Genome Wide Association Studies Applications Polygenic Risk Score (PRS)

21/09/2023

Alper Bülbül

Outline

- Data Collection and Genotyping
- Quality Control
- Imputation
- Association
- Fine Mapping
- Meta Analysis
- Variant annotation / Enrichment or gene-set analysis
- Causality
- PRS analysis

Data Collection and Genotyping

Data Collection

- Study cohorts with genotype and phenotype.
  - UK Biobank
  - TOPMed
  - BioBank Japan
  - Million Veteran Programme
- Avoid collider bias.

Genotyping

- Microarrays to capture **common variants**
- Whole-genome sequencing (WGS)
- Whole-exome sequencing (WES)

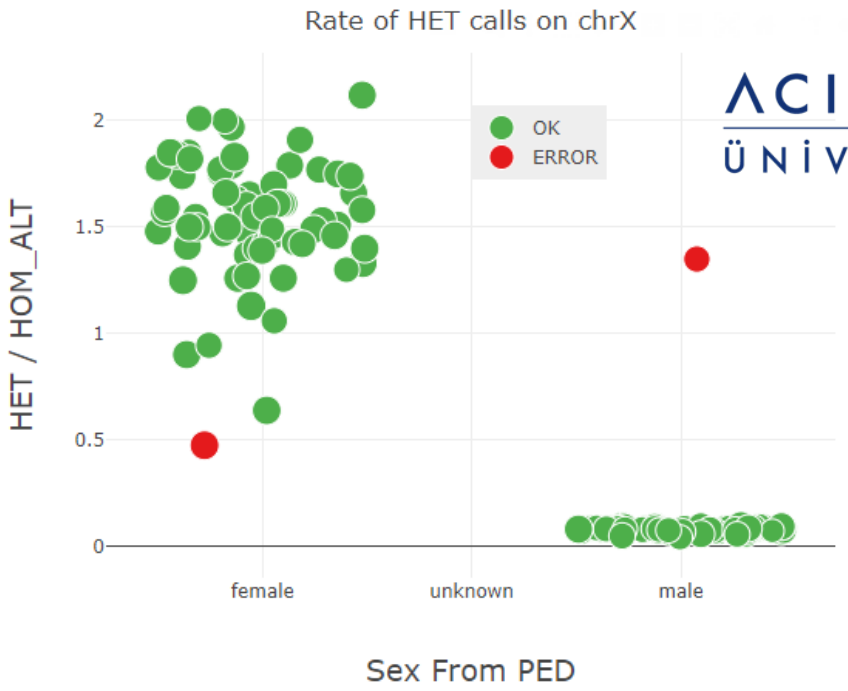
Quality Control

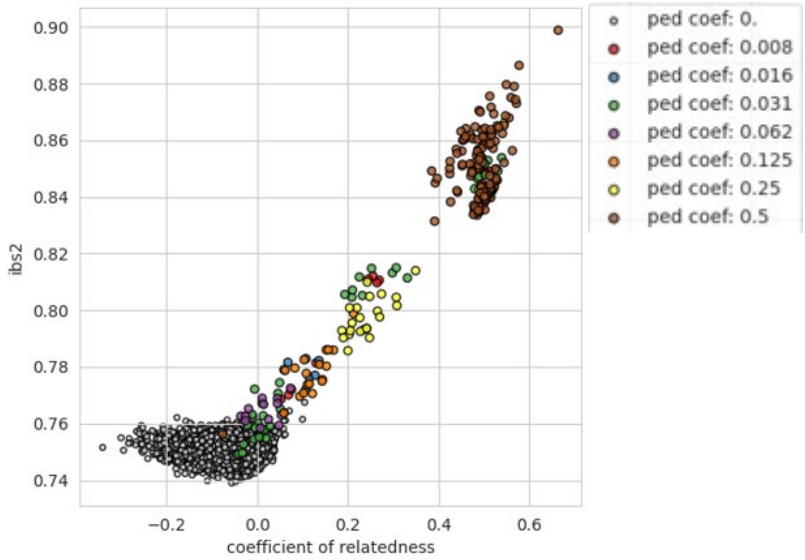
SNP QC

- For Genotype Call Rate Quality Check (Hardy–Weinberg)
- Minor Allele Frequency (More Than 0.1) (**PLINK**)

Sample QC

- Sex Check (X chromosome HET/HOM_ALT) (Female: 0.5>, Male 0.5<)
- Relatedness Check (**Peddy**)

