DEPARTMENT OF PSYCHOLOGY
THE UNIVERSITY OF TEXAS AT AUSTIN

University of Colorado **Boulder**

Institute for Behavioral Genetics

GenomicSEM: A Novel Method for Modeling Multivariate Genetic Architecture

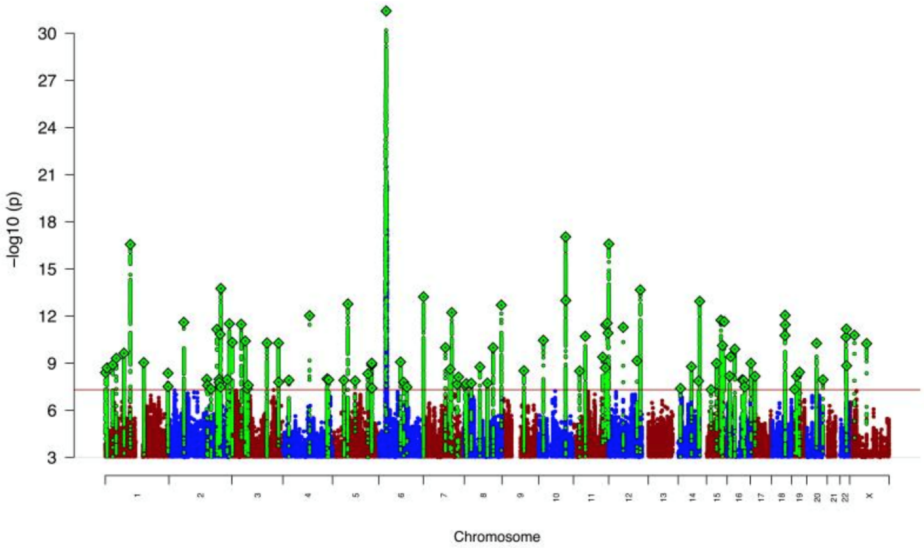
Presented by:

Andrew D. Grotzinger

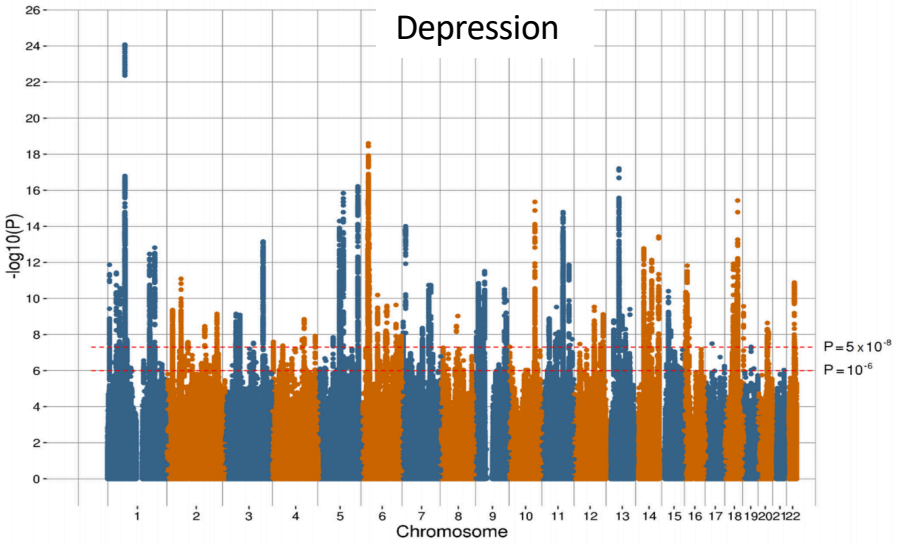
Paper: Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D, Ip, H. F., McIntosh, A. M., Deary, I. J., Koellinger, P. D., Harden, K. P., **Nivard, M. G.**, & **Tucker-Drob, E. M.** (in press). **Genomic SEM provides insights into the multivariate genetic architecture of complex traits.** *Nature Human Behaviour.*

Link to paper: rdcu.be/bvn7t

Schizophrenia



Depression



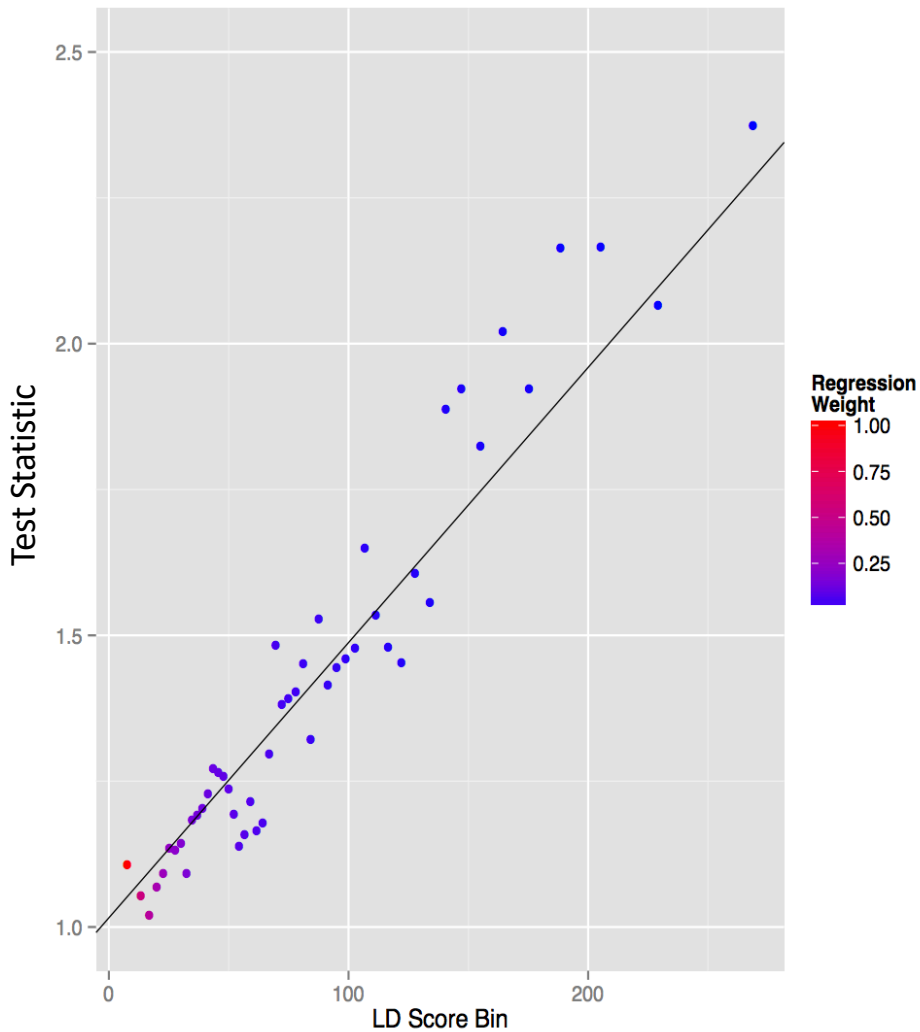
Traits are highly polygenic, so not simply a matter of identifying ~5 overlapping genes

An atlas of genetic correlations across human diseases and traits

Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price & Benjamin M Neale

Nature Genetics **47**, 1236–1241 (2015) | Download Citation

Estimates genetic correlations between samples with varying degrees of sample overlap using publicly available data



- To estimate SNP Heritability:
  - Regress GWAS test statistic against LD Scores for all SNPs (not just significant ones)
- To estimate Genetic Correlation:
  - Regress product of GWAS test statistics for two different phenotypes against LD Scores

Pervasive (Statistical) Pleiotropy Necessitates Methods for Analyzing Joint Genetic Architecture

Analysis of shared heritability in common disorders of the brain

The Brainstorm Consortium*†

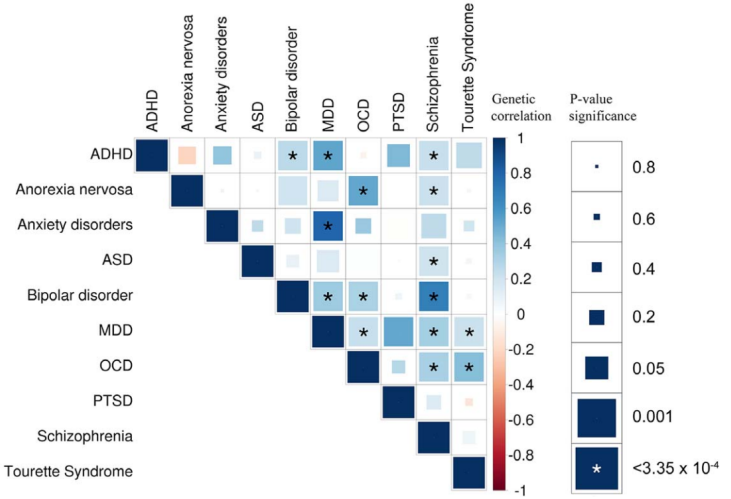


Fig. 1. Genetic correlations across psychiatric phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly different from zero after Bonferroni correction.

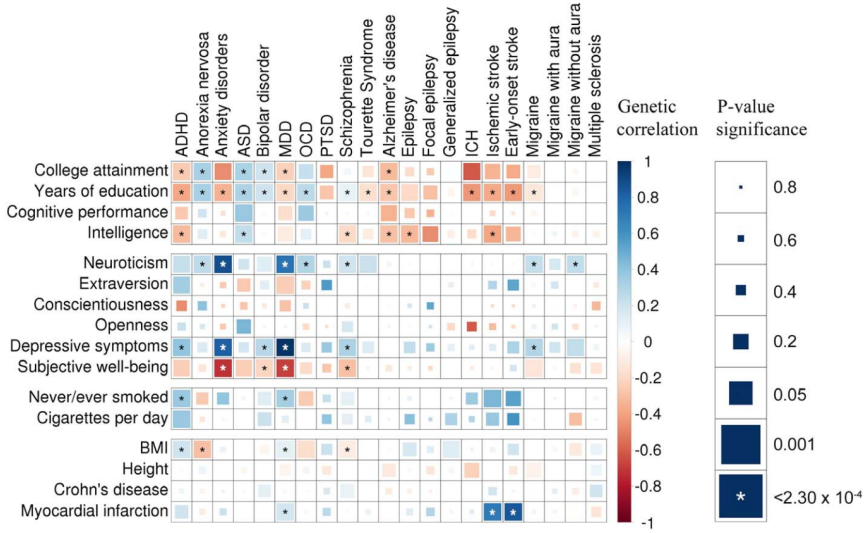


Fig. 4. Genetic correlations across brain disorders and behavioral-cognitive phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly different from zero after Bonferroni correction.

The Brainstorm Consortium, *Science* **360**, 1313 (2018) 22 June 2018

Background

- Genome-wide methods are clearly suggestive of both high polygenicity and pervasive pleiotropy

- **Genetic correlations as data to be modeled, not simply results by themselves**
  - What data-generating process gave rise to the correlations?

Genomic SEM

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0566-x

Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

Andrew D. Grotzinger[1*], Mijke Rhemtulla[2], Ronald de Vlaming[3,4], Stuart J. Ritchie[5,6], Travis T. Mallard[1], W. David Hill[5,6], Hill F. Ip[7], Riccardo E. Marioni[5,8], Andrew M. McIntosh[5,9], Ian J. Deary[5,6], Philipp D. Koellinger[3,4], K. Paige Harden[1,10], Michel G. Nivard[7,11] and Elliot M. Tucker-Drob[1,10,11]

He says hoi

Nivard

Tucker-Drob

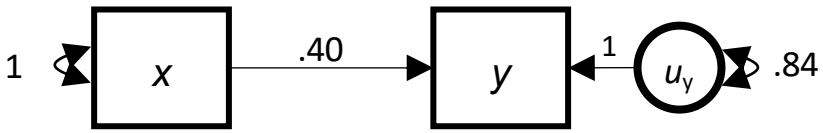
Our solution: GenomicSEM

- Apply structural equation model to estimated genetic covariance matrices
  - Moves past family-based methods by allowing user to examine traits that could not be measured in the same sample

- Genomic SEM provides flexible framework for estimating limitless number of structural equation models using multivariate genetic data from GWAS summary statistics
  - Can be applied to sum stats with varying and unknown degrees of overlap

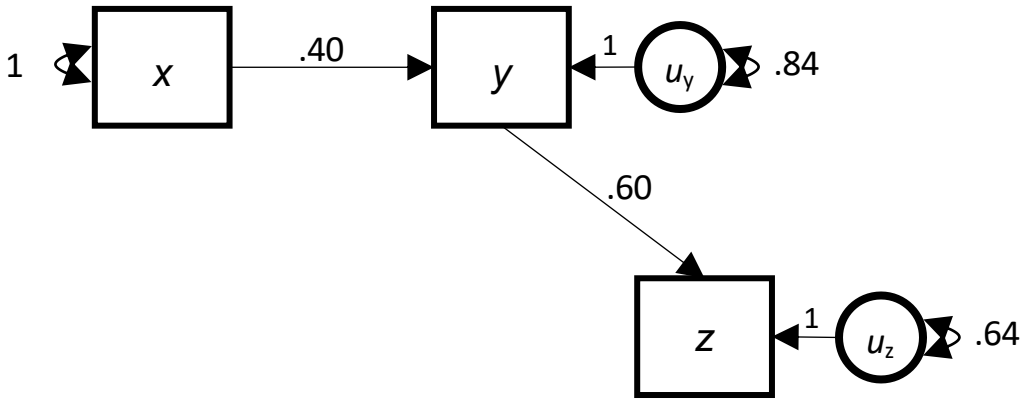
Short Primer on Structural Equation Modeling (SEM)

Imagine we knew the generating causal process



$y = .40\,x + u_y$ $\qquad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

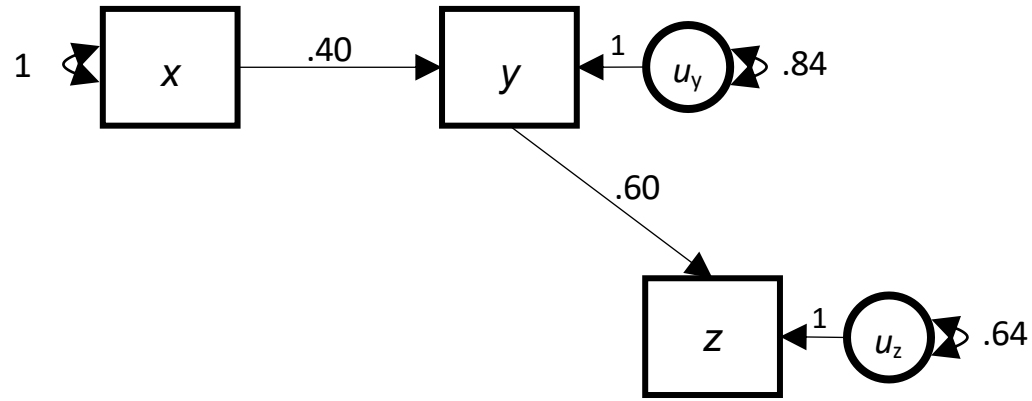
Imagine we knew the generating causal process



