University of Colorado **Boulder**

**Institute for Behavioral Genetics**

# Multivariate Concepts

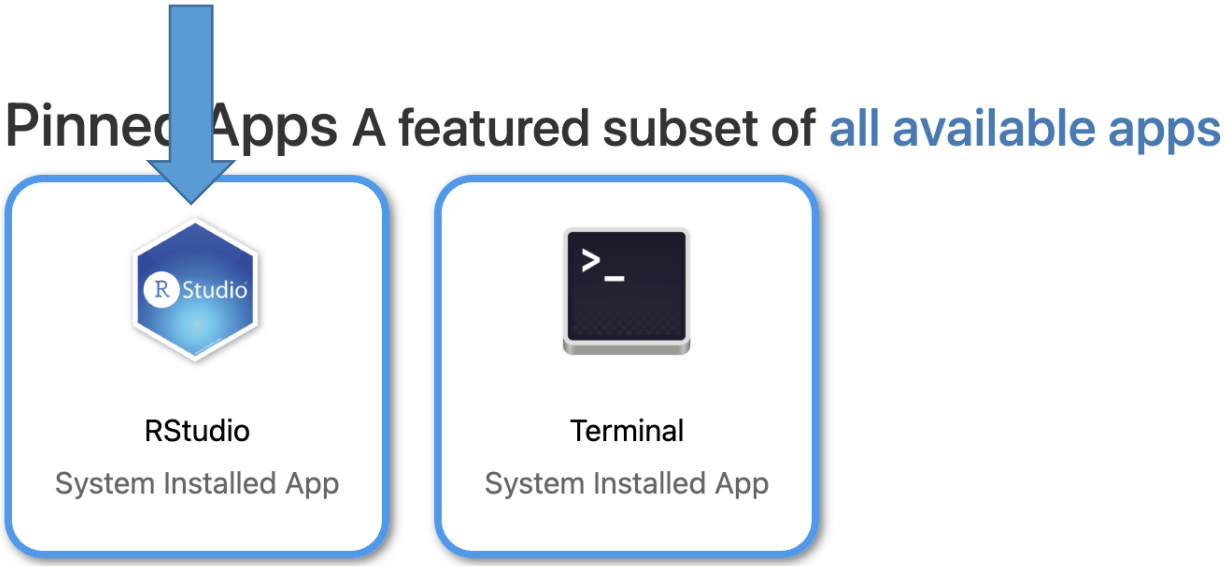
## ISG Workshop 2026

Andrew Grotzinger
June 9th, 2026

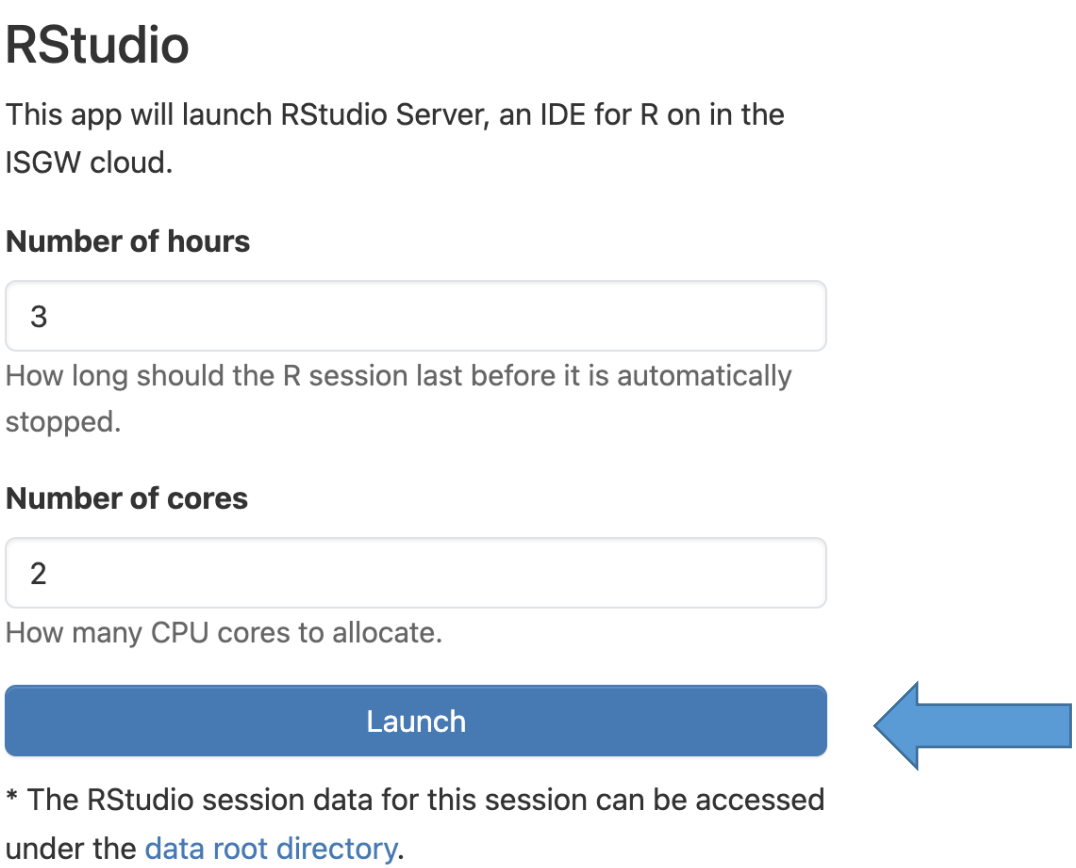
# Let’s start by going to:

https://workshop-ood.colorado.edu

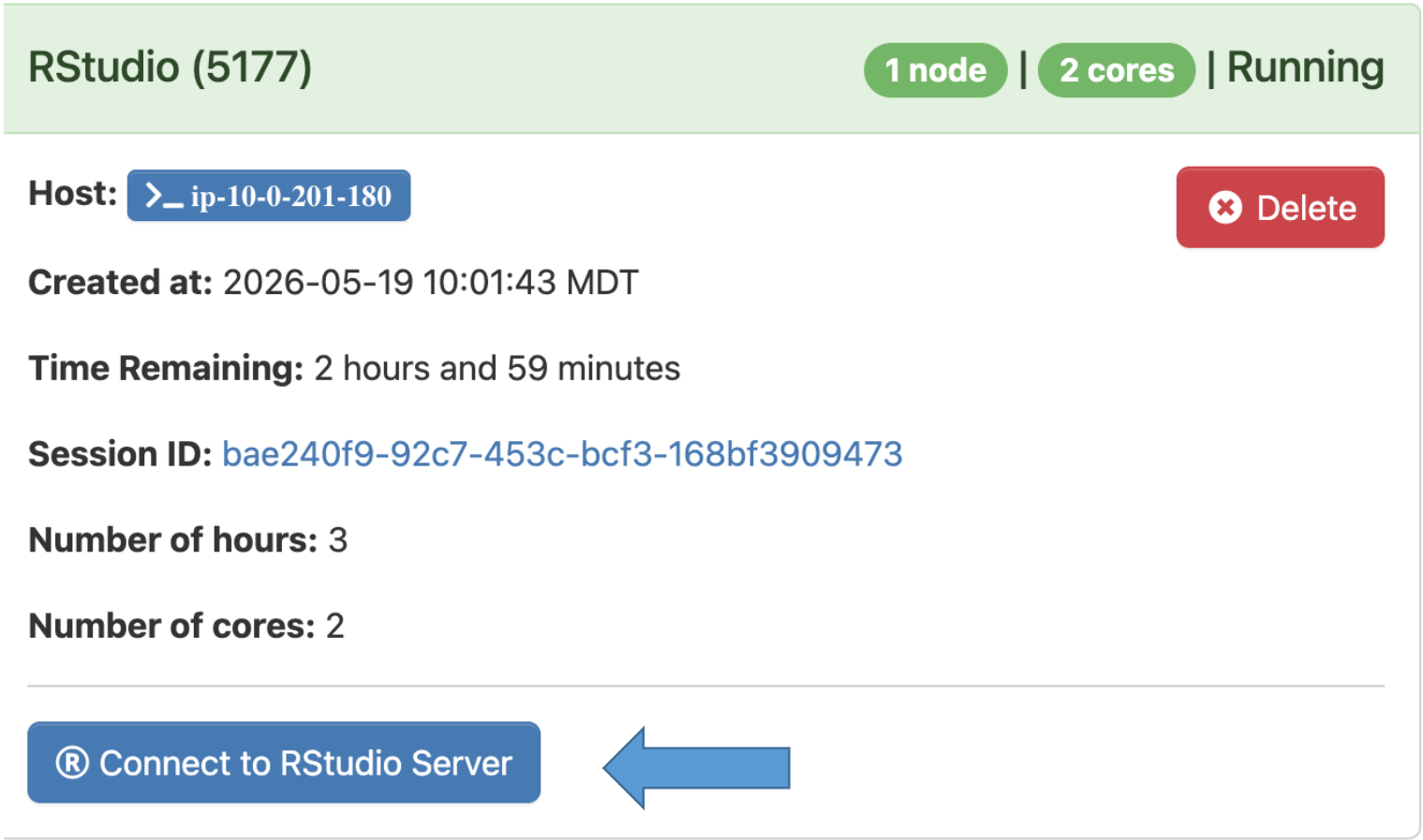
Then launch RStudio



# Request 3 hours with 2 cores

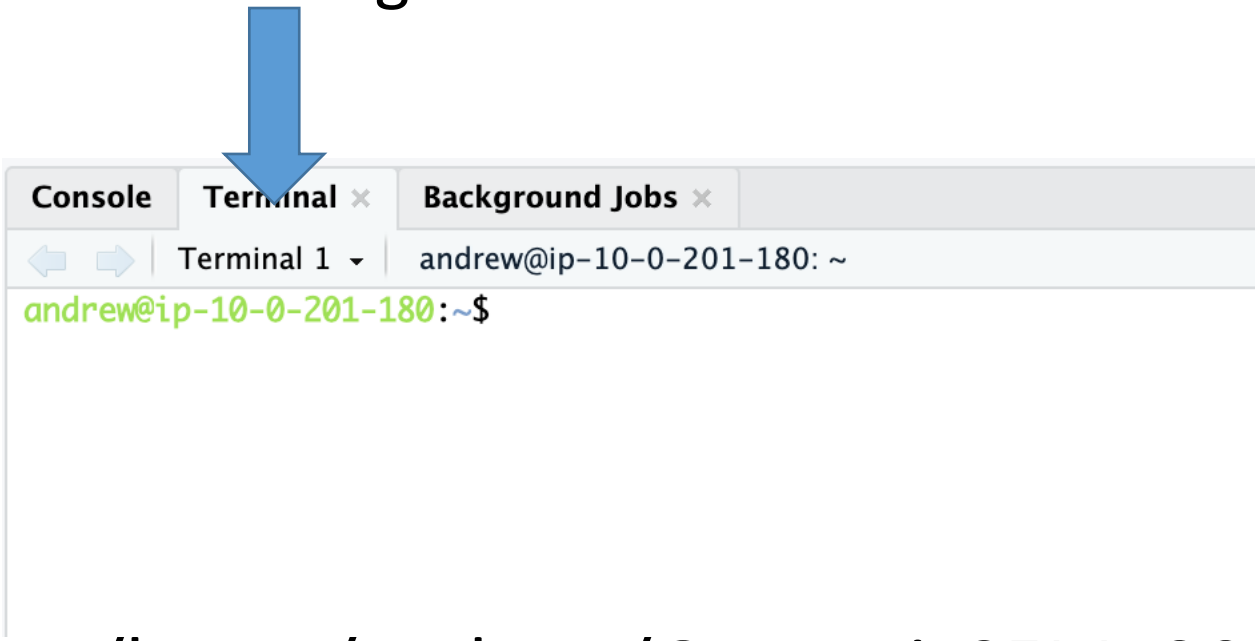


# Connect to Rstudio Server



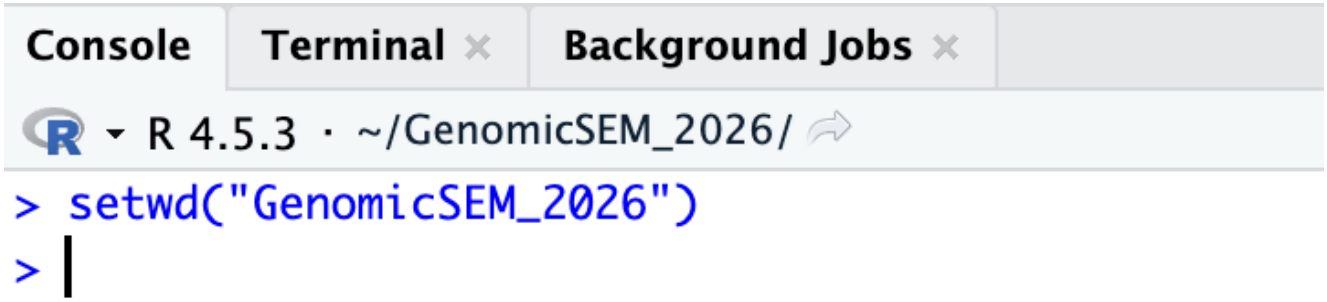
# Now copy over the materials

And clicking on the terminal tab.

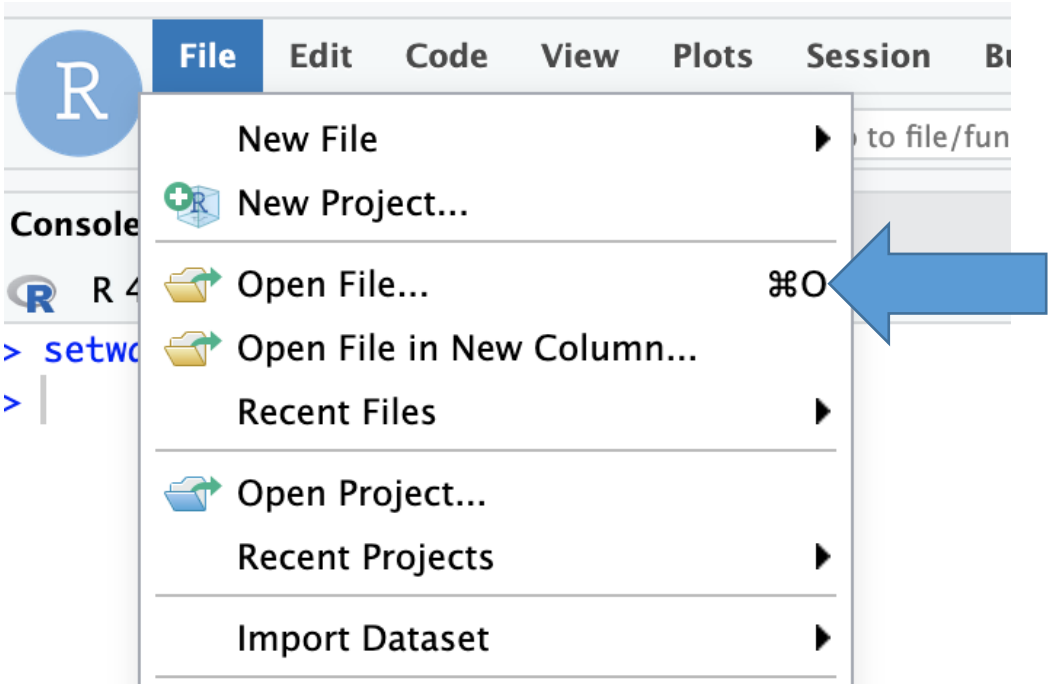



```
cp -r /home/andrew/GenomicSEM_2026  .

cd GenomicSEM_2026/
```

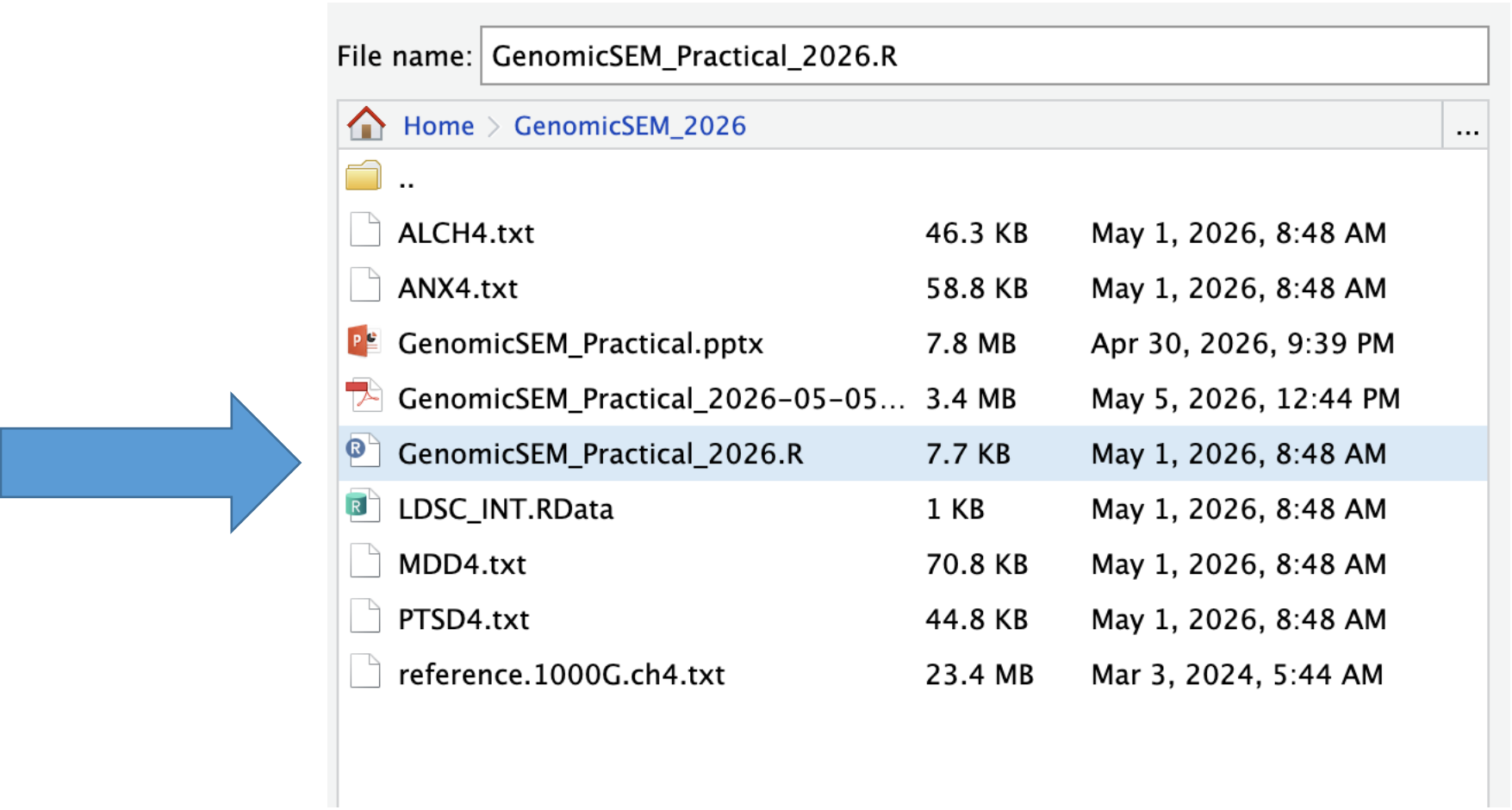
# Now let's go over to the console and *setwd* for this new folder you just copied