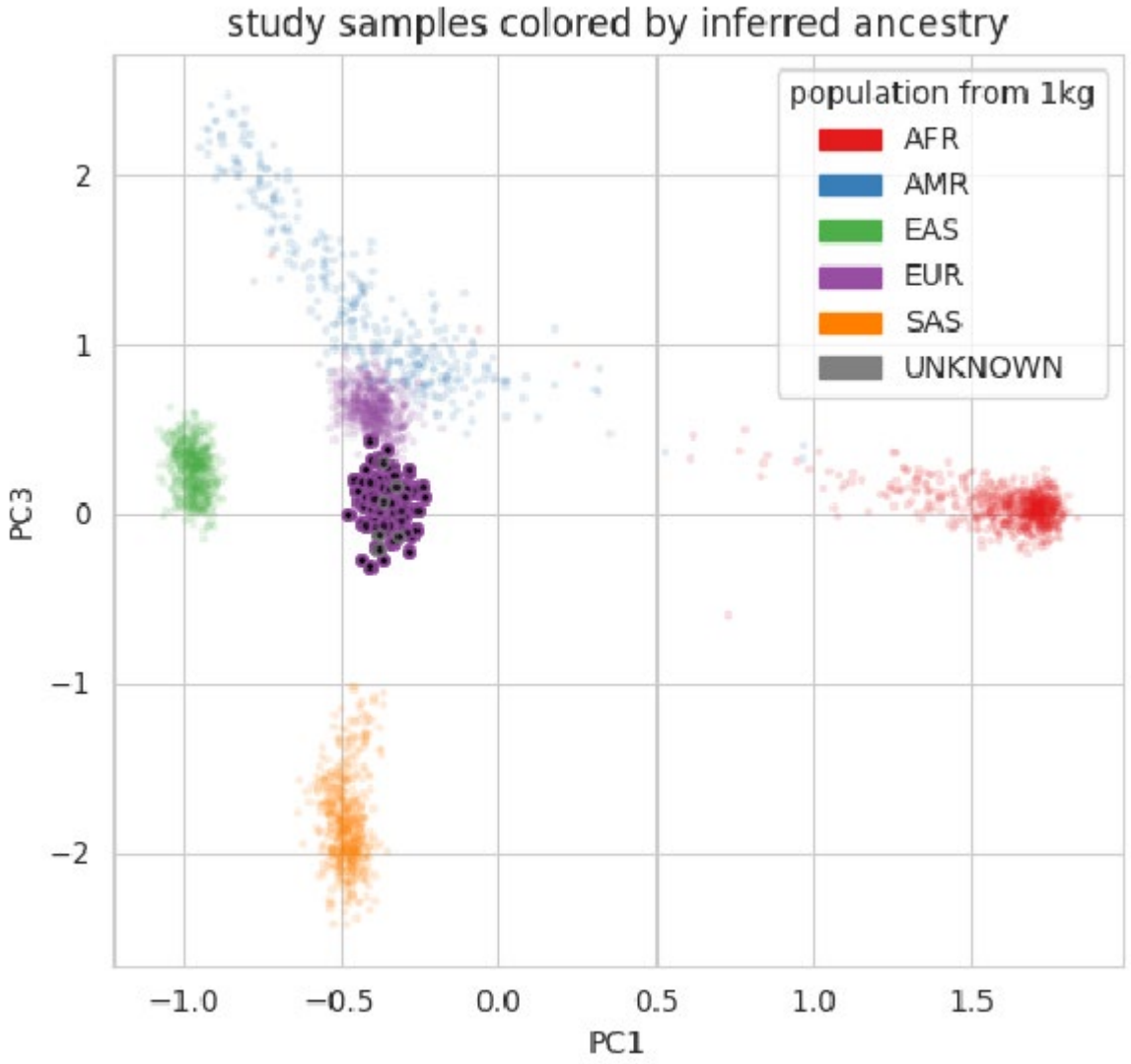




Quality Control

Sample QC

- Population Stratification
- Ancestry check for samples



Imputation

- Missing Genotype imputed according to selected ancestry
- **Minimac 4** (pre-phasing for faster calculation)
- **Eagle** Haplotype Phasing
- R^2 < 0.3

NIH National Heart, Lung, and Blood Institute | BioData CATALYST
TOPMed Imputation Server

- 97,256 reference samples
- 308,107,085 genetic variants

Reference panel sample size	Imputation accuracy (mean r^2)		
	MAF <0.1%	MAF 0.1–1%	MAF >1%
1000	0.41	0.64	0.96
10000	0.69	0.84	0.98
20000	0.79	0.89	0.99

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2014). minimac2: faster genotype imputation. *Bioinformatics*, *31*(5), 782-784.

Association

BOLT-LMM

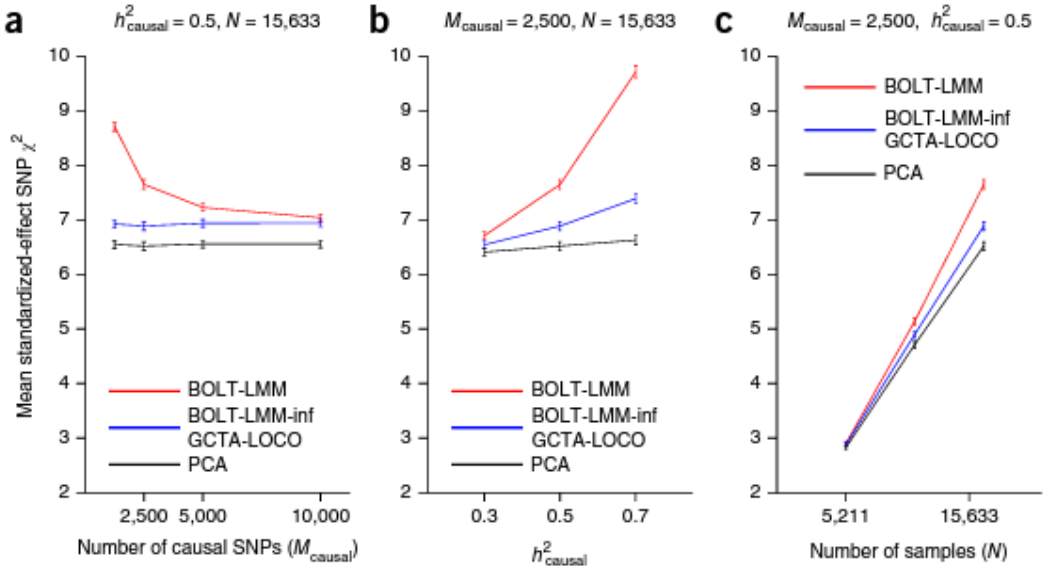
- Bayesian mixed-model association testing
- **More than 5,000 samples**

- log OR = β / (μ * (1 - μ)), where μ = case fraction.