$y = .40\,x + u_y$ $\qquad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

$z = .60\,y + u_z$ $\qquad$ $u_z \sim (0,.64)$

Imagine we knew the generating causal process



Implied covariance matrix
in the population

$\text{cov}(x,y,z)_{pop}$ =

1.00		
.40	1.00	
.24	.60	1.00

$y = .40\,x + u_y$ $\qquad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

$z = .60\,y + u_z$ $\qquad$ $u_z \sim (0,.64)$

In practice, we only observe the sample data, and we propose a model

observed covariance matrix
in a sample

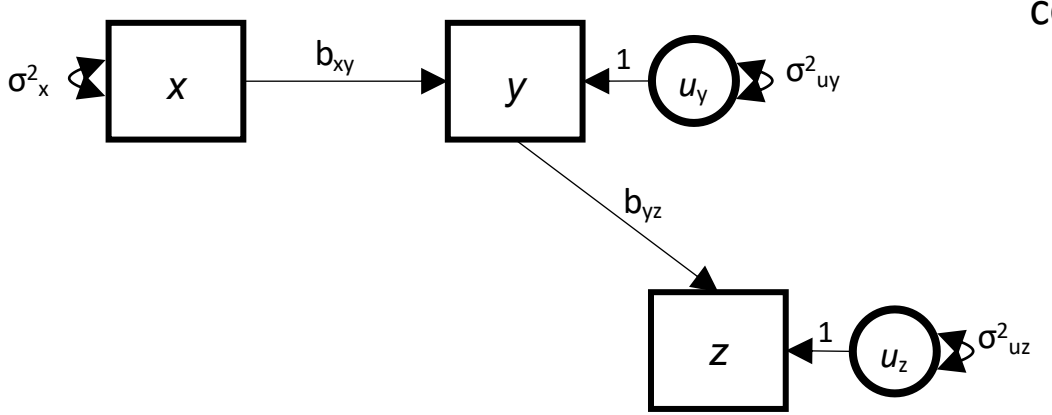
.94		
.33	1.02	
.27	.62	1.02

≈

covariance matrix
in population

1.00		
.40	1.00	
.24	.60	1.00

For the proposed model, estimate parameters from the data, and evaluate model fit to the data



$cov(x,y,z)_{sample}$ =

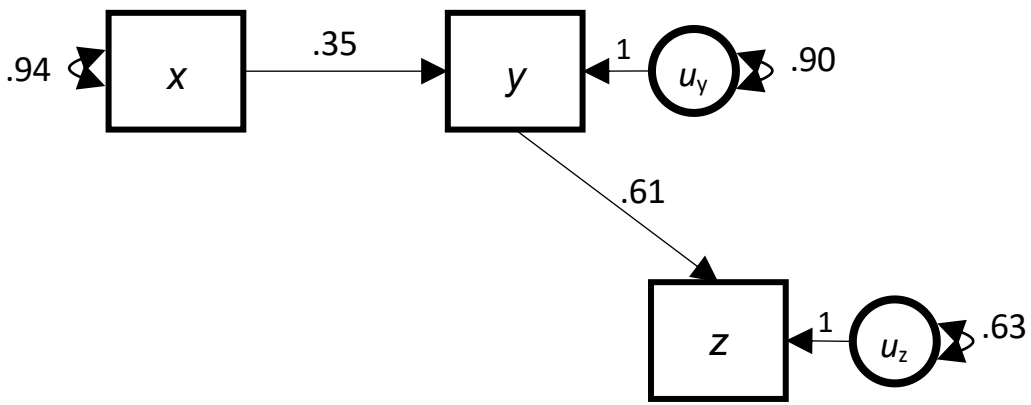
.94		
.33	1.02	
.27	.62	1.02

6 unique elements in the covariance matrix being modeled

5 free model parameters

1 degree of freedom (df)

For the proposed model, estimate parameters from the data, and evaluate model fit to the data



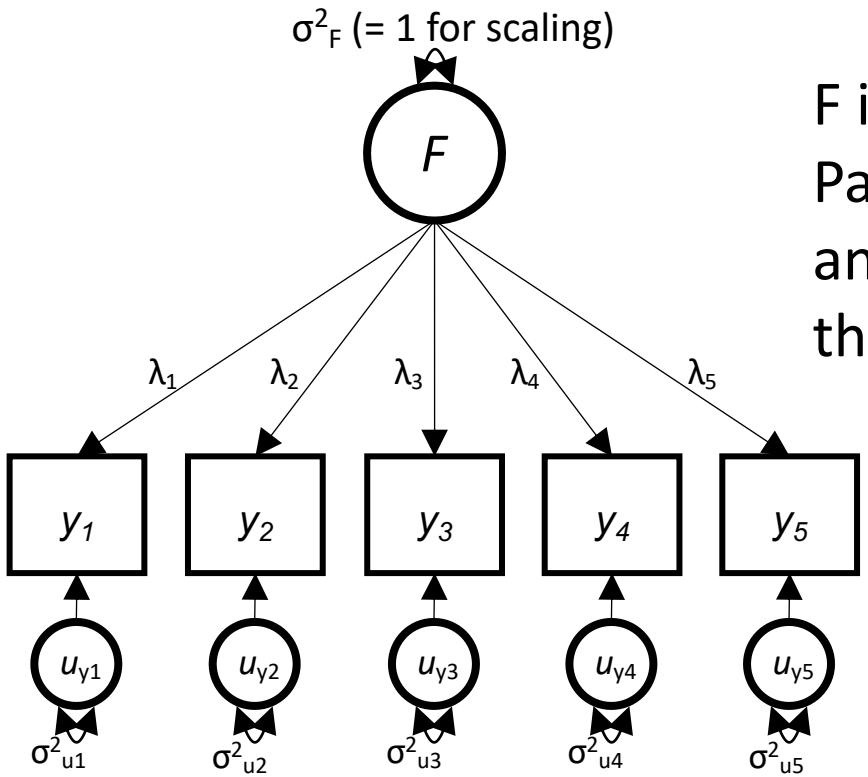
$\text{cov}(x,y,z)_{\text{sample}}$ =

.94		
.33	1.02	
.27	.62	1.02

$\text{cov}(x,y,z)_{\text{implied}}$ =

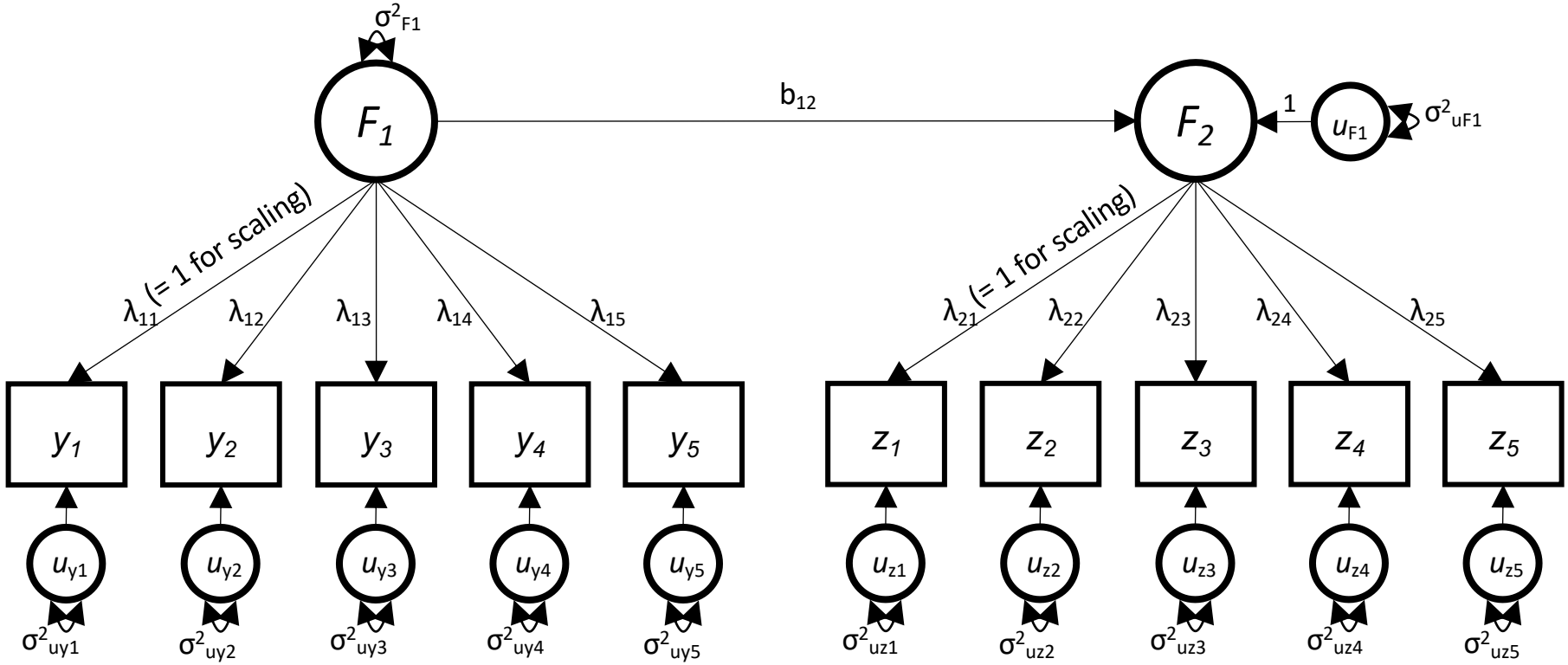
.94		
.33	1.03	
.20	.63	1.00

The model that we fit may include some variables for which we do not observe data



F is unobserved.
Parameters are estimated from,
and fit is evaluated relative to,
the sample covariance matrix for y_1-y_k.

The model that we fit may include some variables for which we do not observe data



Genomic SEM uses these principles to fit structural equation models to genetic covariance matrices derived from GWAS summary statistics using 2 Stage Estimation

- Stage 1: Estimate Genetic Covariance Matrix and associated matrix of standard errors and their co-dependencies
  - We use LD Score Regression, but any method for estimating this matrix (e.g. GREML) and its sampling distribution can be used
- Stage 2: Fit a Structural Equation Model to the Matrices from Stage 1

Fitting Structural Equation Models to GWAS-Derived Genetic Covariance Matrices

Start with GWAS Summary Statistics for the Phenotypes of Interest

- No need for raw data
- No need to conduct a primary GWAS yourself: Download them online!
  - sumstats for over 3700 phenotypes have been helpfully indexed at http://atlas.ctglab.nl/
  - sumstats for over 4000 UK Biobank phenotypes are downloadable at http://www.nealelab.is/uk-biobank

CHR	SNP	BP	A1	A2	INFO	OR	SE	P	Nca	Nco	MAF
8	rs62513865	101592213	T	C	0.957	1.01461	0.0153	0.3438	59851	113154	0.07330
8	rs79643588	106973048	A	G	0.999	1.02122	0.0136	0.1231	59851	113154	0.09200
8	rs17396518	108690829	T	G	0.980	1.00331	0.0080	0.6821	59851	113154	0.43500
8	rs6994300	102569817	A	G	0.466	0.88126	0.4243	0.7658	16823	25632	0.00556
8	rs138449472	108580746	A	G	0.734	0.97181	0.0598	0.6320	41253	79756	0.00852
8	rs983166	108681675	A	C	0.991	0.99144	0.0080	0.2784	59851	113154	0.43200

Stage 1 Estimation: Multivariable LDSC

Create a genetic covariance matrix, S: an "atlas of genetic correlations"

Diagonal elements are (heritabilities)

$$S = \begin{bmatrix} h_1^2 & & & \\ \sigma_{g1,g2} & h_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{g1,gk} & \sigma_{g2,gk} & \cdots & h_k^2 \end{bmatrix}$$

Off-diagonal elements are coheritabilities

Stage 1 Estimation: Multivariable LDSC

Also produced is a second matrix, V, of squared standard errors and the dependencies between estimation errors

$$V_S = \begin{bmatrix} SE(h_1^2)^2 & & & & & & \\ \mathrm{cov}(h_1^2, \sigma_{g1,g2}) & SE(\sigma_{g1,g2})^2 & & & & & \\ \vdots & \vdots & \ddots & & & & \\ & & & SE(\sigma_{g1,gk})^2 & & & \\ & & & & \ddots & & \\ & & & & & SE(h_j^2)^2 & \\ & & & & & & \ddots \\ & & & & & & SE(\sigma_{gj,gk})^2 \\ & & & & & & & SE(h_k^2)^2 \end{bmatrix}$$

Diagonal elements are squared standard errors of genetic variances and covariances

Off-diagonal elements are dependencies between estimation errors used to directly model dependencies that occur due to sample overlap from contributing GWASs

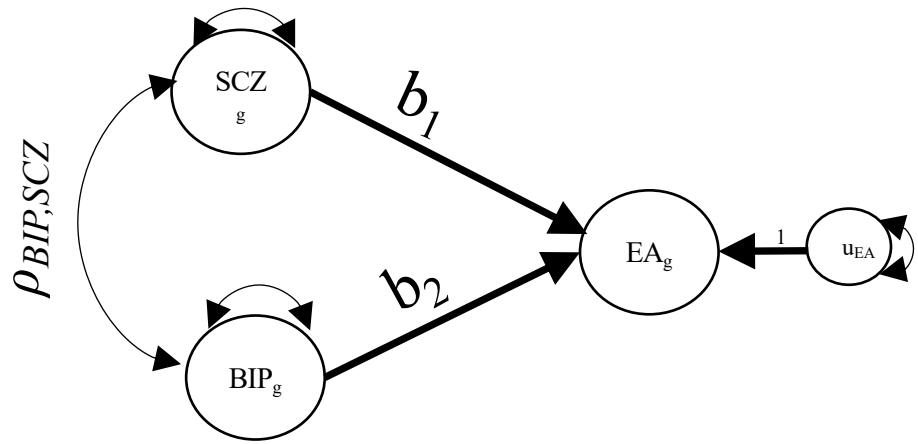
Stage 2 Estimation: Specify the SEM

Example: Genetic multiple regression

$$EA_g = b_1 \times SCZ_g + b_2 \times BIP_g + u$$

S =

SCZ		
.67	BIP	
.11	.30	EA

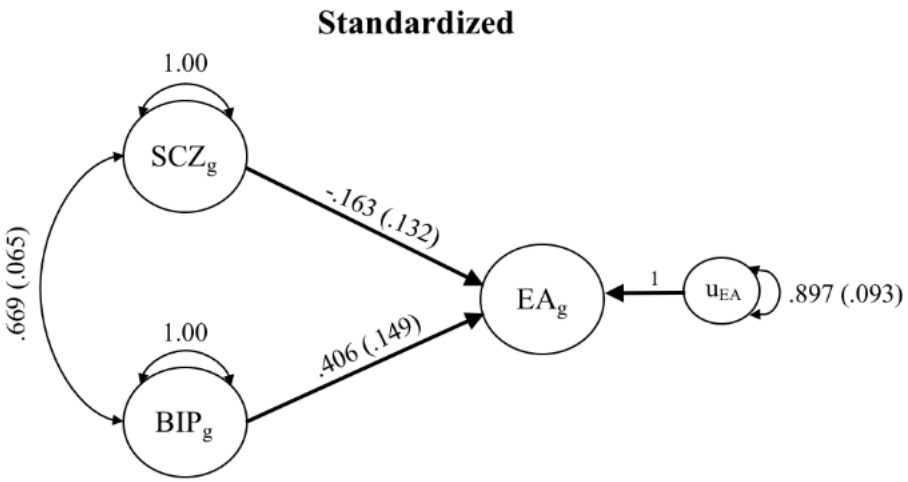


(df = 0, model parameters are a simply a transformation of the matrix)