# Now let's go to File at the top and open the .R script with the commands we will be running



# Open the R script "GenomicSEM_Practical_2026.R"



# Line 2 of R code: the Qualtrics link

https://qimr.az1.qualtrics.com/jfe/form/SV_6fjh8RI0EWpqhoi

# Practical outline

I. Background and Sales Pitch

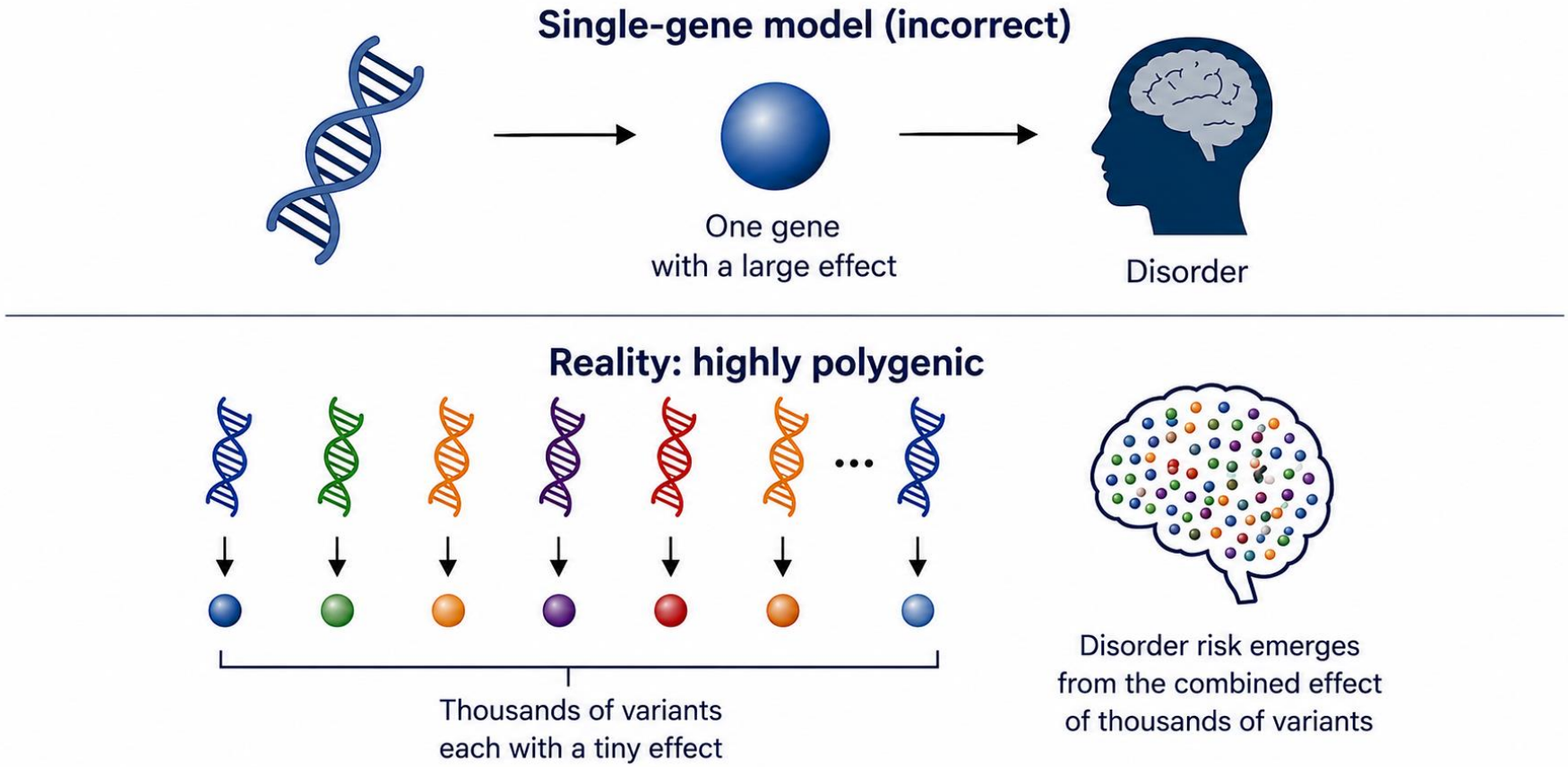
II. Estimating genome-wide models

III. Estimating multivariate GWAS in Genomic SEM

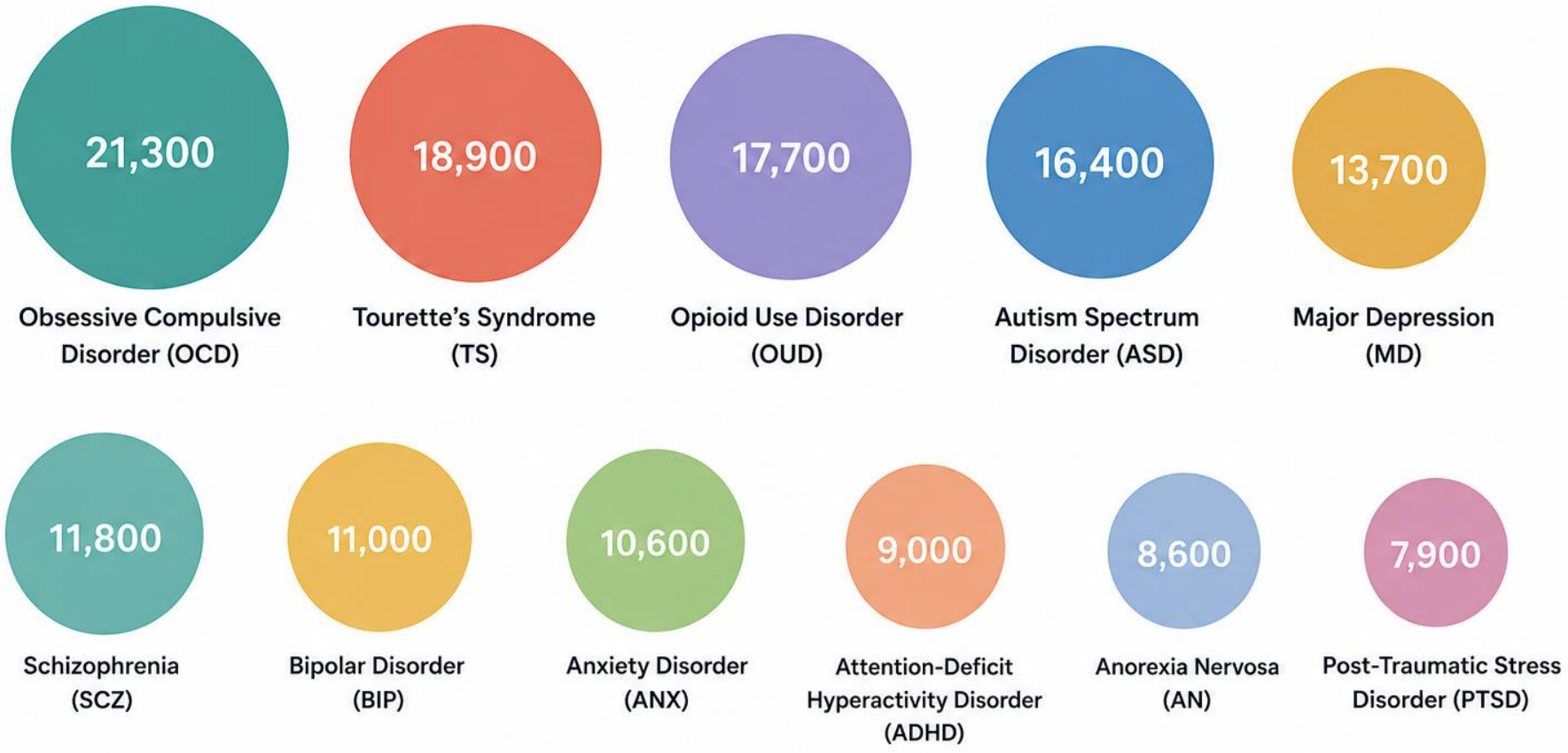
IV. More practice with genome-wide models

# I. Background and Sales Pitch

# Psychiatric Risk: More than one gene

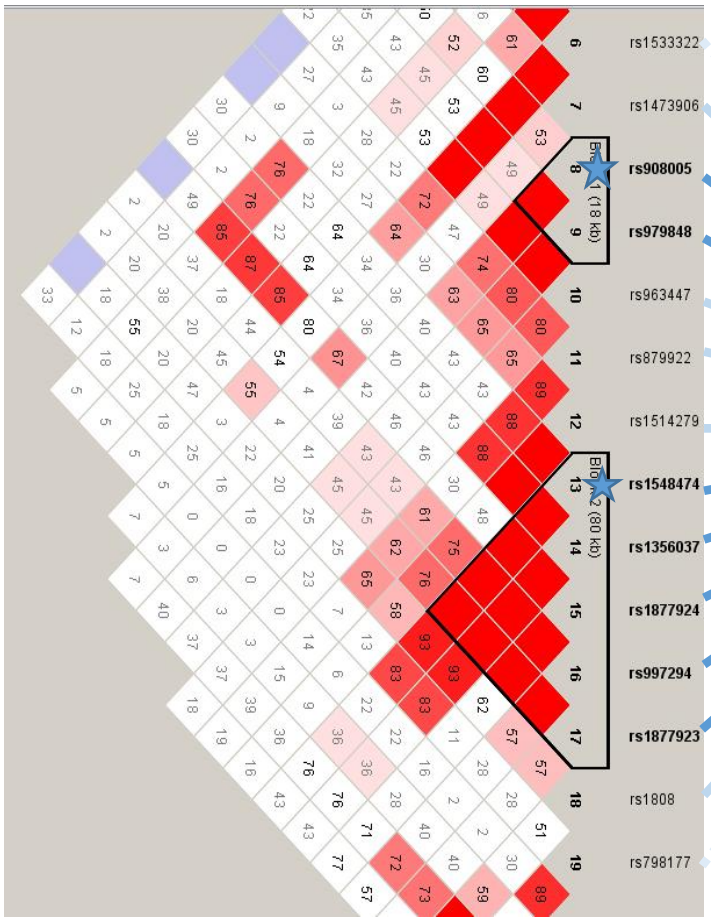


# Estimated Number of Genetic Variants Affecting Psychiatric Risk




Human complex traits are highly polygenic.
This necessitates statistical methods for mapping genetic overlap

**Toy Example:** 2 True Causal Variants



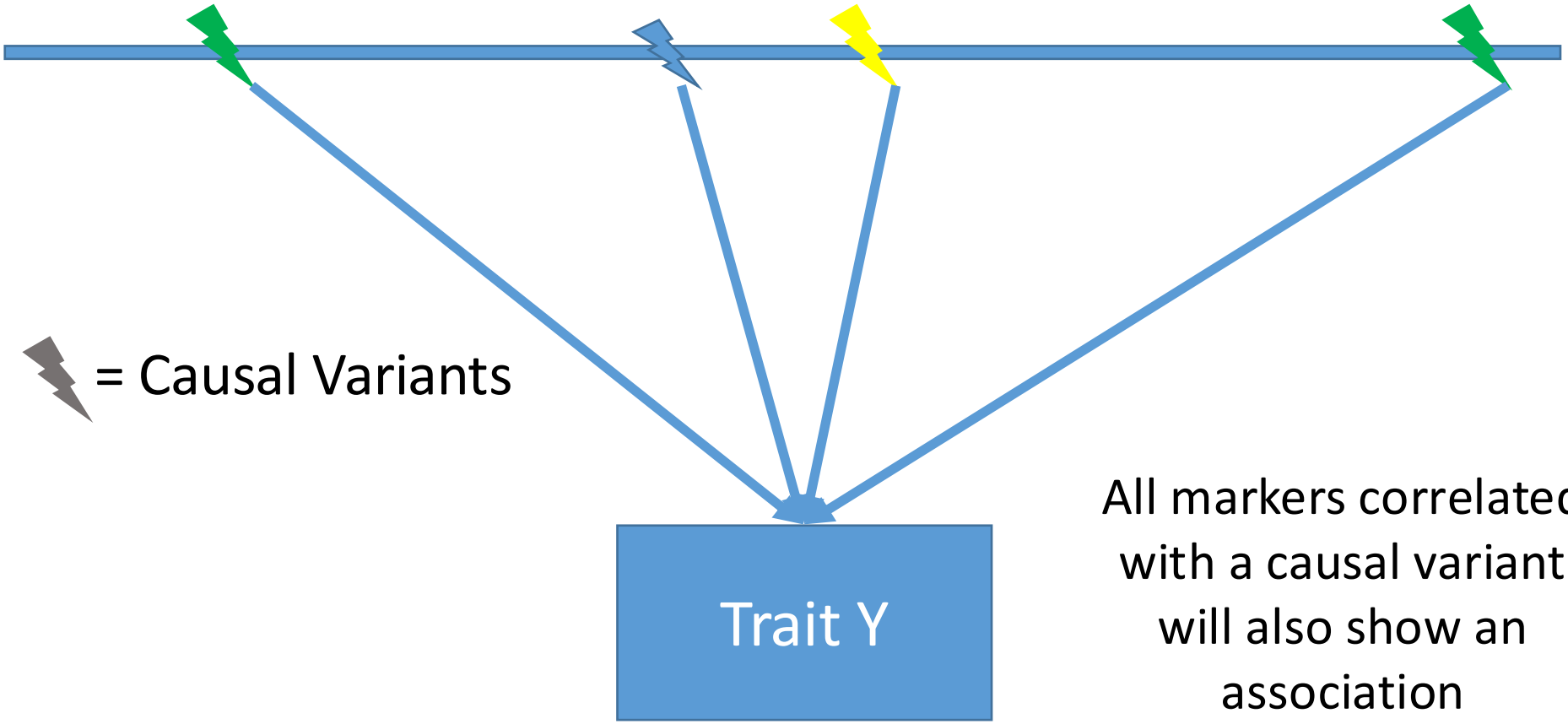

Phenotype Y

$$\ell_j = \sum_i r_{ij}^2$$

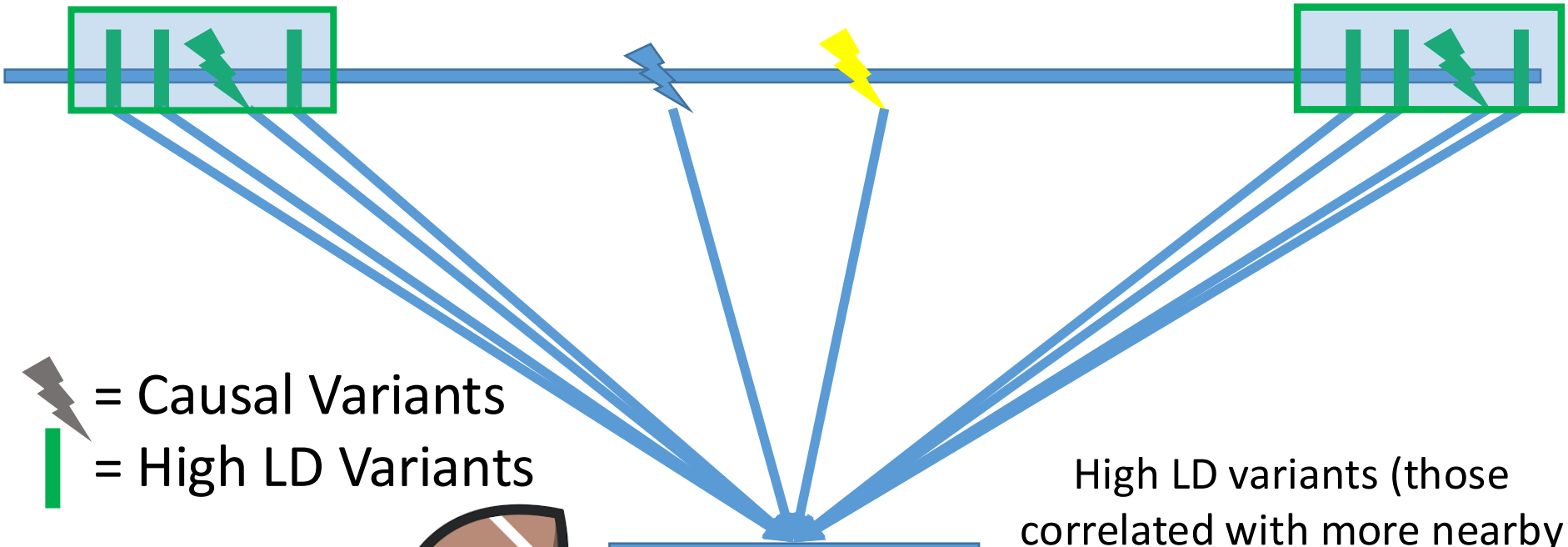
Imagine that there are two causal variants in the population.

All variants in LD with those variants are going to be estimated as having an affect on the phenotype, Y

The LD-score of a variant is the sum of $r^2$ across SNPs.




