University of Colorado **Boulder**

Institute for Behavioral Genetics

Multivariate GWAS in Genomic SEM

Presented by:

Andrew D. Grotzinger

Four Primary Steps

1. Munge the summary statistics ***(munge)***
2. Run LD-Score Regression to obtain the genetic covariance and sampling covariance matrices ***(ldsc)***
3. Prepare the summary statistics for multivariate GWAS ***(sumstats)***
4. Run the multivariate GWAS ***(commonfactorGWAS; userGWAS)***

These two steps mirror that for models without SNP effects and need not be run again for the same traits

Github Example: P-factor

Using GWAS sumstats for:

- Schizophrenia (Pardiñas et al., 2018); *N* = 105,318
- Bipolar Disorder (Sklar et al., 2011); *N* = 16,731
- Major Depressive Disorder (Wray et al., 2018); *N* = 173,005
- PTSD (Duncan et al., 2017); *N* = 9,537
- Anxiety (Otowa et al., 2016); *N* = 17,310

Step 1: *munge*

Munge: convert raw data from one form to another

Converts to Z-statistics

Aligns to same reference allele

Restricts to hapmap3 SNPs

Example Munge .log file for MDD

```
Munging file: MDD2018_ex23andMe.txt
Interpreting the SNP column as the SNP column.
Interpreting the A1 column as the A1 column.
Interpreting the A2 column as the A2 column.
Interpreting the OR column as the effect column.
Interpreting the INFO column as the INFO column.
Interpreting the P column as the P column.
Interpreting the  NCA  column as the N_CAS (sample size for cases) column.
Interpreting the  NCO  column as the N_CON (sample size for controls) column.
Interpreting the MAF column as the MAF (minor allele frequency) column.
As the file includes both N_CAS and N_CON columns, the summation of these two columns will be used as the total sample
size
Merging file: MDD2018_ex23andMe.txt with the reference file: w_hm3.snplist
13554550 rows present in the full MDD2018_ex23andMe.txt summary statistics file.
12343318 rows were removed from the MDD2018_ex23andMe.txt summary statistics file as the rs-ids for these rows were not
present in the reference file.
The effect column was determined to be coded as an odds ratio (OR) for the MDD2018_ex23andMe.txt summary statistics
file. Please ensure this is correct.
4 row(s) were removed from the MDD2018_ex23andMe.txt summary statistics file due to the effect allele (A1) column not
matching A1 or A2 in the reference file.
100533 rows were removed from the MDD2018_ex23andMe.txt summary statistics file due to INFO values below the designated
threshold of 0.9
12128 rows were removed from the MDD2018_ex23andMe.txt summary statistics file due to missing MAF information or MAFs
below the designated threshold of 0.01
1098567 SNPs are left in the summary statistics file MDD2018_ex23andMe.txt after QC.
I am done munging file: MDD2018_ex23andMe.txt
The file is saved as mdd.sumstats.gz in the current working directory.
```

Step 2: *ldsc*

Computes the genetic covariance (*S*) and sampling covariance (*V*) matrix discussed in Michel's video

Note that it is best to practice to pause here and fit the "base" model using the *usermodel* or *commonfactor* functions before trying to run multivariate GWAS to make sure your model fits well and doesn't produce warnings/errors

Step 3: *sumstats*

As with munge, makes sure that the same allele is the reference allele in all cases.

The coefficients and their SEs are then further transformed such that they are scaled relative to unit-variance scaled phenotypes.

How this rescaling occurs will depend on both the scale of the outcome and how the GWAS was run

sumstats arguments

- **files**: The name of the summary statistics files. This should be the same as the name of the files used for the munge function in Step 1 and the files should be in the same listed order used for the ldsc function in step 2.
- **ref**: The reference file used to calculate SNP variance across traits
- **trait.names**: The names of the traits in the order that they are listed for the files.
- **se.logit**: Whether the SEs are on a logistic scale.
- **OLS**: Whether the phenotype was a continuous outcome analyzed using an observed least square (OLS; i.e., linear) estimator.

sumstats arguments

- **linprob**: Whether the phenotype was a dichotomous outcome analyzed using an OLS estimator
- **prop**: In order to perform the LPM conversion above from OLS betas prop takes the proportion of cases over the total sample size (range: 0 - 1).
- **N**: A user provided N listed in the order the traits are listed for the files argument needed for LPM and OLS transformations.
- **info.filter**: The INFO filter to use. The package default is 0.6.
- **maf.filter**: The MAF filter to use. The package default is 0.01.
- **keep.indel**: Whether insertion deletions (indels) should be included in the output. Default = FALSE.
- **parallel**: Whether the function should be run in parallel. Default = FALSE.
- **cores**: When running in parallel, whether you want the computer to use a certain number of cores.