- BETA(β)*:* effect size from BOLT-LMM approximation to the infinitesimal mixed model

- https://storage.googleapis.com/broad-alkesgroup-public/BOLT-LMM/BOLT-LMM_manual.html

1. Determine a high-confidence set of SNPs (e.g., based on R^2 or INFO score)
2. SNPs in PLINK format.
3. Use PLINK to LD prune to ~500K SNPs (--indep-pairwise 50 5 *r2thresh*)
4. Run BOLT-LMM using the final hard-called SNPs



Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., ... & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, *47*(3), 284-290.

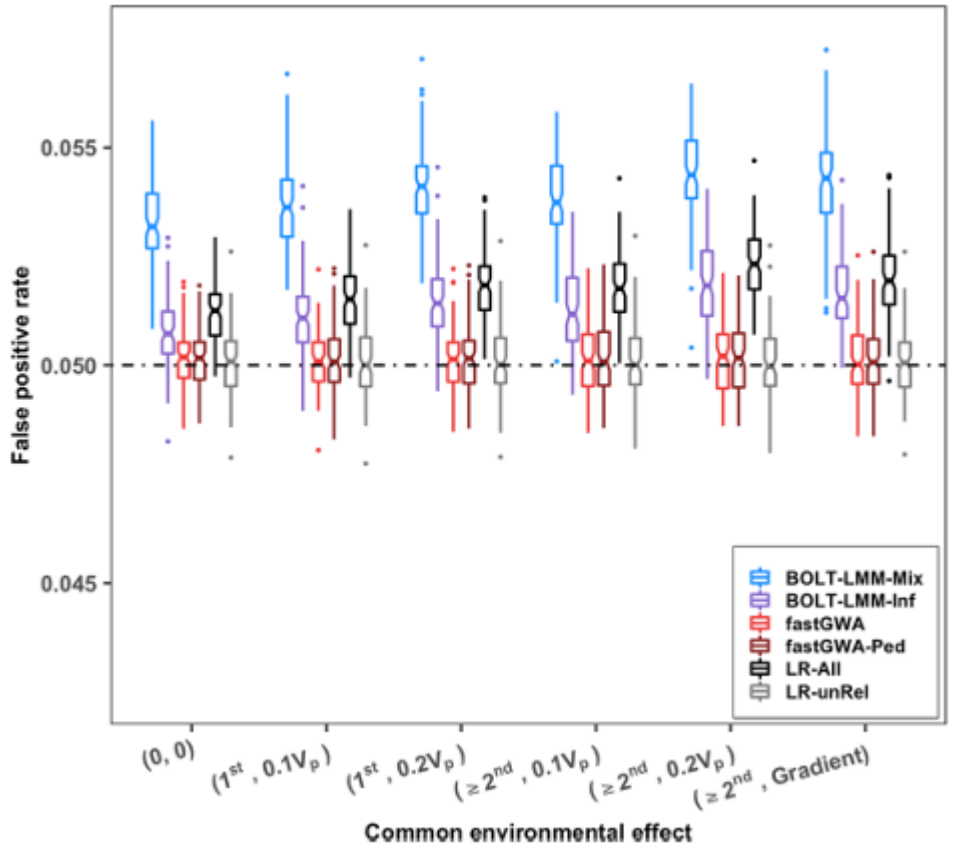
Association

fastGWA: A fast MLM-based Genome-Wide Association tool

- Mixed linear model (MLM)
- **Related Samples**

- y = Xsnp*βsnp + Xc*βc + g + e

- **y**: phenotypes,
- **Xsnp**: genotype variables of a variant
- **βsnp**: Effect size
- **Xc**: fixed covariates (for example, sex, age, and the first few PCs)
- **βc**: corresponding coefficients
- **g**: total genetic effects captured by pedigree relatedness
- **e**: vector of residuals

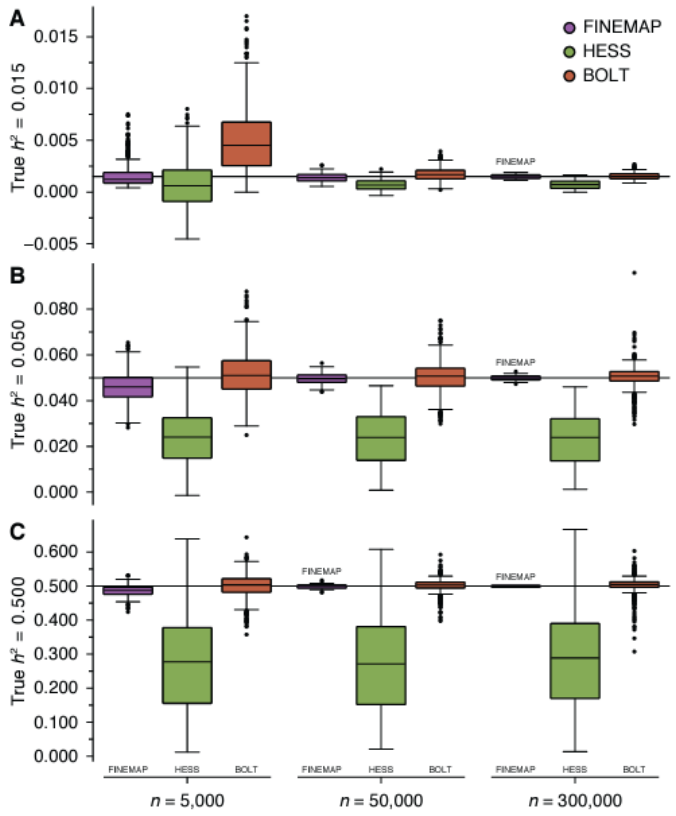


Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics*, *51*(12), 1749-1755.

Fine Mapping

FINEMAP

- estimate the **effect sizes** and **regional heritability**
- **Shotgun Stochastic Search**



Benner, C., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2018). Refining fine-mapping: effect sizes and regional heritability. *BioRxiv*, 318618.