All markers correlated with a causal variant will also show an association

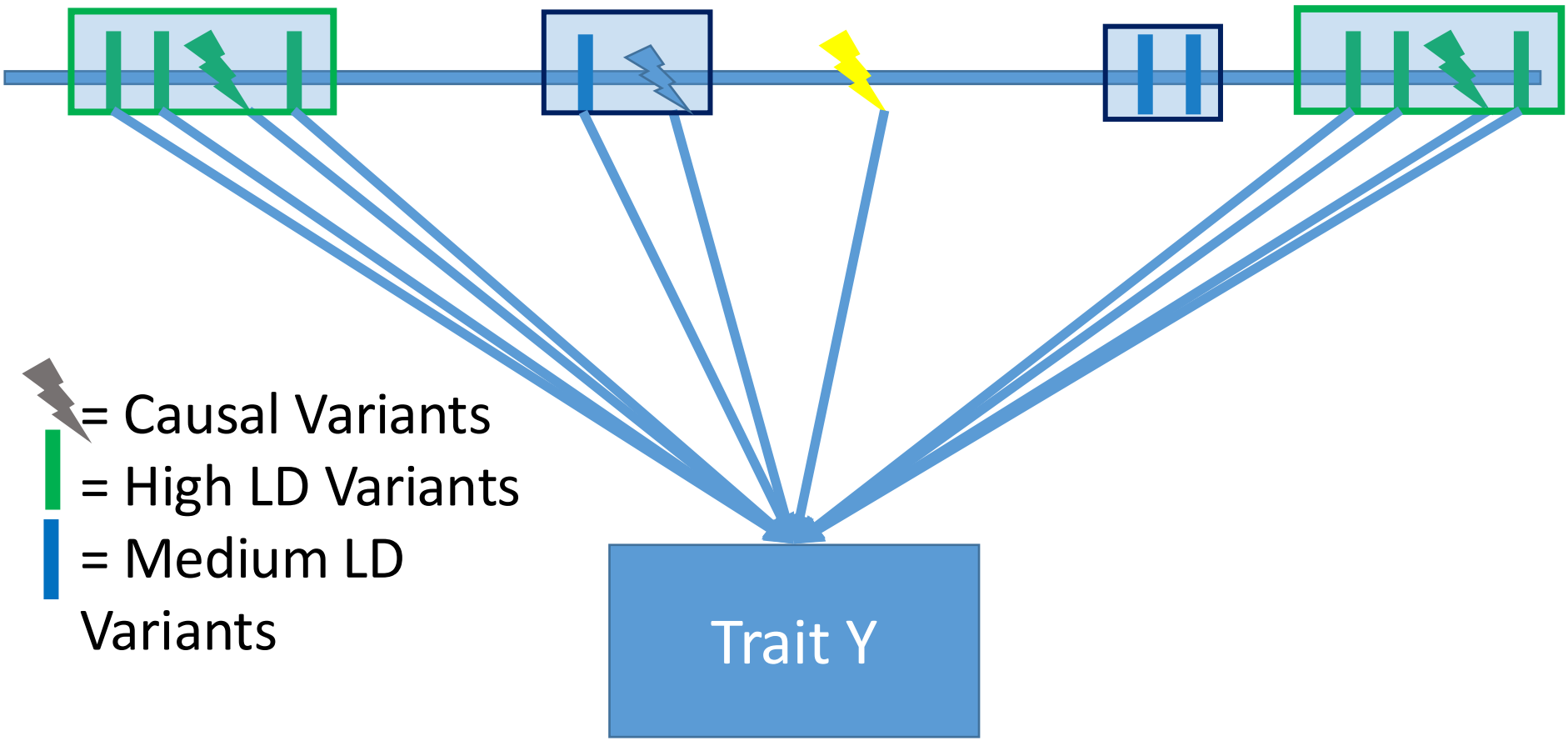


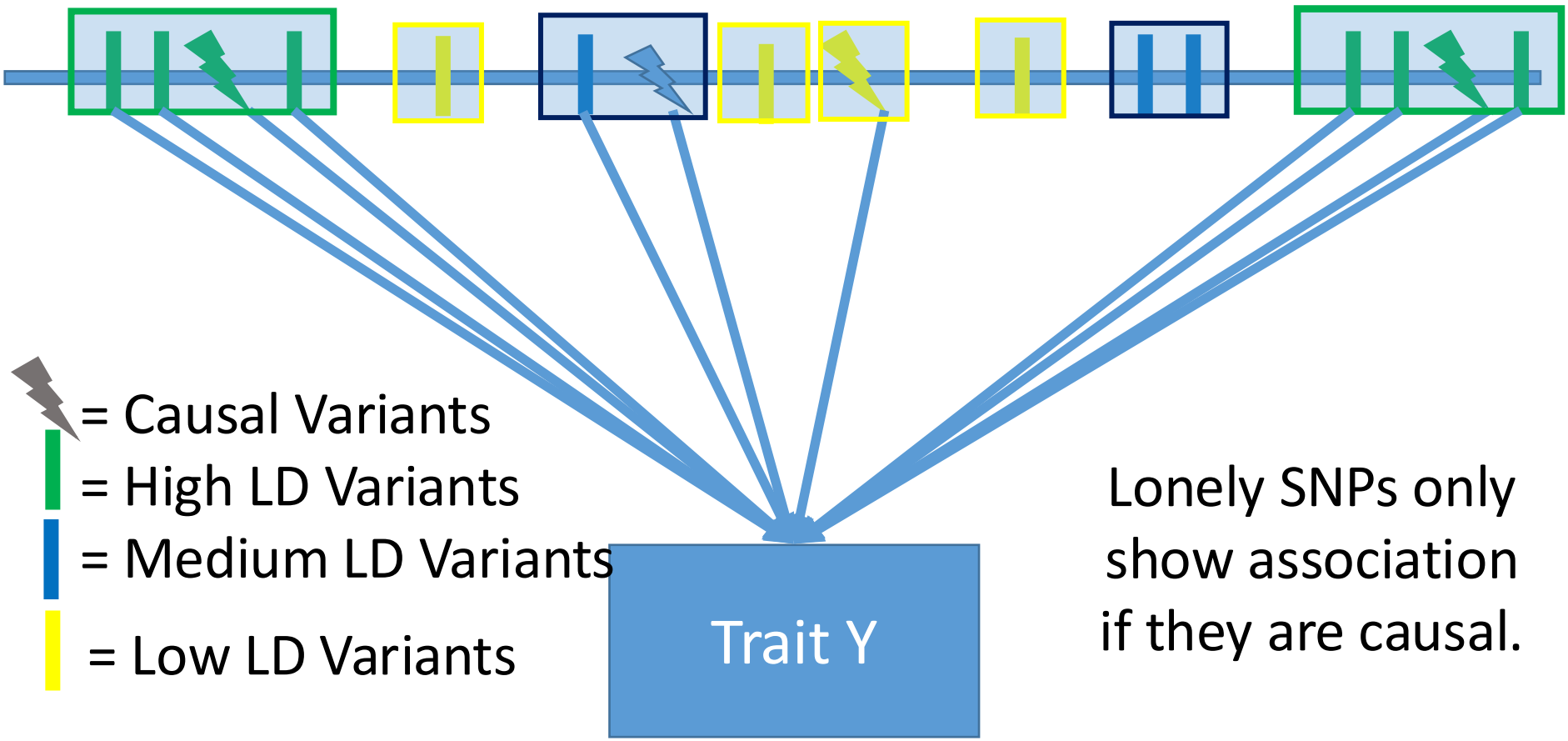
= Causal Variants

= High LD Variants

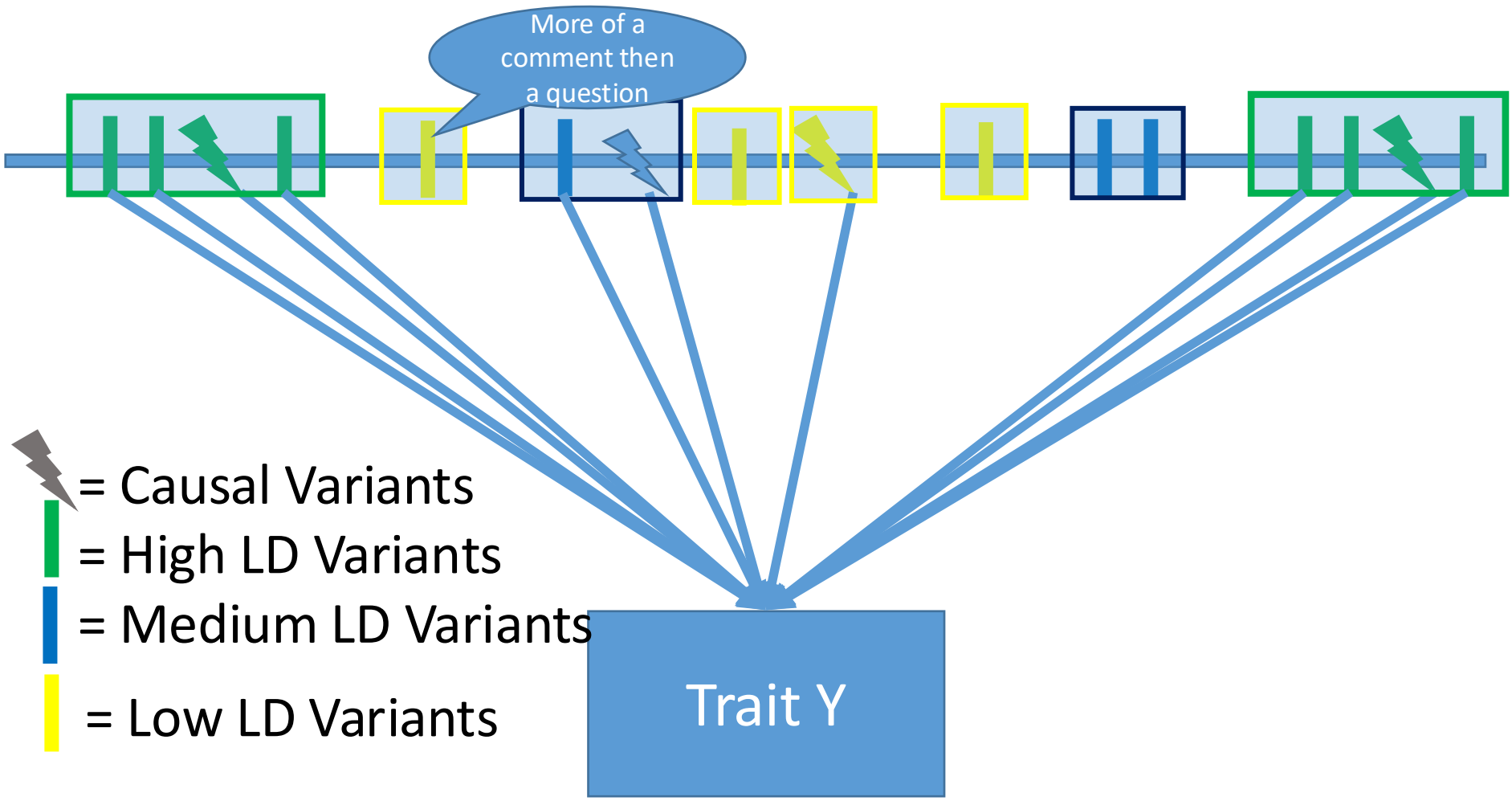
Trait Y

High LD variants (those correlated with more nearby SNPs) tend to pick up on signal because they are more likely to be correlated with causal variants


= Causal Variants
= High LD Variants
= Medium LD Variants
Trait Y



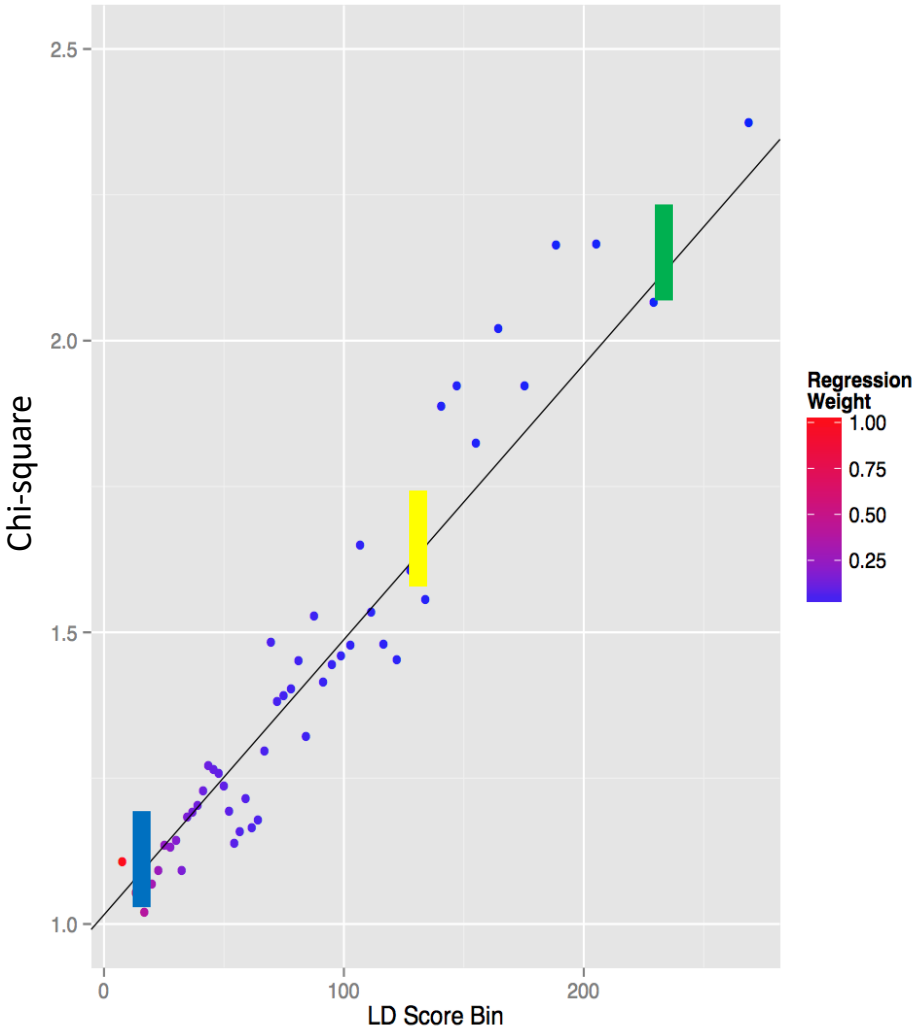

Lonely SNPs only show association if they are causal.


More of a comment then a question
= Causal Variants
= High LD Variants
= Medium LD Variants
= Low LD Variants
Trait Y

**Because traits are *highly* polygenic:**
Variants with higher LD are more likely to pick up on effect of causal variant

*Estimated Heritability* = strength of association between LD and SNP effect size

LD-scores are going to be more strongly related to effect size for traits with stronger genetic signal (i.e., higher heritability)




- To estimate SNP Heritability:
  - Regress GWAS chi-square against LD Scores for all SNPs (not just significant ones)

- Unbiased by sample overlap or cryptic population stratification
  - These only affect the intercept, but not the slope (which is what we back out heritability from)

$$E\left[\chi^2 \middle| \ell_j\right] = \frac{Nh^2}{M}\ell_j + N\mathrm{a} + 1$$

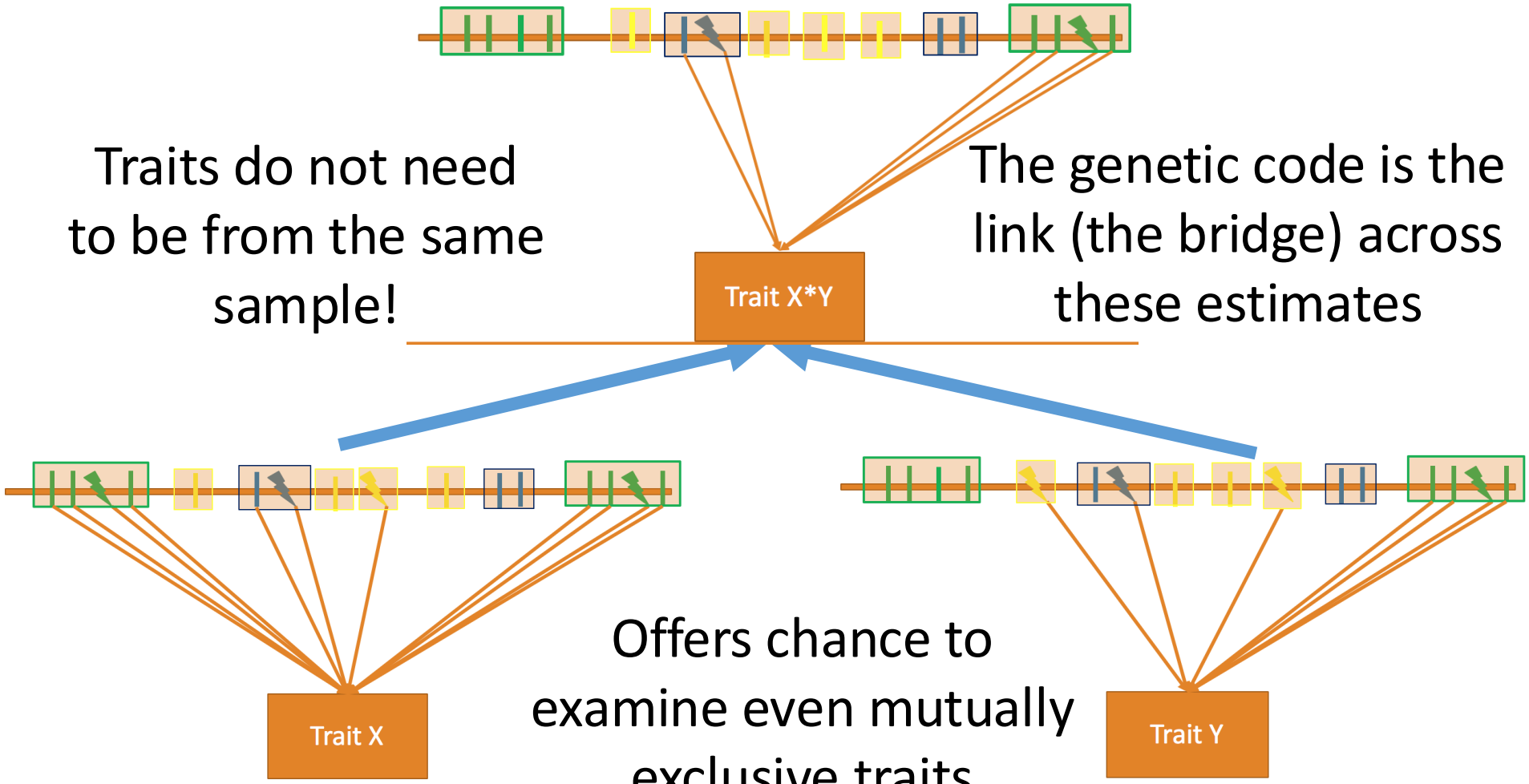
If you have two traits that have the exact same GWAS estimates (same p-value) but these sample sizes:

Trait 1: N = 50,000
Trait 2: N= 100,000

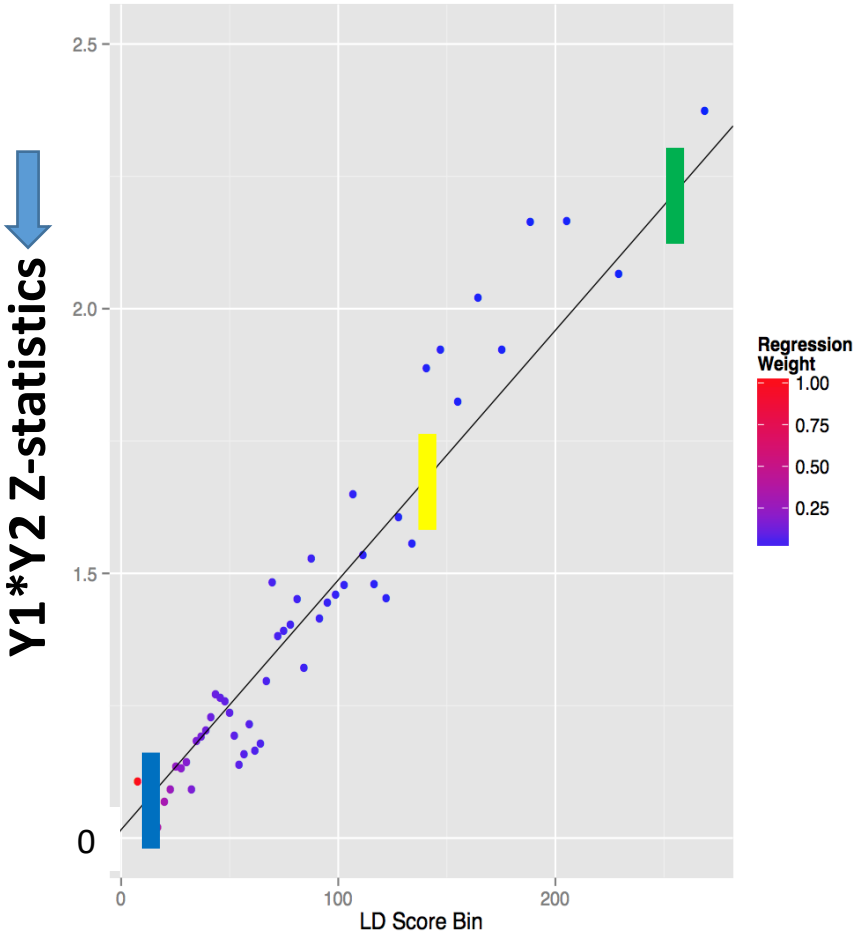
Which is more heritable?

