LD Score Regression
Part II: Genetic Correlation

International Statistical Genetics Workshop

IBG, University of Colorado Boulder

Wednesday, 5th March 2025

MadhurBain Singh

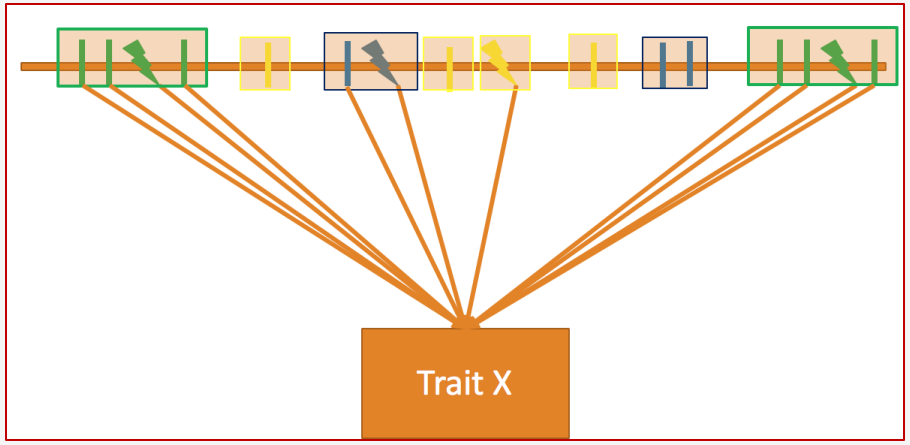
singhm18@vcu.edu

Virginia Institute for Psychiatric and Behavior Genetics

Virginia Commonwealth University, Richmond, VA

Many thanks to Benjamin Neale and Andrew Grotzinger.