RESULTS

```
$results
   lhs op rhs Unstand_Est  Unstand_SE STD_Genotype STD_Genotype_SE    STD_All
1   EA  ~ SCZ -0.09464024 0.076689510   -0.1630718      0.13214140 -0.1630718
2   EA  ~ BIP  0.32300380 0.118183679    0.4063490      0.14867882  0.4063490
3  SCZ ~~ BIP  0.12229827 0.011879865    0.6694887      0.06503310  0.6694887
10 SCZ ~~ SCZ  0.25020062 0.017482875    1.0000000      0.06987543  1.0000000
11 BIP ~~ BIP  0.13337232 0.013696265    1.0000000      0.10269196  1.0000000
12  EA ~~  EA  0.07559303 0.007863554    0.8970141      0.09331176  0.8970141
```

$EA_g = -.163 \times SCZ_g + \mathbf{.406} \times BIP_g + u$



Example 2: Model Comparisons for Neuroticism

- 12 neuroticism traits from round 1 of Neale Lab UKB GWAS
- Goal = use model fit indices to compare common factor, two-factor, and three-factor model

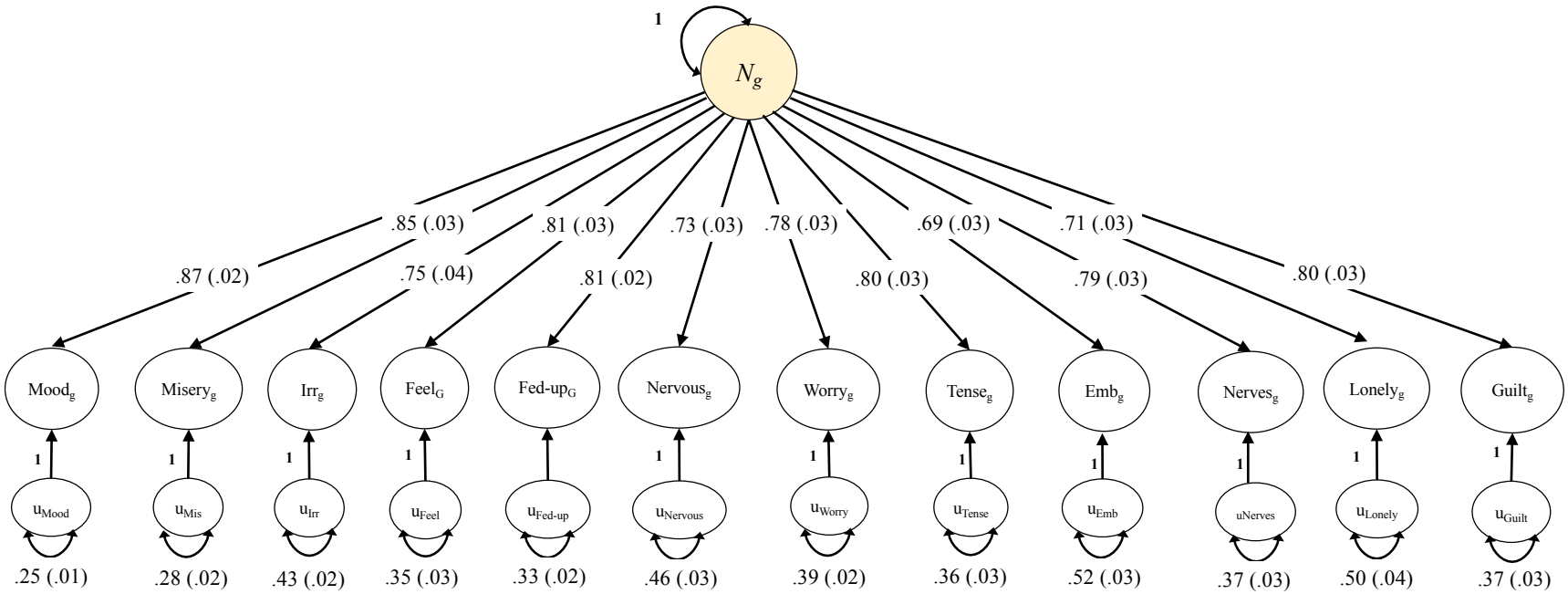
Genetic Correlation Matrix

	mood	misery	irritability	hurt	fedup	nervous	worry	tense	embarass	nerves	lonely	guilt
mood	1											
misery	0.86	1										
irritability	0.77	0.69	1									
hurt	0.67	0.63	0.52	1								
fedup	0.89	0.87	0.69	0.6	1							
nervous	0.52	0.48	0.44	0.67	0.45	1						
worry	0.6	0.56	0.52	0.72	0.5	0.86	1					
tense	0.66	0.54	0.7	0.62	0.55	0.76	0.75	1				
embarass	0.44	0.53	0.44	0.72	0.43	0.69	0.69	0.53	1			
nerves	0.62	0.59	0.47	0.63	0.57	0.87	0.75	0.77	0.59	1		
lonely	0.75	0.77	0.57	0.54	0.8	0.37	0.38	0.53	0.35	0.53	1	
guilt	0.64	0.66	0.57	0.66	0.6	0.56	0.64	0.61	0.77	0.59	0.55	1

