Multi-trait GWAS Simulator

User Manual

Heather F. Porter & Paul F. O'Reilly

multitraitgwas@gmail.com

MRC Social, Genetic and Developmental Psychiatry Centre,
Institute of Psychiatry, Psychology & Neuroscience,
King's College London, UK

Contents

1 Background **3**

2 Software program **4**

3 R packages **4**

4 Quick start **5**

5 Simulating multivariate data **7**

5.1 Phenotypes with direct SNP effects 7

5.2 Phenotypes with indirect SNP effects 8

5.3 Case-control phenotypes 8

5.4 Output 8

6 Multivariate method comparison **9**

6.1 Methods 9

6.2 **S1:** fixed genetic effects and phenotypic correlations 9

6.3 **S2:** uniform genetic effects and phenotypic correlations 9

6.4 **S3:** phenotypic correlations reflective of genetic effects 10

6.5 **S4a:** real data informed phenotypic correlations 10

6.6 **S4b:** real data informed genetic effects and phenotypic correlations 11

6.7 Additional Options 11

6.8 Output 12

1 Background

Many genetic analyses are now performed on multiple phenotypes jointly, e.g. to increase discovery power in genome-wide association studies (GWAS), for sub-phenotype and cross-phenotype prediction using polygenic risk scores or to estimate genetic correlations among traits. However, estimating the power of the related methods and comparing their performance is challenging due to the complexity of the biological network introduced when considering multiple phenotypes.

We have produced a simulation framework in order to address this - the framework, summarised below, models this network explicitly, enabling the vast model space of different genetic effects on a set of multiple phenotypes of varying degrees of correlation to be explored systematically.

We present our framework as an R executable program for generating multi-trait GWAS data corresponding to the simulation framework, and comparing new and existing multi-trait GWAS methods across a wide range of scenarios.

The simulation framework incorporates five different scenarios that explore different combinations of genetic effects and phenotypic correlations:

S1: genetic effects and phenotypic correlations varied systematically

S2: genetic effects and phenotypic correlations sampled from uniform distribution

S3: genetic effects reflect phenotypic correlations

S4a: fixed genetic effects, phenotypic correlations sampled from real data

S4b: genetic effects and phenotypic correlations sampled from real data

This user manual describes how to run the software program that implements this simulation framework; examples are provided for each function.

If you use this software in any published work, please cite our comparison study paper. For questions regarding this software or the comparison study, please contact us:
multitraitgwas@gmail.com

2 Software program

The software program folder **Multi_Trait_GWAS_Software** contains four R scripts and 14 data files. The main R script used for the comparisons is **MultiTraitGWAS.R**; the other R scripts are dependencies of this script, and should be stored in the same location in order to run the software. The data files are only required for the final simulation scenario (S4b). They contain effect size and phenotypic correlation data that is used to inform the simulations. These files should be stored in the same location as the R scripts.

You will also need to download the software corresponding to four of the methods if you wish to perform comparisons with them. For S_{Het} and S_{Hom} (download CPASSOC here) make sure that the FunctionSet.R script is in the same directory as MultiTraitGWAS.R. The executable files for BIMBAM (download here) and SNPTEST (download here) should be downloaded, renamed **bimbam** and **snptest** respectively, and stored in the same directory as MultiTraitGWAS.R. This software has been tested for compatibility with SNPTEST v2.5.1.

This software is written in R, but should be executed from the terminal window. The examples in this user manual give the commands to execute from the terminal.

N.B. This software makes use of BASH scripting, and can currently only be run on UNIX/Linux machines.

3 R packages

This software requires the following R packages and their dependencies:

- abind
- aod
- batch
- combinat
- epitools
- lmtest
- meta
- MultiPhen
- zoo

Make sure that the packages are installed in the directory where the software package files are stored.