Traits do not need to be from the same sample!

The genetic code is the link (the bridge) across these estimates




Offers chance to examine even mutually exclusive traits




Co-heritability (genetic covariance) = strength of association between LD and product of effect sizes for two different phenotypes (Y1*Y2)

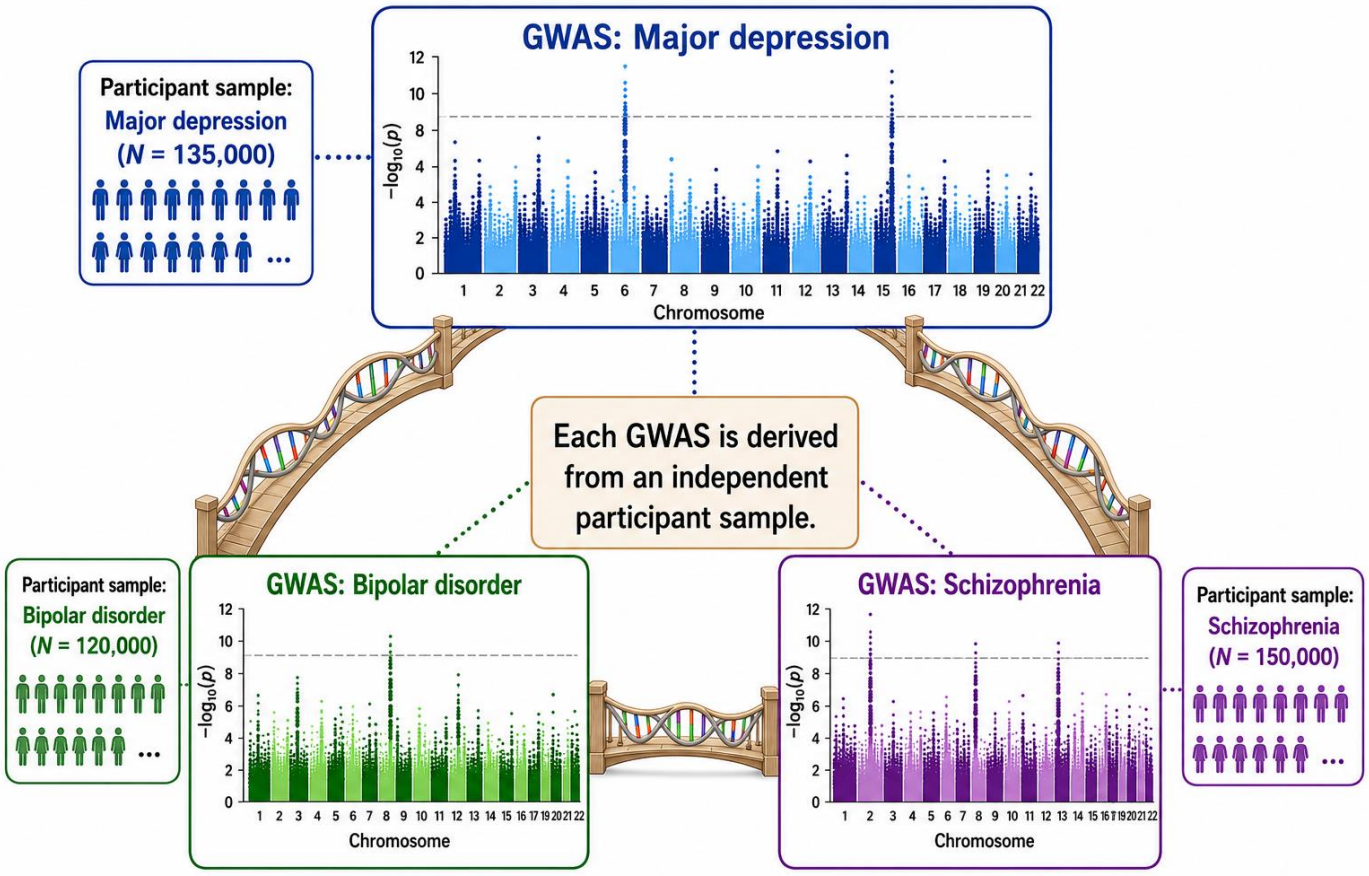
$$E\left[z_{1j}z_{2j}\,|\ell_j\right] = \frac{\sqrt{N_1N_2}\;\rho_g}{M}\ell_j + \sqrt{N_1N_2}a + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

# LD-Score Regression

Estimates genetic correlation based on meta-analytic GWAS data

**<u>Can come from independent participant samples</u>**

GWAS data is readily available




Bulik-Sullivan et al., 2015

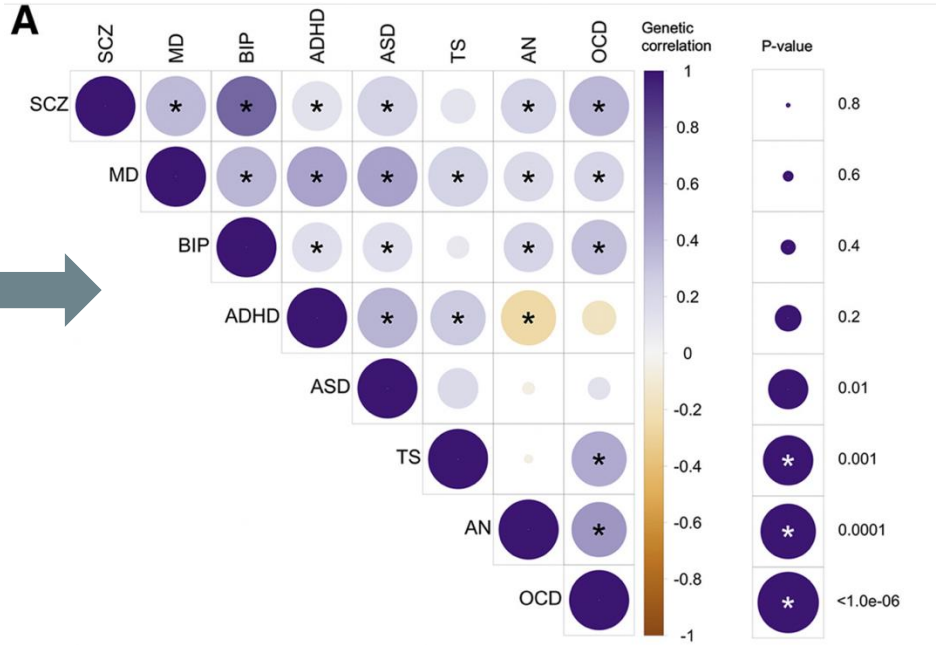
# Pervasive Overlap Necessitates Methods for Analyzing Genetic Co-Morbidity

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

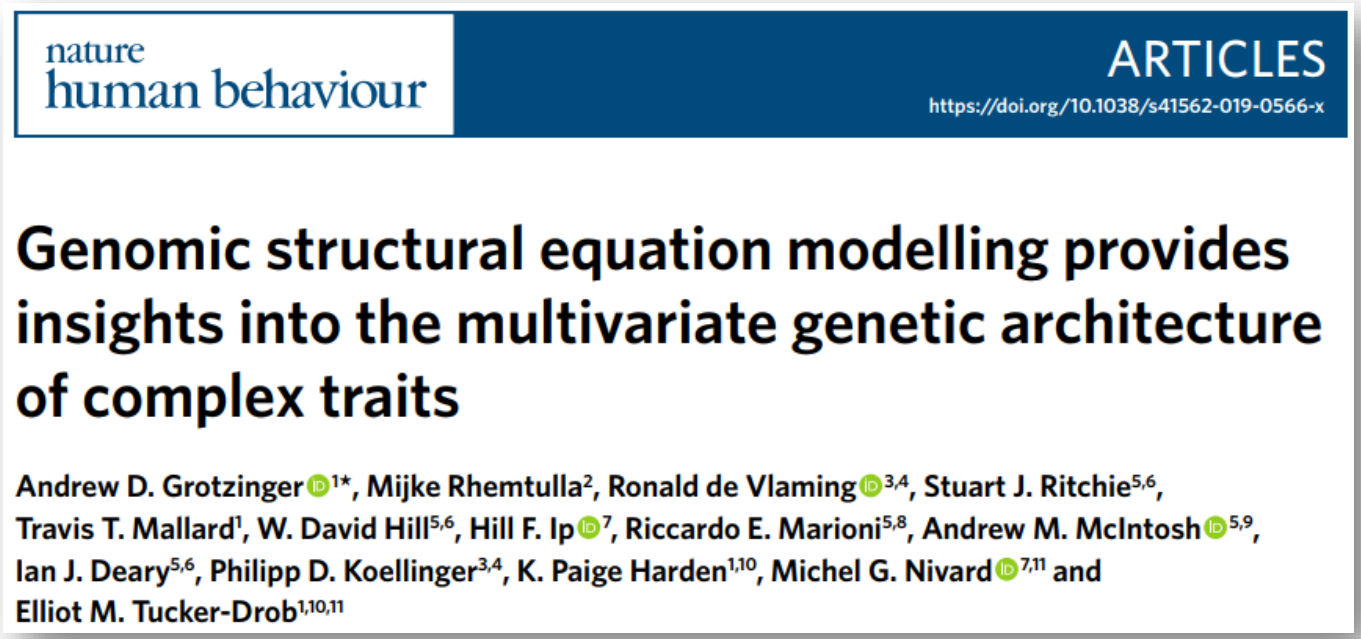
*Nature Genetics* **47**, 1236–1241 (2015) | Download Citation

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




Bulik-Sullivan et al., 2015; Lee et al., 2019

# Genomic SEM



nature human behaviour

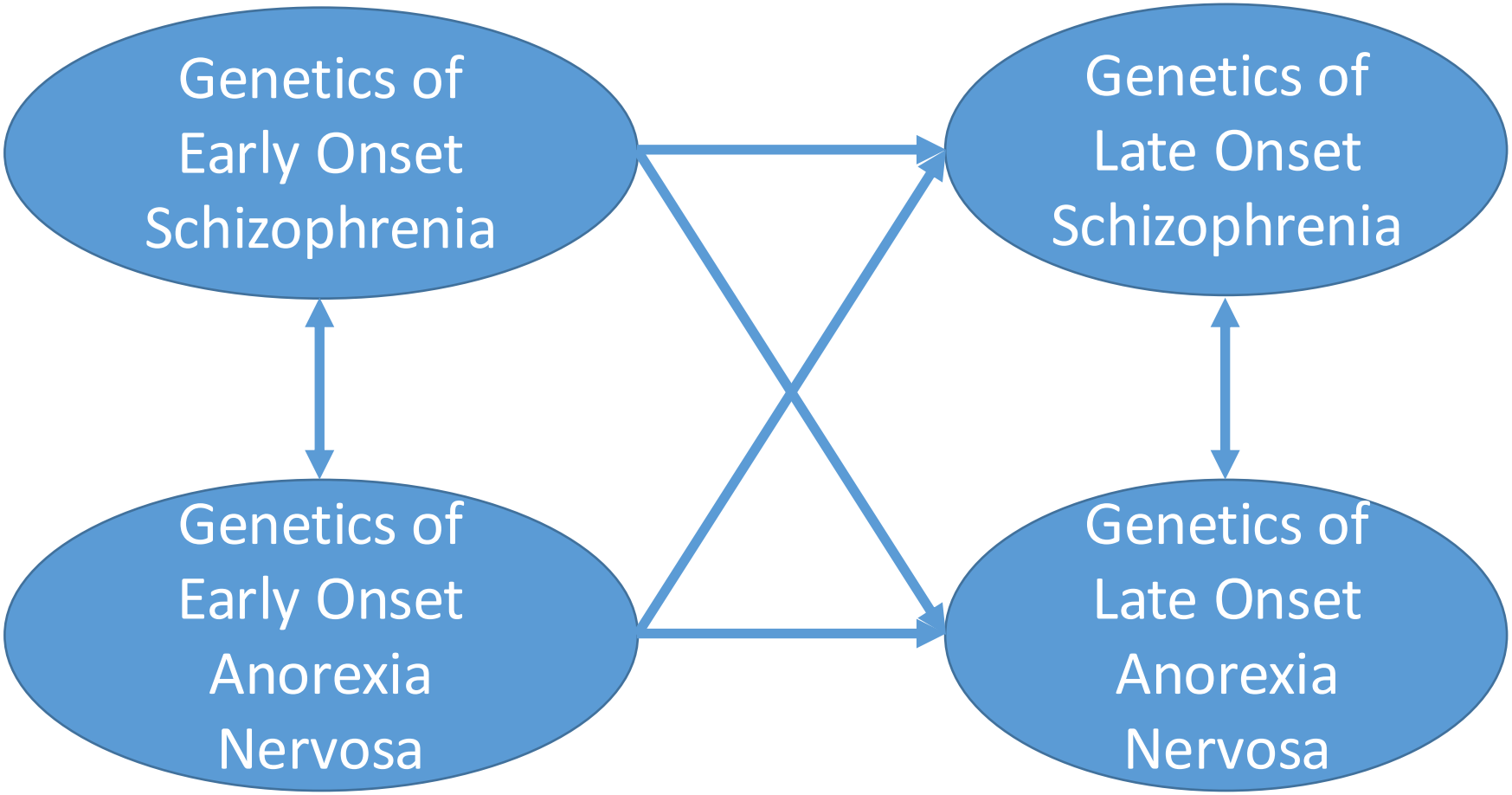
ARTICLES

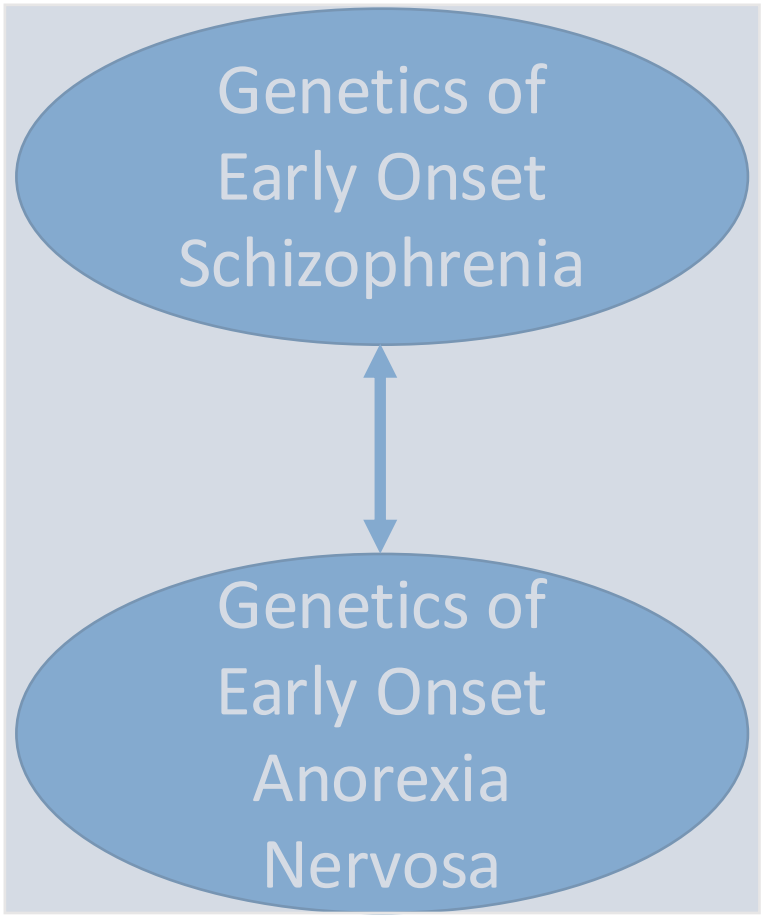
https://doi.org/10.1038/s41562-019-0566-x

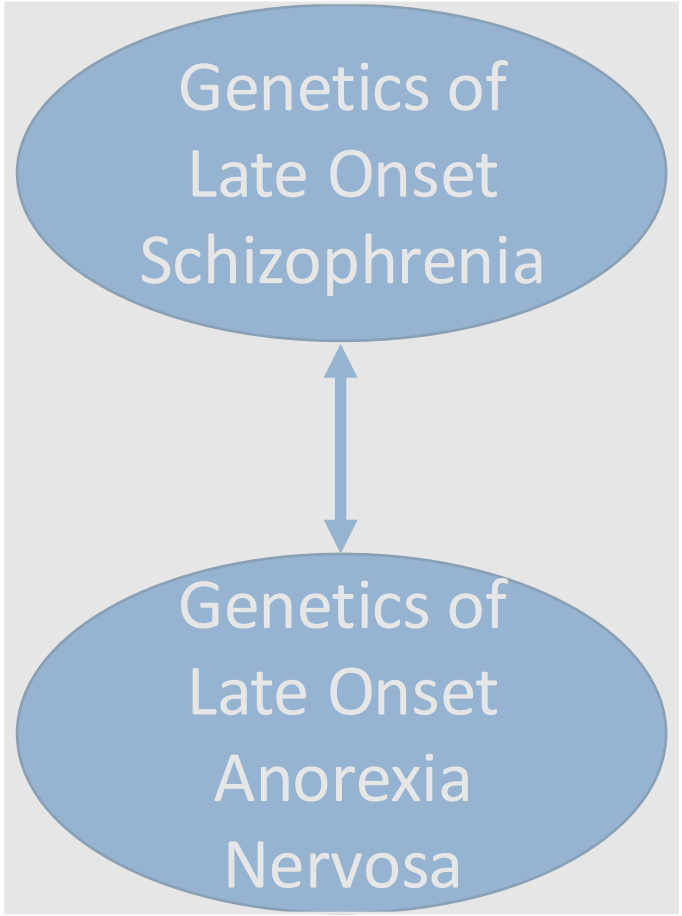
## Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits