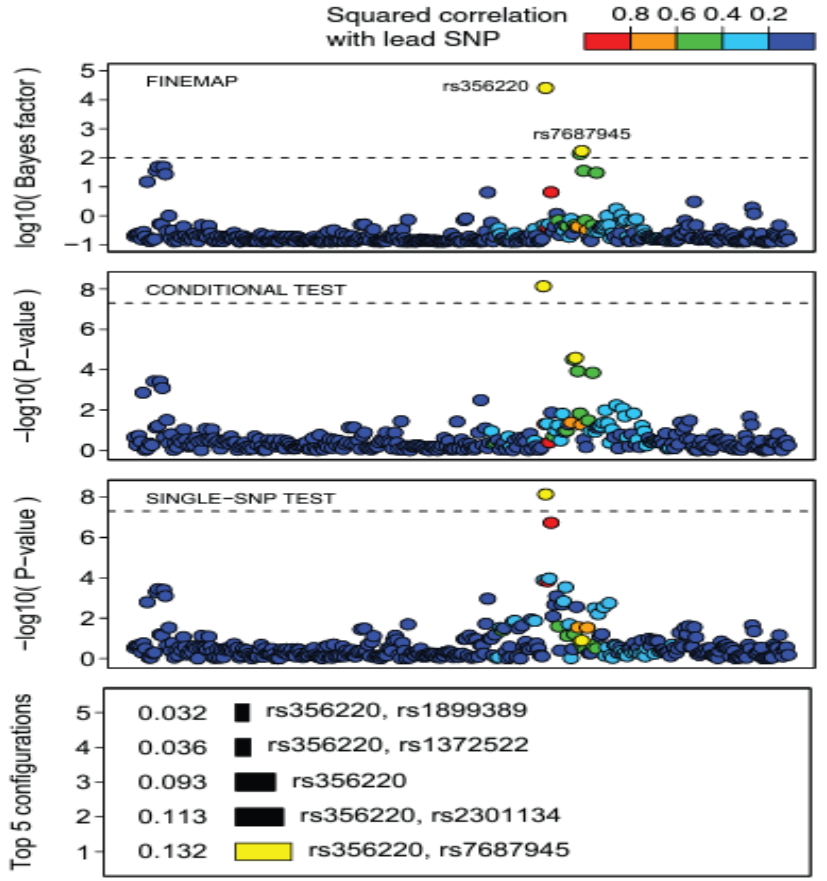


Fig. 6. Fine-mapping of 4q22/*SNCA* region associated with Parkinson's disease. Associated SNPs rs356220 and rs7687945 are highlighted by ○ and their configuration by ▭. Dashed lines correspond respectively to a single-SNP Bayes factor of 100 and *P*-value of 5×10^{-8}. Squared correlations are shown with respect to rs356220

Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., & Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, *32*(10), 1493-1501.

Meta Analysis

GWAMA

- Much of the genetic variation remains unexplained.
- Detect further novel loci is through meta analysis
  - same population
  - increasing the sample size over any individual study

Table 2: Comparison of software packages for genome-wide meta-analysis of association summary statistics.

Software package	METAL	MetABEL	META	GWAMA
Pre-processing of GWA analysis files	No	*ABEL	SNPTEST	SNPTEST, PLINK
Strand flipping for aligning effect directions	Yes	Yes	Yes	Yes
Fixed effect analysis	Yes	Yes	Yes	Yes
Random effect analysis	No	No	Yes	Yes
Heterogeneity statistics (Cochran's *Q* statistic, I^2)	*Q*	No	*Q*, I^2	*Q*, I^2
Automated genomic control for population structure	Yes	Yes	Yes	Yes
Graphical visualisation of meta-analysis results	No	Forest plot	No	Separate scripts for Manhattan and QQ plots

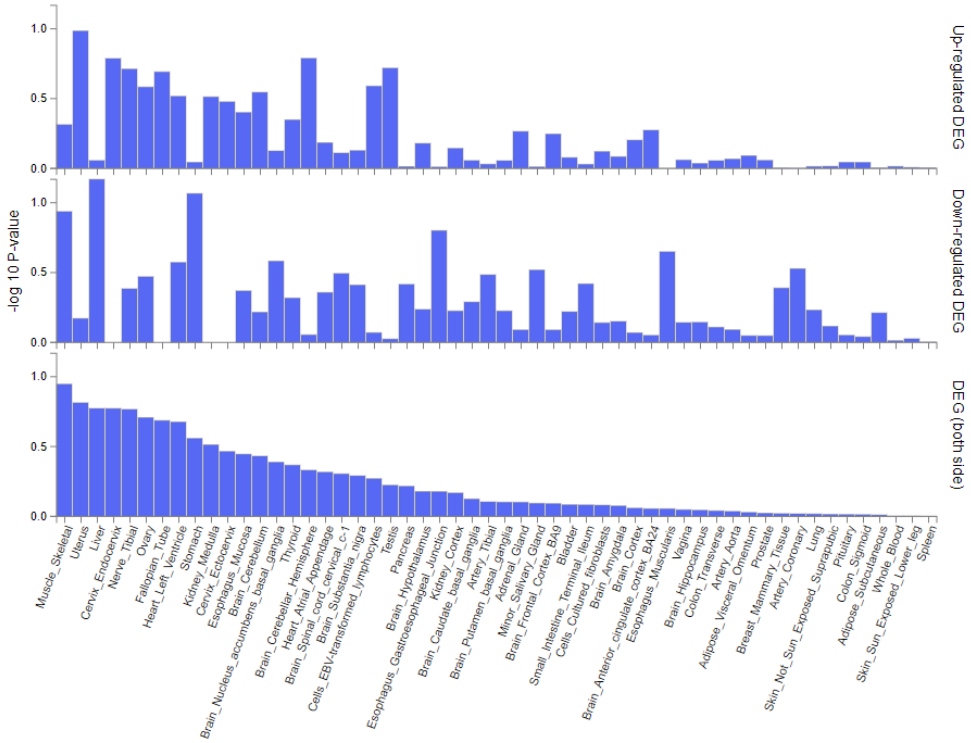
Mägi, R., & Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC bioinformatics*, *11*, 1-6.