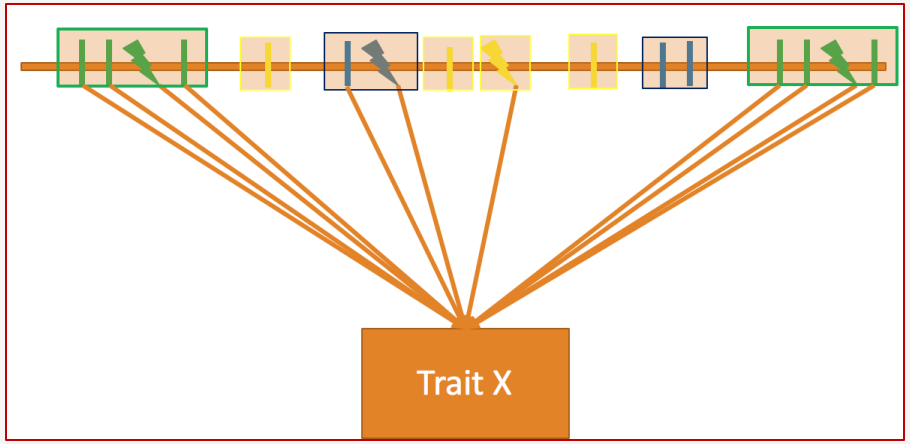
Research Question 1

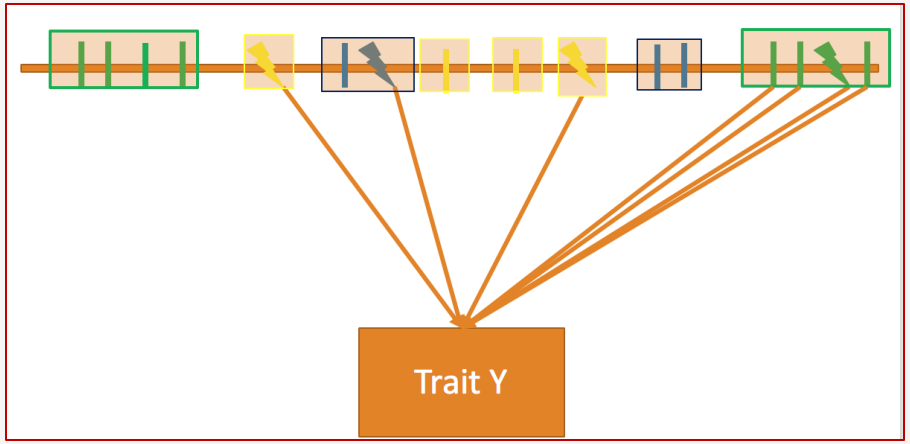


Proportion of phenotypic *variance*

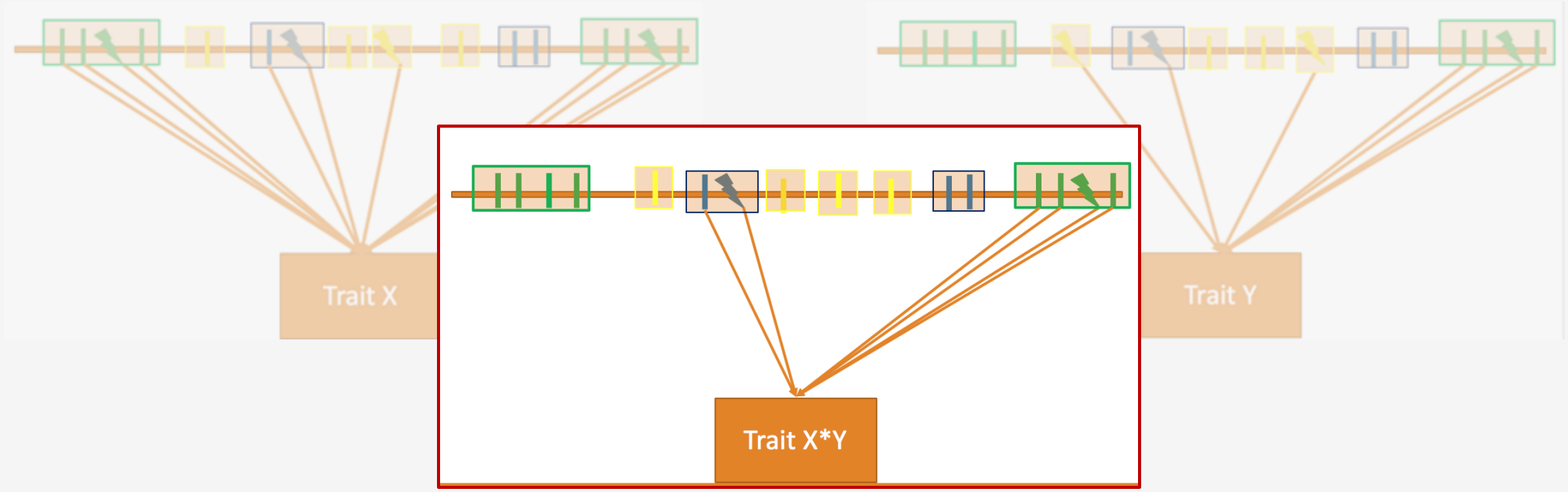
Explained by genome-wide common genetic variants

