Polygenic Risk Scores (PRSs) based on summary statistics: lassosum2 tutorial

Evangelina López de Maturana & Oscar González-Recio

Genome-wide prediction

Outline of the practical session

- We will use simulated test data to examine the PRS's performance
- The genotypes of 811 SNPs of n = 1,150 individuals were simulated
- Outcome: breast cancer risk
- We will follow the next steps:
  1. Install bigsnpr
  2. Download the example data
  3. Create the files for input
  4. SNP selection
  5. PRS construction
  6. Examine results

Day4.lassosum2.R

Lassosum, Lassosum2

- Lassosum is a method for constructing PRSs using summary statistics, often obtained from GWAS meta-analysis, while accounting for Linkage Disequilibrium (LD) (Mak et al, 2017)
- Lassosum2 (Privé et al, 2022), provides a new method for updating the coefficients
- It employs penalized regression (elastic-net) and has two hyperparameters:
    - Lambda: regularization parameter controlling the inclusion of SNPs, where higher values select fewer SNPs
    - Delta: adjusts the estimated LD matrix to align the problem with elastic-net regression
- This method is implemented in the R package bigsnpr (https://privefl.github.io/bigsnpr/reference/snp_lassosum2.html)

File formats for input data

- Three files are needed for the analysis:
  - A file containing the summary statistics
  - An LD matrix computed from individuals from the same genetic ancestry:
    - LD panels from the reference panels from Hapmap3+
  - Individual-level data for tuning hyper-parameters and testing the models

1. Summary statistics file format

- It contain the result of the GWAS analyses (or GWAS metaanalysis)
- It has the following columns:
  - rsid: rs id or SNP identifier.
  - chr: Chromosome number.
  - pos: SNP position.
  - a0: Reference allele.
  - a1: Alternative allele.
  - beta: Regression coefficient.
  - beta se: Regression coefficient standard deviation.
  - p: Regression p-value.

1. Summary statistics file format

- We will use the summary statistics of SNPs from 21 chromosome (http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST004001-GCST005000/GCST004988/)
- Outcome is breast cancer risk, ncases = 137045 and ncontrols = 119078

- It is necessary to compute the effective size:

    - Case-control studies: $\mathrm{neff} = \frac{4}{\frac{1}{ncases} + \frac{1}{ncontrols}}$

    - Continuous traits: $\mathrm{neff} = \mathrm{n}$

2. Reference panel LD matrix

- The authors recommend using a set of HapMap3+ variants when using imputed genotypes
- If not, they suggest manually computing the LD matrix
- Hapmap3+: LD matrix can be downloaded from
  - https://figshare.com/articles/dataset/LD reference for HapMap3 /21305061

3. Test data (PLINK format)

- We will use simulated test data to examine PRS performance, genotypes of 811 SNPs of Ncases=448 and ncontrols=702
    - The format is a fileset of three different files that must accompany each other and have the same file prefix:
        - .bed file: binary file that contains genotype information
        - .bim: it contains variant information, Chromosome code (either an integer, or 'X'/'Y'/'XY'/'MT'; '0' indicates unknown) or name, Variant identifier, Position in morgans or centimorgans, Base-pair coordinate (1-based), Allele 1 (usually minor), Allele 2 (usually major)
        - .fam: (FID, IID, ID_father, ID_mother, Sex (1' = male, '2' = female, '0' = unknown), Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control))

Steps

- Intersect the SNPs with the test set
- Restrict the PRS construction to the SNPs present in the test data
- Preparation of LD matrix based on the SNPs available in the test file
- Join the betas with SNPs in the test set
- Select LD matrix for those SNPs
- Run the lassosum2 function to compute beta coefficients for polygenic risk scores (`snp_lassosum2(corr, df_betas)`) using different values for lambda and delta
- Extract grid parameters from the lassosum2 result
- Compute the polygenic risk score (PRS) for each individual based on SNPs
- Compute the AUC (Area Under the Curve) for each PRS
- Select the best grid of parameters based on AUC
- Get the SNPs with non-zero coefficients for PRS calculation
- Get the effect sizes (coefficients) of the selected SNPs