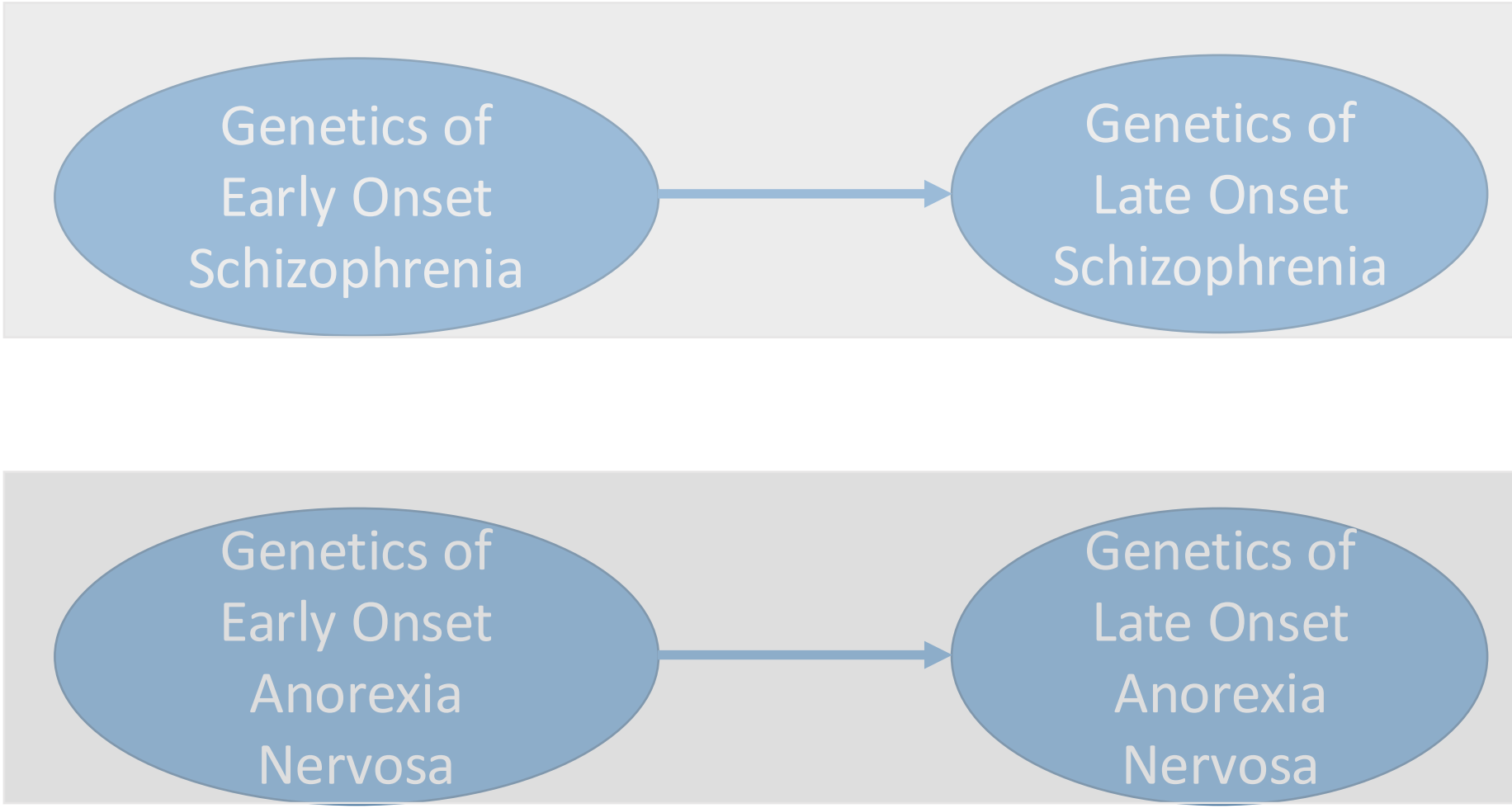
Andrew D. Grotzinger[1*], Mijke Rhemtulla[2], Ronald de Vlaming[3,4], Stuart J. Ritchie[5,6], Travis T. Mallard[1], W. David Hill[5,6], Hill F. Ip[7], Riccardo E. Marioni[5,8], Andrew M. McIntosh[5,9], Ian J. Deary[5,6], Philipp D. Koellinger[3,4], K. Paige Harden[1,10], Michel G. Nivard[7,11] and Elliot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

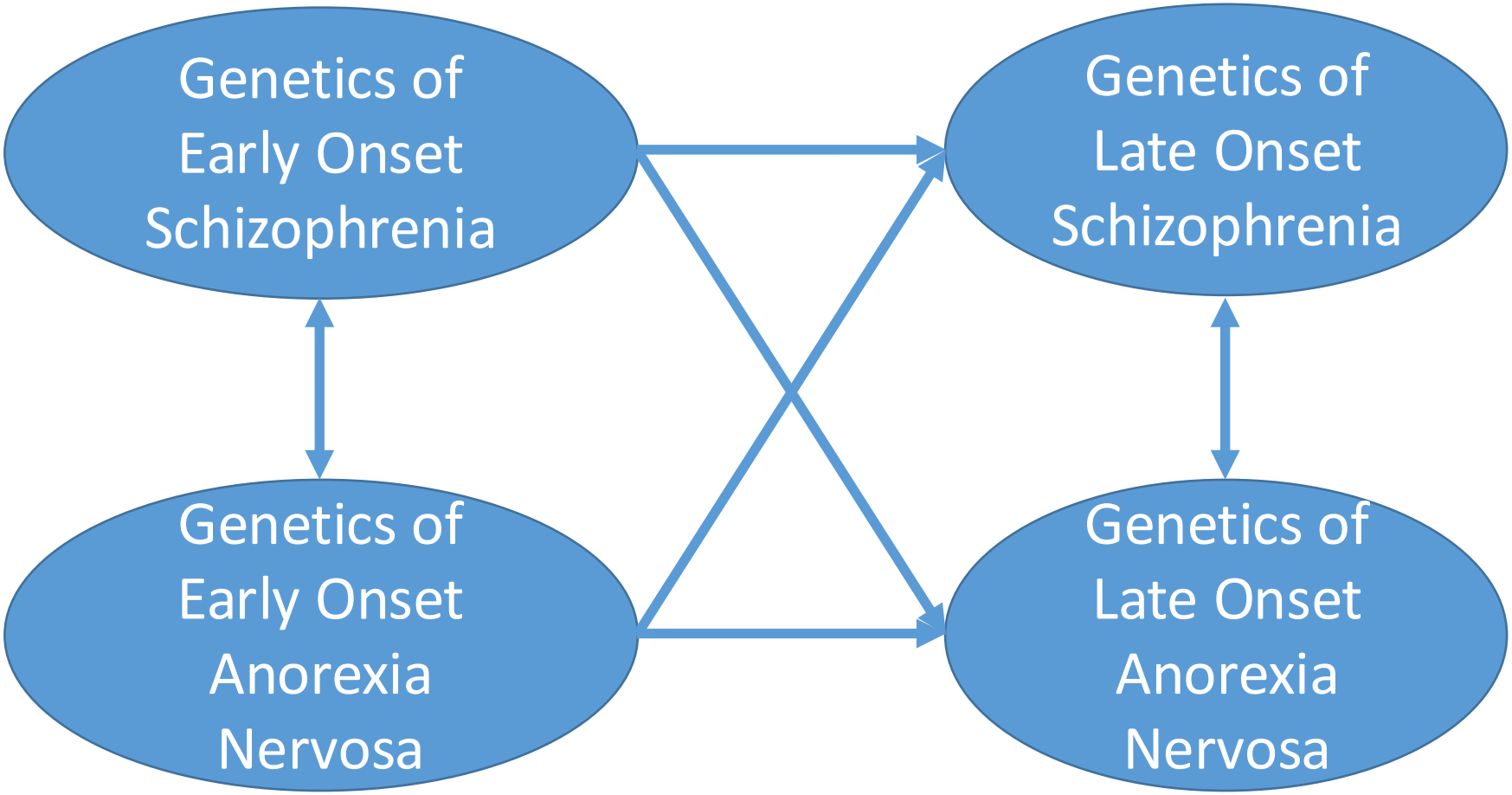
**Genomic SEM provides flexible framework for estimating limitless number of structural equation models using multivariate genetic data from GWAS summary statistics for even mutually exclusive traits**


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

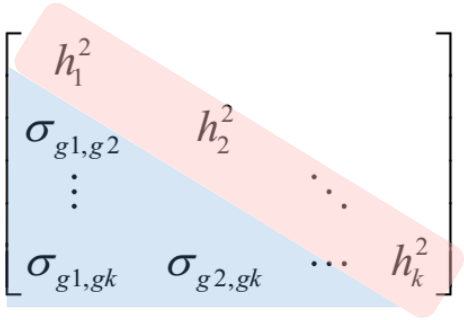
# Genomic SEM Uses 2 Stage Estimation

1. Use LD-score regression to estimate genetic covariances
2. Fit a structural equation model to the matrices from Stage 1

# Stage 1 Estimation

Create a genetic covariance matrix

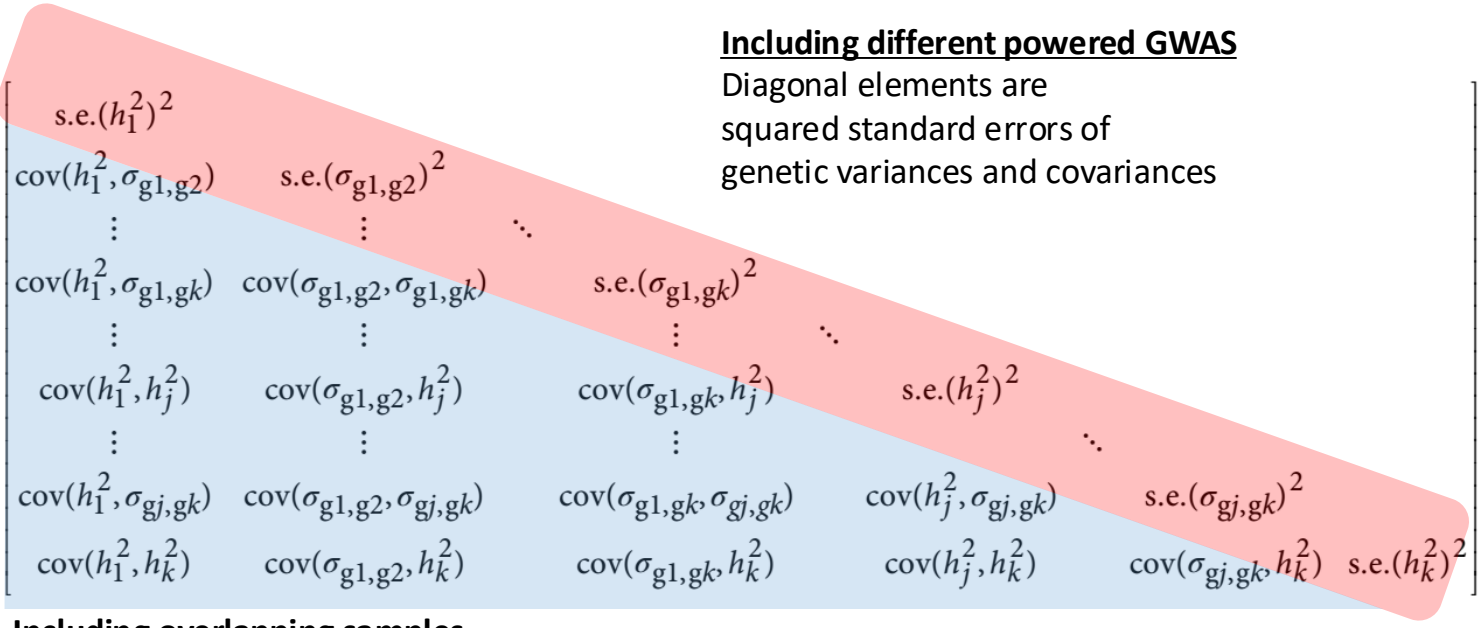
Diagonal elements are heritabilities




Off-diagonal elements are genetic covariances (genetic correlations when standardized)

# Stage 1 Estimation

Also produced is a corresponding sampling covariance matrix



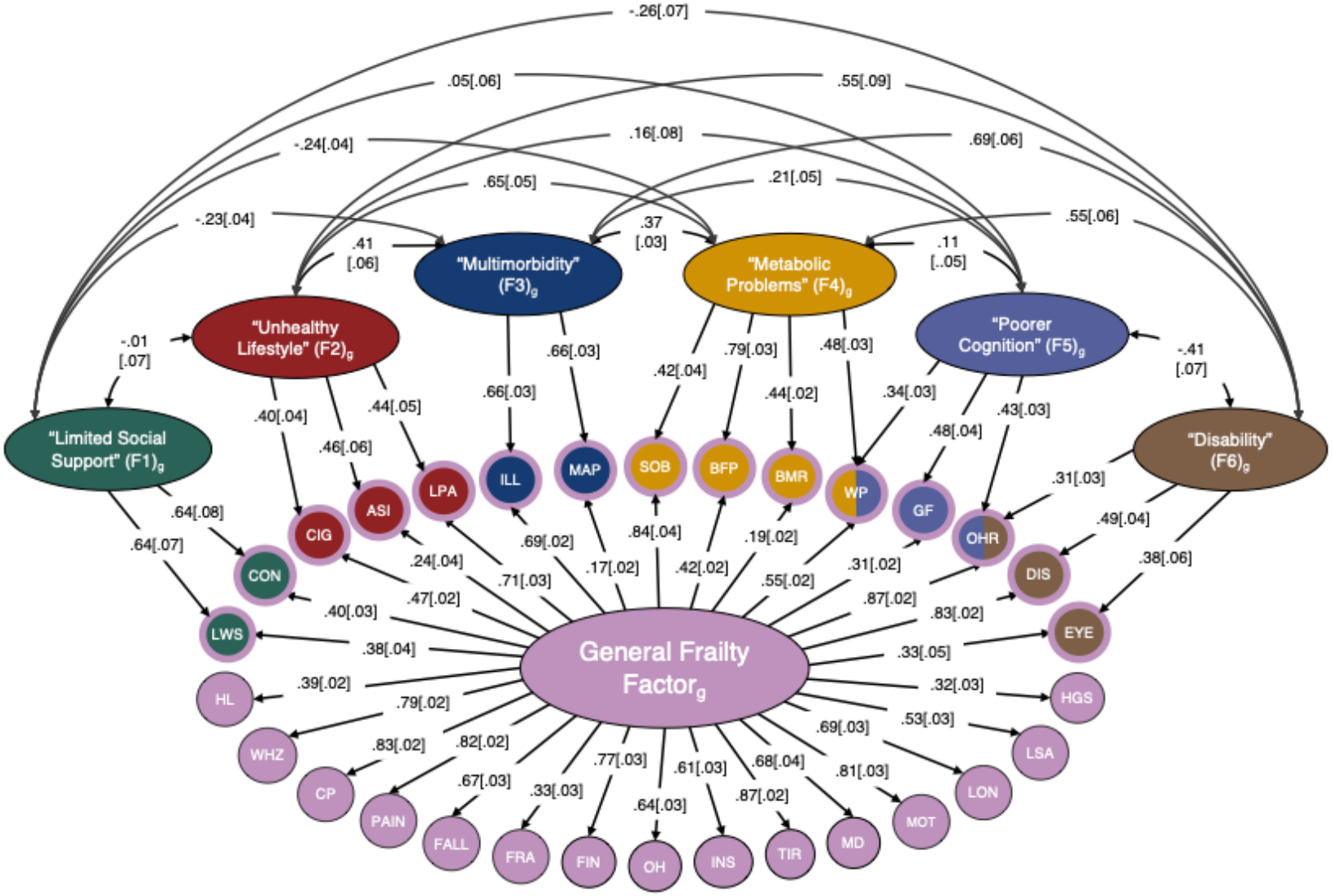

**Including different powered GWAS**

Diagonal elements are squared standard errors of genetic variances and covariances

**Including overlapping samples**

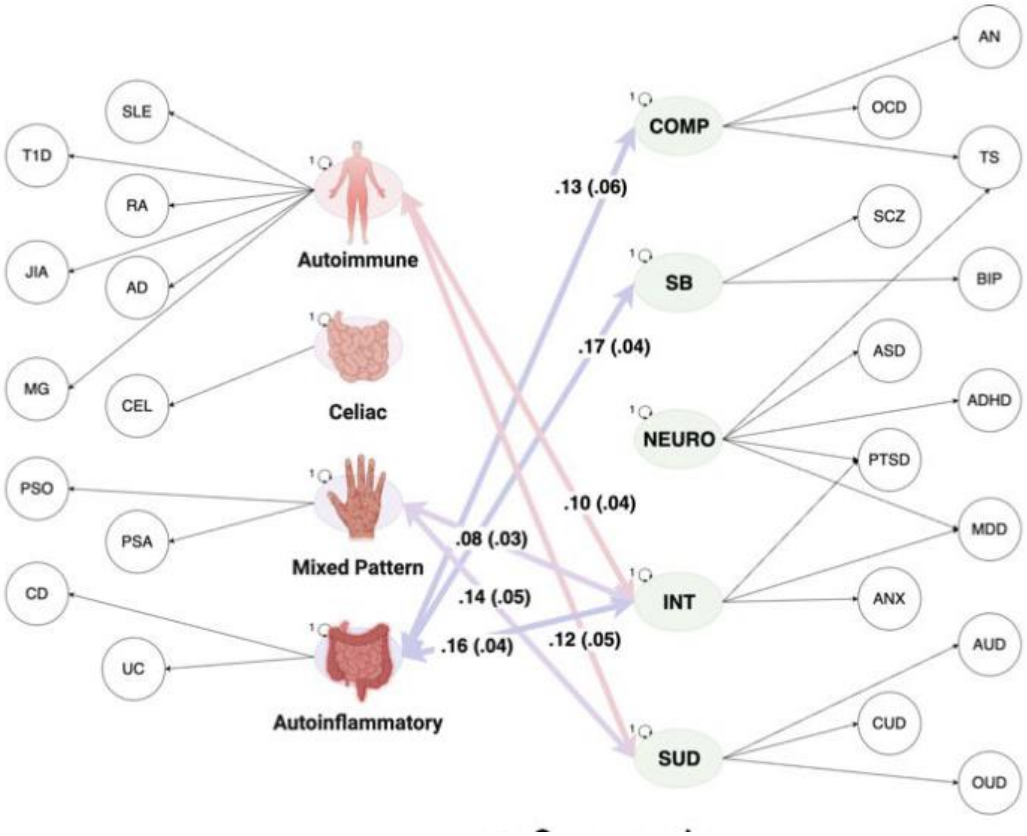
Off-diagonal elements are dependencies between estimation errors used to directly model dependencies that occur due to sample overlap