-0.1 0.01 0.12 0.23 0.34 0.45 0.56 0.67 0.78 0.89 1

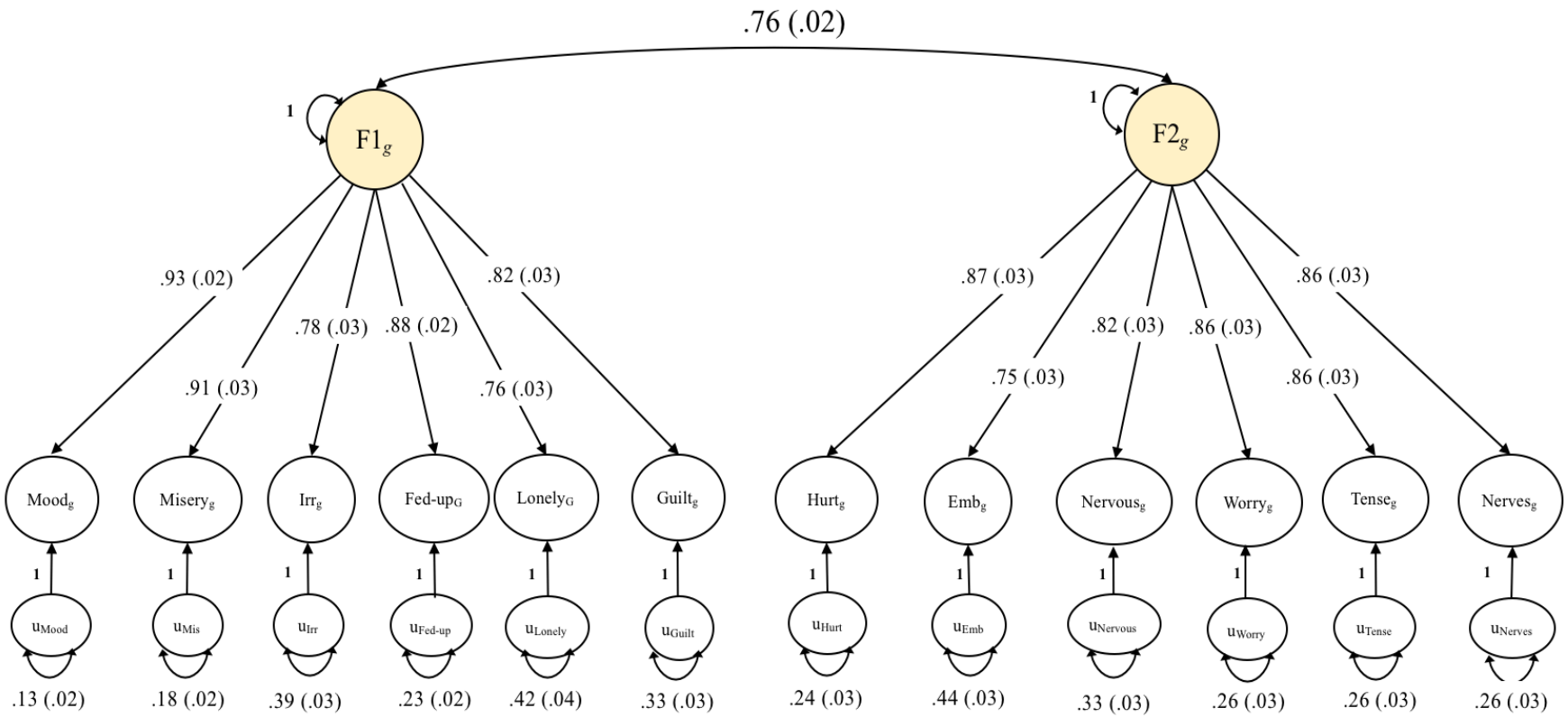
Model 1: Common Factor Model

chisq	df	p_chisq	AIC	CFI	SRMR
4884.104	54	0	4932.104	0.8933184	0.1095286



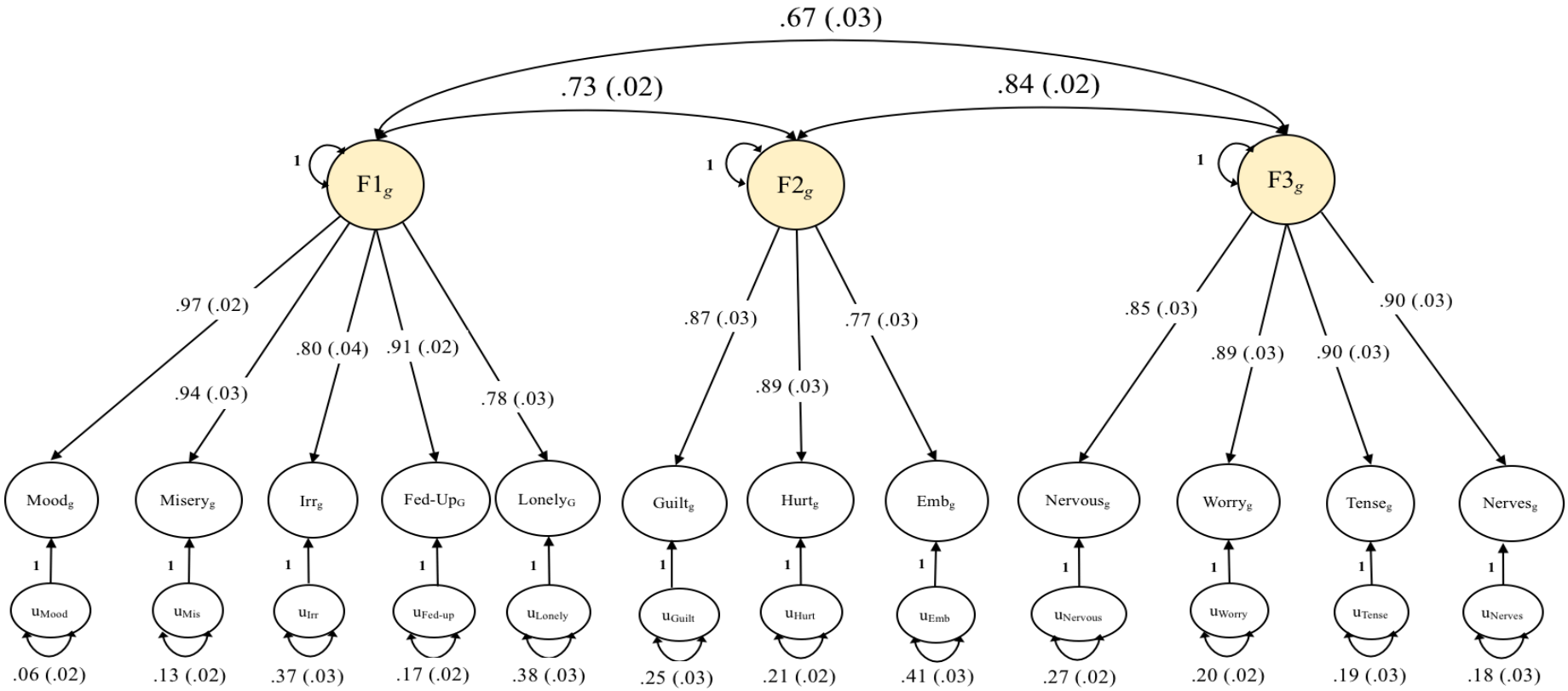
Model 2

chisq	df	p_chisq	AIC	CFI	SRMR
2758.176	53	0	2808.176	0.9402513	0.0766612

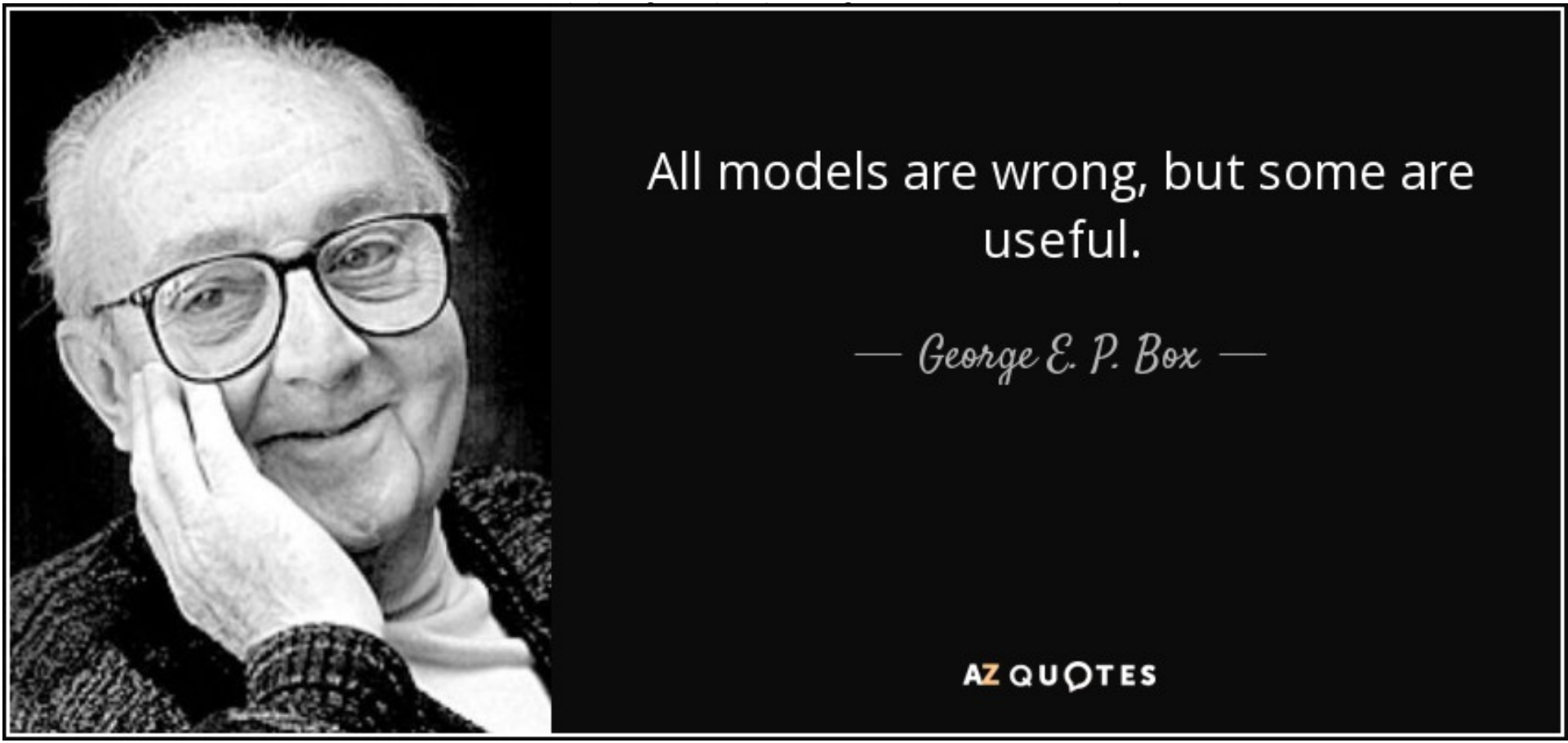


Model 3

chisq	df	p_chisq	AIC	CFI	SRMR
1879.308	51	0	1933.308	0.9596185	0.05733665



Comparison of Model Fit Indices



Incorporating Genetic Covariance Structure into Multivariate GWAS Discovery

Example: the *p* factor as a GWAS target

The American Journal of
Psychiatry
175th Year of Publication

REVIEWS AND OVERVIEWS
Mechanisms of Psychiatric Illness

All for One and One for All: Mental Disorders in One Dimension

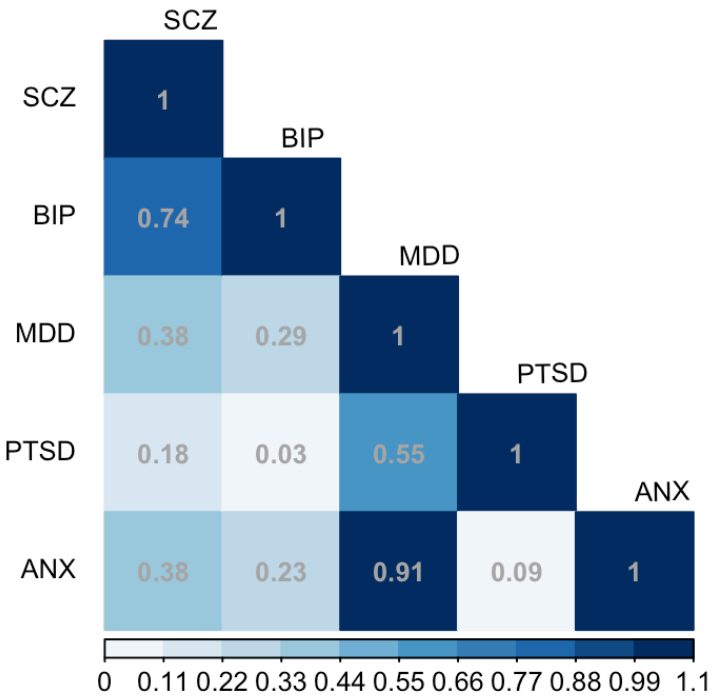
Avshalom Caspi, Ph.D., Terrie E. Moffitt, Ph.D.

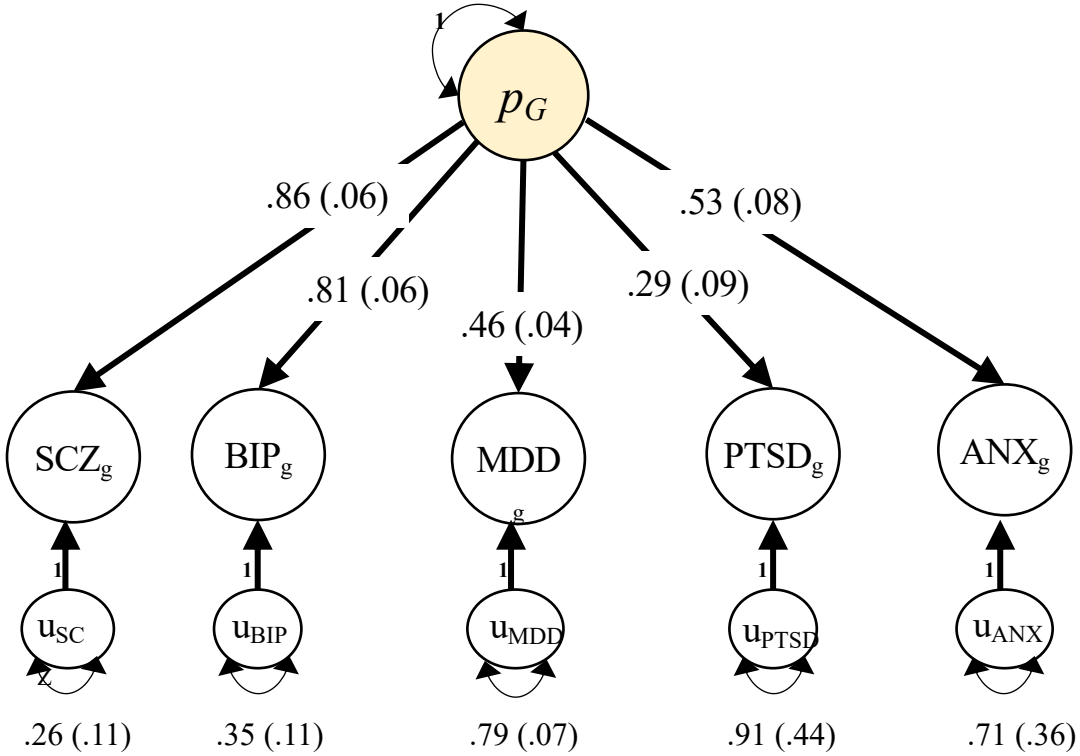
The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders?

Clinical Psychological Science
2014, Vol. 2(2) 119–137

DOI: 10.1177/2167702613497473
cpx.sagepub.com
SAGE

Genetic Correlation Matrix





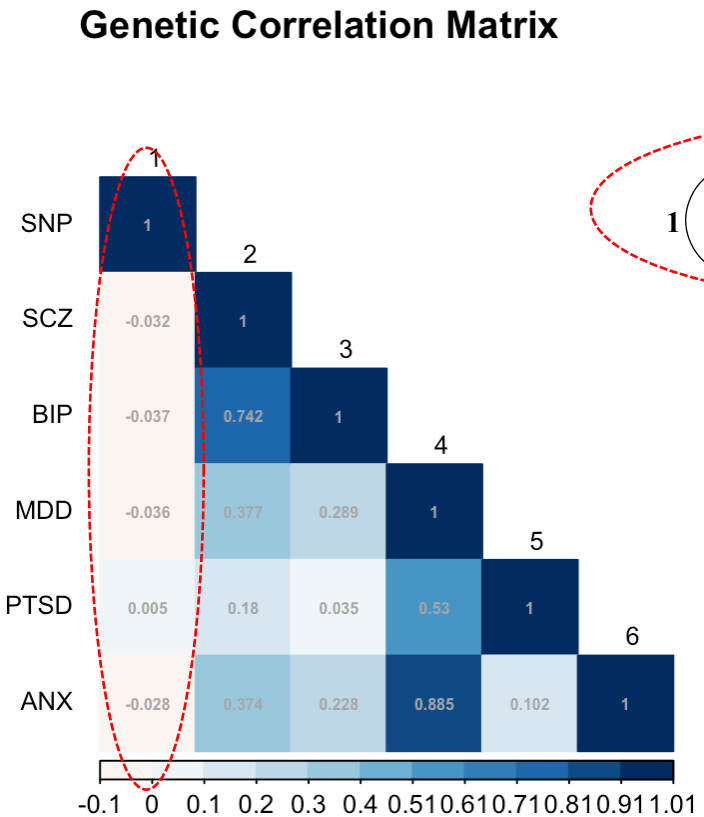
Add SNP Effects to the "Atlas"

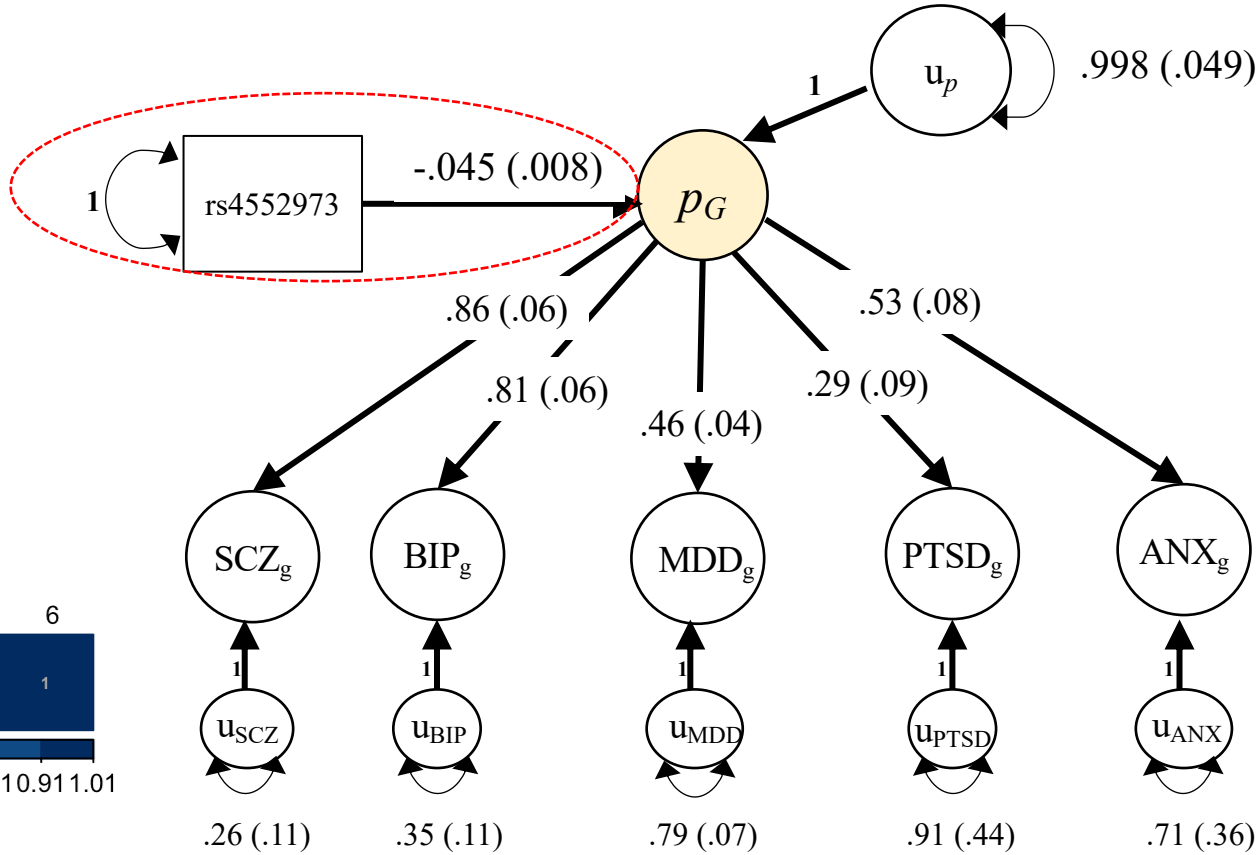
$$S_{\text{Full}} = \begin{bmatrix} \sigma^2_{\text{SNP}} & & & & & \\ \sigma_{\text{SNP,g1}} & h_1^2 & & & & \\ \sigma_{\text{SNP,g2}} & \sigma_{\text{g1,g2}} & h_2^2 & & & \\ \sigma_{\text{SNP,g3}} & \sigma_{\text{g1,g3}} & \sigma_{\text{g2,g3}} & h_3^2 & & \\ \vdots & \vdots & & & \ddots & \\ \sigma_{\text{SNP,g}k} & \sigma_{\text{g1,g}k} & \sigma_{\text{g2,g}k} & \sigma_{\text{g3,g}k} & \cdots & h_k^2 \end{bmatrix}$$