SNP Heritability


Trait X


Trait Y

Research Question 2

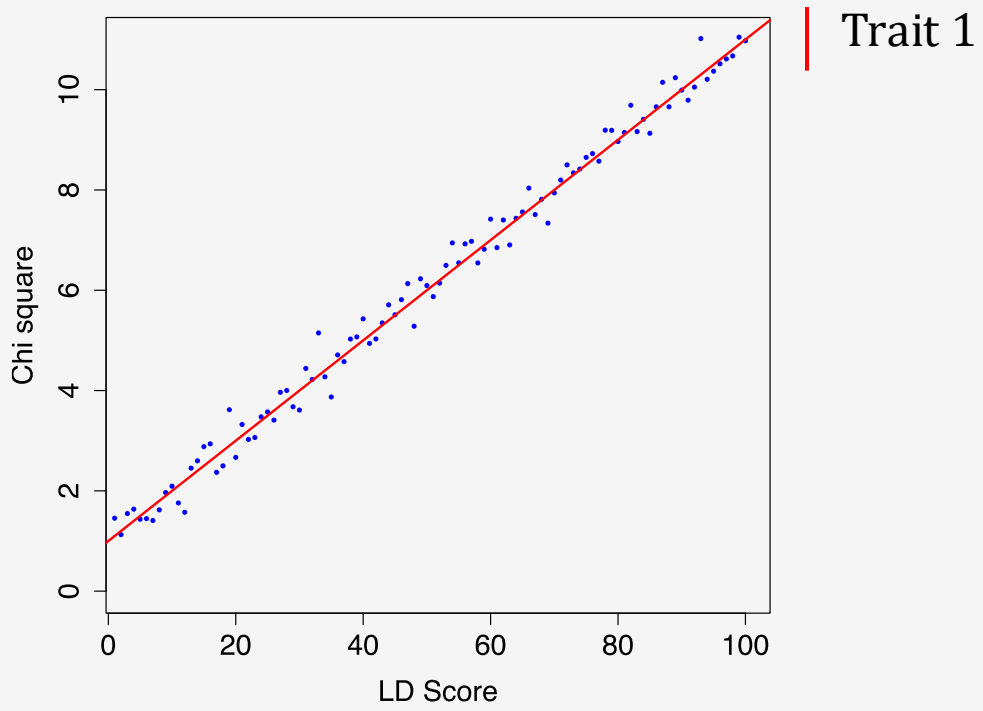


Proportion of phenotypic *covariance* of two traits

Explained by genome-wide common genetic variants

Genetic Covariance

LD Score Regression
SNP Heritability (Recap)

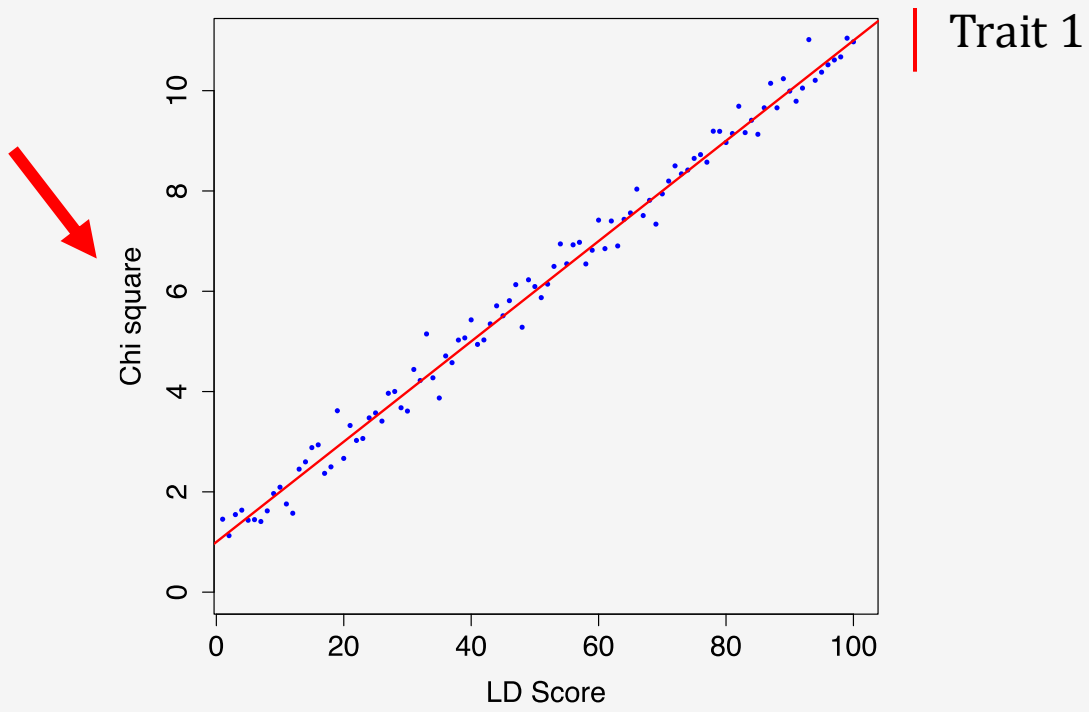


Slope estimates heritability

Regress GWAS chi-square on LD Scores

Across all SNPs (not just significant ones)