# Stage 2: Estimate the model




Foote et al., 2025; *Nature Genetics*

# Psychiatric Factors Relate to Immune Function



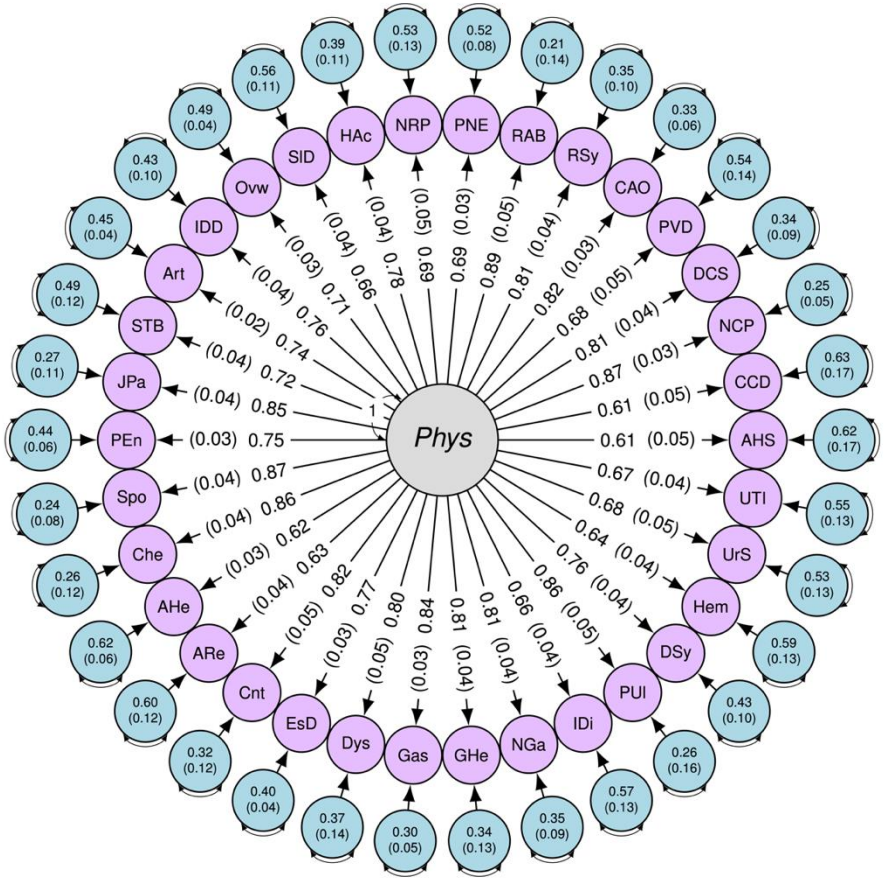

Sophie Breunig

Different psychiatric relationships depend on whether immune conditions are adaptive (autoimmune) or innate (autoinflammatory)

Breunig et al., 2025; *Translational Psychiatry*

# Common Factor of Medical Traits

Jeremy Lawrence



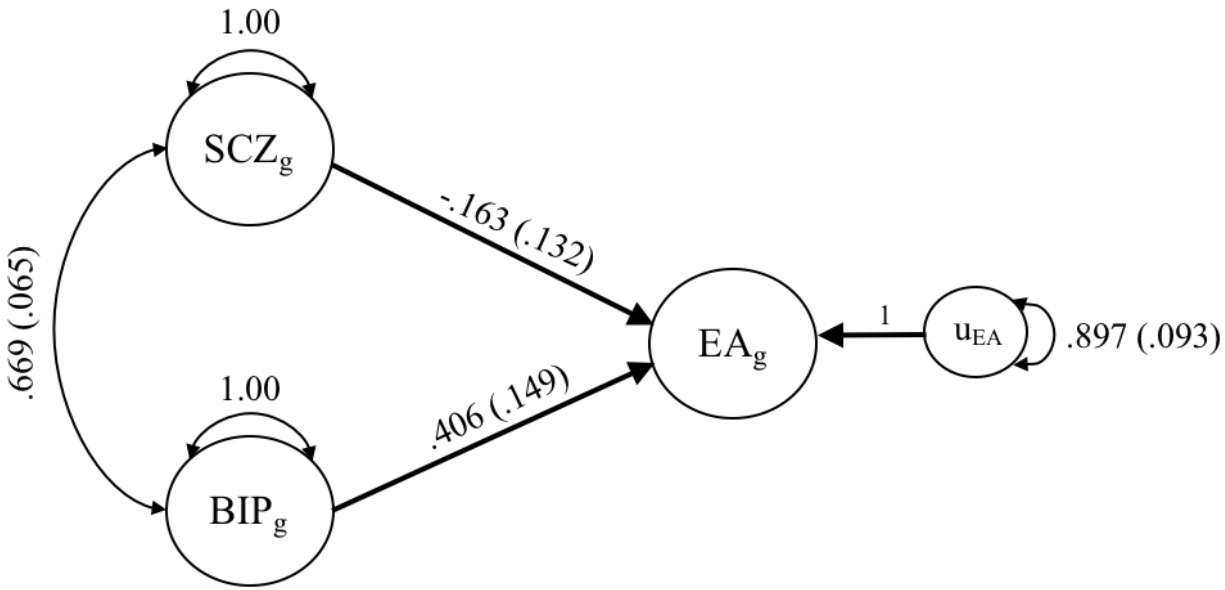

**Common factor reveals majority of genetic signal is shared across range of medical traits from seven domains:** neurological, respiratory, circulatory, digestive, endocrine/metabolic, genitourinary, musculoskeletal, and cancers

Lawrence et al., 2025; *Nature Communications*

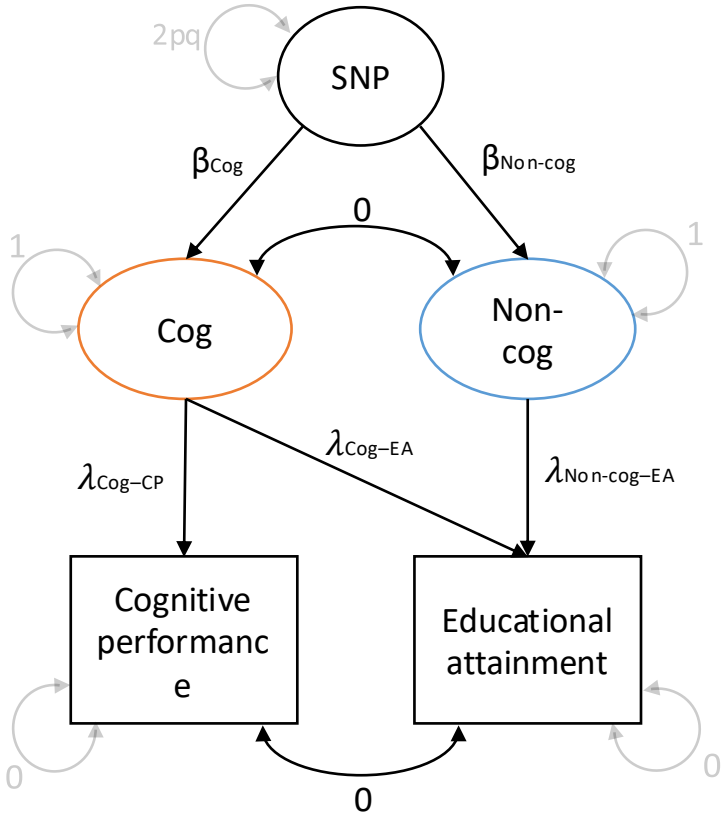
# Genetic Multiple Regression

| | | |
|---|---|---|
| SCZ | | |
| .67 | BIP | |
| .11 | .30 | EA |

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



# GWAS-by-subtraction




nature genetics
ARTICLES
https://doi.org/10.1038/s41588-020-00754-2
Check for updates

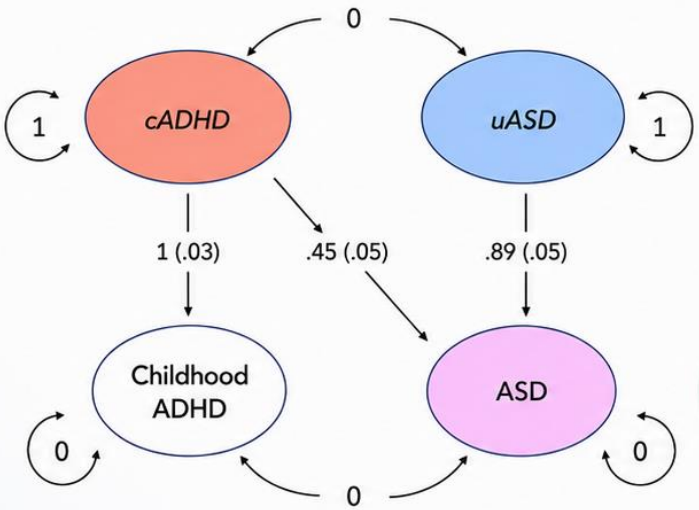
## Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction

Perline A. Demange[1,2,3,20], Margherita Malanchini[4,5,6,20], Travis T. Mallard[6], Pietro Biroli[7], Simon R. Cox[8], Andrew D. Grotzinger[6], Elliot M. Tucker-Drob[6,9], Abdel Abdellaoui[1,10], Louise Arseneault[5], Elsje van Bergen[1,3], Dorret I. Boomsma[1], Avshalom Caspi[5,11,12,13], David L. Corcoran[12], Benjamin W. Domingue[14], Kathleen Mullan Harris[15], Hill F. Ip[1], Colter Mitchell[16], Terrie E. Moffitt[5,11,12,13], Richie Poulton[17], Joseph A. Prinz[12], Karen Sugden[11], Jasmin Wertz[11], Benjamin S. Williams[11], Eveline L. de Zeeuw[1,3], Daniel W. Belsky[18,19,21] ✉, K. Paige Harden[6,21] ✉ and Michel G. Nivard[1,21] ✉

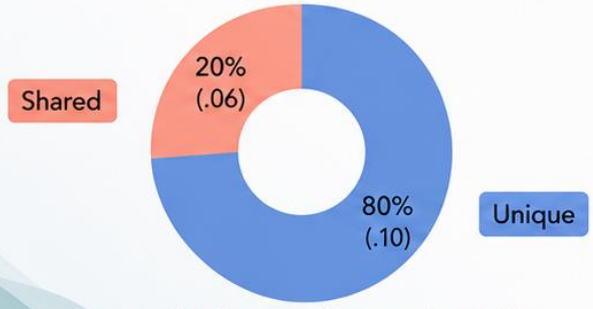
# Removing ADHD from Autism

Luke Schaffer

A.

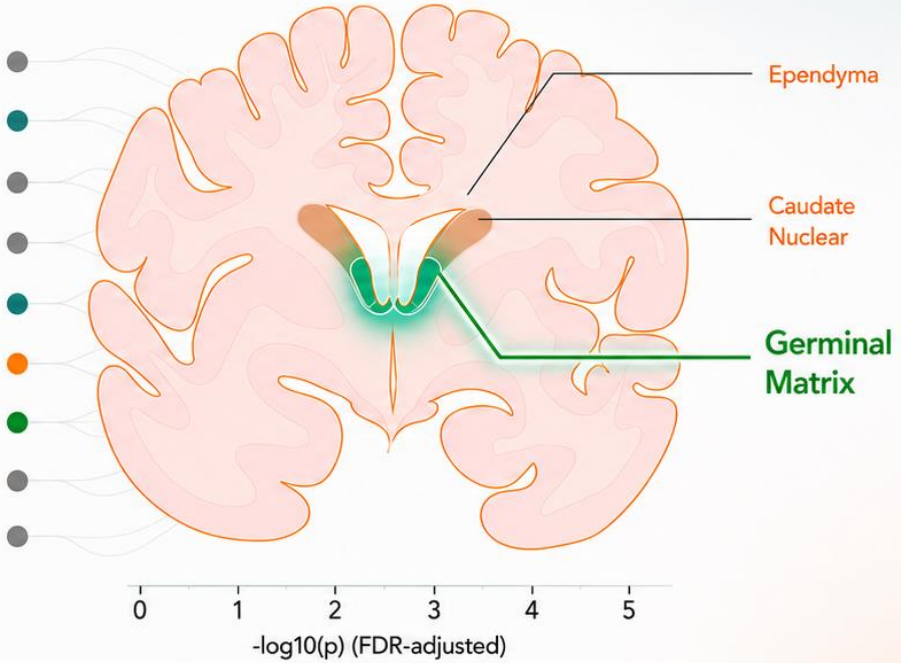



B.




~1/3 ASD cases also meet for ADHD

## Stratified Genomic SEM Enriched Functional Annotations



Article | Published: 11 November 2025

# A genome-wide analysis of the shared genetic risk architecture of complex neurological and psychiatric disorders

Olav B. Smeland, Gleda Kutrolli, Shahram Bahrami, Vera Fominykh, Nadine Parker, Julian Fuhrer, Guy F. L. Hindley, Linn Rødevand, Piotr Jaholkowski, Markos Tesfaye, Pravesh Parekh, Torbjørn Elvsåshagen, Andrew D. Grotzinger, Nils Eiel Steen, Der M. Dale, Alexey A. Shadrin, Oleksandr Fre

*Nature Neuroscience* **28**, 2439–2450 (2(



# Mapping the genetic lands disorders

Andrew D. Grotzinger, Josefin Werme, Wouter Bicks, Qiuyu Guo, Michael P. Margolis, Brandon J Joanna M. Biernacka, Ole A. Andreassen, Verneri Ditte Demontis, Howard J. Edenberg, Stephen V. Gelernter, Alexander S. Hatoum, Anxiety Dis Consortium, Attention-Deficit/Hyperactivity Genomics Consortium, Autism Spectrum Di Consortium, Bipolar Disorder Working Grou Working Group of the Psychiatric Genomics the Psychiatric Genomics Consortium, Nico Obsessive-Compulsive Disorder and Touret Consortium, Post-Traumatic Stress Disorde Schizophrenia Working Group of the Psychi Working Group of the Psychiatric Genomics

*Nature* **649**, 406–415 (2026) | Cite this ar

# n of the human skeletal form

.NARASIMHAN +8 authors Authors Info & Affiliations

Article | Open access | Published: 14 November 2022

# Genome-wide association and multi-trait analyses characterize the common genetic architecture of heart failure

Michael G. Levin, Noah L. Tsao, Pankhuri Singhal, Chang Liu, Ha My T. Vy, Ishan Paranjpe, Joshua D. Backman, Tiffany R. Bellomo, William P. Bone, Kiran J. Biddinger, Qin Hui, Ozan Dikilitas, Benjamin A. Satterfield, Yifan Yang, Michael P. Morley, Yuki Bradford, Megan Burke, Nosheen Reza, Brian Charest, Regeneron Genetics Center, Renae L. Judy, Megan J. Puckelwartz, Hakon Hakonarson, Atlas Khan, ... Scott M. Damrauer + Show authors

*Nature Communications* **13**, Article number: 6914 (2022) | Cite this article

# elucidates the genetic architecture of shared components of the metabolic syndrome

Sanghyeon Park, Soyeon Kim, Beomsu Kim, Dan Say Kim, Jaeyoung Kim, Yeeun Ahn, Hyejin Kim, Minku Song, Injeong Shim, Sang-Hyuk Jung, Chamlee Cho, Soohyun Lim, Sanghoon Hong, Hyeonbin Jo, Akl C. Fahed, Pradeep Natarajan, Patrick T. Ellinor, Ali Torkamani, Woong-Yang Park, Tae Yang Yu, Woojae Myung & Hong-Hee Won

*Nature Genetics* **56**, 2380–2391 (2024) | Cite this article

# Selling the product

Now

# Selling the product

I know what the grad students in the audience must be thinking

# Selling the product

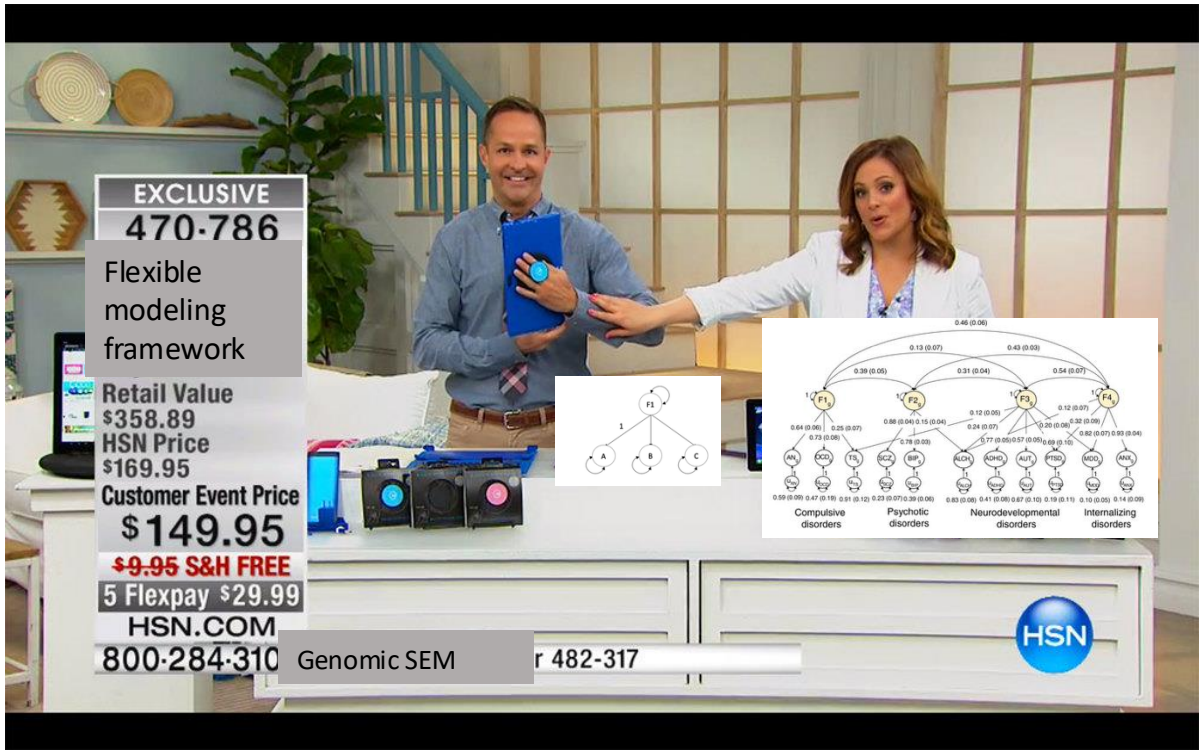
You're selling us an amazing tool to flexibly model genetic overlap across ***even mutually exclusive traits.***

But I can barely pay rent and my bar tab, there's no way I can afford this

# Selling the product

But you know what…

# Selling the product




I like this group and I'm feeling generous so I'm going to slash our prices just for you

# Selling the product

Ya'll, at these prices we are practically (literally) giving the product away:

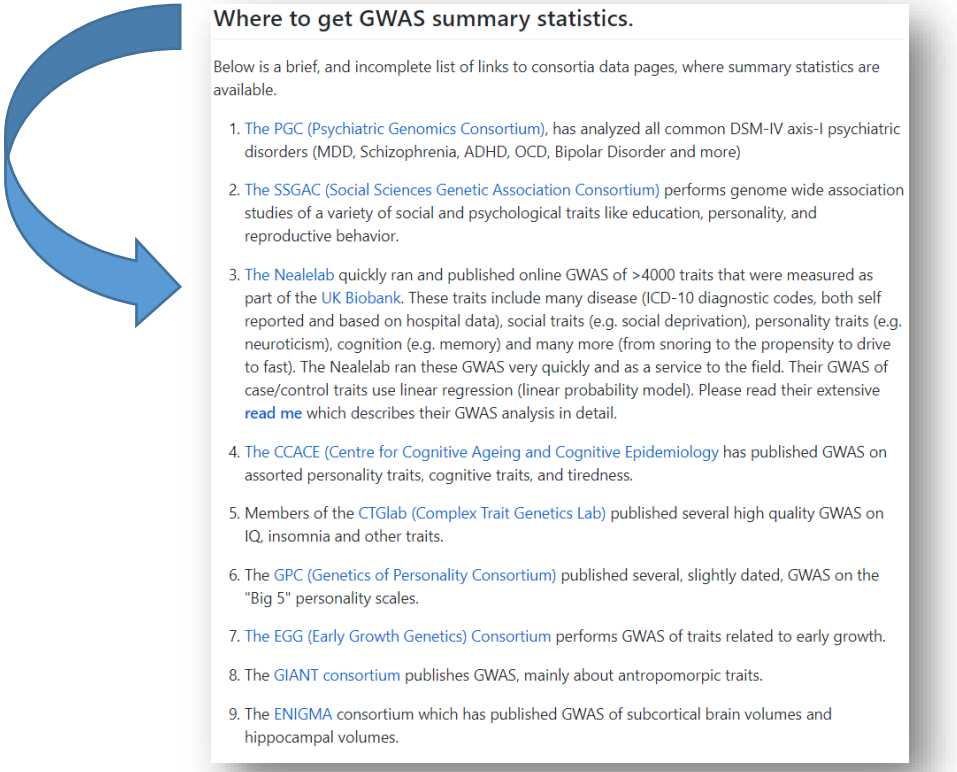
$0 down
$0 monthly payment
No annual fees

# Selling the product

Alright, but what about the data? That costs money doesn’t it?

