MetaCCA Implementation

Gurdeep Singh Bhambra

August 2022

About

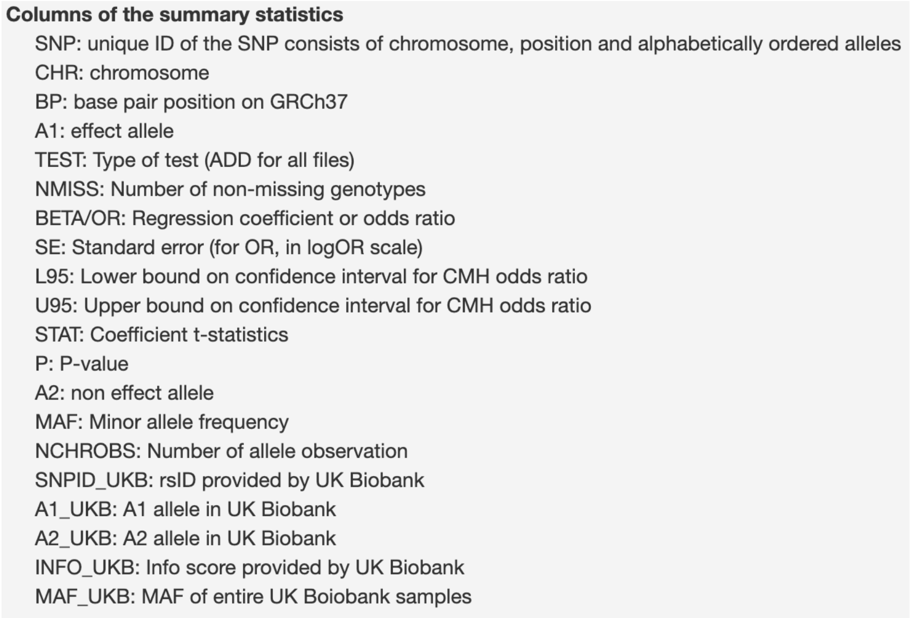
This document guides through the meta-analysis of systolic, diastolic and pulse rate phenotypes using metaCCA. The code files and results along with this guide provide all the information required to check and reproduce the results.

1 Data Preparation

1.1 GWAS Summary Data

The following GWAS summary datasets were downloaded from the atlas website (`https://atlas.ctglab.nl`):

1. – Pulse Rate (Atlas ID: 3188, UKB Phenotype ID: 102): `https://atlas.ctglab.nl/ukb2_sumstats/f.102.0.0_res.EUR.sumstats.MACfilt.txt.gz`
2. – Diastolic Blood Pressure (Atlas ID: 3379, UKB Phenotype ID: 4079): `https://atlas.ctglab.nl/ukb2_sumstats/f.4079.0.0_res.EUR.sumstats.MACfilt.txt.gz`
3. – Systolic Blood Pressure (Atlas ID: 3380, UKB Phenotype ID: 4080): `https://atlas.ctglab.nl/ukb2_sumstats/f.4080.0.0_res.EUR.sumstats.MACfilt.txt.gz`



Columns of the summary statistics

SNP: unique ID of the SNP consists of chromosome, position and alphabetically ordered alleles
CHR: chromosome
BP: base pair position on GRCh37
A1: effect allele
TEST: Type of test (ADD for all files)
NMISS: Number of non-missing genotypes
BETA/OR: Regression coefficient or odds ratio
SE: Standard error (for OR, in logOR scale)
L95: Lower bound on confidence interval for CMH odds ratio
U95: Upper bound on confidence interval for CMH odds ratio
STAT: Coefficient t-statistics
P: P-value
A2: non effect allele
MAF: Minor allele frequency
NCHROBS: Number of allele observation
SNPID_UKB: rsID provided by UK Biobank
A1_UKB: A1 allele in UK Biobank
A2_UKB: A2 allele in UK Biobank
INFO_UKB: Info score provided by UK Biobank
MAF_UKB: MAF of entire UK Boiobank samples

Figure 1: Columns of the summary statistics files

These datasets were part of UK biobank data accessed in 2019 which were analysed and their results were shared on the atlas website along with the GWAS summary statistics [1]. The original UK biobank study can be traced back using the UKB phenotype id for each phenotype. The GWAS summary statistics were calculated based on the first visit and first run of the study. There were 361,411 participants where systolic, diastolic and pulse rates were recorded. The columns and their definition are as per the atlas documentation in figure-1.

Each of the phenotypes had 10,534,620 SNPs. Common SNPs among all three phenotypes were filtered which resulted in a total of 10,534,620 SNPs.

2 Implementation

This section documents what different code files do and the order in which they were executed. All the code files have comments in them referring to what it does. Python, R and bash programming languages were used and are switched from one another based on ease of use or computational advantage. Each of the sub-sections is in the order in which the code was executed. Most of the code was according to Osborne's scripts[2] and metacca scripts[3].

2.1 Pre-Processing GWAS Datasets

The "Preprocessing" file contains the code for pre-processing these datasets, merging the datasets and creating a SXY file with all the SNPs, a SXY file with filtered SNPs based on minor allele frequency (MAF) greater than 1% and a vcf file with the MAF filtered SNPs.

All the data was clean and there were no duplicates. The SNP ids across the datasets completely overlapped.

2.2 Preparing SYY Data

After creating SXY data containing all the common SNPs, the SYY phenotype correlation matrix was computed according to the code in the "SYY" file.

Systolic and diastolic were relatively in strong correlation compared to pulse rate.

2.3 Preparaing SXX Data

To calculate the genotype-genotype correlation matrix, the SXY and 1000 genome data were used. The 1000genome data for the European population and human genome build GRCh37/hg-19 data was downloaded from the VEGAS2 website (`https://vegas2.qimrberghofer.edu.au/`). "xx_mat" file documents the step-by-step approach in calculating the final SXY data annotated according to the reference data. The bash script uses two more python scripts which were used for filtering the SNPs across SXY, SXX and VCF files.

2.4 Gene Annotation

The last section of the "xx_mat" file contains the code used for gene annotation. The gene annotation was according to human genome build, GRCh37/hg19. Only the SNPs with a single gene were used for creating the SXY, and SXX data.

2.5 Applying MetaCCA

The final ”metaCCA” file has the code for applying the metaCCA model along with Bonferroni correction. The significant genes were filtered with a 0.05 threshold and saved in the ”metcca_results_pBonf_sig” file. There were about 860 genes that passed through the p-value threshold.

3 References

[1] Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, van der Sluis S, Andreassen OA, Neale BM, Posthuma D. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019 Sep;51(9):1339-1348. doi: 10.1038/s41588-019-0481-0. Epub 2019 Aug 19. Erratum in: Nat Genet. 2020 Mar;52(3):353. PMID: 31427789.

[2] Osborne A. CCA_scripts. `https://github.com/AmyJaneOsborne/CCA_scripts`

[3] Anna Cichonska, Juho Rousu, Pekka Marttinen, Antti J. Kangas, Pasi Soininen, Terho Lehtimäki, Olli T. Raitakari, Marjo-Riitta Järvelin, Veikko Salomaa, Mika Ala-Korpela, Samuli Ripatti, Matti Pirinen, metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis, Bioinformatics, Volume 32, Issue 13, 1 July 2016, Pages 1981–1989, `https://doi.org/10.1093/bioinformatics/btw052`