4 Quick start

To get a feel for the software program, first try simulating some multivariate data using the examples below. In the simplest case, the parameters you will need to specify are:

- sample size (**n**)
- number of SNP replicates (**m**)
- minor allele frequency of the SNP (**MAF**)
- number of phenotypes (**k**)
- genetic effects as a percentage of phenotypic variance explained (**v**)
- phenotype correlation matrix (**cor.mat**)

The simulated SNPs are independent; the **m** parameter specifies the number of replicates of a SNP to simulate. The more replicates, the less variance in the power estimates. The phenotype data are simulated multivariate normal variables, with multivariate normal error term whose variance is the phenotype correlation matrix. This means that the simulated phenotype data correlation is not exactly as specified, but the same to within 2.d.p.

To simulate data on 2 phenotypes with correlation 0.45 and phenotypic variance explained of 0.2% on phenotype 1 and 0.3% of phenotype 2 for 1000 individuals and 100 replicates of a SNP with minor allele frequency 0.1, use the following command:

```
R --file=./MultiTraitGWAS.R -q --args n 1000 m 100 k 2 MAF 0.1 v 0.002,0.003
cor.mat 0.45 strand data
```

A folder (**data_output_date_time**) is created, containing 100 SNP data files (**snp_data_i**) and 100 phenotype data files (**pheno_data_i**). There is also the **data_output_log** file that details the parameters and the run time for the simulation.

Example first 10 lines of a SNP data file:

```
rs1
2
1
0
1
0
0
0
1
0
0
```

Example first 10 lines of a phenotype data file:

```
Pheno1              Pheno2
0.826669320195689   0.42083503810129
0.195993062417846   -0.328777224652241
0.13776496562821    0.0259098766867293
0.456580144739775   -1.2848771006169
```

```
-1.19281347343584      0.572394773681447
0.173851560588671      -0.245754462108122
-0.224286770332517     -1.47187854347651
1.79286516350429       1.80735748486297
-1.1057183510766       -0.183142822421996
0.37684498650171       0.229370868124437
```

Example log file:

```
######################################################################


Multi-trait GWAS Software

Heather F. Porter, Paul F. O'Reilly
If you use this software in published work, please cite:


######################################################################

Date: 27-08-2015
Time: 11.58.09 BST

######################################################################

Sample size: 1000
SNPs: 100
MAF: 0.1
Phenotypes: 2
Effect sizes: 0.002,0.003
Correlation(s): 0.45

######################################################################

Run time: 1.18 seconds

######################################################################
```

There are more available options for simulating multivariate data (see the **Simulating multivariate data** section), but in all cases the output structure remains the same. The input parameters are always specified in the same way provided you know the name of the parameter and its input form, which is detailed in the **Simulating multivariate data** section.

5 Simulating multivariate data

This software program can generate simulated SNP and multivariate phenotype data. SNP and phenotype data can be generated according to the following parameters:

- sample size (**n**)
- number of SNP replicates (**m**)
- minor allele frequency of the SNP (**MAF**)
- number of phenotypes (**k**)
- genetic effect vector of phenotypic variance explained (**v**)
- phenotype correlation matrix (**cor.mat**)
- direct effects (default: **direct** = **TRUE**)
- number of indirect effects (**n.indirect**)
- case-control (default: **cc** = **FALSE**)
- case-control phenotypes (**which.cc**)
- case prevalence (**prev**)

5.1 Phenotypes with direct SNP effects

To generate phenotypes with direct SNP effects you need to specify the number of individuals (**n**), number of SNPs (**m**), and number of phenotypes (**k**) for which to generate the data, as well as the minor allele frequency of the SNPs (**MAF**), the genetic effect vector (**v**) of phenotypic variance explained, and the phenotypic correlation matrix (**cor.mat**) in the form $c_{1,2}, c_{1,3}, \ldots, c_{1,k}, \ldots, c_{k-2,k}, c_{k-1,k}$ for correlation matrix:

$$\begin{pmatrix} 1 & c_{1,2} & \ldots & c_{1,k} \\ c_{2,1} & 1 & \ldots & c_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k,1} & c_{k,2} & \ldots & 1 \end{pmatrix}$$

i.e. the upper triangular entries ($\neq 1$) of the correlation matrix from left to right, from the first to the last row.

