Genetic basis of caffeine consumption

Background

- Genome Wide Association Studies (GWAS) help elucidate genetic variants that are associated with traits/phenotypes.
- In this project, you will be identifying the genetic determinants of caffeine consumption with a simulated dataset. You will follow through a GWAS pipeline and critically assess the methods and results of each step.
- At the end of this project, you will gain a deeper understanding of GWAS methods as well as the factors that come into play when performing this type of research.

Project Overview

Sections:

1. Setting up python environment
2. SNP-level filtering (call rate and MAF)
3. GWAS (with and without covariates)
4. GWAS at the population level
5. GWAS meta-analysis
6. Functional analysis
7. (Optional) Fisher's method from a mathematical perspective

Total Points: 24 (scaled to 6)

Final report: Submit a ZIP file where we can reproduce your results. This includes your report as a jupyter notebook in both the **html** and **ipynb** format.

Fisher’s Method: Meta Analysis

1. P-values are uniformly distributed between 0 and 1 = U(0,1) unless there is an effect
2. Transformation of U = U(0,1) --> - 2 * ln(U) = Chi2(df=2)
3. Combining p-values from different group is equivalent to -2 * sum(log(p1) + log(p2) + ….)
4. Combining p-values per group (n) follows: Chi2(df=2*n)
5. Null hypothesis Chi2: no effect, Alt hypothesis: at least 1 group has an effect

Few Comments

- Your input files will need to be cleaned before you can analyse them. See the project description for further details.
- You do not need to complete the optional task - you can achieve full marks without it.
- The structure of your report will also factor into your final score - keep things tidy!
- If you are not sure about anything, ask for help!