Project1 :
DNA sequence Analysis

What is Human Genome ?
Complete set of genetic instructions found in the human being and is composed of DNA which carries all information needed to build and maintain the body. It contains many DNA sequences organized into multiple chromosomes.

The genome is composed of :
46 chromosomes: 23 pairs (22 autosomes + X/Y sex chromosomes) of chromosomes.
Each chromosome is a separate DNA sequence (a long DNA molecule). So, in a ***diploid*** human cell(somatic cell), there are **46 DNA molecules**, each with its own sequence.
Genome encodes around 20 k protein coding genes, regulatory and non-coding genes.

What is Chromosome?

A chromosome is a long DNA molecule (genes + non coding regions) which carries part of the genetic material of the organism as each is a portion of the total genome. It is composed of :

- **genes** : Segments of DNA that code for the proteins(Histones).
- **regulatory regions** : control how and when genes are expressed.
- **Non-coding DNA** : Includes introns, repetitive sequences and structural elements.
- **Centromeres and telomeres** : Important for chromosomes stability and replication.

DNA molecule Vs. DNA sequence ?
DNA Molecule :
A physical double stranded helix with sugar-phosphate backbone, which is one long polymer of nucleotides.It can composed of millions of base pairs. It includes structural features like lenses, regulatory regions etc.

DNA sequence :
Textual representation of the DNA molecule, which is used ini biological computation and genomics. DNA sequence is the linear string of nucleotides/ bases(A, T, G, C) that makes up the molecule. It is one continuous sequence that contains many biological regions or features within the sequence such as gene, exons, introns.
- Unknown bases are denoted as **“N”.**
- *codons* : Triplets that code for the amino acids.
- *exons* : coding regions.
- *Introns* : non-coding regions.

It is stored in FASTA files, databases, and used for analysis, such as GC content calculation, ORF detection, Transcription/translation, Sequence alignment.

Structure of the DNA molecule

- DNA/ RNA is double stranded and have two distinct ends. The ends are labeled according to the C atoms in the sugar moelecule (deoxyribose or ribose) of the each neucleotide.
- DNA is double stranded, which two strands are complimentary and anti parallel.
- DNA follows spsecific base pairing rules **(A->T , T->A , G->C, C-> G)**
- Forward (coding) strand runs from **5'-> 3'** (eg: 5′ — A T G C G T A — 3′)
- reverse complement(template) strand runs from **3'->5'** (eg: 3′ — T A C G C A T — 5′)

What is 5' prime -> 3' prime mean ?

- **5′ (five-prime) end** → has a phosphate group attached to the 5′ carbon of the sugar.
- **3′ (three-prime) end** → has a hydroxyl group (-OH) on the 3′ carbon of the sugar.
- So, when DNA is synthesized, new nucleotides are always added to the 3′ end, never to the 5′ end.Hence, DNA (and RNA) grows from 5′ → 3′ direction.

DNA replication

- Copy entire DNA molecule before cell division, so that each daughter cell get a complete genome.
- **Helicase** unwinds the double strand & **DNA polymerase** adds comlpementary bases to the 3' of each strand and synthesizes new strands.
- Both strands are read in opposite directions but synthesized in the 5'-> 3' direction.

Molecule 1: Old forward strand (5′→3′)
- New complementary reverse strand (3′→5′) --> Lagging strand made of fragments.

Molecule 2: Old reverse strand (3′→5′)

- New complementary forward strand (5′→3′) --> Leading strand continuous.

Transcription

- Process of buding a mRNA(messednger RNA) by using the template strand of the DNA (**RNA polymerase** binds to promoter of the DNA strand) .
- RNA is syntehseized in the 5'->3' direction.
- Its sequence is identical to codinig DNA strand, except that thymine (T) in DNA is replaced by uracil (U) in RNA.

Transaltion

- Process When the riboseomes binds to the mRNA at the start codon(AUG) and build a protein using amino acids.
- Uses mRNA which is complementary to the DNA templpate strand and identical to the DNA coding strand (DNA isn't directly used).
- mRNA reads in 5'->3' direction.

