ALZHEIMER

In the first part of the code, instructions are provided regarding the file locations. The genetic information provided by ADNI will be utilized, including ADNI1 GWAS, ADNI WGS+omni, ADNI GO/2 GWAS (both datasets), and ADNI3 GWAS (both datasets). Additionally, information will be provided on reference datasets to download and how to manipulate them to make them most usable for the analysis.

- all_phase3.pgen.zst
- all_phase3.pvar.zst
- phase3_corrected.psam
- All_20180423.vcf.gz
- Homo_sapiens.GRCh37.dna.primary_assembly.fa.gz
- 00-All.vcf.gz
- ADNIMERGE.csv
- AD_sumstats_Jansenetal_2019sept.txt.gz

QUALITY CONTROL

For each ADNI dataset, the following analyses are performed:

1) **upload_pheno_AD.R :** This script merges two input files (ADNIMERGE.csv and .fam file), identifies rows with Alzheimer's Disease status ("AD", "LMCI", and "EMCI"), checks for sample IDs in both files, updates phenotypes accordingly, and renames the RID column.

2) **sex_missing.R :** This R script performs quality checks on genetic data using PLINK outputs (--check-sex, --missing). It checks for sample sex consistency and the proportion of missing genotypes. Valid samples are identified and saved separately for further analysis.

3) **perl HRC-1000G-check-bim.pl :** This tool checks for consistency of strand, alleles, positions, Ref/Alt assignments, and frequencies between your SNPs and the HRC panel. It produces a set of PLINK commands to update or remove SNPs based on the results of these checks. Running this tool also requires a file containing the allele frequencies in your genotype file. The file is modified according to the instructions in https://rpubs.com/maffleur/452627.

4) The .vcf file is subsequently analyzed using bcftools sort, bcftools index, bcftools plugin using as a reference file riferimento All_20180423.vcf.gz and Homo_sapiens.GRCh37.dna.primary_assembly.fa.gz, again bcftools sort, and bcftools index to perform bcftools norm --check-ref.

5) The resulting .vcf.gz file is finally split into 22 files, one for each chromosome.

IMPUTATION

For this process, the Michigan Imputation Server was used with the following files:

- Select: Run -> Genotype Imputation (Minimac4)
- Reference Panel: 1000G Phase 1 v5
- Input files: VCF.gz
- Array Build: GRCh37/hg19
- rsq Filter: NULL
- Phasing: Eagle2
- Population: EUR
- Mode: Quality Control & Imputation

POST IMPUTATION

The Michigan Imputation Server returns a .zip file for each chromosome uploaded, which needs to be opened using the password sent via email by the platform.
Following that, several post-imputation quality control steps are performed. In particular:
1) Extract the INFO metric from the VCF file for all variants. Generate a list of variants with an INFO metric less than 0.3. Utilize VCFtools to exclude those variants based on their position.
2) For non-header lines, the script verifies whether the length of the third field exceeds 16000 characters. If it does, the field is truncated to 16000 characters.
3) The .vcf files are analyzed using bcftools sort and bcftools index before being merged into a single file with bcftools concat, with the following parameters: (-allow-overlaps, --remove-duplicates).
4) Conversion to .bim/.bed/.fam format is carried out using PLINK2. Additionally, filters are applied: --double.id, --snps-only, --max-alleles 2, --maf 0.05, --geno 0.1, --hwe 5e-7.
5) **post_imputation_fam.R :** It updates the SEX column in the .fam file based on the information from a file .fam got before imputation. Additionally, it modifies the names of RID and FID columns in the .fam file. Samples that are labeled as "Not Hisp/Latino" in ADNIMERGE.csv are removed from the analysis. Phenotypes are updated based on specific criteria from ADNIMERGE.csv, and the .fam file is accordingly modified.
6) a questo punto tutti i dataset di ADNI vengono uniti e successivamente viene fatto un --recover-var-ids force partial utilizzando come riferimento il file 00-All.vcf.gz.
7) infine vengono fatti ulteriori filtri QC sul file finale utilizzando --maf 0.05 e --geno 0.1

POPULATION ANCESTRY

1) To perform the analysis, we followed the guidelines outlined in the following link: https://cran.r-project.org/web/packages/plinkQC/vignettes/AncestryCheck.pdf. The file is then manipulated to make it consistent with all_phase3. The two files are merged.
2) The resulting PCA is used in the R function present in the file **evaluate_check_ancestry.R :** It computes the maximum Euclidean distance (maxDist) of the European reference samples from this centre. All study samples whose Euclidean distance from the centre falls outside the circle described by the radius r=europeanTh* maxDist are considered non-European and their IDs are returned as failing the ancestry check.
Subsequently, individuals not belonging to the European population are excluded.
3) **PCA-plot.R :** creates and saves 2D and 3D plots.
4) PerformLD-pruning with --indep-pairwise 1000 50 0.1
5) Remove the subjects with relatedness > 0.1

PRSice-2

1) **pheno_covariate.R :** This R script reads a .fam file and extracts FID, RID, SEX columns. It also reads a PCA.eigenvec file to extract principal components (PCs) and merge them with the covariate information and then save as "covariate.txt" with the first six columns. Additionally, the script creates a new phenotype file by extracting specific columns from the .fam file and saves it as "recodedpheno.txt".
2) **Base_file.R :** This script reads a file named used as base file in PRSice-2, filters out data within a specific range of base pair values on chromosome 19, adds information for two

SNPs related to APOE-e4 and APOE-e2, and saves the processed data into multiple output files.
3) In PRSice-2, the files previously obtained are inputted pheno_covariate.R e Base_file.R (--cov, --pheno), --base. In addition, the commands are specified --clump-kb 1000, --clump-r2 0.1, --target and --all-score.
4) **PRSice.plot.R :** This script performs an analysis of polygenic scores (PRS) using PRSice-2 output files and recoded phenotype data. It generates density plots for PRS1, PRS2, and the best PRS, distinguishing between healthy and diseased individuals.
PRS1 is referred to PRS at Pt 0.01, while PRS2 is referred to PRS at Pt 0.5 extrapolated from .all.score. The resulting plots are saved as PNG files for further examination.

PRS PERFORMED IN PLINK

The PRS was calculated using PLINK, employing a method described in https://choishingwan.github.io/PRS-Tutorial/plink/ .
1) One way of approximately capturing the right level of causal signal is to perform clumping, which removes SNPs in ways that only weakly correlated SNPs are retained but preferentially retaining the SNPs most associated with the phenotype under study.
2) From the result, that containing the index SNPs after clumping, we can extract the index SNP ID.
3) Create a file containing SNP IDs and their corresponding P-values and a file containing the different P-value thresholds for inclusion of SNPs in the PRS.
4) We assign to each filter: --q-score-range, --score, --extract their respective files.
5) **clumping.fit.R :** This script conducts a polygenic risk score (PRS) analysis using PLINK output files and phenotype data. It first reads in the phenotype file, principal components (PCs), and covariates. Then, it calculates the null model R-squared and iterates through different p-value thresholds to calculate PRS R-squared, coefficients, standard errors, and p-values. Finally, it identifies the best PRS result and saves the results to output files.
6) **plink.barplot.R :** This script reads in the results of a PRS analysis and generates a bar plot to visualize the relationship between different p-value thresholds and the corresponding PRS R-squared values. The plot is customized using ggplot2 and saved as "clumping.bar.png".
7) **plink.plot.R :** This script reads in PRS data and phenotype information, renames the columns, merges the files, and generates a density plot of polygenic scores stratified by phenotype. The resulting plot is saved as "Plink_0.5.png".