# Where to get FREE summary statistics

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



**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.


https://www.ebi.ac.uk/gwas/

BioBank Japan PheWeb (PheWeb.jp)

https://pheweb.jp/

FINNGEN

https://www.finngen.fi/en/access_results

Pan-UK Biobank

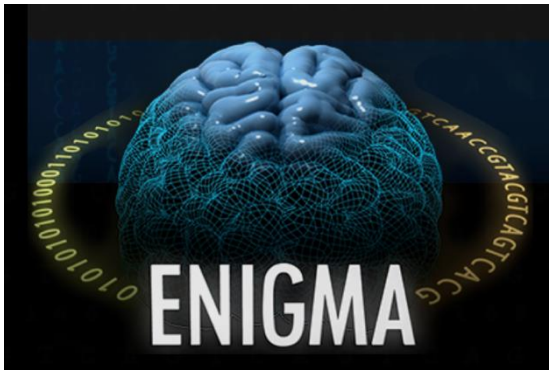
Pan-ancestry genetic analysis of the UK Biobank

https://docs.google.com/spreadsheets/d/1AeeADtT0U1AukliiNyiVzVRdLYPkTbruQSk38DeutU8/edit#gid=268241601

PGC

Psychiatric Genomics Consortium

https://pgc.unc.edu/for-researchers/download-results/




https://enigma.ini.usc.edu/research/download-enigma-gwas-results/

# II. Estimating Genome-wide Models

# 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***)

# Lab

**Using GWAS summary statistics for:**

- Major Depressive Disorder (Cases = 170,756; Controls = 329,443; Howard et al., 2019)
- Anxiety Disorders (Cases = 31,977; Controls = 82,114; Purves et al., 2020)
- Alcohol use disorder (Cases = 8,485; Controls = 20,272; Walters et al., 2018)
- PTSD (Cases = 2,424; Controls = 7,113; Duncan et al., 2018)

We are going to start with **Step 3**: Estimating the genome-wide model

# How to specify a model in lavaan

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)

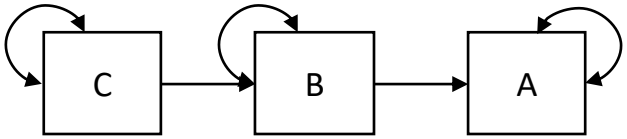
**Allows you to use model constraints for this parameter:**

a > .001

# Let’s make that a bit more specific

Model1 <- “ A ~ B

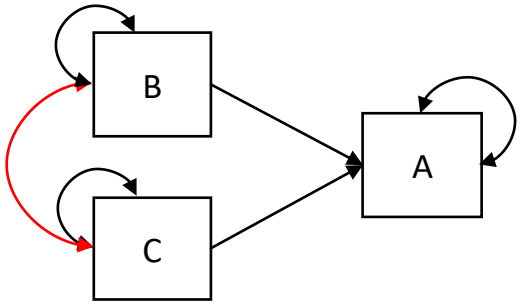
B ~ C”




Model2 <- “ A ~ B

A ~ C

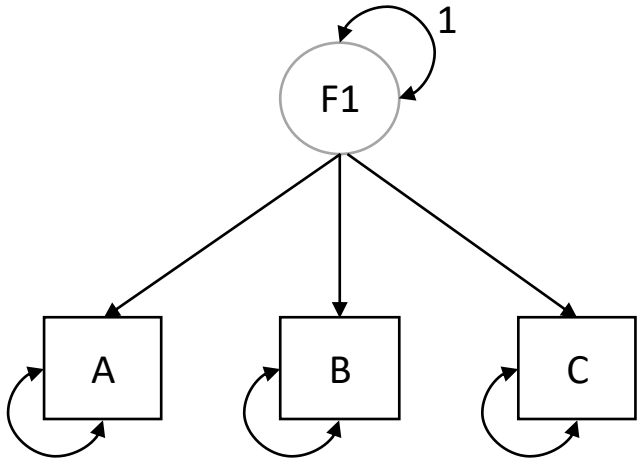
B ~~ C”



# Identifying the factor option 1: Unit variance identification

Model3 <- " F1 =~ NA*A + B + C

F1 ~~ 1*F1"




Note that the above is the same as running

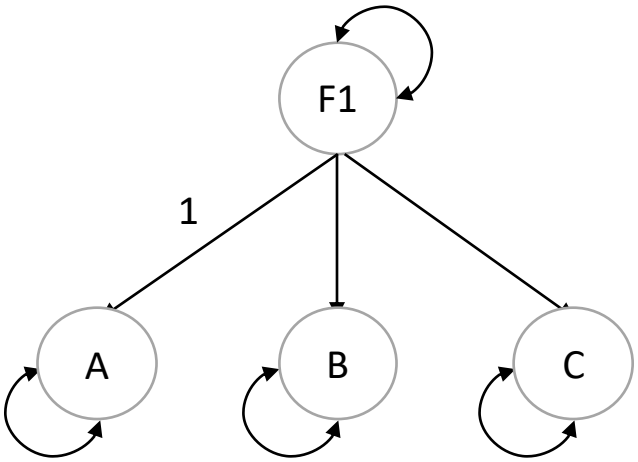
Model3 <- " F1 =~ A + B + C"

And setting the argument std.lv = TRUE

Where std.lv denotes standardized latent variables

# Identifying the factor option 2: Unit loading identification

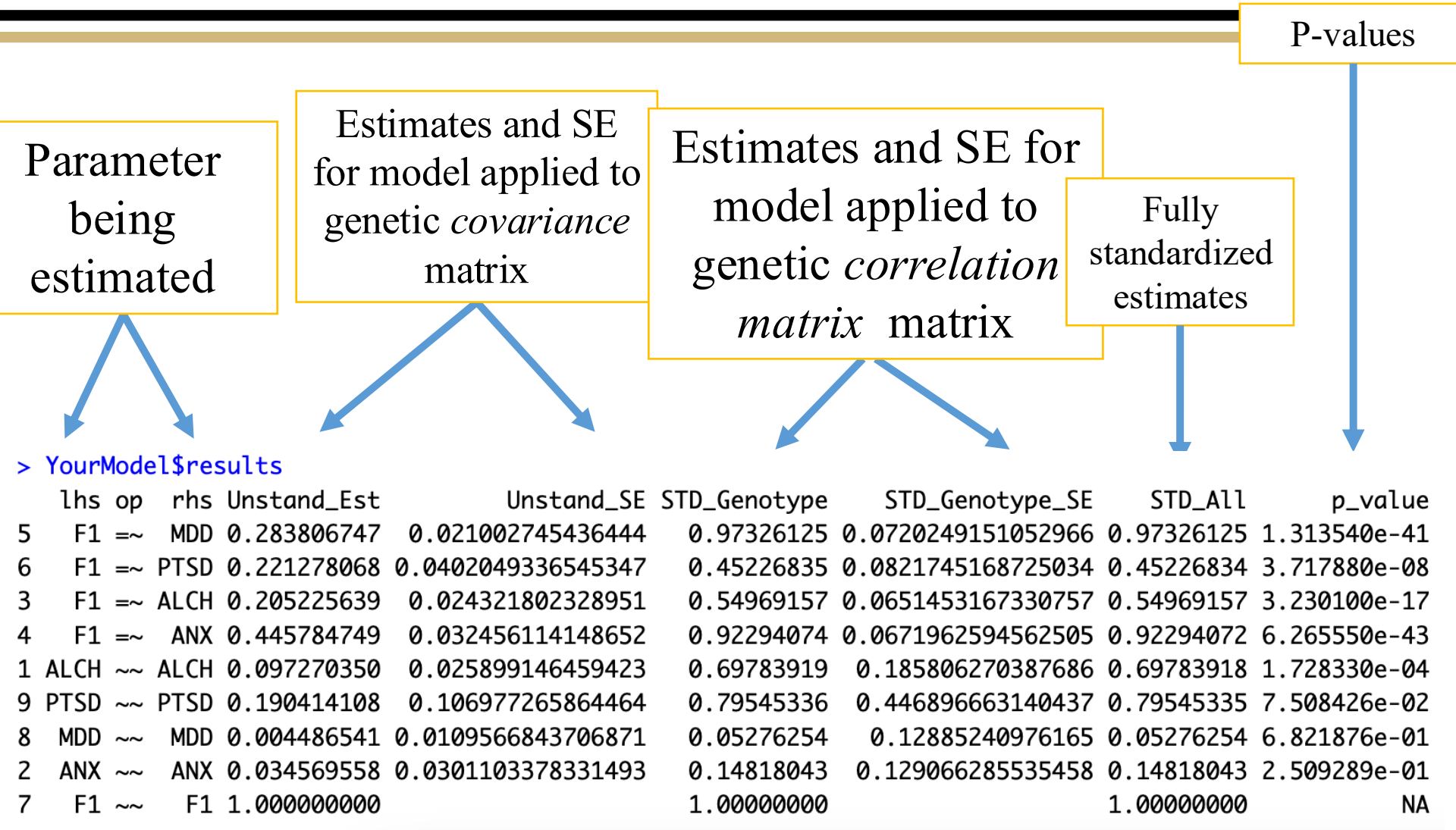
Model3 <- " F1 =~ 1*A + B + C"



# Specifying Arguments in *usermodel*

```
#load in the LDSC object made in step 2 above
load("LDSC_INT.RData")

#Takes two necessary arguments:
#1. covstruc: the output from multivariable ldsc
covstruc<-LDSC_INT

#2. model: the user specified model
#here we specify a common factor model
INT.model<-"F1=~MDD+PTSD+ALCH+ANX"

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

#Run your model
IntResults <- usermodel(covstruc=covstruc, model=INT.model,std.lv=std.lv)
```

# IntResults$results




```
> YourModel$results
   lhs op  rhs Unstand_Est         Unstand_SE STD_Genotype    STD_Genotype_SE    STD_All      p_value
5   F1 =~  MDD 0.283806747  0.021002745436444   0.97326125 0.0720249151052966 0.97326125 1.313540e-41
6   F1 =~ PTSD 0.221278068 0.0402049336545347   0.45226835 0.0821745168725034 0.45226834 3.717880e-08
3   F1 =~ ALCH 0.205225639  0.024321802328951   0.54969157 0.0651453167330757 0.54969157 3.230100e-17
4   F1 =~  ANX 0.445784749  0.032456114148652   0.92294074 0.0671962594562505 0.92294072 6.265550e-43
1 ALCH ~~ ALCH 0.097270350  0.025899146459423   0.69783919  0.185806270387686 0.69783918 1.728330e-04
9 PTSD ~~ PTSD 0.190414108  0.106977265864464   0.79545336  0.446896663140437 0.79545335 7.508426e-02
8  MDD ~~  MDD 0.004486541 0.0109566843706871   0.05276254   0.12885240976165 0.05276254 6.821876e-01
2  ANX ~~  ANX 0.034569558 0.0301103378331493   0.14818043  0.129066285535458 0.14818043 2.509289e-01
7   F1 ~~   F1 1.000000000                     1.00000000                    1.00000000           NA
```

# IntResults$modelfit

| chisq | df | p_chisq | AIC | CFI | SRMR |
|---|---|---|---|---|---|
| 1.283452 | 2 | 0.526383 | 17.28345 | 1 | 0.03621695 |

- **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

Let’s go to the code and fill out the Qualtrics questions for genome-wide models

# III. 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 *(**userGWAS**)*

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

# Step 3: *sumstats* example code Note that the traits need to be in the same order as for LDSC

```
#1. files = the name of the summary statistics file
##**note that these are a drastically reduced subsets of SNPs for the practical ONLY
##*that reflect a selected set of chromosome 4 variants
##*Also note that these need to be in the same order as your ldsc object
files<-c("ALCH4.txt", "PTSD4.txt", "MDD4.txt", "ANX4.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 for chromosome 4
#the full reference set is available on our github
ref <- "reference.1000G.ch4.txt"

#3. trait.names = the name of the traits to use for column labeling
trait.names<-c("ALCH","PTSD", "MDD", "ANX")

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

#5. linprob: whether it was a binary outcome that was analyzed as continuoue -or-
#it is a file with only Z-statistics. This is true for ALCH for our data
linprob<-c(T,F,F,F)

#6. sample size. This is only needed for continuous outcomes or outcomes where linprob is TRUE
#we do not provide sample size for ALCH as it is already a column in the GWAS data
N<-c(NA,NA,NA,NA)

#run sumstats putting these pieces together
INT_sumstats<-sumstats(files=files,ref=ref,trait.names=trait.names,se.logit=se.logit,linprob=linprob,N=N)
```

**Flowchart on github to help you figure out arguments for sumstats**

# The linear probability model argument for ALCH

*sumstats* needs a beta and SE to perform analyses, but sometimes the summary data only has a Z-statistic (or it was a binary outcome analyzed as continuous).

In these cases, linprob is set as TRUE for that trait and the sum of effective sample size is needed

| | CHR | SNP | BP | A1 | A2 | Z | P | Weight |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | rs10799799 | 23973601 | T | G | -0.190 | 0.8495 | 21857.28 |
| 2 | 1 | rs200328701 | 57729103 | CT | C | -0.278 | 0.7807 | 18324.09 |
| 3 | 1 | rs75245025 | 151716030 | C | G | 0.710 | 0.4779 | 22791.88 |
| 4 | 1 | rs146090341 | 211396305 | T | G | -0.092 | 0.9265 | 18999.23 |
| 5 | 1 | rs188551793 | 215407827 | A | T | 0.321 | 0.7482 | 11611.03 |
| 6 | 1 | rs2490387 | 237279677 | T | C | -1.216 | 0.2239 | 20769.08 |

# Examine ALCH_PTSD_MDD_ANX_sumstats.log

```
The preparation of 4 summary statistics for use in Genomic SEM began at: 2026-05-19 10:54:59.358452
Please note that the files should be in the same order that they were listed for the ldsc function
Reading in reference file
Applying MAF filer of 0.01 to the reference file.
All files loaded into R!

Preparing summary statistics for file: ALCH4.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 Z column as the effect column.
Interpreting the P column as the P column.
Interpreting the WEIGHT column as the N column.
0 rows were removed from the ALCH4.txt summary statistics file due to entries that were duplicated for rsID.
These are removed as they likely reflect multiallelic variants.
Merging file: ALCH4.txt with the reference file: reference.1000G.ch4.txt
1000 rows present in the full ALCH4.txt summary statistics file.
0 rows were removed from the ALCH4.txt summary statistics file as the rsIDs for these SNPs were not present in the reference file.
The effect column was determined NOT to be coded as an odds ratio (OR) for the ALCH4.txt summary statistics file based
on the median of the effect column being close to 0.
An transformation used to back out logistic betas for binary traits is being applied for: ALCH4.txt
No INFO column, cannot filter on INFO, which may influence results
1000 SNPs are left in the summary statistics file ALCH4.txt after QC and merging with the reference file.
```

# Step 4: *userGWAS* example code

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

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

#3. model = the model to be run
#going to troubleshoot estimated ov variances are negative for 4 SNPs
#by adding model constraint for all residuals to be above 0
model<-"F1=~MDD+PTSD+ALCH+ANX
F1~SNP"
```

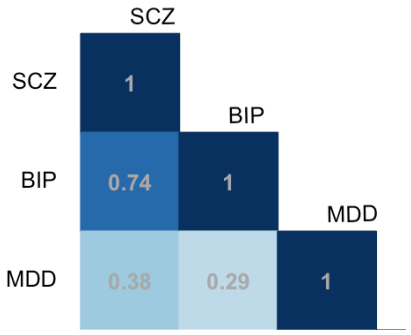
# Step 4: *userGWAS* example code

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

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

#6. Q_SNP = optional argument specifying whether you want
#the heterogeneity index calculated for your factors
Q_SNP<-TRUE

#run the multivariate GWAS below
INT_GWAS<-userGWAS(covstruc=covstruc, SNPs=SNPs, model=model, sub=sub, parallel=parallel, Q_SNP=Q_SNP)
```