LD Score Regression
SNP Heritability

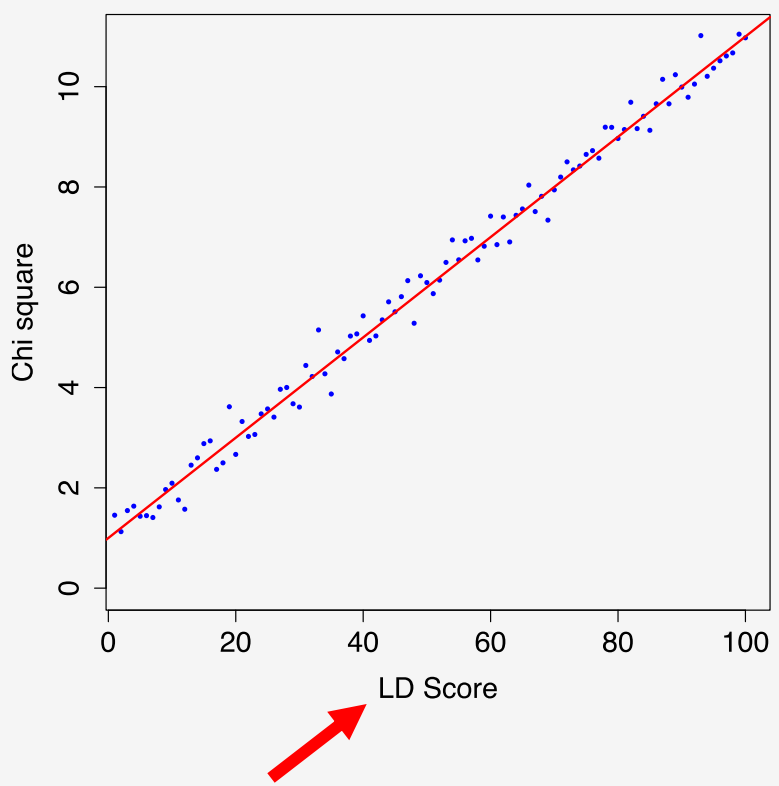


The expectation for the GWAS χ^2 given the LD score for the SNP j

$$E\left[{\chi_j}^2 \mid \ell_j\right] = \frac{Nh^2}{M}\ell_j + Na + 1$$

Note: ${\chi_j}^2$ is equivalent to $(Z_j)^2$.

LD Score Regression
SNP Heritability



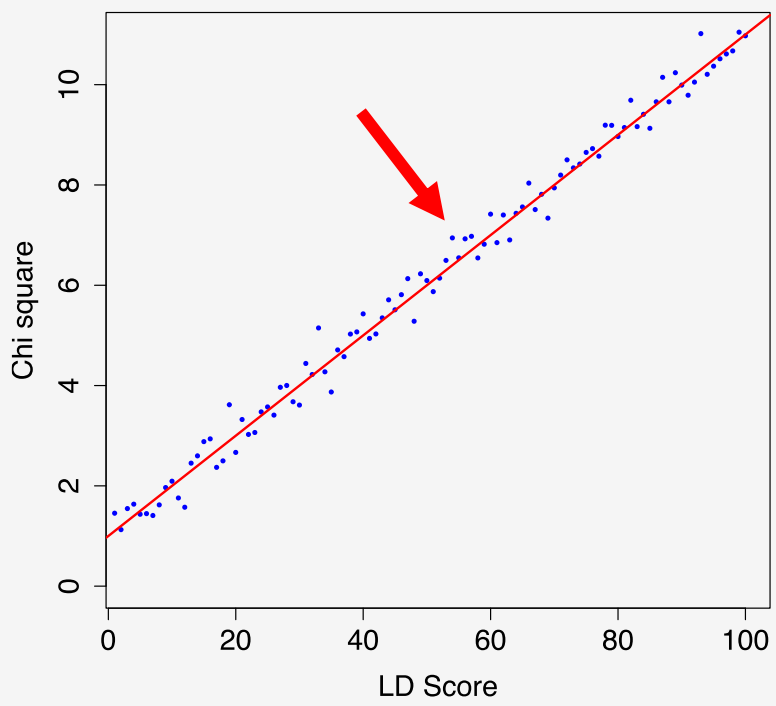
$$E[\chi_j^2 | \ell_j] = \frac{Nh^2}{M}\ell_j + Na + 1$$

<u>The Independent Variable</u>

The LD score of SNP *j*.