Variant annotation

FUMA GWAS (Functional Mapping and Annotation of Genome-Wide Association Studies) and **MAGMA**

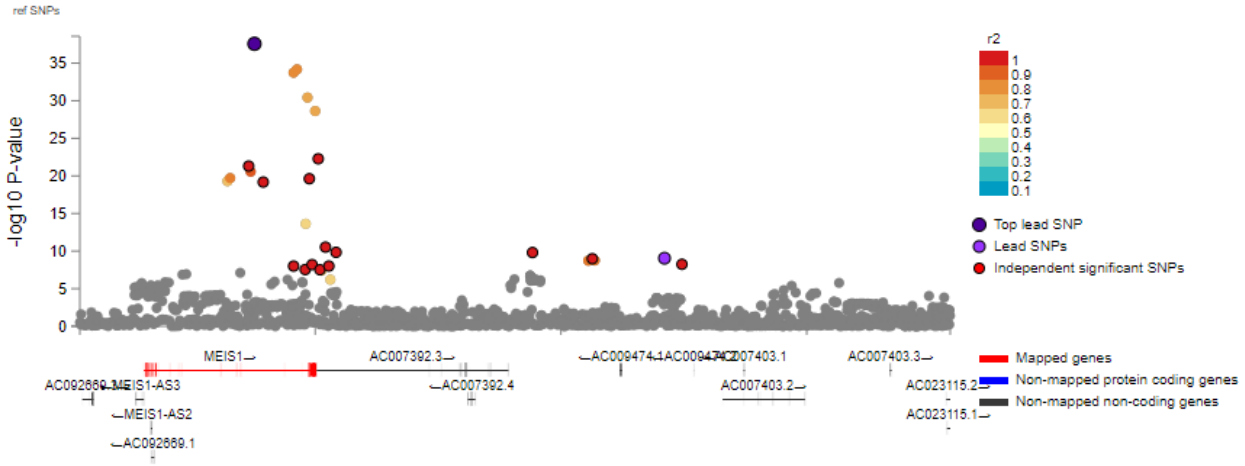
https://fuma.ctglab.nl/

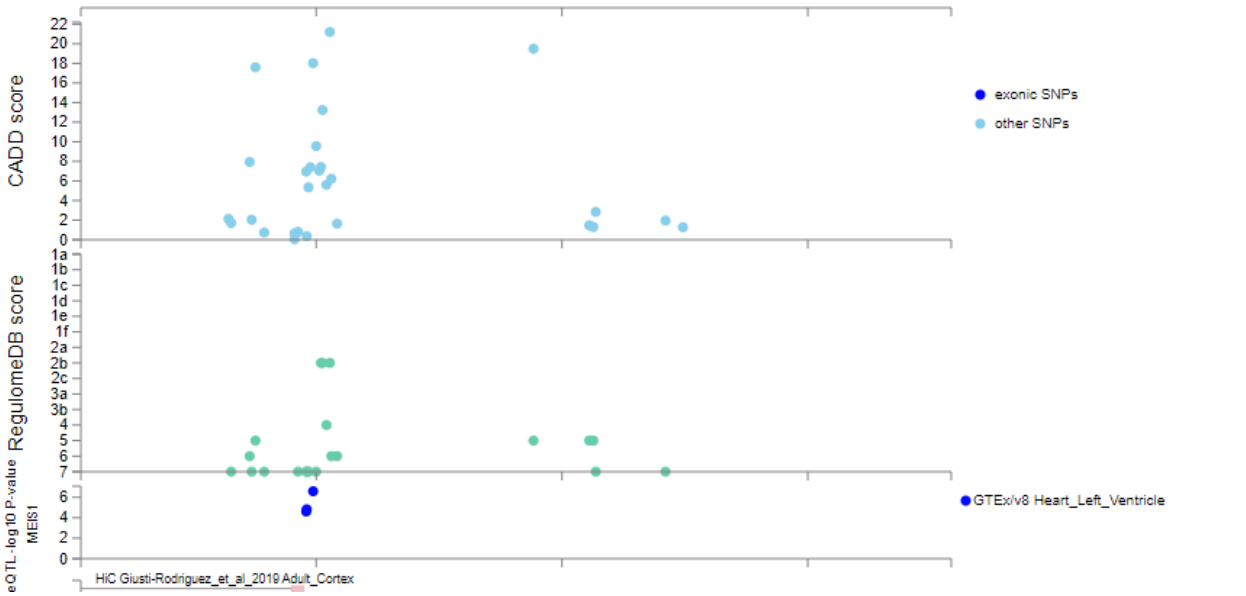
GTEx v8 54 tissue types



Poor Sleep Quality

Regional plot

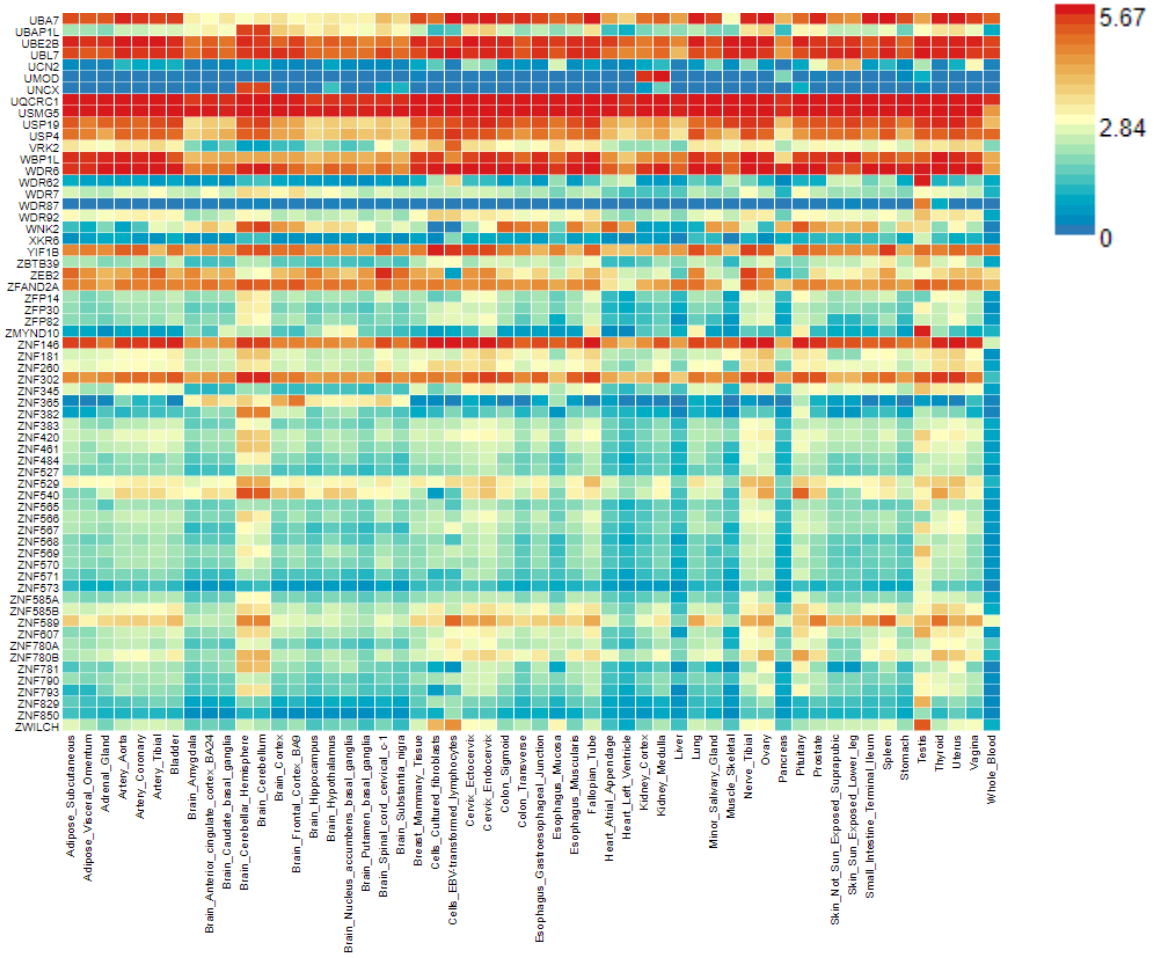