Example *sumstats* .log file

```
Please note that the files should be in the same order that they were listed for the ldsc function
Preparing summary statistics for file: pgc.bip.full.2012-04.txt
Interpreting the SNPID column as the SNP column.
Interpreting the A1 column as the A1 column.
Interpreting the A2 column as the A2 column.
Interpreting the OR column as the effect column.
Interpreting the INFO column as the INFO column.
Interpreting the SE column as the SE column.
Interpreting the PVAL column as the P column.
Merging file: pgc.bip.full.2012-04.txt with the reference file: reference.1000G.maf.0.005.txt
2427220 rows present in the full pgc.bip.full.2012-04.txt summary statistics file.
73841 rows were removed from the pgc.bip.full.2012-04.txt summary statistics file as the rsIDs for these SNPs were not
present in the reference file.
The effect column was determined to be coded as an odds ratio (OR) for the pgc.bip.full.2012-04.txt summary statistics
file based on the median of the effect column being close to 1. Please ensure the interpretation of this column as an OR
is correct.
2794 rows were removed from the pgc.bip.full.2012-04.txt summary statistics file due to effect values estimated at
exactly 0 as this causes problems for matrix inversion necessary for later Genomic SEM analyses.
311 row(s) were removed from the pgc.bip.full.2012-04.txt summary statistics file due to the effect allele (A1) column
not matching A1 or A2 in the reference file.
18 row(s) were removed from the pgc.bip.full.2012-04.txt summary statistics file due to the other allele (A2) column not
matching A1 or A2 in the reference file.
122061 rows were removed from the pgc.bip.full.2012-04.txt summary statistics file due to INFO values below the
designated threshold of 0.6
Performing transformation under the assumption that the effect column is either an odds ratio or logistic beta (please
see output above to determine whether it was interpreted as an odds ratio) and the SE column is a logistic SE (i.e., NOT
the SE of the odds ratio) for: pgc.bip.full.2012-04.txt
2228195 SNPs are left in the summary statistics file pgc.bip.full.2012-04.txt after QC and merging with the reference
file.
```

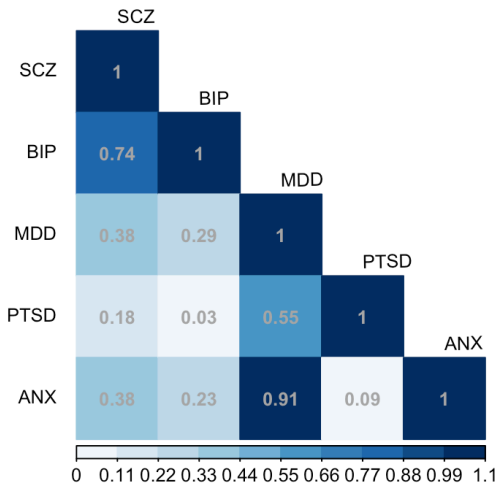
Step 4a: *commonfactorGWAS*

Automatically specifies a common factor model where the SNP predicts the common factor

Behind the scenes

- GenomicSEM GWAS functions automatically combine output from Steps 2 and 3
- Creates as many covariance matrices as there are SNPs across traits