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 3 MAF 0.3 v 0.005,0.001,
0.002 cor.mat 0.25,-0.3,0.01 strand data
```

The above command generates SNP data for 100 SNPs, 5000 individuals, and minor allele frequency 0.3. From these SNPs phenotype data is generated for 3 phenotypes where the SNP explains 0.5%, 0.1% and 0.2% of the variance in phenotype respectively, with phenotypic correlation matrix:

$$\begin{pmatrix} 1 & 0.25 & -0.3 \\ 0.25 & 1 & 0.01 \\ -0.3 & 0.01 & 1 \end{pmatrix}$$

5.2 Phenotypes with indirect SNP effects

In addition to the parameters specified above, you will need to set **direct** equal to FALSE, and specify **n.indirect**, the number of phenotypes to have indirect effects (by default, the first phenotype always has a direct SNP effect, so **n.indirect** $\in [0, k-1]$)

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 3 MAF 0.3 v 0.005,0.001,
0.002 cor.mat 0.25,-0.3,0.01 direct FALSE n.indirect 1 strand data
```

The above command generates data for 3 phenotypes, 5000 individuals, 100 SNPs, minor allele frequency 0.3, 0.5% variance explained from SNP to the first phenotype, 0.1% variance explained from SNP to the second phenotype, and 0.2% variance explained from the second phenotype to the third phenotype. The phenotypic correlation matrix is the same as above.

N.B. The first **k - n.indirect** phenotypes have direct SNP effects and the last **n.indirect** phenotypes have indirect SNP effects, with the indirect SNP effect on phenotype i being induced by phenotype $i-1$, for $i \in [k - n.indirect + 1, k]$.

5.3 Case-control phenotypes

In addition to the parameters specified above for direct/indirect effects, to generate case-control data you will need to set **cc** equal to TRUE, specify which phenotypes should be case-control (**which.cc**), and the prevalence of the cases (**prev**).

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 3 MAF 0.3 v 0.005,0.001,0
cor.mat 0.2,0.6,-0.05 direct TRUE cc TRUE which.cc 1,3 prev 0.01,0.05
strand data
```

The above command generates data for 3 phenotypes, 100 SNPs, 5000 individuals, minor allele frequency 0.3, and SNP variance explained of 0.5%, 0.1% and 0% for the three phenotypes respectively. The first and third phenotypes are case-control (the second is quantitative), with case prevalences of 1% and 5% respectively. The phenotypic correlation matrix is:

$$\begin{pmatrix} 1 & 0.2 & 0.6 \\ 0.2 & 1 & -0.05 \\ 0.6 & -0.05 & 1 \end{pmatrix}$$

5.4 Output

A folder with name in the format **data_output_date_time** is created containing m SNP data files (**snp_data_i** for $i \in [1, m]$), and m phenotype data files (**pheno_data_i** for $i \in [1, m]$). Each SNP data file contains a single column representing a single SNP, and each phenotype data file contains k columns representing each phenotype.

6 Multivariate method comparison

6.1 Methods

The current list of methods included in this software program are:

- TATES
- min-P
- S_{Het}
- S_{Hom}
- CCA
- MANOVA
- MultiPhen
- Combined-PC
- BIMBAM
- SNPTEST

N.**B**. Use the exact names of the methods that appear above when selecting which methods to compare using the software program. For S_{Het} and S_{Hom} you should refer to the methods as SHet and SHom.

6.2 S1: fixed genetic effects and phenotypic correlations

To run comparisons for scenario S1 of the simulation framework, along with the sample size (**n**), number of SNPs (**m**), minor allele frequency (**MAF**), number of phenotypes (**k**), and phenotypic variance explained (**v**), you need to specify:

- methods to compare (default: **methods** = ALL)
- phenotypic correlations (default: **cor** $= -0.9, -0.8, \ldots, 0.9$)

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 2 MAF 0.3 v 0.005,0.001
cor -0.6,0.2,0.1 methods min-P,CCA strand S1
```

The above command performs the S1 comparison for the methods min-P and CCA for two phenotypes, 100 SNPs, 5000 individuals, minor allele frequency 0.3, phenotypic variance explained by SNP of 0.5% and 0.1% respectively, and for phenotypic correlations ranging from -0.6 to 0.2 in increments of 0.1.

