GLGC 2024 Cohort Analysis Plan

Created by: Jacqueline S. Dron
Created: July 22, 2023
Last updated: December 8, 2023

Diagram Key

- Predefined process
- Use script or document from GLGC repository
- Cohort-specific method
- Data set
- Results to upload
- Intermediate results
- GLGC Meta-Data tab input: cohort_summaries
- GLGC Meta-Data tab input: technical_summaries
- GLGC Meta-Data tab input: phenotype_summaries

Set Up

Start

Download GLGC Pipeline repository

Set up folder directories

Install required dependancies

Do you already have imputed genotypes?

No

Yes

If you are working with whole genome sequence data, please follow the instructions provided in the Analysis Plan for the preparation of genetic data. Otherwise, you can refer to this workflow for "Sample and Variable Preparation" and "GWIS".

Genotype and Sample QC

Array genotypes

Array information

Examples:
Call rate
Sex discordance
Inbreeding coefficient
Heterozygosity
Relatedness
Missing phenotypes
Known lipid disorders

Sample QC

Variant QC

Examples:
MAF
Call rate
HWE P-value

Variant and sample QC information

QC PLINK bfiles (BED/BIM/FAM)

Sample and Variable Preparation

Assign individuals to population groups

Method to define and assign population groups

Study samples

Sample breakdown by sex and population group

Is this a case-control cohort?

Yes

Prepare phenotype variables separately for cases and controls

No

Are there youths and adults?

Yes

Prepare phenotype variables separately for youths and adults

No

Are there fasting measurements?

Yes

For LDL-C and TG analyses, prepare phenotype variables separately

No

Remove phenotypic outliers and convert units if appropriate for HDL-C, LDL-C, TG, age, and BMI

Prepare covariates

Regenerate new PCs for the final set of study samples

Methods for PC generation

PCs for each population group

Complete phenotype set

Provide summaries for each outcome and exposure

Preapre a phenotype file for each analysis group

Phenotype files for each analysis group

Imputation

Download TOPMed reference file and update file name

Check genotype data prior to imputation

or

optional

Ancestry-specific allele frequency check

Method for ancestry-specific allele check

Prepare files for imputation

Autosome .VCF files

Imputation with TOPMed refernece panel

R2 filter: 0.3
Phasing: Eagle v2.4
Population: TOPMed panel

chr*.dose.vcf.gz

chr*.info.gz

*.log

qcreport.html

Post-imputation variant pruning

Were multiple batches run?

Yes

Merge imputation batches

No

chr*.imputed.poly.filtered.vcf.gz

GWIS

Does the cohort have family structures?

No

Prepare genetic input

or

Only **one** file type is needed to run GEM. We provide a script that takes the processed imputation files and outputs .BGEN files.

chr*.pgen

chr*.bgen (preferred)

chr*.bfiles (BED/BIM/FAM)

Complete the 4.1_groups-for-analysis_WORKBOOK.xlsx

Run GWIS using GEM for each analysis group

Summary stats for each analysis group

Yes

Prepare genetic input

or

chr*.gds

chr*.bgen

The provided scripts work assuming .GDS files are being used.

Complete the 4.1_groups-for-analysis_WORKBOOK.xlsx

Generate a GRM

Generate the null model for each analysis group

Run GWIS using MAGEE for each analysis group

Summary stats for each analysis group

Specify GWIS parameters and analysis groups

Assess quality of summary statistics

Result Upload

Apply naming convention to all files for upload

Confirm appropriate directory structure and file location

Upload results to SFTP site

Complete post-analysis Google survey

Finish