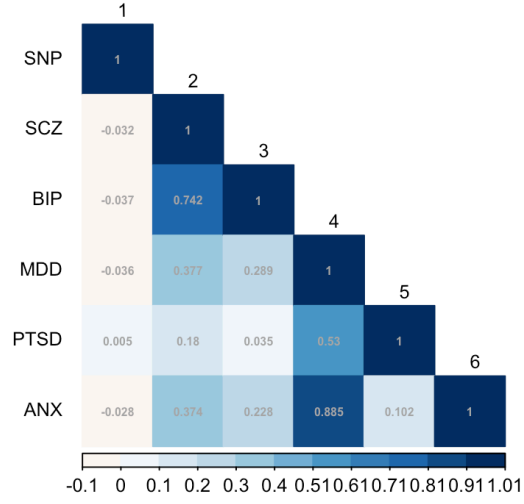
Step 2: Run `ldsc`



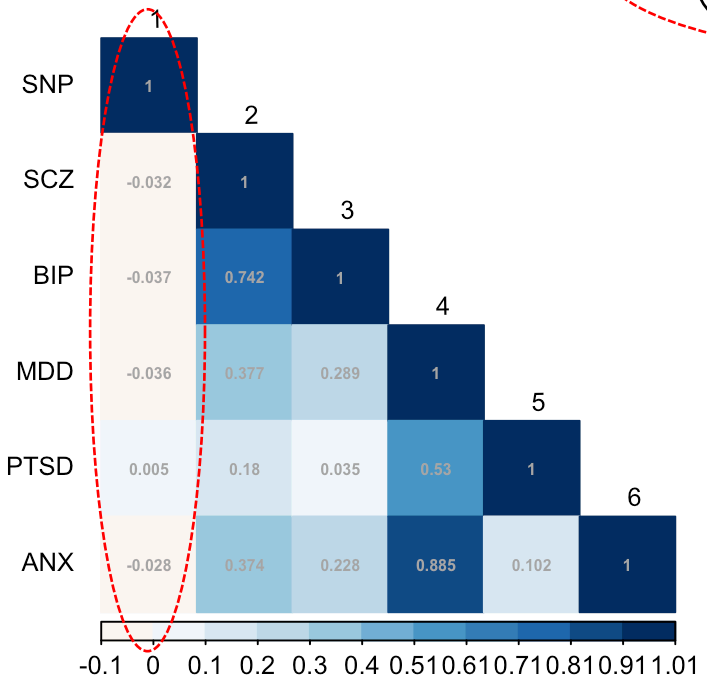
+

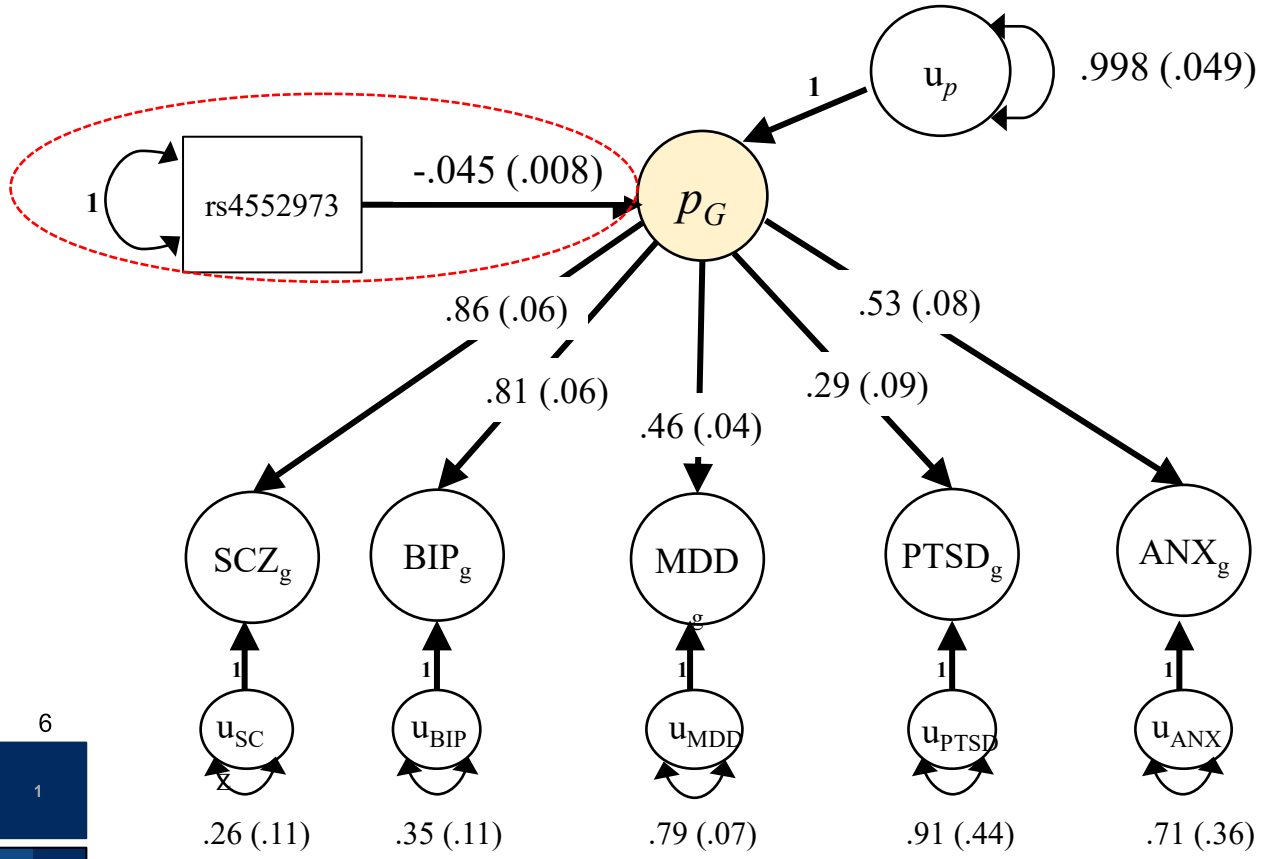
=

Step 3: Run `sumstats` GWAS functions `combine the two`



Expanded S Matrix





commonfactorGWAS arugments

- **covstruc**: The output from LDSC.
- **SNPs**: The output from sumstats.
- **estimation**: Whether the models should be estimated using "DWLS" or "ML" estimation. The package default is "DWLS".
- **cores**: How many computer cores to use in parallel processing. The default is to use 1 less core than is available in the local environment.
- **toler**: What tolerance level to use for matrix inversions. This is only something that needs to be of concern if warning/error messages are produced to the effect of "matrix could not be inverted".

