# Genetic Risk Scores for Type 2 Diabetes Using Python

## 1 Background:

Type 2 Diabetes (T2D) is a metabolic disorder with a significant genetic component. This project focuses on calculating genetic risk scores (GRS) based on identified SNPs associated with T2D from GWAS studies.

## 2 Objective:

To write a Python program to calculate GRS for T2D for 1000 individuals, utilizing provided genotypic data and GWAS information. This includes data handling, computation of risk scores, visualization, and identification of high-risk individuals.

## 3 Genetic Risk Score:

The formula for calculating the Genetic Risk Score (GRS):

$$GRS = \sum(\log(\text{Odds Ratio}) \times \text{Allele Count})$$

In this formula:

- $\sum$ denotes the summation across all SNPs.
- log(Odds Ratio) is the natural logarithm of the odds ratio associated with each SNP's risk allele.
- Allele Count is the number of risk alleles (0, 1, or 2) present in an individual's genotype for each SNP.

## 4 Provided Data:

1. GWAS Information for T2D:

| rsID | Risk Allele | Odds Ratio |
|---|---|---|
| rs7903146 | T | 1.4 |
| rs1801282 | G | 1.3 |
| rs5219 | A | 1.2 |
| rs4402960 | G | 1.5 |
| rs13266634 | C | 1.6 |

2. Genotypic Data for 1000 Individuals:

The genotypes.tsv table of 1000 individuals, where each row represents an individual, and each column represents a SNP.

Here are the top five rows of the genotypic data table:

| individual_id | rs13266634 | rs1801282 | rs4402960 | rs5219 | rs7903146 |
|---|---|---|---|---|---|
| 1 | CA | AA | AA | TT | AA |
| 2 | CA | GA | AA | TT | AA |
| 3 | CC | AA | AA | TT | AA |
| 4 | AA | GG | AA | TT | TT |
| 5 | CA | AA | AA | AT | AA |

3. Example Calculation for First Individual:

For `individual_id` 1, suppose the genotypes are:

- `rs13266634`: CA
- `rs1801282`: AA
- `rs4402960`: AA
- `rs5219`: TT
- `rs7903146`: AA

The GRS is calculated as follows:

- `rs13266634` (C, 1.6): 1 risk allele, score contribution = $\log(1.6)$
- `rs1801282` (G, 1.3): 0 risk alleles, score contribution = 0
- `rs4402960` (G, 1.5): 0 risk alleles, score contribution = 0
- `rs5219` (A, 1.2): 0 risk alleles, score contribution = 0
- `rs7903146` (T, 1.4): 0 risk alleles, score contribution = 0

Thus, the total GRS = $0 + 0 + 0 + 0 + \log(1.6) \approx 0.47$

## 5 Tasks:

- Data Preprocessing
- Risk Score Calculation
- Statistical Analysis and Visualization
- Identification of High-Risk Individuals

## 6 Deliverables:

1. Google Colab Notebook: Create a Google Colab notebook that includes all parts of the project:
    - Data Loading and Preprocessing.
    - Risk Score Calculation.
    - Statistical Analysis and Visualization. Use Python libraries like Matplotlib and Seaborn to visualize the distribution of genetic risk scores. Include histograms, density plots, or other relevant visualizations.
    - Identification of High-Risk Individuals: Write code to identify individuals within the top 5% of GRS.
    - Discussion: Use markdown cells to interpret your findings, discuss potential implications for T2D risk prediction, and possible limitations of this approach.
2. Comprehensive Report within the Notebook: Alongside the code, your Colab notebook should include a detailed report. Use markdown cells to document:
    - The methodology used.
    - The results obtained from the analysis.
    - Visualizations of the data.
    - A discussion section covering the interpretation of results, implications, and limitations.
    - Any conclusions drawn from the project.
3. Shareable Link to the Notebook: Once your project is complete, ensure that the Google Colab notebook is shareable and accessible. Provide a link to the notebook as part of your submission.