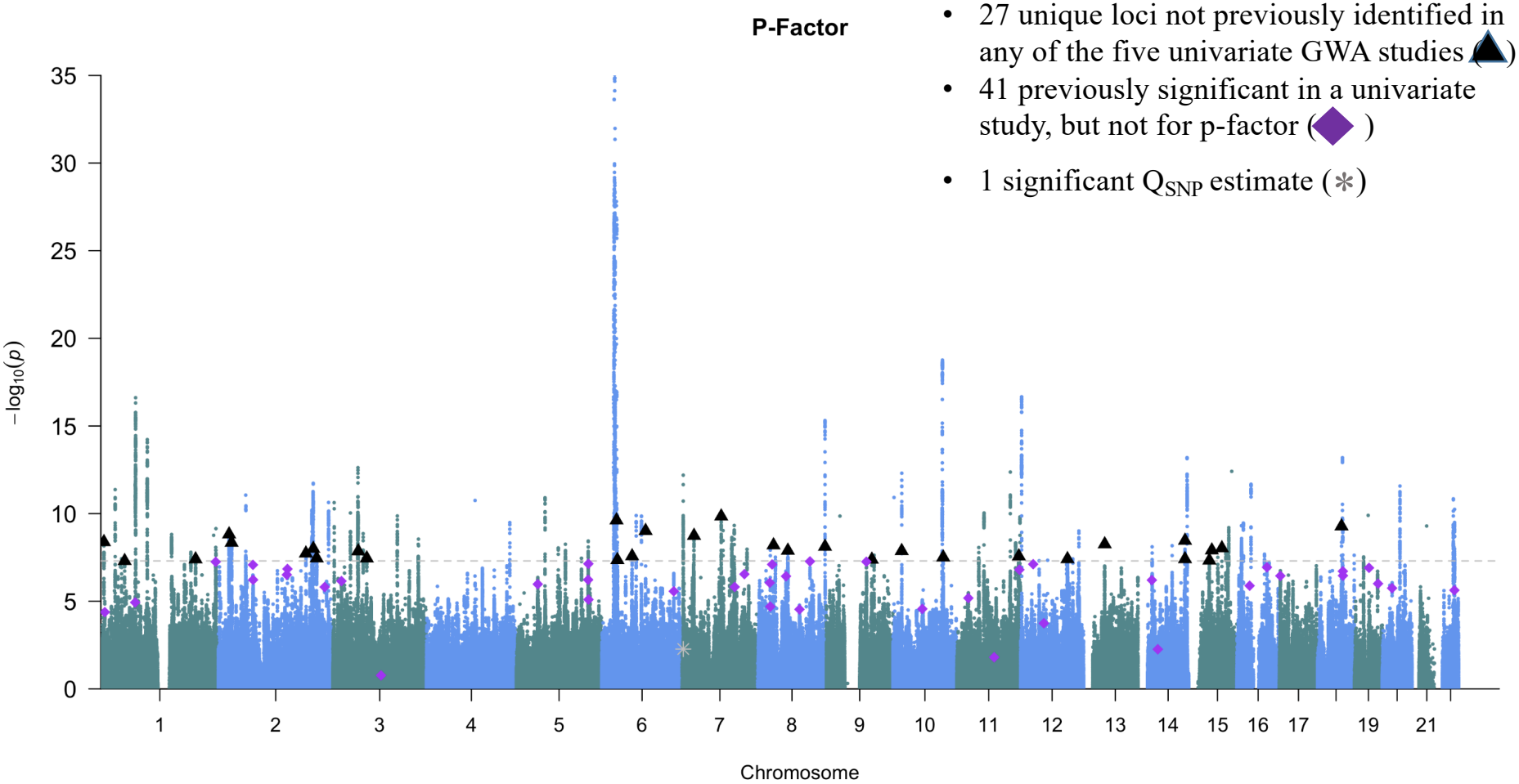
Genetic Covariances from LDSC

Betas from GWAS sumstats scaled to covariances using MAFs

GWAS of a Latent Factor

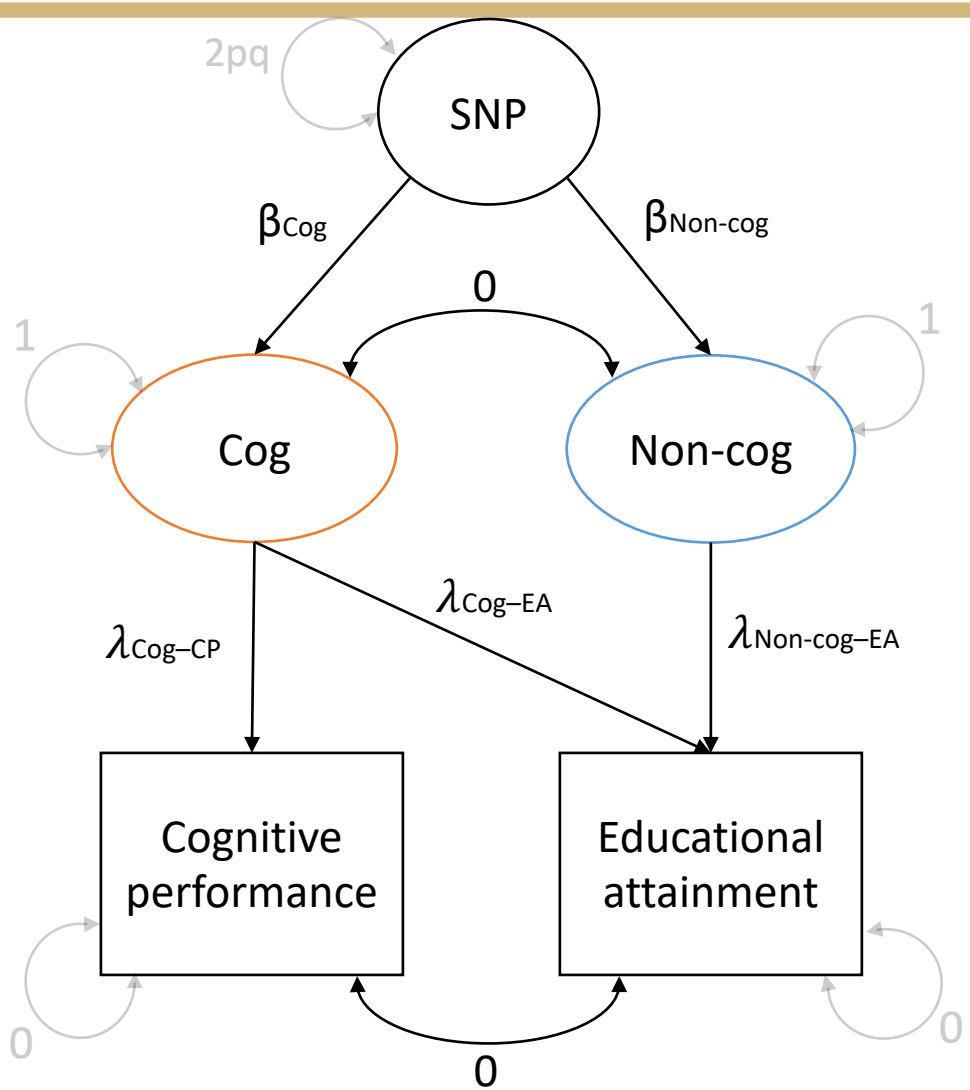






- 128 lead SNPs
- 27 unique loci not previously identified in any of the five univariate GWA studies (▲)
- 41 previously significant in a univariate study, but not for p-factor (◆)
- 1 significant Q_{SNP} estimate (*)

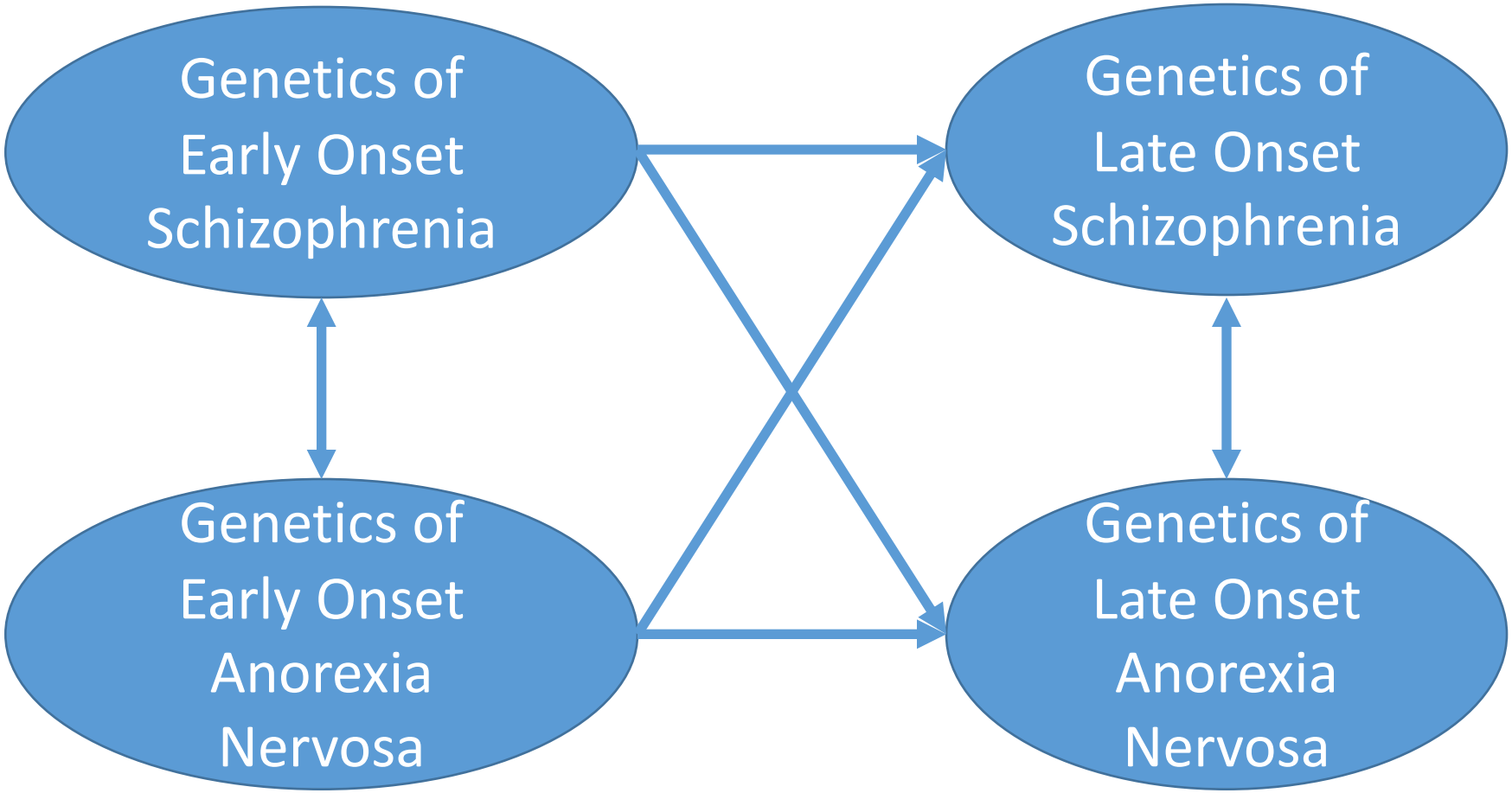
GWAS by subtraction

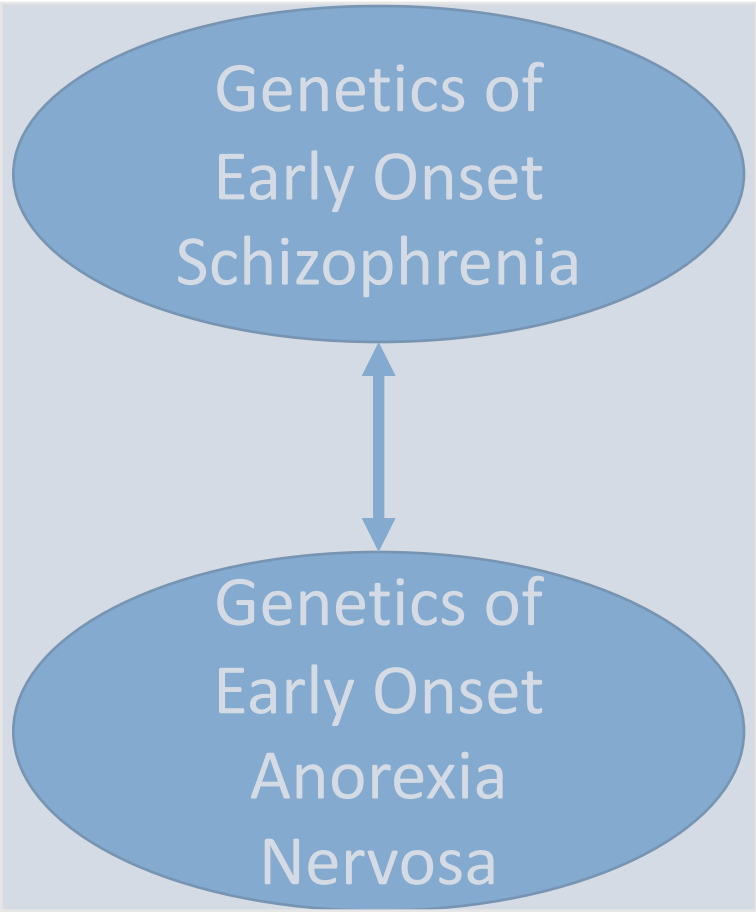


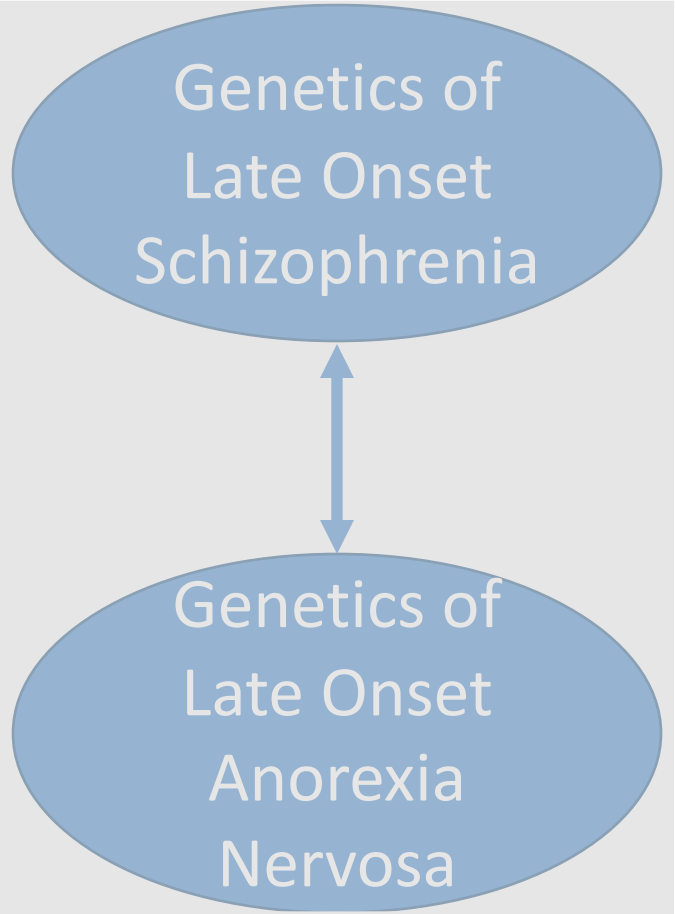
Even if you are not interested in genetics:

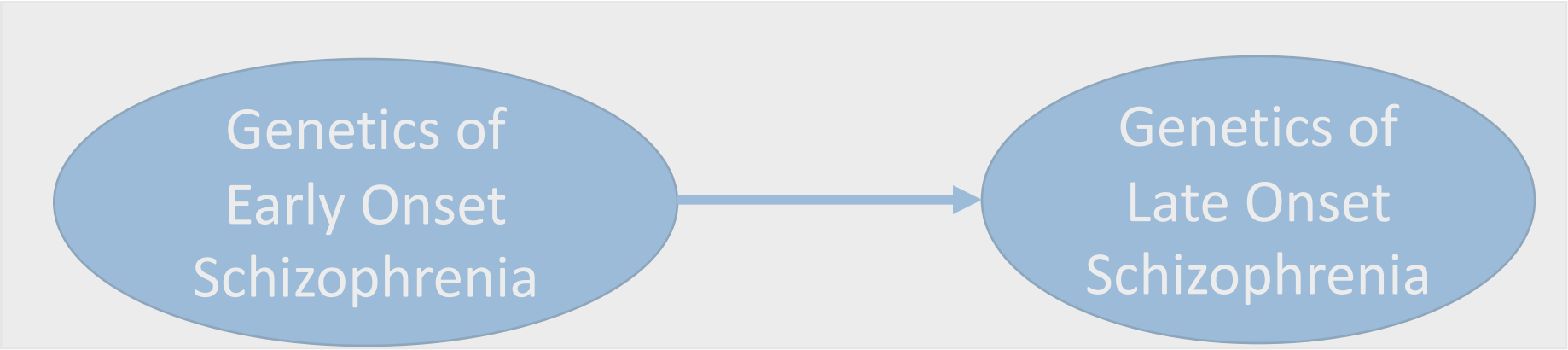
Can now examine systems of relationships between a wide array of (rare) traits