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

LD Score Regression
SNP Heritability



Slope estimates heritability

$$E[\chi_j^2 | \ell_j] = \frac{Nh^2}{M}\ell_j + Na + 1$$

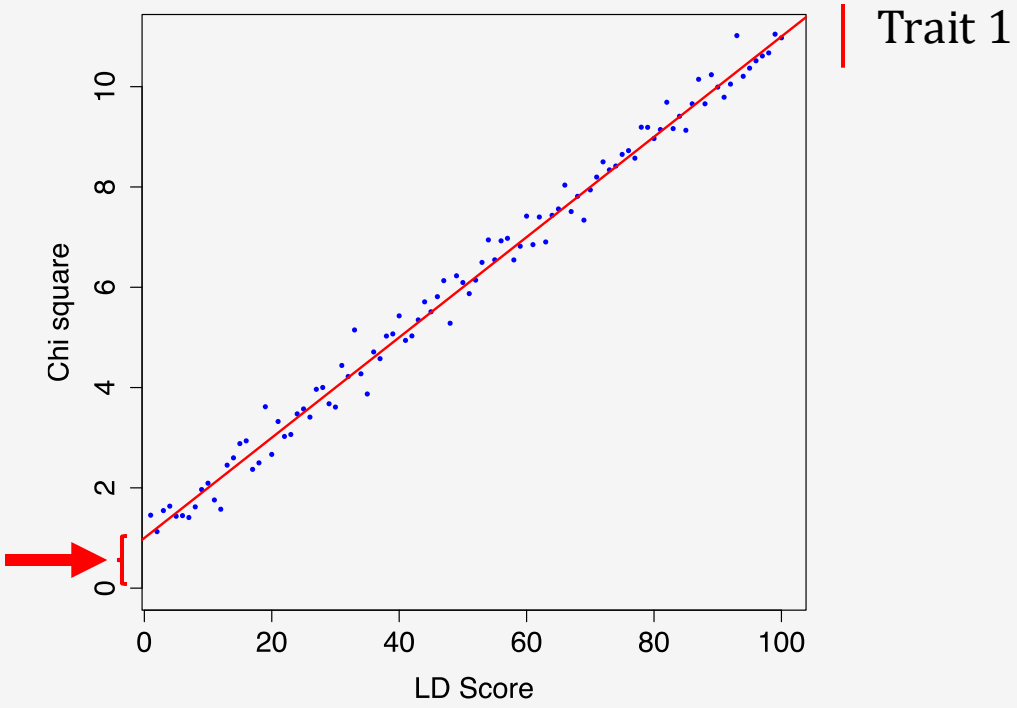
The Regression Slope

h^2 = SNP-based heritability

N = Sample size

M = Number of SNPs used in the LDSC analyses

LD Score Regression
SNP Heritability



Slope estimates heritability

$$E[\chi_j^2 | \ell_j] = \frac{Nh^2}{M}\ell_j + Na + 1$$