6.3 S2: uniform genetic effects and phenotypic correlations

To run comparisons for scenario S2 of the simulation framework you will need to specify the sample size (**n**), number of SNPs (**m**), minor allele frequency (**MAF**), number of phenotypes (**k**) and methods to compare (**methods**). The phenotypic variance explained and phenotypic correlations are sampled from the uniform distribution.

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 2 MAF 0.3 methods min-P,
CCA strand S2
```

The above command runs the S2 comparison on the methods min-P and CCA, for two phenotypes, 100 SNPs, 5000 individuals, and minor allele frequency 0.3.

6.4 S3: phenotypic correlations reflective of genetic effects

To run comparisons for scenario S3 of the simulation framework, along with the sample size (**n**), number of SNPs (**m**), minor allele frequency (**MAF**), number of phenotypes (**k**), phenotypic variance explained (**v**), and methods (**methods**), you need to specify:

- genetic correlations (default: $\mathbf{gen.cor} = 0.6, 0.2, 0.05$)

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 2 MAF 0.3 v 0.005,0.001
gen.cor 0.6,0.2,0.05 methods min-P,CCA strand S3
```

The above command compares the min-P and CCA methods for two phenotypes, 100 SNPs, 5000 individuals, minor allele frequency 0.3, and phenotypic variance explained by SNP of 0.5% and 0.1% respectively. The genetic correlations are 0.6, 0.2 and 0.05, meaning that if the phenotypic variance explained is the same for both phenotypes, their correlation is 0.6, if the variance explained is different but neither are zero, the pairwise phenotypic correlation is 0.2, and if one variance explained is zero, the pairwise phenotypic correlation is 0.05. For example, if $\mathbf{k} = 4$ and $\mathbf{v} = 0.005, 0.005, 0.001, 0$, then the phenotypic correlation matrix is given by:

$$\begin{pmatrix} 1 & 0.6 & 0.2 & 0.05 \\ 0.6 & 1 & 0.2 & 0.05 \\ 0.2 & 0.2 & 1 & 0.05 \\ 0.05 & 0.05 & 0.05 & 1 \end{pmatrix}$$

6.5 S4a: real data informed phenotypic correlations

Scenario S4a samples the phenotypic correlations from a mixture-Gaussian distribution informed by real data - to run comparisons for S4a, you need to provide the sample size (**n**), number of SNPs (**m**), minor allele frequency (**MAF**), number of phenotypes (**k**), phenotypic variance explained (**v**), and the methods to compare (**methods**).

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 100 k 2 MAF 0.3 v 0.005,0.001
methods min-P,CCA strand S4a
```

The above command compares the min-P and CCA methods for two phenotypes, 100 SNPs, 5000 individuals, minor allele frequency 0.3, and phenotypic variance explained by SNP of 0.5% and 0.1% respectively.

6.6 S4b: real data informed genetic effects and phenotypic correlations

Scenario S4b in addition uses real data to inform the SNP effects (variance explained) - to run comparisons for S4b, you should provide the sample size (**n**), number of SNPs (**m**), number of phenotypes (**k**), and the methods to compare (**methods**).

N.B. Currently the maximum number of phenotypes for this scenario is 12, but this will likely increase in the future when more summary data is incorporated.

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 1000 k 12 methods min-P,CCA
strand S4b
```

The above command compares the min-P and CCA methods for 12 phenotypes, 5000 individuals, and 1000 SNPs under the S4b framework (for more details see our paper).

In addition, S4b allows methods to be compared on combinations of the 12 phenotypes. To do this, set **k** equal to any positive integer less than 12.