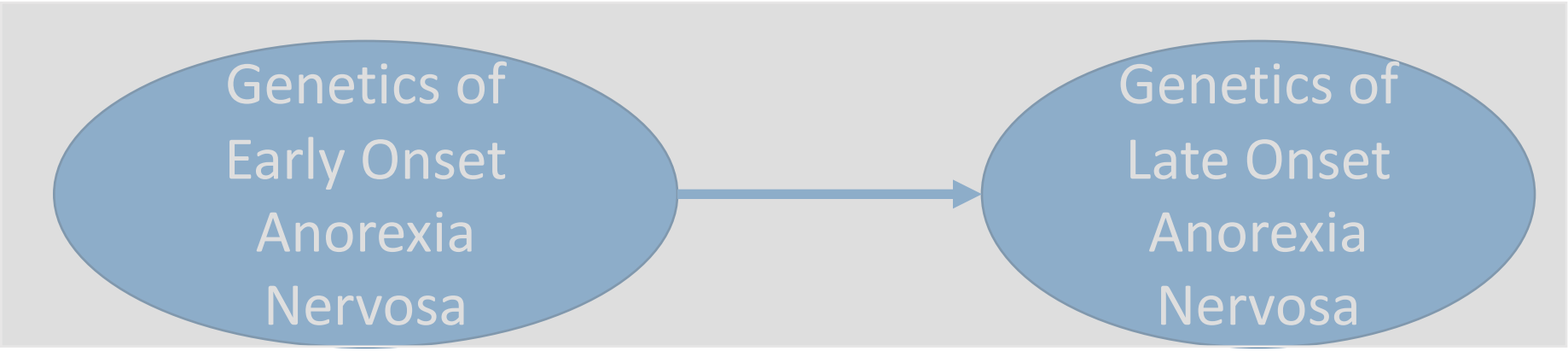
that could not be measured in the same sample

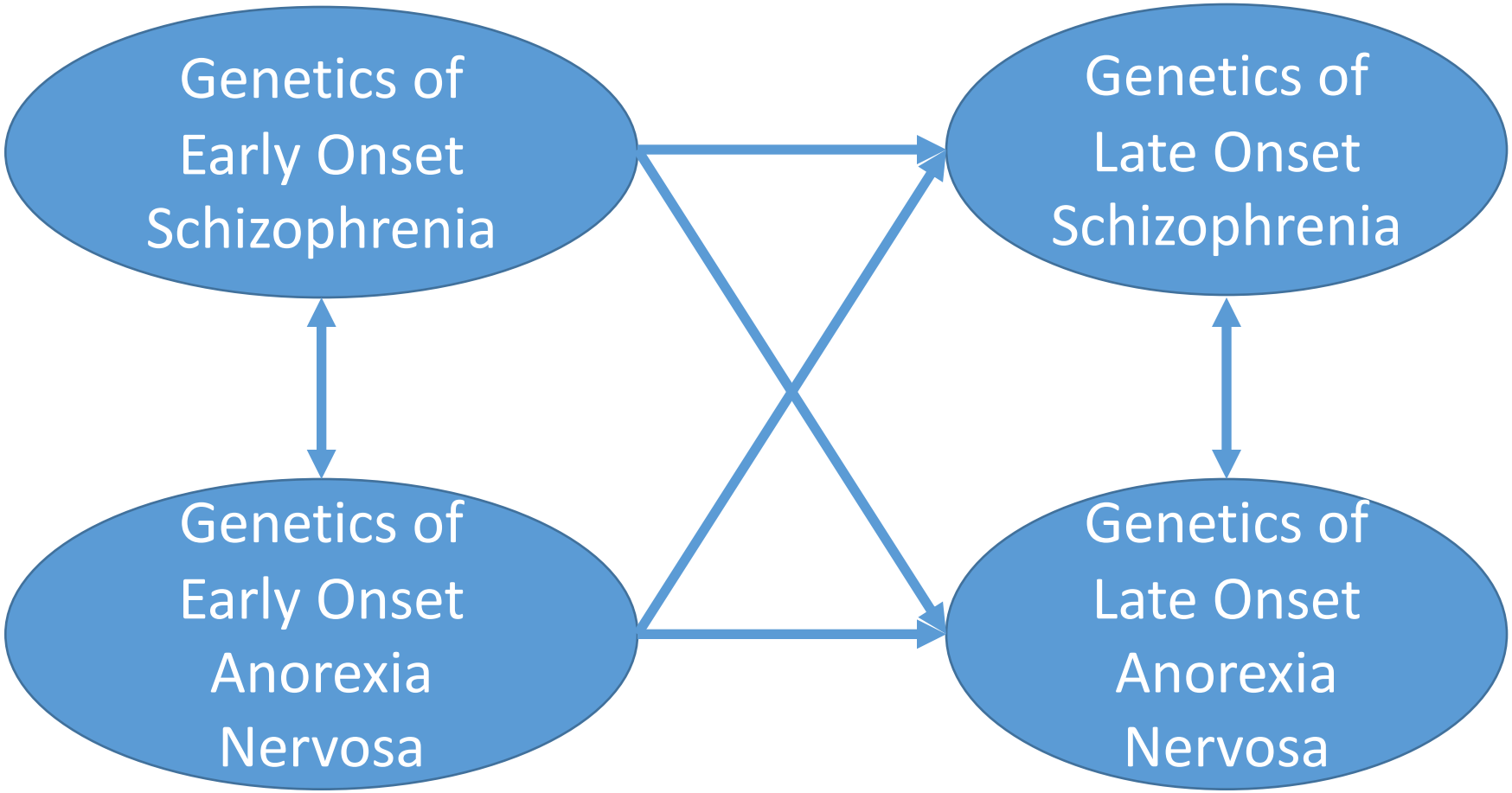

Genetics of Early Onset Schizophrenia
Genetics of Late Onset Schizophrenia
Genetics of Early Onset Anorexia Nervosa
Genetics of Late Onset Anorexia Nervosa


Genetics of
Early Onset
Schizophrenia
Genetics of
Early Onset
Anorexia
Nervosa


Genetics of
Late Onset
Schizophrenia
Genetics of
Late Onset
Anorexia
Nervosa


Genetics of Early Onset Schizophrenia
Genetics of Late Onset Schizophrenia


Genetics of Early Onset Anorexia Nervosa
Genetics of Late Onset Anorexia Nervosa


Genetics of Early Onset Schizophrenia
Genetics of Late Onset Schizophrenia
Genetics of Early Onset Anorexia Nervosa
Genetics of Late Onset Anorexia Nervosa

Psychological Medicine

Genetic heterogeneity in self-reported depressive symptoms identified through genetic analyses of the PHQ-9

Jackson G. Thorp (a1), Andries T. Marees (a1) (a2), Jue-Sheng Ong (a3), Jiyuan An (a3) ...

DOI: https://doi.org/10.1017/S0033291719002526

Published online by Cambridge University Press: 18 September 2019

nature genetics

Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use

Mengzhen Liu, Yu Jiang, [...] Scott Vrieze

Nature Genetics **51**, 237–244 (2019) | Download Citation

Molecular Psychiatry

Genomic prediction of cognitive traits in childhood and adolescence

A. G. Allegrini, S. Selzam, K. Rimfeld, S. von Stumm, J. B. Pingault & R. Plomin

Molecular Psychiatry **24**, 819–827 (2019) | Download Citation

Genetic stratification of depression by neuroticism: revisiting a diagnostic tradition

Psychological Medicine

Mark J. Adams[1], David M. Howard[1,2], Michelle Luciano[3,4], Toni-Kim Clarke[1], Gail Davies[3,4], W. David Hill[3,4], 23andMe Research Team[5], Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium†, Daniel Smith[6], Ian J. Deary[3,4], David J. Porteous[7] and Andrew M. McIntosh[1,3]

Received: 25 April 2019
Revised: 1 August 2019
Accepted: 5 September 2019

Practical

Practical outline

I. Initial considerations
II. Estimating common factor models
III. Estimating user specified model
IV. Estimating multivariate GWAS in Genomic SEM

I. Initial Considerations

Start with GWAS Summary Statistics for the Phenotypes of Interest

- No need for raw data
- No need to conduct a primary GWAS yourself: Download them online!

Example of the top of a summary statistics file

CHR	SNP	BP	A1	A2	INFO	OR	SE	P	Nca	Nco	MAF
8	rs62513865	101592213	T	C	0.957	1.01461	0.0153	0.3438	59851	113154	0.07330
8	rs79643588	106973048	A	G	0.999	1.02122	0.0136	0.1231	59851	113154	0.09200
8	rs17396518	108690829	T	G	0.980	1.00331	0.0080	0.6821	59851	113154	0.43500
8	rs6994300	102569817	A	G	0.466	0.88126	0.4243	0.7658	16823	25632	0.00556
8	rs138449472	108580746	A	G	0.734	0.97181	0.0598	0.6320	41253	79756	0.00852
8	rs983166	108681675	A	C	0.991	0.99144	0.0080	0.2784	59851	113154	0.43200

Where to get summary statistics

- List lots of resources on the Genomic SEM Wiki: https://github.com/MichelNivard/GenomicSEM/wiki/2.-Important-resources-and-key-information

What you need to know about GWAS before you get started

1. A genome wide association study (GWAS) boils down to a linear regression of a phenotype (y) on a genetic variant, usually a single nucleotide polymorphism (x). This regression results in a parameter estimate (beta), test statistic (Z or t) for each SNP, and information that can be used to determine with respect to which allele the effect size is computed. When available for a considerable portion of all SNPs, this information is sufficient to compute the heritability of the traits and genetic correlation between traits. This information is also sufficient to fit structural equation models to the genetic covariance between several traits.

2. You need the full or very lightly cleaned summary statistics generated from a GWAS, so if the authors provide summary statistics only for the top 5.000 SNPs, or even the top 100.000 "pruned" SNPs this is not sufficient. Often if you get in touch with the authors, they have a mechanism for you to obtain the full summary statistics. Sometimes this may involve you agreeing not to identify the participants in their study. Sometimes you may need to sign some documents.

3. You need to know whether the GWAS was a logistic regression, or a linear regression. Note that not all case/control studies use logistic regression. This is because logistic regression can be computationally prohibitive if sample sizes are huge. When a dichotomous outcome (e.g. a case/control trait) is analyzed using a linear regression, this is called a "linear probability model" and it is strictly speaking misspecified. The function `sumstats` does know how to deal with this scenario, and please see the package help for instructions. The package also can deal with a GWAS of a continuous trait being analyzed using linear regression (use the `OLS` flag in `sumstats` to indicate which GWAS are of continuous traits), or a case/control traits analyzed using logistic regression (the default in `sumstats`). Another issue is the use of "linear mixed models" (LMM) in GWAS. These models are used to guard against populations stratification, and

Where to get GWAS summary statistics.

Below is a brief, and incomplete list of links to consortia data pages, where summary statistics are available.

1. The PGC (Psychiatric Genomics Consortium), has analyzed all common DSM-IV axis-I psychiatric disorders (MDD, Schizophrenia, ADHD, OCD, Bipolar Disorder and more)

2. The SSGAC (Social Sciences Genetic Association Consortium) performs genome wide association studies of a variety of social and psychological traits like education, personality, and reproductive behavior.

3. The Nealelab quickly ran and published online GWAS of >4000 traits that were measured as part of the UK Biobank. These traits include many disease (ICD-10 diagnostic codes, both self reported and based on hospital data), social traits (e.g. social deprivation), personality traits (e.g. neuroticism), cognition (e.g. memory) and many more (from snoring to the propensity to drive to fast). The Nealelab ran these GWAS very quickly and as a service to the field. Their GWAS of case/control traits use linear regression (linear probability model). Please read their extensive **read me** which describes their GWAS analysis in detail.

4. The CCACE (Centre for Cognitive Ageing and Cognitive Epidemiology has published GWAS on assorted personality traits, cognitive traits, and tiredness.

5. Members of the CTGlab (Complex Trait Genetics Lab) published several high quality GWAS on IQ, insomnia and other traits.

6. The GPC (Genetics of Personality Consortium) published several, slightly dated, GWAS on the "Big 5" personality scales.

7. The EGG (Early Growth Genetics) Consortium performs GWAS of traits related to early growth.

8. The GIANT consortium publishes GWAS, mainly about antropomorpic traits.

9. The ENIGMA consortium which has published GWAS of subcortical brain volumes and hippocampal volumes.

Things to know before getting started

1. Be sure you are using summary statistics calculated within a single ethnic population
   - Example: PTSD on PGC web-site
2. Be sure to use LD scores that match the ethnic population in sum stats
3. Typically advisable to only include summary statistics from a GWAS with N >= 10,000

PTSD	Download All
All participants	Download AA
African Americans (AA)	Download EA
European Ancestry (EA)	Download FAA
Female African American(FAA)	Download FEA
Female European (FEA)	Download FTE
Female Trans Ethnic (FTE)	Download MAA
Male African American (MAA)	Download MEA
Male European Ancestry (MEA)	Download MTE
Male Trans Ethnic (MTE)	

Things to know before getting started

4. GenomicSEM allows for varying and unknown degrees of sample overlap
- The user does *not* need to know the specific levels of overlap

5. Multivariate GWAS in Genomic SEM uses listwise deletion
- If certain summary statistics have low genomic coverage this will affect the number of SNPs available for all included traits

6. Make sure you are not using a pruned list of summary statistics (e.g., the top 5,000 hits)

Things to know before getting started

7. Both the `munge` and `sumstats` functions in GenomicSEM use sample size to perform necessary conversions. Sample size from summary statistics file or provided by the user.

In order to produce accurate results, this should be the **total** sample size for all included traits.

Be wary of:

a. Summary statistics that report the effective samples
b. Publicly available summary statistics that exclude certain cohorts (e.g., 23andMe).

