**Scripts for Multi-Ancestry GWAS of Cardiovascular Traits**

This repository contains a collection of MATLAB, R, and BASH scripts used to perform a comprehensive genetic analysis of cardiovascular traits in individuals of African (AFR) and European (EUR) ancestry. The analysis pipeline includes phenotypic comparisons, genome-wide association studies (GWAS), multi-trait meta-analysis using JASS, heritability and genetic correlation estimation with LDSC and LAVA, and post-GWAS analyses such as novelty assessment and functional annotation using FUMA and LDLink.

**Data Availability**

Please note that the primary individual-level phenotype and genotype data used in this study were obtained from the **UK Biobank Resource**. These data are subject to a Material Transfer Agreement and are not publicly available. Access can be obtained by approved researchers through a formal application to the UK Biobank.

* **UK Biobank Access Application:** <https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>

The following publicly available reference datasets were used in this analysis:

* **1000 Genomes Project:** Used as a reference panel for linkage disequilibrium (LD).
  + **Link:** <https://www.internationalgenome.org/data>
* **GWAS Catalog:** Used for novelty assessment of identified SNPs.
  + **Link:** <https://www.ebi.ac.uk/gwas/>
* **LDSC LD Scores:** Standard European LD scores and HapMap3 SNPs list.
  + **Link:** [https://alkesgroup.broadinstitute.org/LDSCORE/](https://www.google.com/search?q=https://alkesgroup.broadinstitute.org/LDSCORE/)
* **LAVA Genomic Loci:** Predefined genomic loci for local genetic correlation analysis.
  + **Link:** [https://ctg.cncr.nl/software/lava/](https://www.google.com/search?q=https://ctg.cncr.nl/software/lava/)

**Script Descriptions**

Here is a detailed description of each script in the analysis pipeline:

**final\_HPC\_heartpipeline.m**

* **Language:** MATLAB
* **Description:** This is the main, overarching script that orchestrates the entire analysis pipeline. It begins by loading and pre-processing the phenotypic data from the UK Biobank (hpc\_ukbiobank\_heart.csv). It performs statistical comparisons of cardiovascular traits (Pulse Rate, Blood Pressure, etc.) between the AFR and EUR ancestry groups, including Welch's t-tests and generating summary statistics. The script also calculates phenotypic correlations, performs functional annotation of SNPs using CADD scores, processes enrichment analysis results, and integrates outputs from other tools like FUMA and JASS to generate final figures and supplementary tables.

**create\_jass\_input.m**

* **Language:** MATLAB
* **Description:** A pre-processing script designed to format GWAS summary statistics for use with the **JASS (Joint Analysis of Summary Statistics)** software. It reads per-trait summary statistics (Beta and Standard Error), calculates Z-scores, and filters for high-quality variants. It then writes the formatted, chromosome-split output files required by JASS for multi-trait analysis.

**get\_LDlink\_GWAS\_Catalog.m**

* **Language:** MATLAB
* **Description:** This script performs novelty assessment for significant SNPs identified in the study. It takes lead SNPs from FUMA output and uses the curl command to query the **LDLink API (ldtrait)**. For each SNP, it identifies known SNP-trait associations in the GWAS Catalog for variants in high linkage disequilibrium (LD). This process helps to determine whether the identified loci are novel or have been previously associated with cardiovascular or other traits in AFR and EUR populations.

**run\_ldsc\_local.sh**

* **Language:** BASH
* **Description:** A wrapper script for running **LD Score Regression (LDSC)**. It automates the two main steps of LDSC:
  1. **Munging:** Formats the input summary statistics for each trait and ancestry using munge\_sumstats.py.
  2. **Analysis:** Estimates SNP-based heritability (h2) for each trait and calculates the genetic correlation (r\_g) across all cardiovascular traits for both AFR and EUR ancestry groups using ldsc.py. The script is configured with the correct paths for ancestry-specific LD score reference panels.

**parallel\_cluster\_lava\_analysis.R**

* **Language:** R
* **Description:** This script performs local genetic correlation analysis using the **LAVA (Local Analysis of [co]Variant Association)** package. It is optimized to run in parallel on a high-performance computing (HPC) cluster managed by SLURM. It processes predefined genomic loci to estimate local SNP-heritability and genetic correlations for the cardiovascular traits within both the AFR and EUR ancestry groups, providing a more granular view of shared genetic architecture.

**process\_mtag\_to\_fuma.sh**

* **Language:** BASH
* **Description:** This is a data-wrangling script that processes the output from a multi-trait analysis. It reads a tab-separated file, calculates a combined Z-score, and estimates the corresponding beta, standard error (SE), and variance of the beta. This creates a summary statistics file that is properly formatted for input into downstream fine-mapping or colocalization tools.

**run\_jass\_project.sh**

* **Language:** Command-Line Call
* **Description:** This file contains an example command for executing the **JASS** software. It demonstrates how to initiate a multi-trait GWAS project by specifying the input data tables (containing Z-scores and covariance information), the phenotypes to be jointly analysed, and the desired output paths for the results file and associated visualizations like Manhattan and quadrant plots.