# Behind the scenes

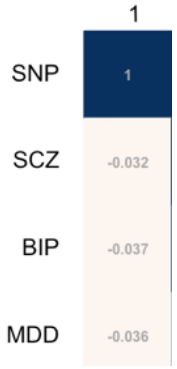
- *userGWAS* combines output from *ldsc* and *sumstats* to be able to specify a model with SNP effects
- Creates as many covariance matrices as there are SNPs across traits
- The measurement model (the factor loadings and correlations) is fixed across SNPs as a default to speed up analysis and improve interpretability (can be turned on/off with *fix_measurement* argument)

**Step 2**: Run *ldsc*



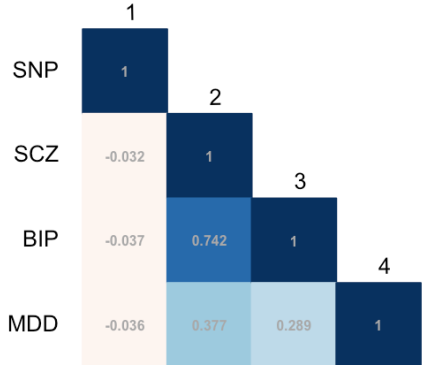

+

**Step 3**: Run ***sumstats***



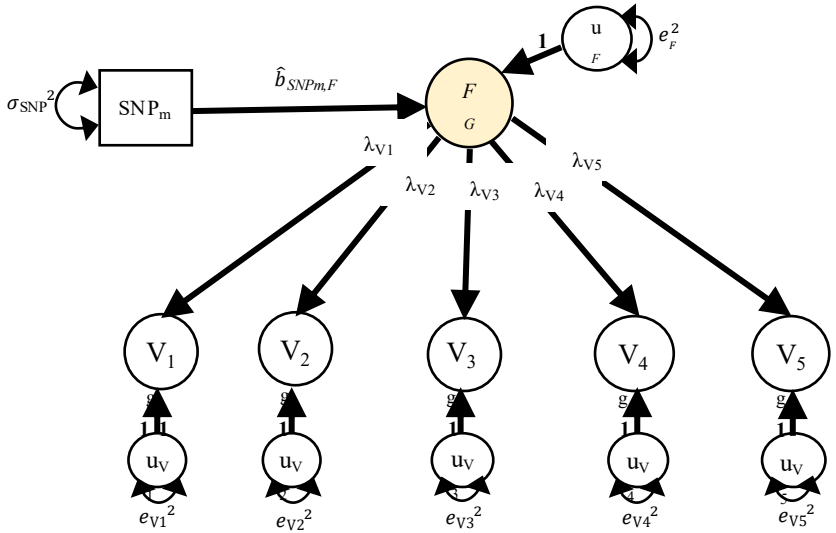

=

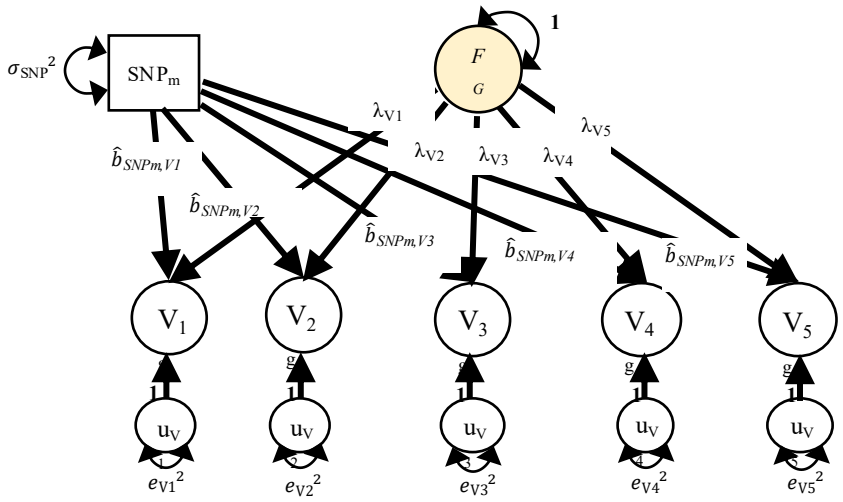
**Step 4:** *userGWAS* combines the two



# Estimates of SNP level heterogeneity ($Q_{SNP}$)

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



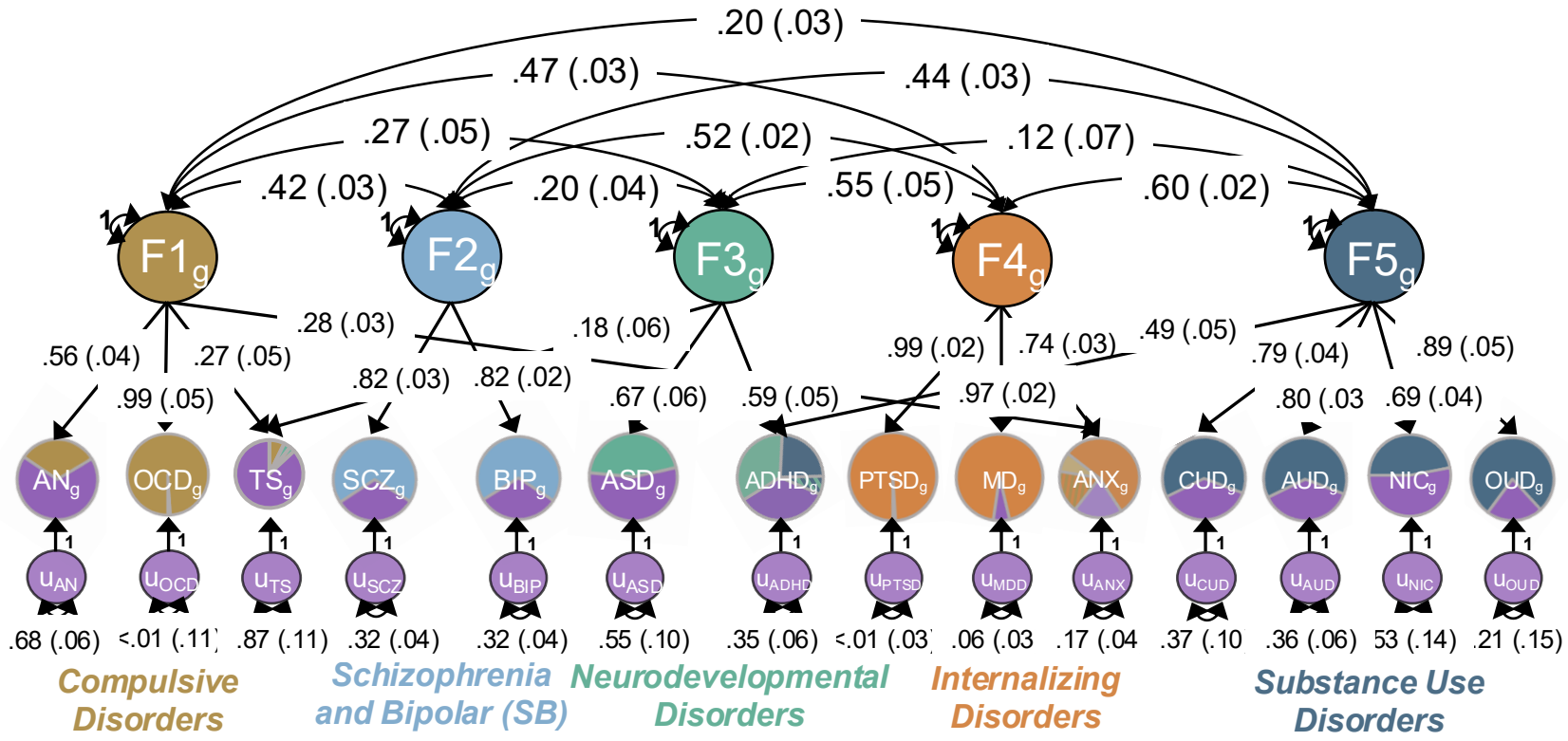



# $Q_{SNP}$ : QC Metric and a Result

- For some projects, you might just use $Q_{SNP}$ to prune your factor GWAS results for SNPs that are unlikely to operate via the factor

- In other instances, your traits might be so highly correlated, but currently considered distinct phenotypes, such that your research question includes investigating what SNPs differentiate your traits

If you have multiple factors, you can get a $Q_{SNP}$ specific metric for each factor. For example, you could have a SNP that fits one factor well, but is very disorder-specific for another factor




Grotzinger et al., 2025

# $Q_{SNP}$ : Better, Faster, Stronger

- It used to be that you had to estimate a separate independent pathways follow-up model.
  - This was slow, error prone, not computationally efficient

- $Q_{SNP}$ is now an argument for *userGWAS* that will automatically calculate this metric for every factor that is predicted by a SNP without estimating a follow-up model

# First five rows of the output

The parameter being estimates (SNP effect on factor)

Beta and SE of SNP effect on factor

Multivariate GWAS *p*-value

| | SNP | CHR | BP | MAF | A1 | A2 | lhs | op | rhs | free | label | est | SE | Z_Estimate | Pval_Estimate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rs10030871 | 4 | 68786 | 0.0765408 | T | C | F1 | ~ | SNP | 6 | | 9.072847e-04 | 0.004121185 | 0.22015141 | 0.8257532 |
| 2 | rs6599368 | 4 | 69567 | 0.0755467 | A | T | F1 | ~ | SNP | 6 | | 1.009838e-03 | 0.004110347 | 0.24568182 | 0.8059285 |
| 3 | rs7678633 | 4 | 69713 | 0.0755467 | G | A | F1 | ~ | SNP | 6 | | 1.051503e-03 | 0.004106863 | 0.25603562 | 0.7979233 |
| 4 | rs13130581 | 4 | 70392 | 0.0725646 | A | G | F1 | ~ | SNP | 6 | | 8.299511e-05 | 0.004146428 | 0.02001605 | 0.9840306 |
| 5 | rs13125929 | 4 | 71566 | 0.0725646 | T | C | F1 | ~ | SNP | 6 | | 4.857205e-04 | 0.004153705 | 0.11693667 | 0.9069102 |

| | chisq | chisq_df | chisq_pval | AIC | Q_SNP | Q_SNP_df | Q_SNP_pval | error | warning |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.905573 | 8 | 0.7676193 | 18.90557 | 3.622121 | 3 | 0.3052654 | 0 | 0 |
| 2 | 4.839514 | 8 | 0.7745834 | 18.83951 | 3.556060 | 3 | 0.3135638 | 0 | 0 |
| 3 | 4.901617 | 8 | 0.7680379 | 18.90162 | 3.618163 | 3 | 0.3057570 | 0 | 0 |
| 4 | 5.029482 | 8 | 0.7544204 | 19.02948 | 3.746031 | 3 | 0.2902263 | 0 | 0 |
| 5 | 5.787667 | 8 | 0.6710025 | 19.78767 | 4.504214 | 3 | 0.2119147 | 0 | 0 |

$Q_{SNP}$ *p*-value

Let’s go to the code and fill out the Qualtrics questions for multivariate GWAS

If you finish early, go onto next slides/section of code to play around more with specifying your own model at the genome-wide level

# Anthropometric traits

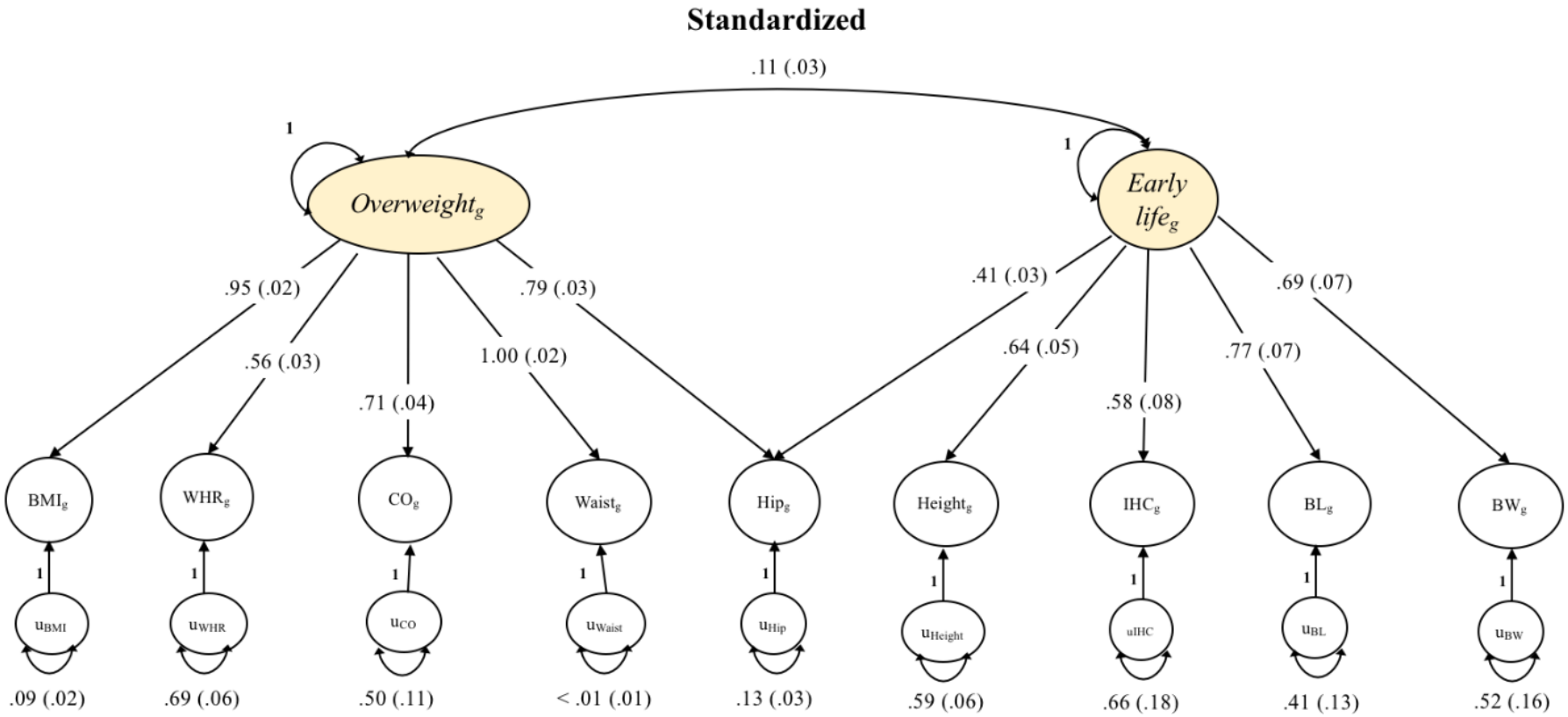
```
###IF theres time
load("Anthro_LDSC.RData")
colnames(anthro$S)

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

# Example model that could be fit to the data (but you should fit your own)



Pre-register your model by writing it down on paper beforehand.

The goal is ***not*** to just pick models until you get “a result”

Rather, we want to test theories or take a documented data-driven approach

# Final Notes

- Parallel processing and MPI for **`userGWAS`** 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!

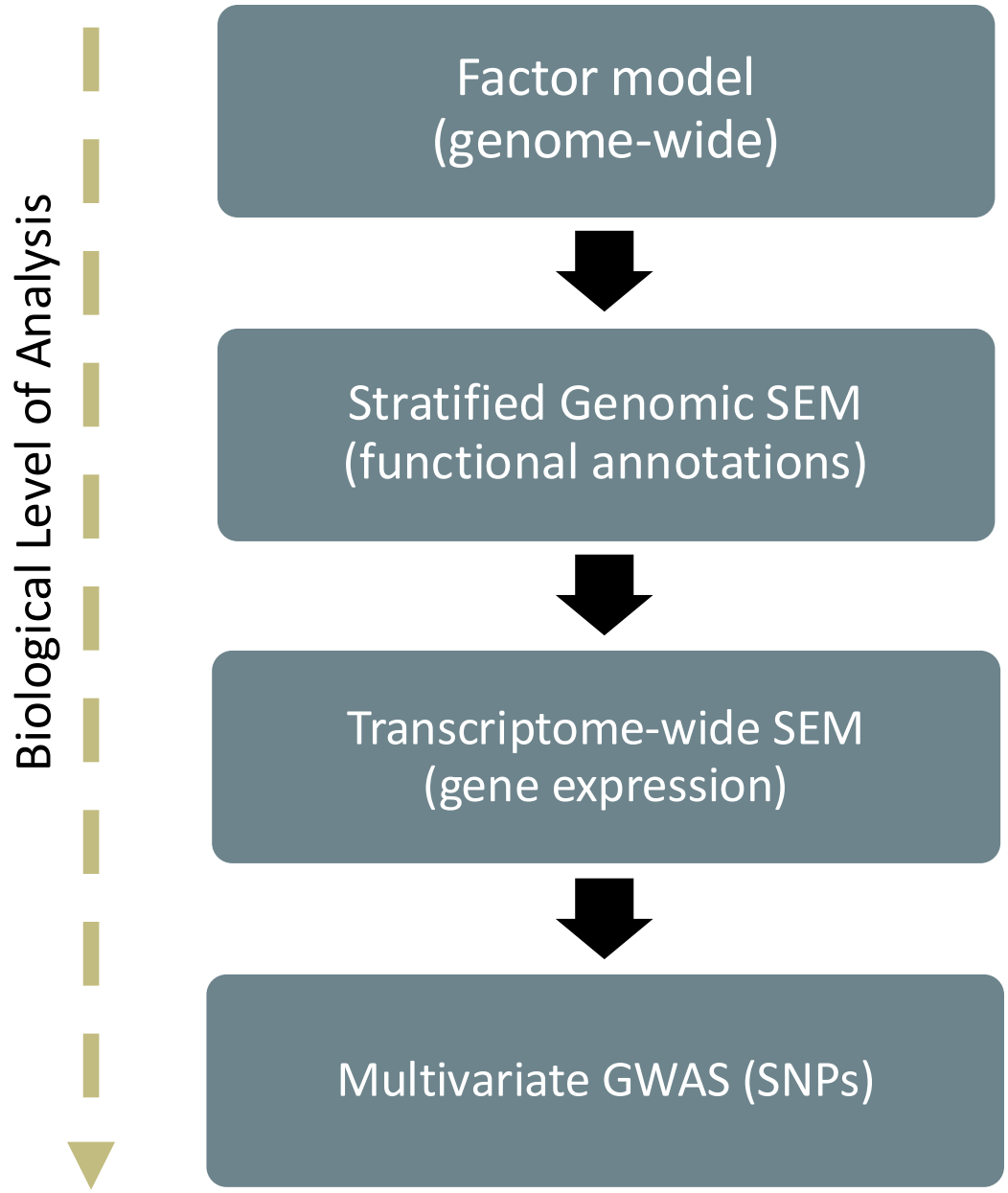
# Overview

**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




We've gone over factor models and multivariate GWAS but two additional sets of functions are available for functional annotations and gene expression analyses

# Resources

- See original paper at:

rdcu.be/bvn7t

- See GitHub at: https://github.com/GenomicSEM/GenomicSEM
- See tutorials at: https://github.com/GenomicSEM/GenomicSEM/wiki