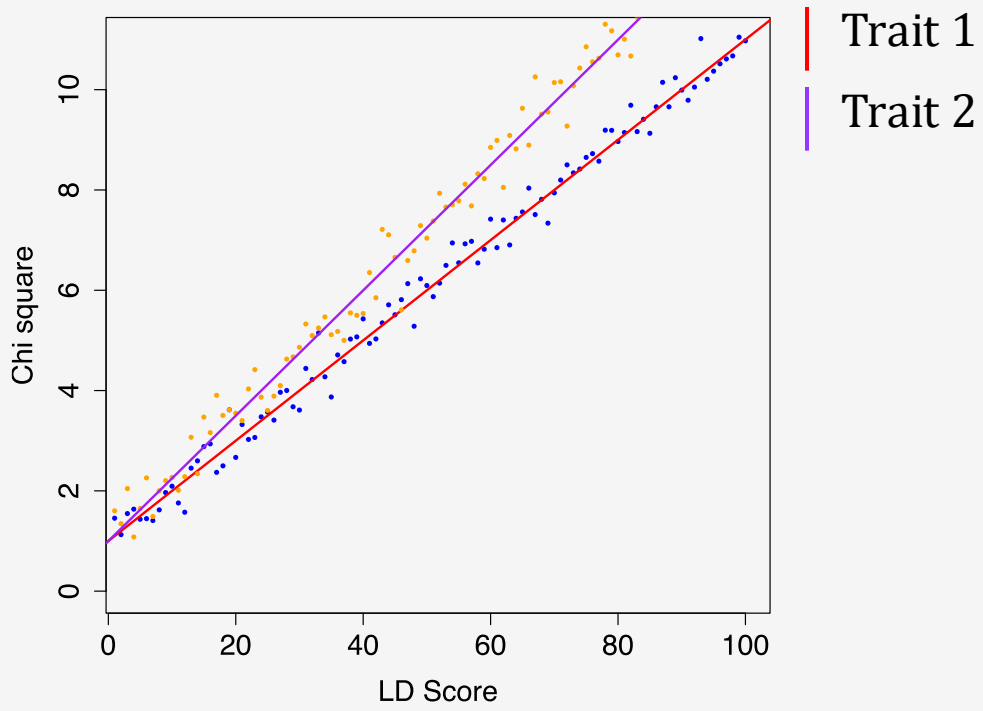
The Intercept

N = Sample size

a = Confounding biases

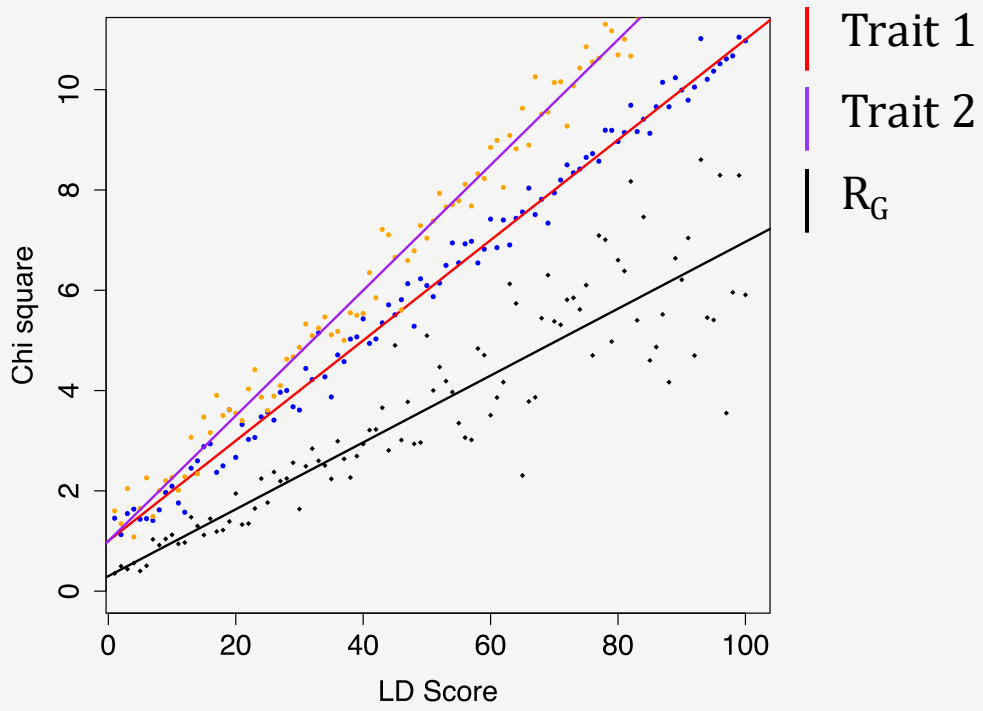
Under no confounding ($a \approx 0$), intercept ≈ 1.

LD Score Regression
SNP Heritability of 2 Traits



For two traits, we can estimate the respective SNP heritability.

LD Score Regression
Genetic Covariance Between 2 Traits



We can also estimate **genetic covariance** ("co-heritability") from the product of the Z-scores for the two traits.