- tRNA molecules bring amino acids matching each codon & Polypeptide chain forms until a stop codon is reached.

In summary, the template strand is the one actively transcribed into RNA, while the coding strand mirrors the RNA sequence and determines the genetic code for protein synthesis.

What is the Human reference genome (Gold Standard)?

Standardized digital blueprint of the human DNA- a composite map of how a typical genome looks like. It is an assembled mosaic from multiple individuals that serves as a baseline for comparison in genetics and genomics research. The mosaic includes **455 top-level sequences,** which include all chromosomes, alternate loci, and unplaced scaffolds.

- It represents the most common DNA sequence found across many people.
- Assembly name : The current version is called GRCh38 (Genome Reference Consortium Human Build 38).
- It includes :
  - Chromosomes included : All 22 autosomes, X and Y sex chromosomes,
  - mitochondrial DNA(MT) : 1 (circular DNA from mitochondria)
  - Alternate Loci : 261 /Alternate representation of the complex regions(MHC, LRC/KIR)
  - unplaced scaffolds : ~169/ sequences not assigned to a specific chromosome.
  - centromere sequence : First time centromeres are part of the reference assembly

- Each sequence is labeled with its chromosome name and contains raw DNA bases (A, T, G, C, N).

The reference genome is ***haploid***(it contains only one copy of the chromosomes). So it contains only 24 canonical chromosomes and their corresponding DNA sequences. But since the ensemble mosaic also contains alternates such as alternate loci, unplaced scaffolds , it contains around total of **455 DNA sequences**.

The reference genome is downloaded from Ensembl or NCBI.

Encoding the metadata about the DNA sequence in Ensembl FASTA

```
from bio import SeqIO
for record in SeqIO.parse(
“Homo_sapiens.GRCh38.dna.primary_assembly.fa”,“fasta”):
   record.description   # Full Rich ID of the sequence
```

1 dna:chromosome chromosome:GRCh38:1:1:248956422:1

- **1** = Chromosome 1
- **dna:chromosome** = type of sequence (DNA from a full chromosome)
- **chromosome:GRCh38** = genome assembly version
- **1:1:248956422:1** = coordinates of the sequence(Chromosome is 1, start position is 1, end poistion or length of chromosome is 248956422, forward strand is 1)

KI270519.1 dna:scaffold scaffold:GRCh38:KI270519.1:1:138126:1 REF

- **KI270519.1 =** genbank accesion ID for scaffold version 1
- **dna:scaffold scaffold =** DNA from a unplpaced scaffold
- **GRCh38 =** genome assembly
- **KI270519.1:1:138126:1 =** metadata (scaffold name, start, end, strand)
- **REF :** Part of reference gennome

The project Goal

The objective of this project is to develop a modular DNA sequence analysis pipeline that operates on biologically meaningful fragments of the human genome. Rather than analyzing entire chromosomes; which span millions of base pairs, the pipeline focuses on targeted regions such as individual genes (e.g., BRCA1, TP53), coding sequences (CDS), exons, or regulatory elements like promoters and enhancers.
By extracting specific genomic intervals from canonical chromosomes, the workflow enables precise computation of GC content, identification of open reading frames (ORFs), transcription of DNA to RNA, and translation of RNA to protein.

This approach supports functional annotation, mutation profiling, and biological interpretation of genomic data, laying the foundation for scalable, reproducible analysis in bioinformatics and genomics research.