II. Estimating Common Factor Models in Genomic SEM

Three Primary Steps

1. Munge the summary statistics (*munge*)

2. Run LD-Score Regression to obtain the genetic covariance and sampling covariance matrices (*ldsc*)

3. Run the model (*commonfactor*)

Munge: convert raw data from one form to another

Lab

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

Step 1: *munge* example code (**done for you**)

```
#STEP 1: MUNGE THE FILES.
#Takes four necessary arguments:
#files = the name of the summary statistics files
files<-c("SCZ_HM3.txt", "BIP_HM3.txt", "MDD_HM3.txt")

#hm3 = the name of the reference file to use for
alligning effects to same ref allele across traits
hm3<-"w_hm3.noMHC.snplist"

#trait.names = names used to create the .sumstats.gz
output files
trait.names<-c("SCZ_HM3","BIP_HM3","MDD_HM3")

#N = total sample size for traits
N<-c(105318, 16731, 173005)

#Run the munge function. This will create three
.sumstats.gz files (e.g., SCZ_HM3.sumstats.gz).
munge(files=files,hm3=hm3, trait.names=trait.names, N=N)
```

Example Munge .log file for bipolar disorder

```
Munging file: BIP_HM3.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 pval column as the P column.
Interpreting the  CEUaf  column as the MAF (minor allele frequency) column.
Merging file: BIP_HM3.txt with the reference file: w_hm3.noMHC.snplist
1063333 rows present in the full BIP_HM3.txt summary statistics file.
0 rows were removed from the BIP_HM3.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 BIP_HM3.txt summary
statistics file. Please ensure this is correct.
203 row(s) were removed from the BIP_HM3.txt summary statistics file due to the effect allele (A1)
column not matching A1 or A2 in the reference file.
1 row(s) were removed from the BIP_HM3.txt summary statistics file due to the other allele (A2)
column not matching A1 or A2 in the reference file.
260291 rows were removed from the BIP_HM3.txt summary statistics file due to INFO values below the
designated threshold of 0.9
44230 rows were removed from the BIP_HM3.txt summary statistics file due to missing MAF
information or MAFs below the designated threshold of 0.01
758608 SNPs are left in the summary statistics file BIP_HM3.txt after QC.
I am done munging file: BIP_HM3.txt
The file is saved as BIP_HM3.sumstats.gz in the current working directory.
```

Step 2: *ldsc* example code (done for you)

```
#STEP 2: RUN LD-SCORE REGRESSION
#Takes four necessary arguments:
#traits = the name of the .sumstats.gz traits
traits<-c("SCZ_HM3.sumstats.gz", "BIP_HM3.sumstats.gz",
"MDD_HM3.sumstats.gz")

#sample.prev = the proportion of cases in sum stats. For
quantitative traits list NA
sample.prev <- c(.39,.45,.35)

#population.prev = lifetime prevalence of the traits
(pull from existing literature)
population.prev <- c(.01,.01,.16)
```

Step 2: *ldsc* example code

```
#ld = folder of LD scores used as predictors in ldsc
ld <- "eur_w_ld_chr/"

#wld = folder of LD scores used as weights in ldsc
(almost always same file as ld)
wld <- "eur_w_ld_chr/"

#trait.names = optional fifth argument to list trait names
trait.names<-c("SCZ", "BIP", "MDD")

#Run the ldsc function
PSYCH_COV<-ldsc(traits=traits, sample.prev=sample.prev,
population.prev=population.prev, ld=ld, wld=wld,
trait.names=trait.names)
```

Populated with ld scores from the same ancestry

Set working directory and load in data!

```
#load in the package
require(GenomicSEM)

#load in the example ldsc objects that we will use for the practical
setwd("")

load("GenomicSEMPractical.RData")
```

Will likely print 24 warnings about replacing previous imports: OK TO IGNORE

Step 3: *commonfactor* example code

```
#STEP 3: ESTIMATE THE COMMON FACTOR MODEL
#requires only one necessary argument:
#covstruc = the output from the ldsc function
covstruc<-PSYCH_COV

#an optional second argument can be provided for the
estimation method
estimation<-"DWLS"

#run the commonfactor model below
PFactor <- commonfactor(covstruc=covstruc,estimation=
estimation)

#Print PFactor results
PFactor$results
```

Pfactor$results

lhs	op	rhs	Unstandardized_Estimate	Unstandardized_SE	Standardized_Est	Standardized_SE
F1	=~	SCZ	0.48075155	0.051804457	0.96942110	0.10446213
F1	=~	BIP	0.38467648	0.040843715	0.74818767	0.07944017
F1	=~	MDD	0.12139721	0.015659093	0.38181492	0.04925052
SCZ	~~	SCZ	0.01481074	0.049701537	0.06022274	0.20209397
BIP	~~	BIP	0.11636844	0.038587291	0.44021521	0.14597353
MDD	~~	MDD	0.08635352	0.007587214	0.85421737	0.07505346

Parameter being estimated

Estimates and SE for model applied to genetic *covariance* matrix

Estimates and SE for model applied to genetic *correlation matrix*

III. Estimate a User-Specified Model

Three Primary Steps

1. Munge the summary statistics (*munge*)

2. Run LD-Score Regression to obtain the genetic covariance and sampling covariance matrices (*ldsc*)

3. Specify and run the model (*usermodel*)

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

How to specify a model

We use the lavaan formula language, slightly extended:

Regression:

A ~ B

(Co)variance:

A ~~ A; A ~~ B

Factor:

F1 =~ A + B + C + D

Fix a parameter:

A ~~ 1*B (the covariance between A and B is 1)

Name a parameter:

A ~~ a*B (the covariance between A and B = parameter label a)

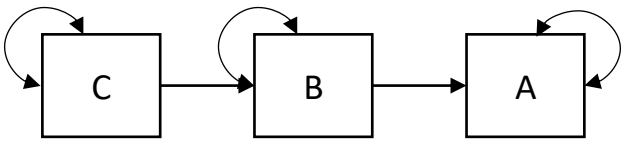
Allows you to use model constraints for this parameter:

a > .001

Lets make that a bit more specific

Model1 <- " A ~ B

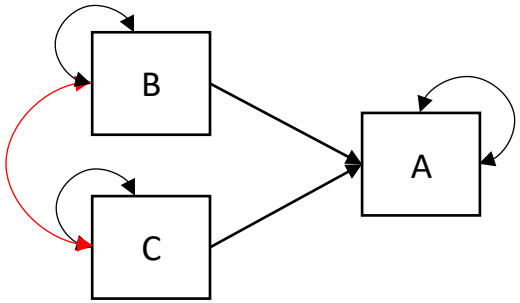
B ~ C"



Model2 <- " A ~ B

A ~ C

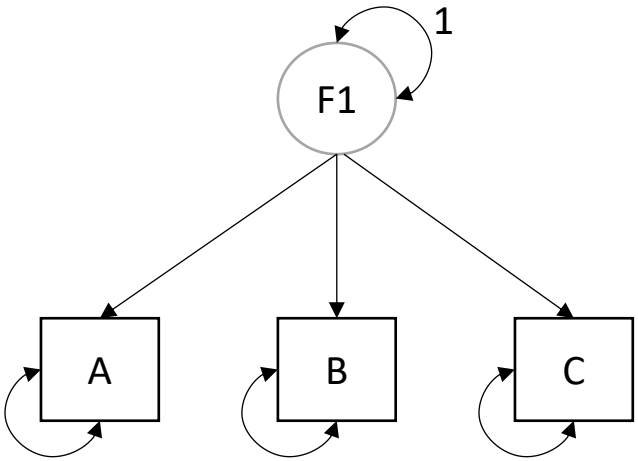
B ~~ C"



Lets make that a bit more specific

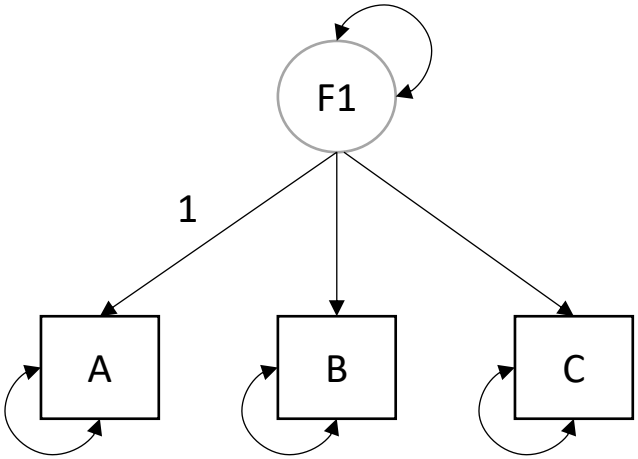
Model3 <- “ F1 =~ NA*A + B + C

F1 ~~ 1*F1”



Lets make that a bit more specific

Model3 <- “ F1 =~ 1*A + B + C”

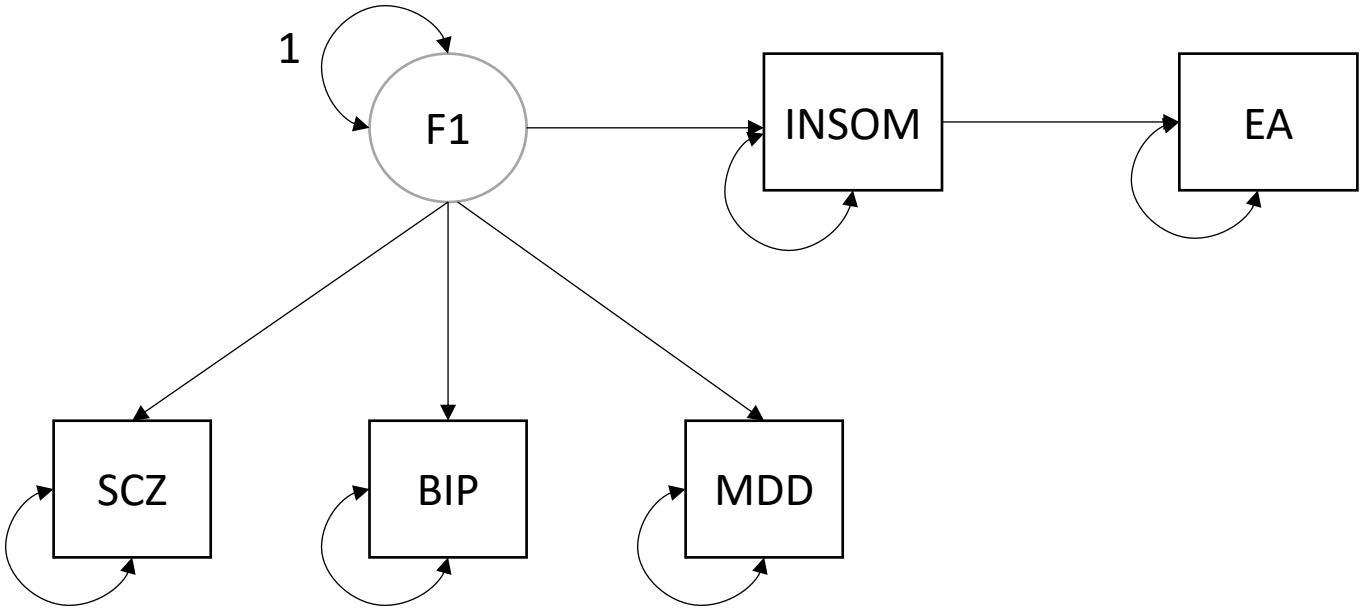


Lab

- **Used 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
  - Educational Attainment (Lee et al., 2019); *N* = 766,035
  - Insomnia (Jansen et al., 2019); *N* = 386,533

My preregistration

```
MY.model<-"F1=~NA*SCZ+BIP+MDD
F1~~1*F1
INSOM~F1
EA~INSOM"
```



Specify Arguments

```
#STEP 3: SPECIFY AND RUN USER MODEL
#Takes two necessary arguments:
#1. covstruc = the output from multivariable ldsc
#in this example = ldsc results for Schizophrenia, Bipolar, MDD, EA, and Insomnia
covstruc<-PRAC_COV

#2. model = the user specified model
MY.model<-"F1=~NA*SCZ+BIP+MDD
F1~~1*F1
INSOM~F1
EA~INSOM"

#estimation = an optional third argument specifying the estimation method to use
estimation<-"DWLS"

#std.lv = optional fourth argument specifying whether variances of latent variables should be set to 1
std.lv=FALSE

#Run your model
YourModel<- usermodel(covstruc=covstruc, model=MY.model,estimation=estimation,std.lv=std.lv)
```

YourModel$results

Parameter being estimated

Estimates and SE for model applied to genetic *covariance* matrix

Estimates and SE for model applied to genetic *correlation matrix* matrix

Fully standardized estimates

lhs	op	rhs	Unstand_Est	Unstand_SE	STD_Genotype	STD_Genotype_SE	STD_All
F1	=~	SCZ	0.47126308	0.049303074480831	0.7915818	0.0612497026504305	0.7915817
F1	=~	BIP	0.38395469	0.0402934507361647	0.8222514	0.0651918967406813	0.8222512
F1	=~	MDD	0.12622644	0.0153598110078367	0.5023109	0.0475709212458088	0.5023109
F1	~~	F1	1.00000000		1.0000000		1.0000000
SCZ	~~	SCZ	0.02232836	0.0471686012314039	0.3733984	0.100780802429095	0.3733984
BIP	~~	BIP	0.11533845	0.0382930314300862	0.3239031	0.134336803656722	0.3239029
MDD	~~	MDD	0.08482525	0.00768672483793622	0.7476836	0.0824663432925547	0.7476837
EA	~~	EA	0.10068370	0.00235251255367696	0.8959785	0.0440255986189161	0.8959782
EA	~	INSOM	-0.50012902	0.0388279618666308	-0.3220857	0.0204667735439688	-0.3225241
INSOM	~~	INSOM	0.04648542	0.00300630225091781	0.9918926	0.0268569076127303	0.9891972
INSOM	~	F1	0.00904497	0.00519588588484769	0.1040779	0.0192029404218759	0.1039364

YourModel$modelfit

chisq	df	p_chisq	AIC	CFI	SRMR
186.8647	5	1.827715e-38	206.8647	0.800525	0.141147

- **chisq:** The model chi-square, reflecting index of exact fit to observed data, with lower values indicating better fit.
  - **df and p_chisq:** The degrees of freedom and p-value for the model chi-square.
- **AIC:** Akaike Information Criterion. Can be used to compare models regardless of whether they are nested.
- **CFI:** Comparative Fit Index. Higher = better. > .90 = acceptable fit; > .95 = good model fit
- **SRMR:** Standardized Room Mean Square Residual. Lower = better. < .10 = acceptable fit; < .05 = good fit

Delete Input for MY.model and run your own!

