Title: Building a Mendelian Randomization Pipeline to find causal estimates between metabolomic exposures and diseases

Project Description:

The project aims to develop a Mendelian Randomization (MR) Pipeline to investigate causal relationships between metabolomic exposures and diseases. Mendelian Randomization is a statistical approach that uses genetic variants as instrumental variables to infer causal effects, minimizing confounding and reverse causation that often affect observational studies. In this study, large-scale metabolomic and genetic datasets will be analyzed to identify potential causal associations with various diseases. The pipeline will involve data preprocessing, instrumental variable selection, causal effect estimation, and statistical validation using multiple MR models. Additionally, the project will integrate sensitivity analyses, including pleiotropy checks, heterogeneity testing, and leave-one-out analyses, to ensure robust causal inference. This project will also focus on translating the R package TwoSampleMR into Python, making Mendelian Randomization more accessible within Python-based bioinformatics workflows. This translation includes harmonizing exposure and outcome datasets, replicating causal effect estimation methods, and implementing visualization tools for MR results. The research will contribute to the understanding of metabolic pathways and their role in disease development, ultimately providing valuable insights for precision medicine, biomarker discovery, and targeted interventions. Additionally, the project offers opportunities to explore different causal inference methods, advanced statistical modeling, and the use of publicly available GWAS summary statistics from resources like UK Biobank and OpenGWAS.

Duties/Activities:

- Work with large-scale metabolomic and genetic datasets, performing data cleaning and transformation using R/Python.
- Handle missing data, normalization, and quality control of genetic and metabolomic variables.
- Identify valid genetic instruments using GWAS summary statistics.
- Implement multiple MR models such as Inverse-Variance Weighted (IVW), MR-Egger, Weighted Median (WM), and other robust methods.
- Conduct sensitivity analyses (e.g., heterogeneity tests, pleiotropy assessments, leave-one-out analysis) to validate MR findings.
- Apply statistical methods to estimate causal effects and assess their significance.
- Visualize MR results using forest plots, scatter plots, and other graphics.
- Analyze causal estimates and generate reports on findings.
- Translate the R package TwoSampleMR into Python, implementing key functionality including data harmonization, causal effect estimation, and sensitivity analyses

Required Skills:

- Programming skills in R or Python. (Preferably both)
- Familiarity with statical methods and knowledge of genetic data is a plus.
- Problem solving and analytical skills.

Learning Opportunities:

- Work on real projects.
- Elevate bioinformatics knowledge.
- Gain experience in using bioinformatics tools relevant to genetics.
- Learn how to work with real-world genetic and metabolomic datasets.
- Potential opportunity to contribute to research publications.
- Understand how genetics, metabolomics, and disease risk factors are interconnected.
- Contribute to creation of a new Python library.
- Develop expertise in statistical modeling and causal inference methods.

Expected Team Size: *2-4*

Mentors

Name: Dr. Mohamed Elshrif email: melshrif@hbku.edu.qa