Meaningful segments of the genes

- A gene (e.g., BRCA1, TP53)
- A promoter or enhancer region
- A coding sequence (CDS) or exon
- A specific genomic interval (e.g., chr17:43,044,294–43,125,482)

Types of genes in the human genome

Category	Description
Protein-coding genes	~20,000–25,000 genes that encode proteins—like BRCA1, TP53, EGFR, etc
Non-coding RNA genes	Genes that produce functional RNA but not proteins (e.g., rRNA, tRNA, miRNA, lncRNA)
Pseudogenes	Gene-like sequences that resemble functional genes but are usually non-functional due to mutations
Regulatory elements	Not genes per se, but crucial for gene expression (e.g., enhancers, silencers, promoters)
Mitochondrial genes	37 genes encoded in mitochondrial DNA, separate from nuclear chromosomes
Immunoglobulin and T-cell receptor genes	Highly variable genes involved in immune response
Olfactory receptor genes	One of the largest gene families—responsible for detecting smells

Functional Categories of the Protein- coding genes

Type	Description	Examples
Tumor suppressors	These genes often prevent the uncontrolled cell growth ,leading the loss function of mutations.	• **TP53** – "Guardian of the genome"; most frequently mutated in cancers • **RB1** – Regulates cell cycle progression • **PTEN** – Inhibits PI3K/AKT signaling • **RB1** – Regulates cell cycle progression • **PTEN** – Inhibits PI3K/AKT signaling • **CDKN2A** – Encodes p16, a cell cycle regulator • **BRCA1 / BRCA2** – DNA repair via homologous recombination
Oncogenes	Gain-of-function mutations or overexpression drive cancer growth.	• **MYC** – Transcription factor promoting proliferation • **KRAS** – GTPase involved in signal transduction; common in pancreatic, lung, colon cancers • **BRAF** – Kinase in MAPK pathway; V600E mutation is a key target • **EGFR** – Receptor tyrosine kinase; mutated in lung and brain cancers • **HER2 (ERBB2)** – Amplified in breast and gastric cancers
Transcription factors	Regulate gene expression; mutations can disrupt cell identity and growth control.	• **SOX2** – Maintains stemness; implicated in glioblastoma • **FOXO1** – Regulates apoptosis and oxidative stress response • **RUNX1** – Key in hematopoiesis; mutated in leukemia • **ETV6** – Fusion partner in various leukemias

Type	Description	Examples
Signal transduction	Transmit growth and survival signals; often hijacked in cancer.	• **EGFR, MAPK1, AKT1, PI3KCA, MET, ALK, RET** – All involved in key signaling cascades • **NOTCH1** – Can act as oncogene or tumor suppressor depending on context
Metabolic enzymes	Altered metabolism supports tumor growth and survival	• **IDH1 / IDH2** – Mutations produce oncometabolites; common in gliomas and AML • **CYP450 family** – Affects drug metabolism and resistance
Structural proteins	Changes affect cell adhesion, migration, and metastasis.	• **COL1A1 / COL3A1** – Collagen genes; altered in tumor microenvironment. • **VIM (Vimentin)** – Marker of epithelial-to-mesenchymal transition (EMT)
Transporters and channels	Regulate ion flow and nutrient uptake; some are altered in cancer.	• **CFTR** – Implicated in pancreatic cancer • **SCN5A / KCNQ1** – Sodium/potassium channels; altered in some tumors

Narrowed Projet Goal

Tumor suppressor genes are a foundational focus in cancer genomics due to their critical role in maintaining cellular integrity and preventing tumor development. These genes—such as TP53, BRCA1, BRCA2, RB1, PTEN, and CDKN2A—are among the most frequently mutated in human cancers, with *TP53 alone altered in over half of all malignancies.*

Loss-of-function mutations in tumor suppressors lead to unchecked cell proliferation, DNA damage accumulates(genomic instability), and evasion of apoptosis, making them clinically actionable targets for prognosis, risk assessment, and therapeutic decision-making. Functionally, these genes regulate essential processes like the cell cycle, DNA repair, and programmed cell death.

Analyzing their DNA sequences involves calculating GC content, detecting open reading frames (ORFs), transcribing to RNA, and translating to protein to study domain structure and function. Further, mutation hotspot mapping (e.g., TP53 exons 5–8), functional annotation, and variant interpretation (such as distinguishing BRCA1 variants of uncertain significance from pathogenic mutations) are key steps in understanding their impact on cancer biology and guiding precision oncology.