PRACTICAL: You Take Control

- As away of preregistering them, write your model down on paper

- Remember five variable names are:

  SCZ, BIP, MDD, EA, INSOM

IV. Multivariate GWAS in Genomic SEM

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

Lab

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

- Pre-subset summary statistics downloaded online to 100 HapMap3 SNPs
  - Not necessary (inadvisable) in practice; pragmatic just for workshop

Step 3: *sumstats* example code

```
#STEP 3: PREPARE SUMMARY STATISTICS FOR MULTIVARIATE GWAS
#Takes four necessary arguments:
#1. files = the name of the summary statistics file
##**note that these are drastically reduced subsets of SNPs for the practical only
files<-c("SCZ_100.txt", "BIP_100.txt", "MDD_100.txt")

#2. ref = the name of the reference file used to obtain SNP MAF
#**note again that this is a drastically reduced subset of SNPs
ref="reference.1000G.subset.txt"

#3. trait.names = the name of the files to be used in
trait.names=c("SCZ","BIP","MDD")

#4. se.logit = whether the standard errors are on an logistic scale
se.logit<-c(T,T,T)

#run the sumstats function below
p_sumstats <- sumstats(files=files,ref=ref ,trait.names=trait.names,se.logit=se.logit)
```

Example sumstats .log file

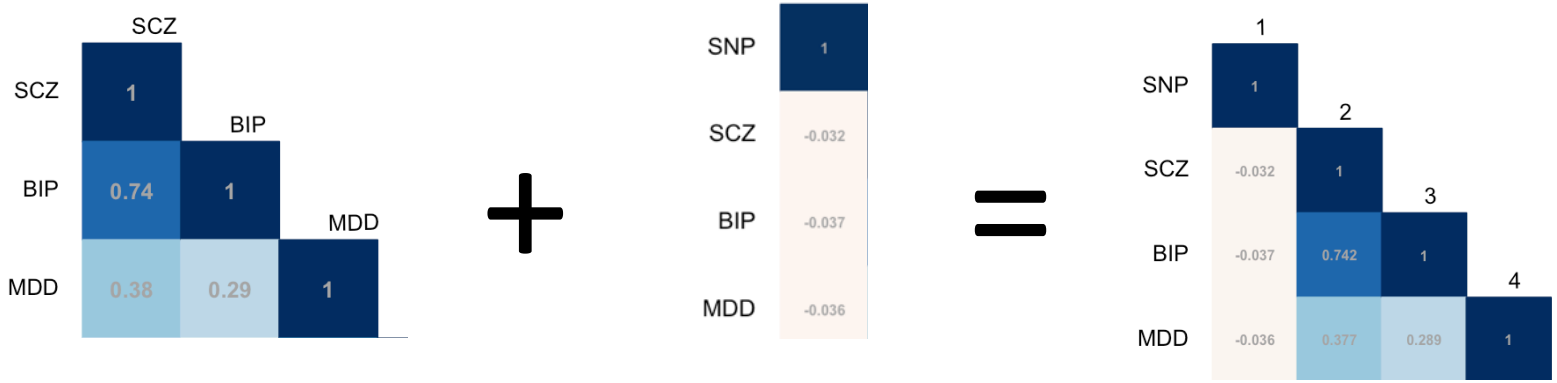
```
The preparation of 3 summary statistics for use in Genomic SEM began at: 2020-03-01 19:22:24
Reading in reference file
Applying MAF filer of 0.01 to the refernece file.
Reading summary statistics for SCZ_100.txt BIP_100.txt MDD_100.txt . Please note that this step
usually takes a few minutes due to the size of summary statistic files.
All files loaded into R!


Preparing summary statistics for file: SCZ_100.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 SE column as the SE column.
Interpreting the P column as the P column.
Merging file: SCZ_100.txtwith the reference file: reference.1000G.subset.txt
100 rows present in the full SCZ_100.txt summary statistics file.
4 rows were removed from the SCZ_100.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 SCZ_100.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.
No INFO column, cannot filter on INFO, which may influence results
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: SCZ_100.txt
96 SNPs are left in the summary statistics file SCZ_100.txt after QC and merging with the
reference file.
```

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

Step 2: Run `ldsc`



Step 4a: *commonfactorGWAS* example code

```
#STEP 4a: RUN THE MULTIVARIATE GWAS
#commonfactorGWAS takes only two necessary arguments
#1. covstruc = the output from the ldsc function
covstruc<-PSYCH_COV

#2. SNPs = output from sumstats function
SNPs<-p_sumstats

#3. estimation = optional third argument specifying estimation method to be used
estimation<-"DWLS"

#4. parallel = optional argument specifying whether it should be run in parallel
#set to FALSE here just for the practical
parallel<-FALSE

#5. SNPSE = optional argument specifying level of SNPSE
SNPSE<-.005

#run the multivariate GWAS below
pfactor_GWAS<-commonfactorGWAS(covstruc=covstruc, SNPs=SNPs, estimation = estimation,parallel=parallel,SNPSE=SNPSE)
```

- To save memory, saves only the effect of the SNP on the common factor

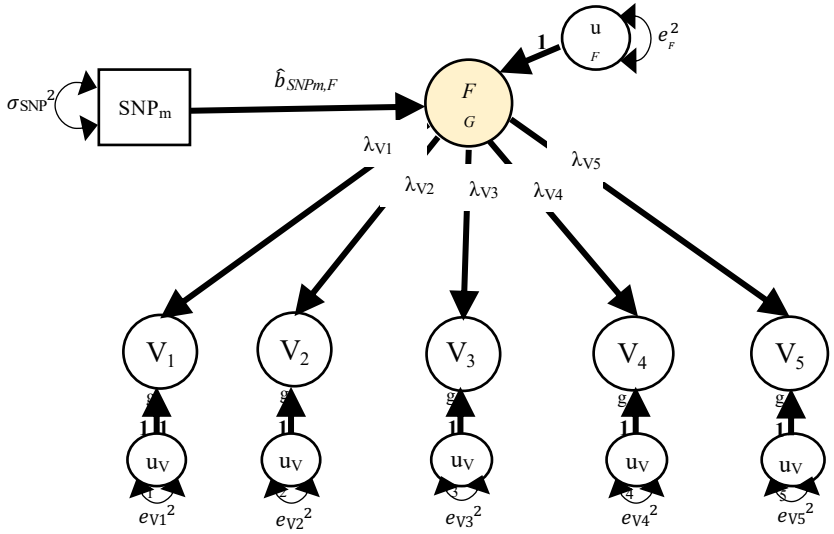
First five rows of the output

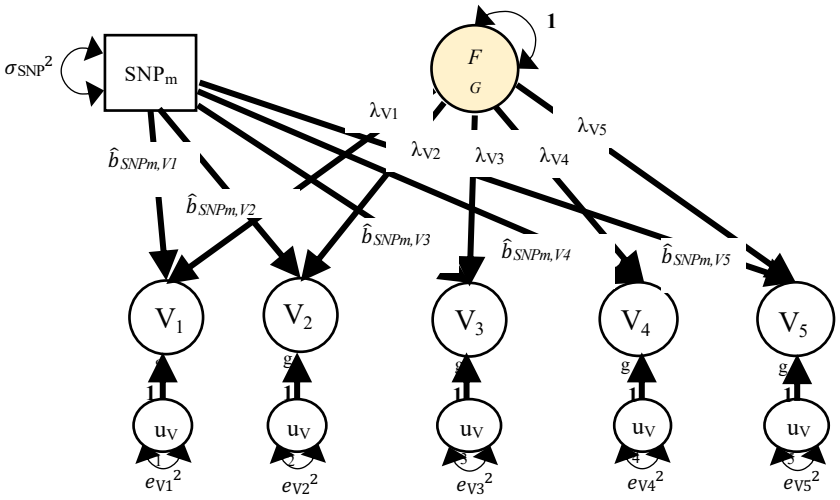
SNP	CHR	BP	MAF	A1	A2	i	lhs	op	rhs	est	se_c	Z_Estimate
rs1000073	1	157255396	0.4165010	A	G	1	F1	~	SNP	4.647717e-05	0.005422914	0.008570516
rs1000050	1	162736463	0.1471170	C	T	2	F1	~	SNP	3.241612e-03	0.007496856	0.432396172
rs1000053	2	12790328	0.0904573	C	T	3	F1	~	SNP	-1.541138e-03	0.009178166	-0.167913549
rs1000016	2	235690982	0.0815109	A	G	4	F1	~	SNP	-2.282467e-04	0.009616981	-0.023733716
rs1000017	2	235691089	0.4671970	C	A	5	F1	~	SNP	2.508369e-04	0.005269484	0.047601800

Pval_Estimate	Q	Q_df	Q_pval	fail	warning
0.9931618	0.3635681	2	0.8337814	0	0
0.6654535	1.1670896	2	0.5579172	0	0
0.8666513	0.7315481	2	0.6936595	0	0
0.9810650	2.6301119	2	0.2684593	0	0
0.9620336	0.3491709	2	0.8398051	0	0

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





Troubleshooting

```
##look at fail messages (0 = good to go)
table(pfactor_GWAS$fail)

#look at warning messages
table(pfactor_GWAS$warning)
```

Step 4b: *userGWAS* example code

```
#STEP 4b: RUN A USER SPECIFIED MULTIVARIATE GWAS
#userGWAS takes three necessary arguments:
#1. covstruc = the output from the ldsc function
covstruc<-PSYCH_COV

#2. SNPs = output from sumstats function
SNPs<-p_sumstats

#3. model = the model to be run
#going to troubleshoot estimated ov variances are negative
#by adding model constraint for all residuals to be above 0
model<-"F1=~SCZ+BIP+MDD
F1~SNP
SCZ~~a*SCZ
BIP~~b*BIP
MDD~~c*MDD
a > .001
b > .001
c > .001"
```

Step 4b: *userGWAS* example code

```
#4. modelchi = optional argument whether you want model chi-square for the individual model
#default = FALSE
modelchi<-FALSE

#5. estimation = optional argument specifying estimation method to be used
estimation<-"DWLS"

#6. sub = optional argument specifying component of model output to be saved
sub<-"F1~SNP"

#7. SNPSE = optional argument specifying value of standard error for SNP
SNPSE<-.005

#8. parallel = optional argument specifying whether it should be run in parallel
#set to FALSE here just for the practical
parallel<-FALSE

#run the multivariate GWAS below
pfactor_GWAS2<-userGWAS(covstruc=covstruc, SNPs=SNPs, model=model,modelchi=modelchi,
                        estimation = estimation,sub=sub,SNPSE=SNPSE,parallel=parallel)
```

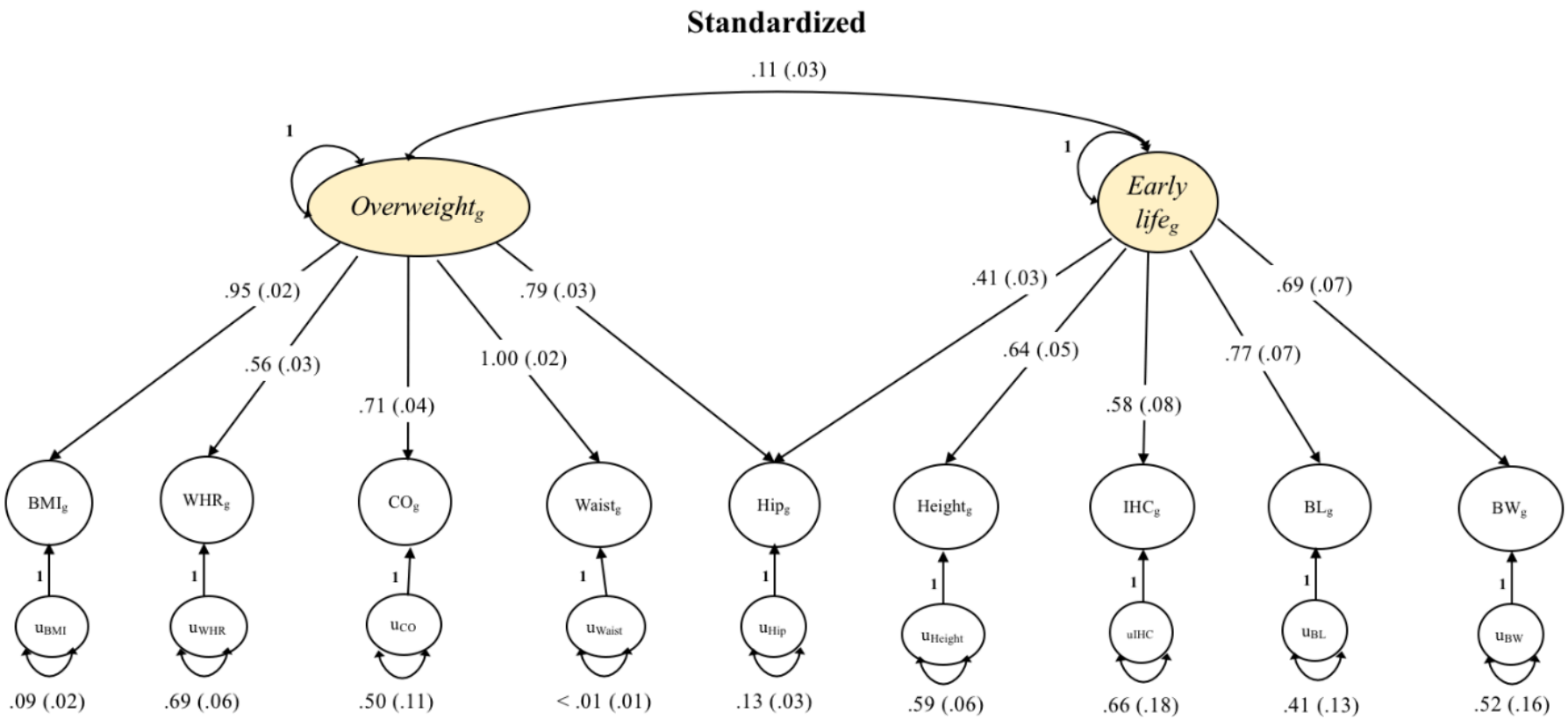
If there's time…
play around with some anthropometric traits

```
###IF theres time
load("Anthro_LDSC.RData")
colnames(anthro$S)

covstruc<-anthro

#2. model = the user specified model
anthro.model<-""

#estimation = an optional third argument specifying the estimation method to use
estimation<-"DWLS"

#std.lv = optional fourth argument specifying whether variances of latent variables should be set to 1
std.lv=FALSE

#Run your model
AnthroModel<- usermodel(covstruc=covstruc, model=anthro.model,estimation=estimation,std.lv=std.lv)
```

Note that you do not need to include all variables in the model

Variable Names

- BMI = Body Mass Index
- WHR = Waist Hip Ratio
- Waist = Waist Circumference
- Hip = Hip circumference
- CO = childhood obesity
- Height = Height
- BL = Birth Length
- BW = Birth Weight
- IHC = Infant Head Circumference


Standardized
.11 (.03)
1
1
Overweight$_g$
Early life$_g$
.95 (.02)
.56 (.03)
.71 (.04)
1.00 (.02)
.79 (.03)
.41 (.03)
.64 (.05)
.58 (.08)
.77 (.07)
.69 (.07)
BMI$_g$
WHR$_g$
CO$_g$
Waist$_g$
Hip$_g$
Height$_g$
IHC$_g$
BL$_g$
BW$_g$
1
u$_{BMI}$
u$_{WHR}$
u$_{CO}$
u$_{Waist}$
u$_{Hip}$
u$_{Height}$
uIHC
u$_{BL}$
u$_{BW}$
.09 (.02)
.69 (.06)
.50 (.11)
< .01 (.01)
.13 (.03)
.59 (.06)
.66 (.18)
.41 (.13)
.52 (.16)

Final Notes

- Parallel 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
- Ideal run-time scenario: split jobs across computing nodes on a cluster and run in-parallel
  - All runs are independent of one another!

Overview

- Genomic SEM is ready for use today!
  - **Ask questions on our google forum**
  - https://groups.google.com/forum/#!forum/genomic-sem-users
- Lots can be done using existing, openly available GWAS summary statistics
- Models are flexible and up to the user
- Use Genomic SEM to derive sumstats for novel phenotypes for use in PGS analyses

Resources

- See paper at: rdcu.be/bvn7t
- See github at: https://github.com/MichelNivard/GenomicSEM
- See tutorials at: https://github.com/MichelNivard/GenomicSEM/wiki

Acknowledgements

- **Elliot M. Tucker-Drob, Michel G. Nivard, Mijke Rhemtulla**
- NIH grants R01HD083613, R01AG054628, R21HD081437, R24HD042849
- Jacobs Foundation
- Royal Netherlands Academy of Science Professor Award PAH/6635
- ZonMw grants 531003014, 849200011
- European Union Seventh Framework Program (FP7/2007-2013) ACTION Project
- MRC grant MR/K026992/1
- AgeUK Disconnected Mind Project