Here,

$$\chi^2 = Z_1 \times Z_2$$

Univariate LDSC

$$E\left[\chi^2 \mid \ell_j\right] = \frac{Nh^2}{M}\ell_j + Na + 1$$

The expectation for the GWAS χ^2 ($= Z^2$) of one trait given the LD score for the SNP *j*

Bivariate LDSC

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

Univariate LDSC: $E\left[\chi^2|\,\ell_j\right] = \frac{Nh^2}{M}\ell_j + Na + 1$

Bivariate LDSC

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

The expectation for the product of Z-statistics of two traits given the LD score for the SNP *j*

Bivariate LDSC

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

The “independent” variable is the same as before
- LD score for a given SNP *j*

Bivariate LDSC

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

ρ_g = Genetic covariance

$\sqrt{N_1N_2}$ = Square root of the sample sizes of trait 1 and trait 2

M = the number of SNPs

Bivariate LDSC

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

Bivariate LDSC intercept

Protects against bias from ***shared*** **population stratification** & **sample overlap**

$\sqrt{N_1N_2}a$ → *Shared* sources of confounding across the two GWASs.

$\frac{\rho N_s}{\sqrt{N_1N_2}}$ → Phenotypic correlation (ρ) among overlapping participant samples

weighted by proportional sample overlap ($\frac{N_s}{\sqrt{N_1N_2}}$)

Bivariate LDSC

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

Bivariate LDSC intercept

Protects against bias from *shared* population stratification & sample overlap

Under no *shared* confounding (a ≈ 0) and no sample overlap ($N_s \approx 0$),

bivariate LDSC intercept ≈ 0.

Bivariate LDSC

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

Note:
We assume identical LD scores for both traits.
So, both GWAS samples must have similar LD structure.

Therefore, only valid for ***within-ancestry*** analyses.

Genetic Correlation

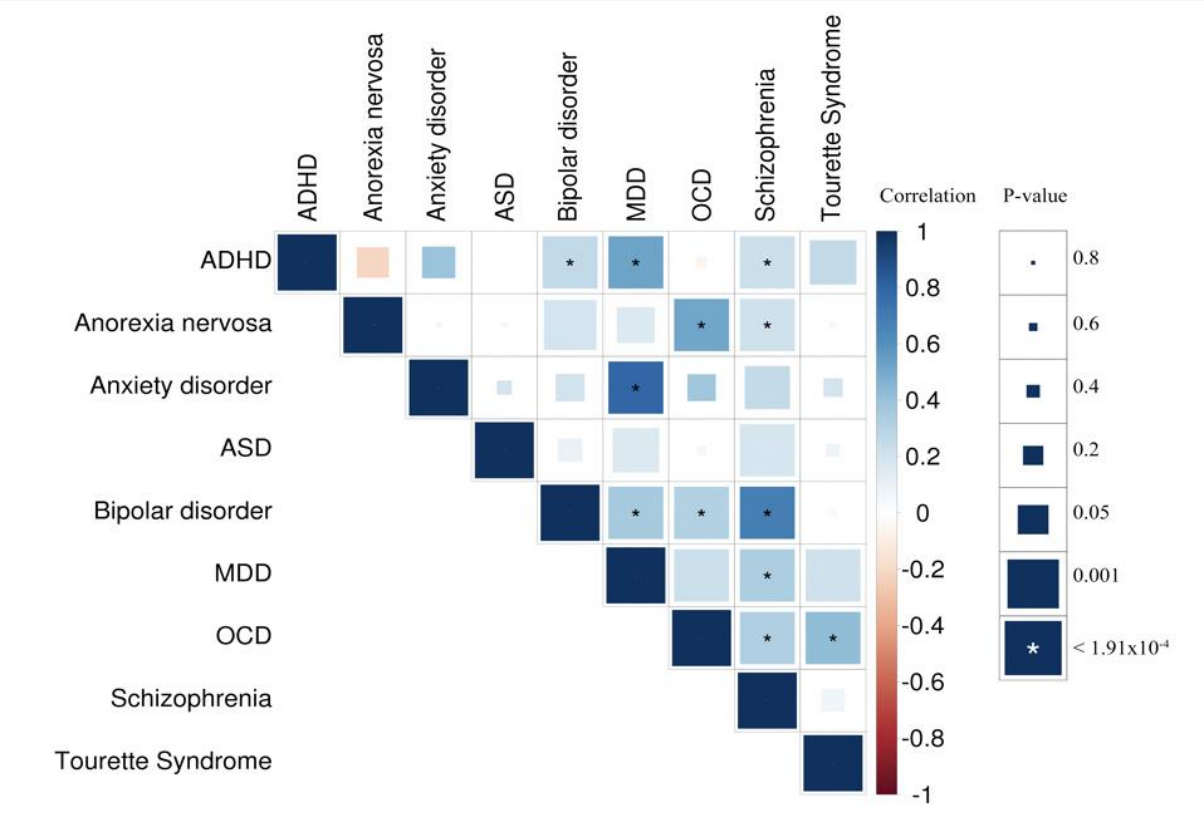
$$E\left[z_{1j}z_{2j}\,|\ell_j\right] = \frac{\sqrt{N_1N_2}\,\boldsymbol{\rho_g}}{M}\ell_j + \sqrt{N_1N_2}a + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