OR

This project focuses exclusively on tumor suppressor genes—key regulators of cell cycle control, DNA repair, and apoptosis(programmed cell death when damage is beyond repair) that are frequently mutated in human cancers. By targeting well-characterized genes such as TP53, BRCA1, BRCA2, RB1, PTEN, and CDKN2A, the pipeline will extract their genomic sequences

from canonical human chromosomes using reference FASTA and GTF files. These sequences will undergo modular analysis including GC content calculation, open reading frame (ORF) detection, transcription to RNA, and translation to protein, enabling functional annotation and domain-level insights. The workflow will also support mutation hotspot mapping (e.g., TP53 exons 5–8) and lay the foundation for variant interpretation in clinical genomics.

Importance —> Tumor suppressors are gatekeepers of:

- Cell cycle checkpoints (e.g., TP53 halts division if DNA is damaged)
- DNA repair pathways (e.g., BRCA1/2 fix double-strand breaks)
- Apoptosis (e.g., TP53 activates death signals when damage is irreparable)

Tumor supressor gene fragments or pseudogenes may appear in some non-canonical chromosomes/ regions due to duplication, alternate haplotypes, or assembly artifacts. For cleaner and reproducable analysis only canonical chromosomes are considred.

How to filter the canonical chromosomes using GTF filtering (pandas)

```
import pandas as pd
canonical_chroms= [str[i] for i in range(1,23)]+[“X”,”Y”,”MT”]


def filter_gtf(input_gtf, output_gtf):
    gtf = pd.read_csv(input_gtf, sep="\t", comment="#", header=None)
    gtf_filtered = gtf[gtf[0].isin(canonical_chroms)]
    gtf_filtered.to_csv(output_gtf, sep="\t", index=False,
header=False)

filter_gtf("Homo_sapiens.GRCh38.110.gtf", "canonical_only.gtf")
```

What is GTF(Gene Transfer Format) ?

Uses for genome annotation where each row to describe the genomic features like genes,exons, transcript, or coding sequence (CDS).

- Each row’s columns would be *chromosome, feature type, genomic cordinates, strand(+ or -), attributes (metadata like gene ID, gene name, transcription ID*).
- It’s structured as tab-delimited text with 9 columns.
- Use for locating genese, extracting cordinates, annotating features.(ex: Find the location of a specific gene such as BRCA, chromosome 17)
- gene level cordinates for specific genes (ex; BRCA1) are only located in the GTF file.

What is FASTA(.fa or .fasta) ?

Uses for genome sequence.
- Contains row neucleotides or protein sesquences.
- Structured like Header line (>) + sequence lines.
- Use for retrieving actual DNA/RNA/protein sequences (ex : Extract BRCA1's DNA sequence from chr17).

- Although *"record.description"* gives full ID, FASTA file doesn't contain any information about specific genes like BRCA1 or TP53 or about which exons or transcripts are involved.

So, to extract a specific gene, should use GTF file to find the cordinates of the specific gene, then slice the sequence from the FASTA file using the above cordiiates. Reverses complement if the strand is "-".

`gtf_filtered['attribute']` : **Column contains the metadata for the entry in the format** of—> `"gene_id "ENSG00000141510"; gene_name "TP53"; gene_type "protein_coding";`

In order to extract the gene name within the ", str.extract( ) function is applied on the attribute using a regex pattern to capture the value defined within the **capture group** .

```
gtf_filtered['attribute'].str.extract('gene_name "( [^ "] +)" ')
```

`SeqIO.parse(fasta_path,"fasta")` —> reads the FASTAA file located at the fasta path and returns an iterator object "SeqRecord" representing the each sequence entry(chromosome, scaffold)

`seqIo.to_dict()` —> converts the iterator to a dictionary, which its key is the header (id) of the Fasta file, values are the "SeqRecord' objects such as `.seq { actual DNA sequence}, .id { identifier], .description[ FASTA header}`

How to save extracted DNA sequences of tumor suppressor genes into a FASTA file ?

```
from Bio.SeqRecord import SeqRecord #SeqRecord class is a container that saves the sequence data along with metadata
From Bio.SeqIO import write #write function writes SeqRecords objects to a specified file

records=[SeqRecord(seq,id=gene, description=" ") for gene, seq in gene_seqs.items()]
# iterates over the key, value which are gene id and seq of the gene_seqs dictionary and creates a SeqRecord object with DNA sequence, id and description. Stores all SeqRecords in a list object.

 write(records, "tumor_suppressors.fa", "fasta") # write in to the FASTA file
```

What is ORF detection?