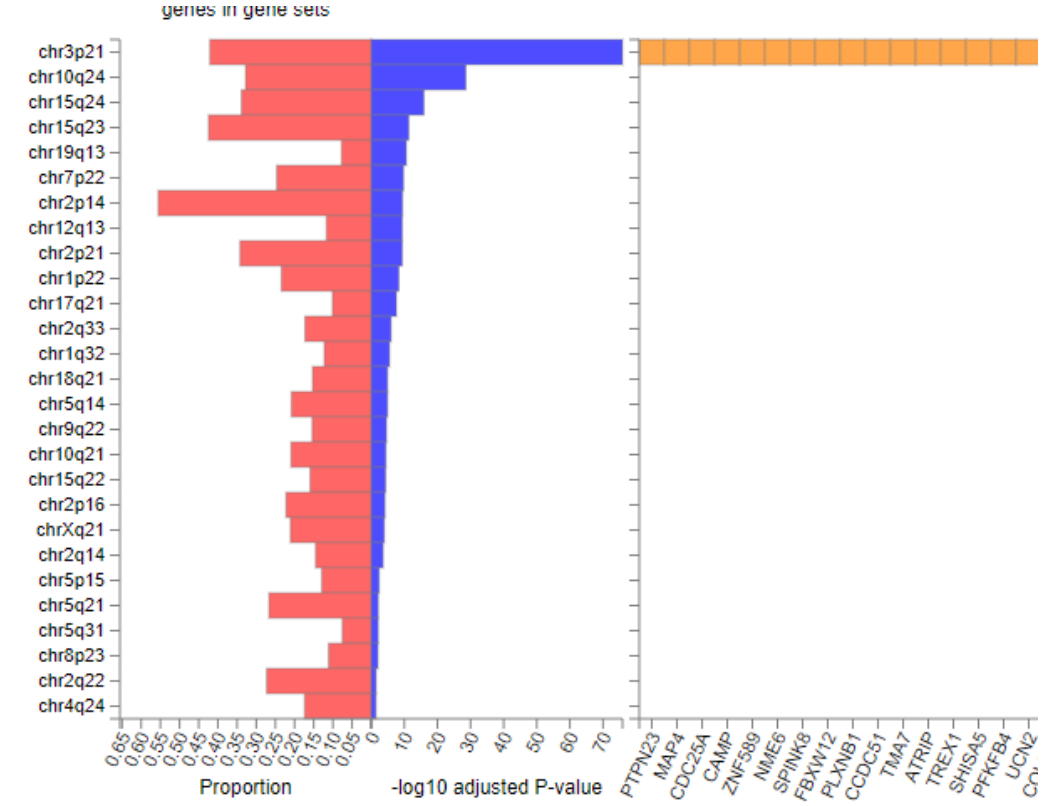
Variant annotation

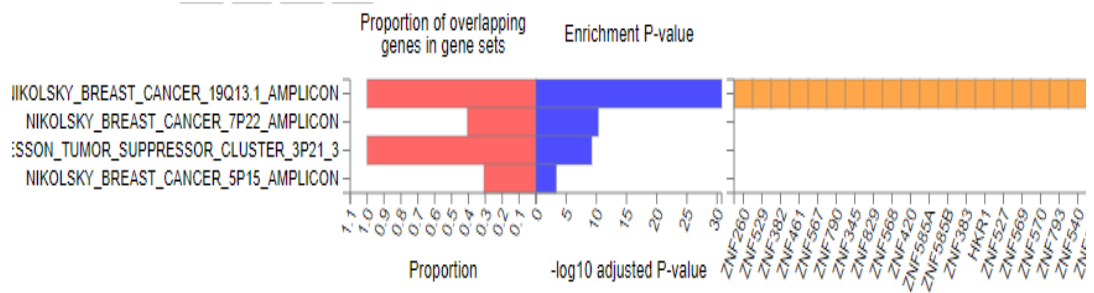
FUMA GWAS

Poor Sleep Quality



Gene expression heatmap

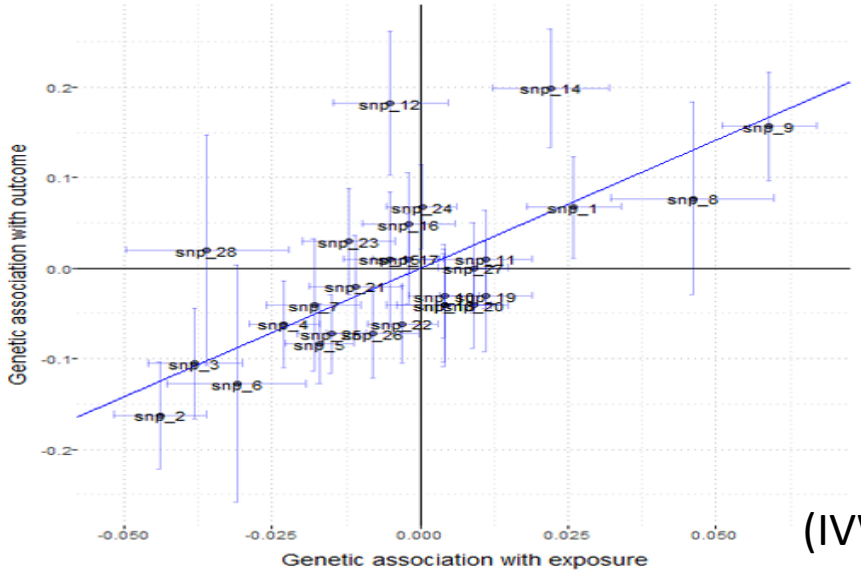




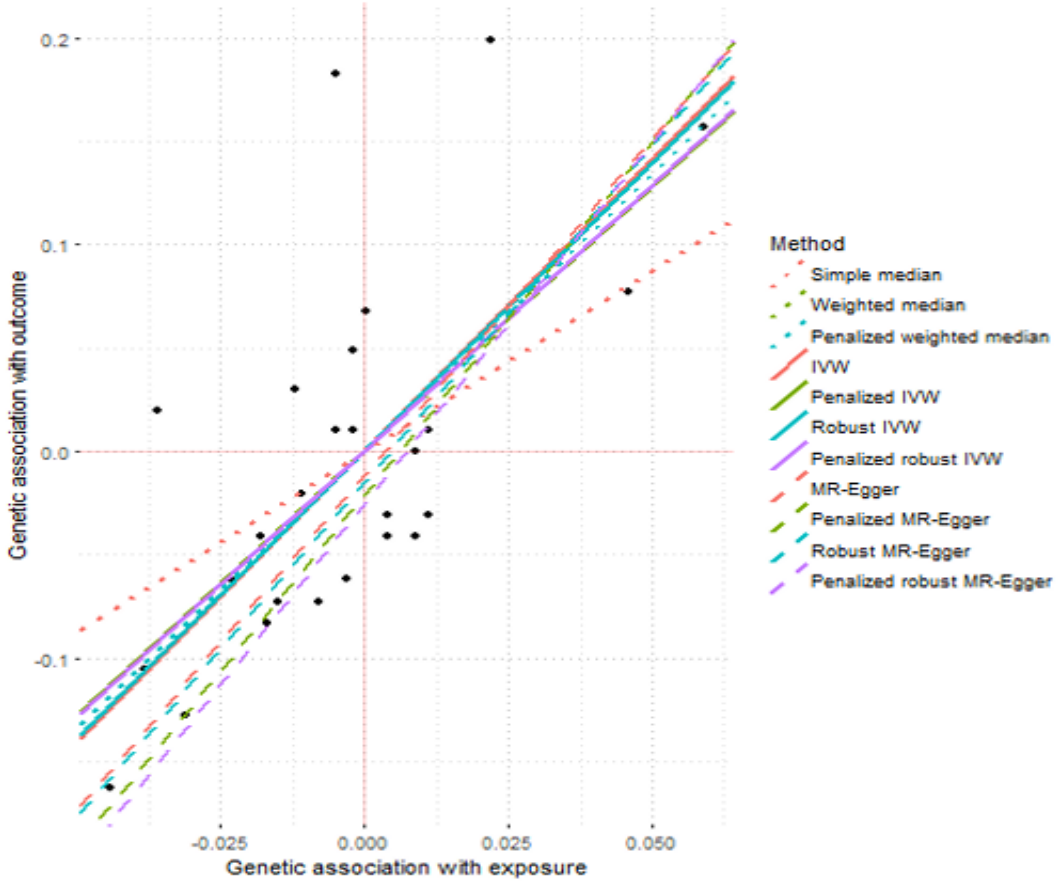
Causality

Mendelian Randomization v0.3.0 R package

- Inverse-variance weighted (IVW) (fixed-effect)
- Median-based
- MR-Egger (random-effects model)
- **Exposure** is a characteristic giving the name of the risk factor, e.g. LDL-cholesterol.
- **Outcome** is a characteristic giving the name of the outcome, e.g. coronary heart disease.
- MR-PRESSO for removing horizontal pleiotropic SNPs