Example:

```
R --file=./MultiTraitGWAS.R -q --args n 5000 m 1000 k 2 methods min-P,CCA
strand S4b
```

The above command performs the S4b comparison on every combination of two phenotypes from the 12, of which there are 66 combinations.

N.B. Since S4b uses real SNP data, comparisons may not be able to be performed on the exact number of SNPs specified without removing SNPs. Instead, comparisons are performed on the closest possible number of SNPs.

6.7 Additional Options

If you would like to perform comparisons across multiple scenarios you should provide the **strand** argument with more than one scenario. For example, to run S1, S4a and S4b, specify:

```
strand S1,S4a,S4b
```

Make sure that you have specified the required parameters for all scenarios.

In addition to obtaining power estimates and P-values/$\log_{10}$ Bayes factors, you can get the software to output the simulated data used in the analysis. To do this, you should specify:

```
data.out TRUE
```

By default, the P-value threshold is set to be the genome-wide significance threshold $P = 5 \times 10^{-8}$, and the $\log_{10}$ Bayes factor threshold is set to be 6. These thresholds can be changed by defining the **p.val.threshold** and **log10.bf.threshold** parameters, for example:

```
p.val.threshold 0.0000005
log10.bf.threshold 4
```

If you want to run the comparisons for all methods except two, say MANOVA and min-P, then you don't need to list all eight other methods for the **methods** parameter. Instead, if you specify:

```
methods -MANOVA,-min-P
```

the comparison will be performed for all methods except MANOVA and min-P.

6.8 Output

A folder with name in the format **strand_output_date_time** is created containing the output files for each comparison analysis performed (if multiple scenarios are selected, a separate folder will be created for each). In each folder, there will be an output log file (**strand_output_log**) which details the parameter settings as well as the run time for the analysis, and the power calculation results file, with one column per method (**strand_power**). For each method compared there will be a separate file containing the P-values or $\log_{10}$ Bayes factors (BFs) for that method, where rows represent SNPs (**strand_p_values_method** or **strand_log10_bayes_factors _method**). For S1, the P-value and $\log_{10}$ BF files have multiple columns representing the different phenotypic correlations, and the rows in the power results file correspond to these correlations.

In addition, for scenario S4b and when $k = 12$, a file is created which details the SNPs used in the comparison, as well as the beta coefficients across all 12 phenotypes and the MAF (**SNP_b_MAF_12**). If **m** is larger than the number of SNPs for these 12 phenotypes, multiple iterations of this comparison will be performed and there will be a file for each iteration, with the iteration number at the end of the filename (**SNP_b_MAF_12_it**, where **it** is the iteration number). When $k \neq 12$, a file is created which details the combinations of phenotypes (**combinations_k**), and one file per combination is created detailing which SNPs were included (**SNP_b_MAF_k_n.comb**, where **n.comb** is the combination number). Again, when **m** is larger than the number of SNPs for that combination, multiple iterations will be performed to estimate the power, and for each combination there will be a file for every iteration performed (**SNP_b_MAF_k_n.comb_it**).

If you requested the simulated data to be output, then a folder called **data** will be created within this directory: this folder will contain **m** SNP data files (**snp_data_i** for $i \in [1, m]$) and **m** phenotype data files (**pheno_data_i** for $i \in [1, m]$). For S4b when $k = 12$, the SNP and phenotype data filenames instead end with **_it_n.snp**, where **n.snp** ranges from 1 to the total number of SNPs for the 12 phenotypes for that iteration. When $k \neq 12$, the SNP and phenotype data filenames end with **_k_n.comb_it_n.snp**, where **n.snp** ranges from 1 to the total number of SNPs for that combination of phenotypes, for that iteration.

N.B. SNP numbers can vary between iterations as they are sampled randomly each time; there is a limit imposed so that there are no more than 20 SNPs associated with any one phenotype in order to not bias the results toward phenotypes with lots of SNP hits.

N.B. For some phenotypes there are as little as two SNPs associated, and so you may find that for some of the combinations of phenotypes there are no SNPs for which there is complete SNP effect data. In this case, there will be no SNP effect data file for that combination.