Open Reading Frame(ORF) is finidng the protein-coding regions in a DNA sequence. ORF detection is finding all continuous regions between the start and stop codons that could acually be translated into proteins. Its a stretch of DNA that;

- Starts with a **start codon** > **ATG(codes for Methionine).**
- Ends with a **stop codon >** one of **TAA, TAG, TGA.**
- In between start and stop codons, it contains sequence of codons**(triplets of bases) for amino acids** sequence
- Must be **in the same reading frame** (read in groups of three bases) with **no stop codon** interrupting it.

What is reading Frame ?

The way how the neucleotides are grouped into codons (triplets of bases).

- Only one frame determines the correct **functional protein** for the gene; that's where the start codon (ATG) and the stop codon (TAA, TAG, or TGA) align properly.
- DNA can read on three possible frames on the forward strand & 3 more frames on the reverse complement strand (not read by ribsomes in transaltion).
- Cells uses signals from promoters and start codons(ATG) to know where to start reading. Once translation starts correctly ribosomes move three bases (codon) at a time ignoring other feames comlpetelty.
- The other reading frames determines meaningless amino acids chains(they exist because ogf how cell is structured but not because cell need them all).
- But there are exemptions, where other reading frames can have functions:

- **Overlapping genes:**
  In viruses and some bacteria, one DNA region can **encode two different proteins** using different reading frames & it saves genome space. (ex: Some bacteriophages and HIV use overlapping ORFs).

- **Regulatory roles:**
  Alternate frames might influence **mRNA structure or stability**, even if they don't code for protein.

- **Frameshifting:**
  Some ribosomes can **intentionally shift** to another frame during translation, producing two proteins from one mRNA. (ex: Certain retroviruses (HIV) use **+1 or -1 frameshifting** to control protein ratios).

Why is detecting ORF important ?

- Tumor supressor genes and oncogens often contain mutations that distrupts the ORF or have alternative ORFs that change the protein product. Accurate ORF detection pinpoint these mutations.
- Help detecting frameshift mutation.Mutations such as insertions and deletions can shift the reading frame , break the correct ORF leading to non functional proteins. ORF detection tools help **locate exactly where** the frame is broken. (eg: A single base deletion in *TP53*

can destroy its open reading frame → p53 protein stops working → uncontrolled cell division → cancer).

- Sometimes mutations and regulatory changes can cause translate to begin at a different ORF which leads truncated or abnormal proteins and harmful functions. By detecting all ORFs, can identify alternaive protein isoforms that only appear in tumor cells and study their progression in the tumor progression or drug resistance.
- ORF detection can help annotate new protein-coding regions and to distinguish between functional genes and pseudogenes.

In overall it helps identify river mutations in coding regionsPredict loss of function in tumor supressors and desig targeted therapies that restore the lost function.

What is GC content ?

GC content is the percentage of neucleotides in the DNA molecule that are either guanine(G) or cytosine(C). GC content affects several biological and technical aspects of DNA:

- **Thermal Stability: G-C pairs form three hydrogen bonds (compared to two in A-T pairs), making GC-rich regions more thermally stable.**
- **Gene Regulation: GC-rich regions are often found in promoters and CpG islands, which are key in regulating gene expression.**
- **Mutation Rates: GC content can influence mutation susceptibility and DNA repair efficiency.**
- **Sequencing and PCR Efficiency: High GC content can complicate DNA amplification and sequencing due to stronger bonding.**

Importance of GC content in the tumor supressor genes ?

