or more chains of amino acids (called polypeptides) whose sequence is encoded in a gene. Some amino acids can be synthesized in the body, but others (essential amino acids) cannot and must be obtained from a person's diet.




3. **File Formats**

    - **FASTA** format:

>Sequence_ID Description

ATGCTAGCTAGCTAGCTA

- **GenBank** format: includes annotation (gene name, organism, etc.).

**Why Genomic Sequences Matter**

- Used for disease gene discovery.
- Comparative genomics: human vs chimp DNA.
- Forensics: DNA fingerprinting.

In 2003, completion of the **Human Genome Project** - "Book of Life" published for the first time.

**Practical Session: Download & Explore a Gene Sequence**

**Objective**: Learn to access and read genomic data.

1. Go to NCBI Gene Database.
2. Search for **"TP53 human"** (tumor suppressor gene).
3. View sequence in **FASTA format**.
4. Switch to **GenBank format**: show annotation.
5. Copy-paste sequence into a text editor: highlight features.

Tasks:

- Find the sequence length (number of base pairs).
- Identify the organism's name in the GenBank entry.

**Quick Review Questions**

1. What is a genome?
2. What's the difference between FASTA and GenBank formats?
3. What base replaces "T" in RNA?

Refs: https://www.cdc.gov/advanced-molecular-detection/about/what-is-genomic-sequencing.html