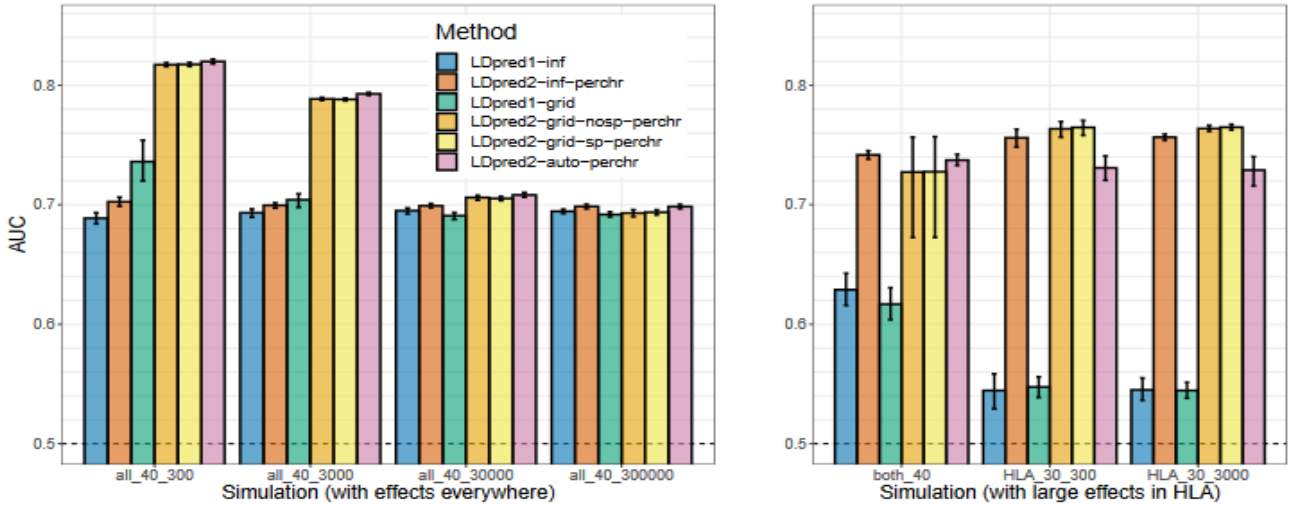
(IVW)



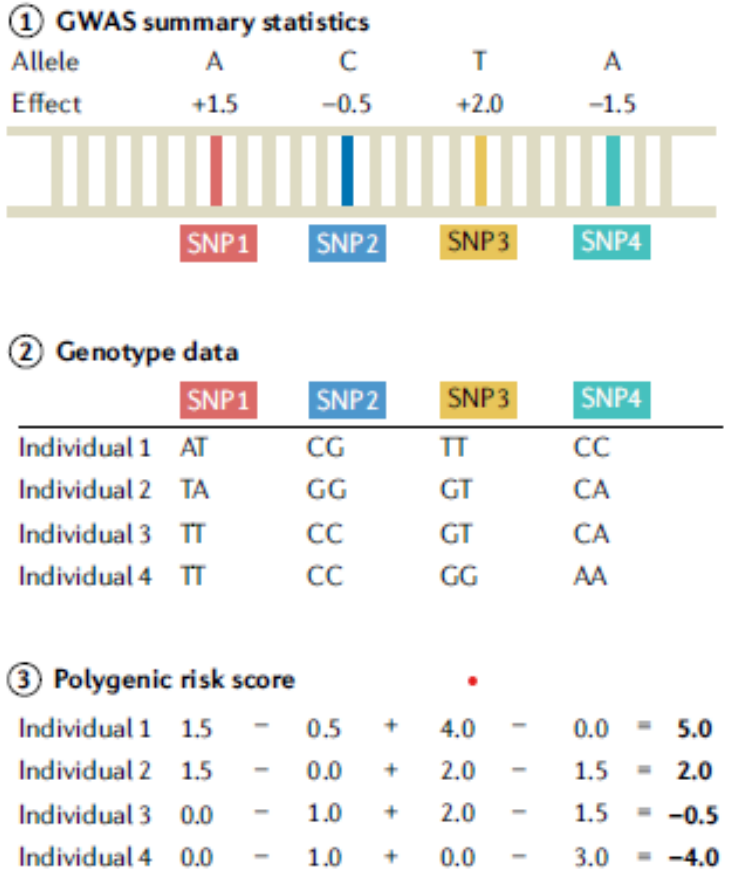
PRS analysis

LDpred2

- Estimation of posterior effect sizes of SNPs using a Bayesian shrinkage approach in **LD blocks**
- More accurate in autoinflammatory disease (RA)
- unstable HLA region because of long-range LD block. LDpred2 more stable
- Hyper parameter tuning for lambda and alpha with grid search



Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2020). LDpred2: better, faster, stronger. *Bioinformatics*, *36*(22-23), 5424-5431.



Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., ... & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59.