- **Mutation Hotspots: GC-rich regions may be more prone to methylation, especially at CpG sites, which can silence gene expression—a common mechanism in cancer.**
- **Epigenetic Regulation: Aberrant methylation of GC-rich promoters in tumor suppressor genes can lead to gene silencing, contributing to tumorigenesis.**
- **Therapeutic Targeting: Knowing GC content helps in designing targeted therapies or diagnostic tools (e.g., PCR primers, probes).**
- **Comparative Genomics: GC content can be used to compare tumor suppressor genes across species or between normal and cancerous tissues**

Key components in gene regulation

promoters

- Region located in the begining of the gene before coding region, which acts like the start signal for the transcription process.
- They are the binding sites for the RNA polymerase and transcription factors that help initiate transcription.
- Types of promoters :
    - Core promoter : contains essential elements like TATA box
    - regulatory Promoter : includes enhancers and modulators that enhance the gene expression.

CpG islands

- Regions in the DNA with higher frequency of CpG sites, where a cytosine(C) is follwoed by a Guanine(G) ; which is linked by a Phosphate hence CpG
- Always found near or within promoters of genes.
- Their role in regulation :
- unmethylated CpG islands : often associate with acive gene expression.
- Methylated CpG isalnds : can silenece gene expression, found in cancer or epigenetic regulation.

What is DNA methylation ?

An epigenetic modification, a chemical chnage to DNA that effects the gene expression without altering the DNA itself. It happens by adittion of the methyl group(-CH3) to the cytosine(C) base in the CpG site. The process is carried by an enzyme called DNA methyltransferases(DNMTs). It affects the followings :

- Gene silencing : methylation in CpG islands within promoter regions can block transcription and effectively turn off the gene.
- Normal function: methylation helps regulate gene expression during development, X-chromosome inactivation, supression of transposable elements.

What is Aberrant Methylation?

Abnormal patterns of DNA methylation that can disrupts the gene function and contribute the diseases such as cancer. Types of abberant methylation:

- **Hypermythelation** : Excessive methylation in the CpG islands, which can silence the tumor supressor genes leading to uncontrolled cell growth. eg : hypermythelation of BRCA1 or p16INK4a in breast and colon cancers.
- **Hypomethylation :** Loss of merhylation across the gene, It can cause genomic instability which leads activation of oncogens and reactivation of transposable elements.

Why myethylation matters in cancer research ?

- Biomarkers: Aberrant methylation patterns can serve as early indicators of cancer.
- Therapeutic targets: Drugs like DNMT inhibitors (e.g., azacitidine) aim to reverse abnormal methylation.
- Personalized medicine: Methylation profiling helps tailor treatments based on individual epigenetic signatures.

Tumorigenesis : Also called as carcinogenesis/ oncogenisis, the biological process on how the normal cells transform into cancer cells, which involves series of genetic and epigenetic changes that distrupt the normal cell regulation. Three main stages of the tumorigenesis :

- Initiation : occurence of geetic mutation caused by carcinogens, radiation or inherited defets.
- promotion: Mutated cell begins to proliferate abnormally(epigenetic changes can silence tumor supressor genes or activate oncogens.
- progression : cells aquire more mutations, become morr abnormal and resistant to cell death forming a malignant tumor.

Oncogens : Mutates or overexpressed genes rgar drive cell division.

Epigenetic modification : A chemical change to DNA or to the proteins (histones) that package DNA, which affects how genes are turned on or off — without changing the DNA sequence itself. Types of epigenetic modifications :

- DNA mytheliation
- Histone Modificaion : DNA wraps around a protein called histones and how tightly it wraps around the proteisn affect the gene activity. Chemical chnages that effects gene activities : acetylation : open DNA and activates gene deactylation : tightens DNA and silences genes Methylation, phosphorylation, ubiquitination, : Can either activate or supress gene depend on the site
- Non coding RNA regulation : Small RNAs (like microRNAs) can block or degrade mRNA, preventing a gene from making its protein.