🧬 Assignment: From GWAS Variants to 3D Genome Interpretation

Goal

This assignment connects **statistical genetics (GWAS)** with **3D genome organization (Hi-C)** to understand how genetic variants contribute to disease risk beyond linear genome annotations.

Context

GWAS identifies disease-associated variants (versions of the same gene with mutations), but most mutations lie in **non-coding regions**. The information available in GWAS data about which genes are affected is usually created algorithmically, based on the location of the mutation (e.g. the nearest gene in the sequence). However, the functional meaning of the mutation often requires **3D chromatin information**, not just linear distance.

Aim

1. Retrieve and interpret real GWAS data via public APIs.
2. Integrate GWAS signals with an artificial Hi-C map to infer **physically affected genes**.

Exercise 1 — GWAS API Exploration & Risk Aggregation

🎯 Objective

Learn how to **retrieve, interpret, and integrate GWAS data** from multiple APIs and quantify **cumulative genetic risk** for a disease.

You will construct a **GWAS summary DataFrame** that will be used in Exercise 2.

🔬 Biological background

GWAS studies report statistical associations between **genetic variants (SNPs)** and diseases. Each association provides:

- A **risk allele** – the position (encoded as rs…) and the nucleotide (A/C/T/G) in this position
- An **odds ratio (OR)** describing how a given nucleotide in this position increases the disease risk
- A **risk allele frequency** – how popular is such mutation in the population

Take into consideration that:

- Associations are distributed across multiple APIs
- Genomic location and allele structure must be resolved separately
- Individual variants have small effects, but **combined risk matters**

🧪 Task 1.1 — Retrieve GWAS associations for a disease

Choose a disease. The possible choices are:

- **Crohn’s disease**
- Type 2 Diabetes
- Rheumatoid Arthritis
- Asthma
- Schizophrenia

You have three APIs to use

- `https://www.ebi.ac.uk/gwas/rest/api/associations` - containing GWAS association for a chosen diseases – it gives you the list of variants given in the form of rsid-nucleotide (e.g. rs312421-C).
- `https://www.ebi.ac.uk/gwas/rest/api/singleNucleotidePolymorphisms/{rsid}` – this can be used to resolve the position of a given rsid – the chromosome and location within
- `https://rest.ensembl.org/variation/human/{rsid}?` – this can be used to find out other nucleotides in the given position, in particular the second most common

Your tasks:

- Query GWAS associations for the chosen disease
- Extract:
  - rsID
  - odds ratio
  - p-value
  - reported risk allele
  - risk allele frequncy (if available)
  - chromosome
  - genomic position
- From the third API identify:
  - all possible alleles at this position
  - minor allele
  - ancestral allele (if available)

📌 **Deliverable:**
A DataFrame with at least:

```
rsid | odds_ratio | pvalue | risk_allele | risk_frequency | chromosome |
position | all_alleles
```

🧪 Task 1.2 — Identify the “most dangerous” SNP set

Define the **most dangerous SNP set** as the subset of SNPs that **maximally increases disease risk** when combined.

Your tasks:

1. Select SNPs with the strongest effect sizes.
2. Compute **accumulated risk** assuming multiplicative odds ratios:

$$OR_{total} = \prod_i OR_i$$

3. Estimate the probability that a randomly chosen individual carries **all risk alleles**:

$$P = \prod_i f_i$$

📌 **Answer the following questions:**

- Which SNPs form the most dangerous set (construct a dataframe or list with position and nucleotide increasing the disease risk)?
- What is the total accumulated odds ratio?
- What is the probability of carrying all risk alleles?
- Is this genotype common or rare in the population?

Exercise 2 — GWAS Meets 3D Genome Organization

🎯 Objective

Use an **artificial Hi-C map** to understand how GWAS variants affect genes through **3D chromatin interactions**.

🧠 Why GWAS alone is not enough

GWAS often reports a **nearest gene**, but:

- Regulatory elements act over long distances
- Enhancers can skip nearby genes
- Linear proximity ≠ functional proximity

Hi-C reveals physical contacts, allowing us to:

- Link SNPs to distant genes
- Identify regulatory targets
- Interpret non-coding risk variants

🧪 Task 2.1 — Generate an artificial Hi-C map

Using the function provided in the assignment materials:

- Input your GWAS DataFrame
- Generate a Hi-C map covering the SNP region

📌 **Deliverable:**

- Hi-C contact matrix
- Bin genomic coordinates
- Short explanation of modelling assumptions

🧪 Task 2.2 — Identify SNP-affected regions in 3D

Your tasks:

- Locate bins corresponding to GWAS SNPs
- Identify bins with **increased contact frequency**
- Determine which genomic regions physically interact with risk loci

📌 **Deliverable:**

- Visualization of Hi-C map (plot)
- Highlight SNP-centered interaction regions

🧪 Task 2.3 — Infer affected genes

Using gene annotations (or simplified assumptions):

- Identify genes located in interacting bins
- Compare with GWAS-reported genes
- Identify **new candidate target genes**

📌 **Answer the following questions:**

- Which genes are implicated by 3D contacts?
- Which genes were not obvious from GWAS alone?
- How does 3D genome organization change interpretation of risk variants?