commonfactorGWAS arugments

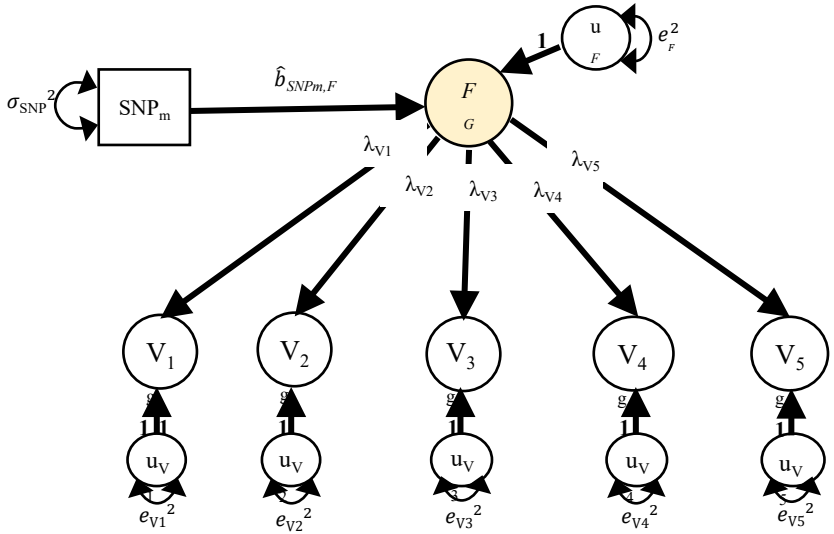
- **parallel**: An optional argument specifying whether you want the function to be run in parallel, or to be run serially.

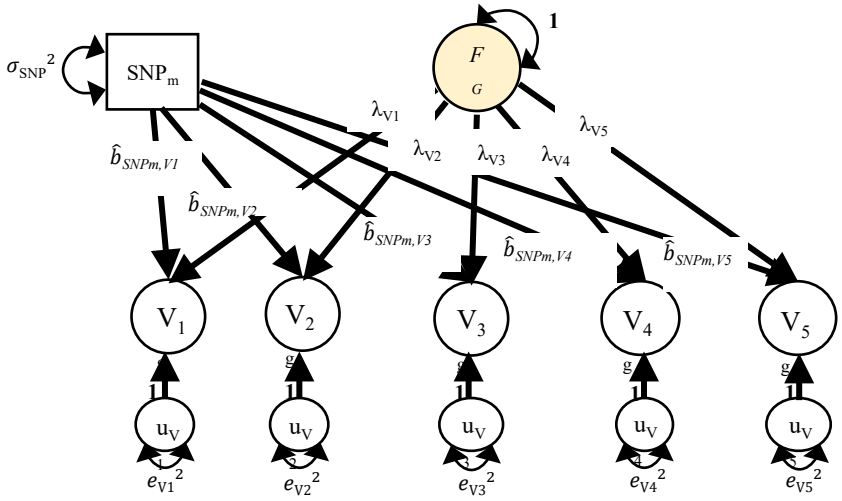
- **GC**: Level of Genomic Control (GC) you want the function to use. The default is 'standard' which adjusts the univariate GWAS standard errors by multiplying them by the square root of the univariate LDSC intercept.

- **MPI**: Whether the function should use multi-node processing (i.e., MPI).

Estimates of SNP level heterogeneity (Q_{SNP})

- Asks to what extent the effect of the SNP operates through the common factor
- χ^2 distributed test statistic, indexing fit of the common pathways model against independent pathways model





Step 4b: *userGWAS*

Allows the user to specify any model including individual SNP effects (e.g., SNP predicting multiple, correlated factors)

userGWAS additional arguments

- **model**: The model that is being estimated (written in lavaan syntax)

- **sub**: An optional argument specifying whether or not the user is requesting only specific components of the model output to be saved.

Run times for this example

1. ***munge:*** *7 minutes 58 seconds*

2. ***ldsc:*** *1 minute 17 seconds*

3. ***sumstats:*** *6 minutes 56 seconds (run in parallel)*

4a. commonfactorGWAS: *17 seconds (run in parallel for 100 SNPs)*

4b. userGWAS: *10 seconds (run in parallel for 100 SNPs)*

Run Time Notes

- Parallel/MPI processing for both **`userGWAS`** and **`commonfactorGWAS`** is available
- Parallel is the same as serial processing, except that it takes an additional cores argument specifying how many cores to use
- MPI takes advantage of multi-node computing environments. Requires that Rmpi already be installed on the computing cluster
- Ideal run-time scenario: split sumstats output across jobs on a cluster and run using MPI
  - All runs are independent of one another!