$$r_g = \frac{\rho_g}{\sqrt{h^2_{Y1}h^2_{Y2}}}$$

The amount of genetic overlap *on the standardized scale*.

Example

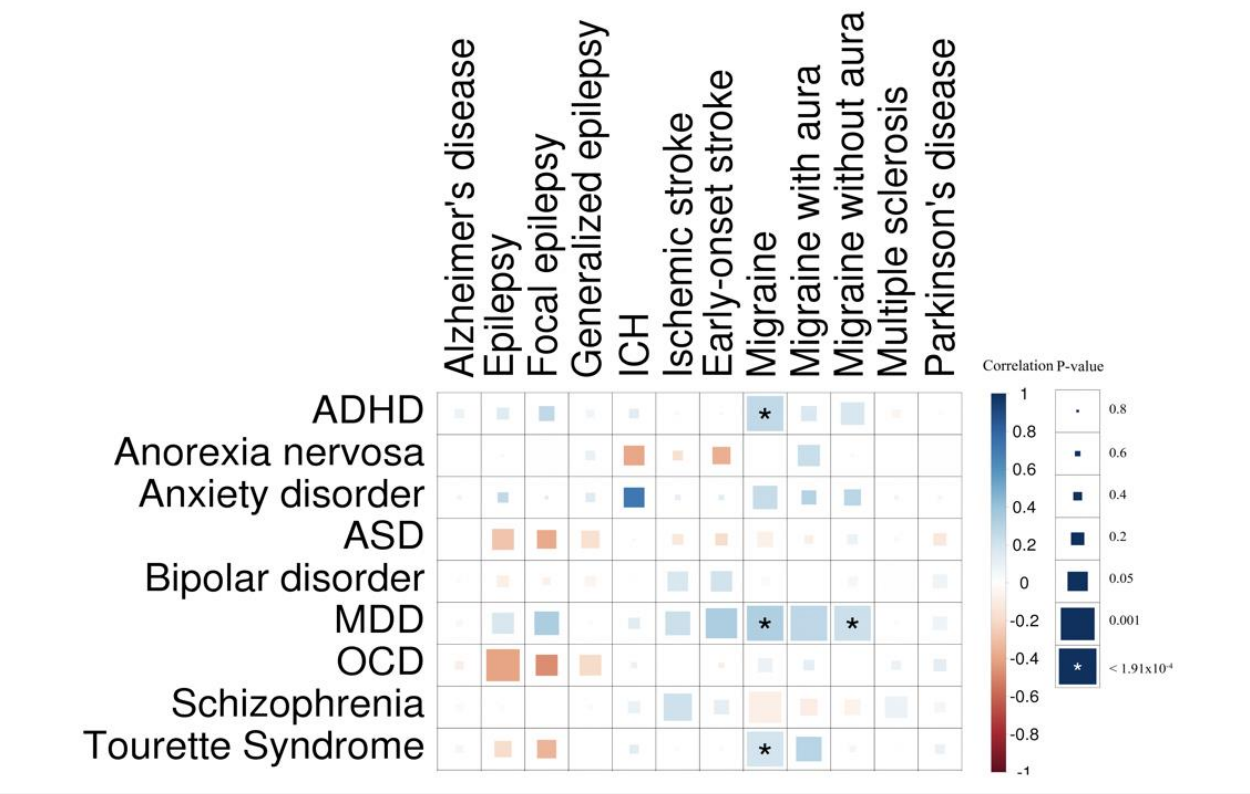
Across Psychiatric Disorders



The Brainstorm Consortium et al. (2018). DOI:10.1126/science.aap8757

Example

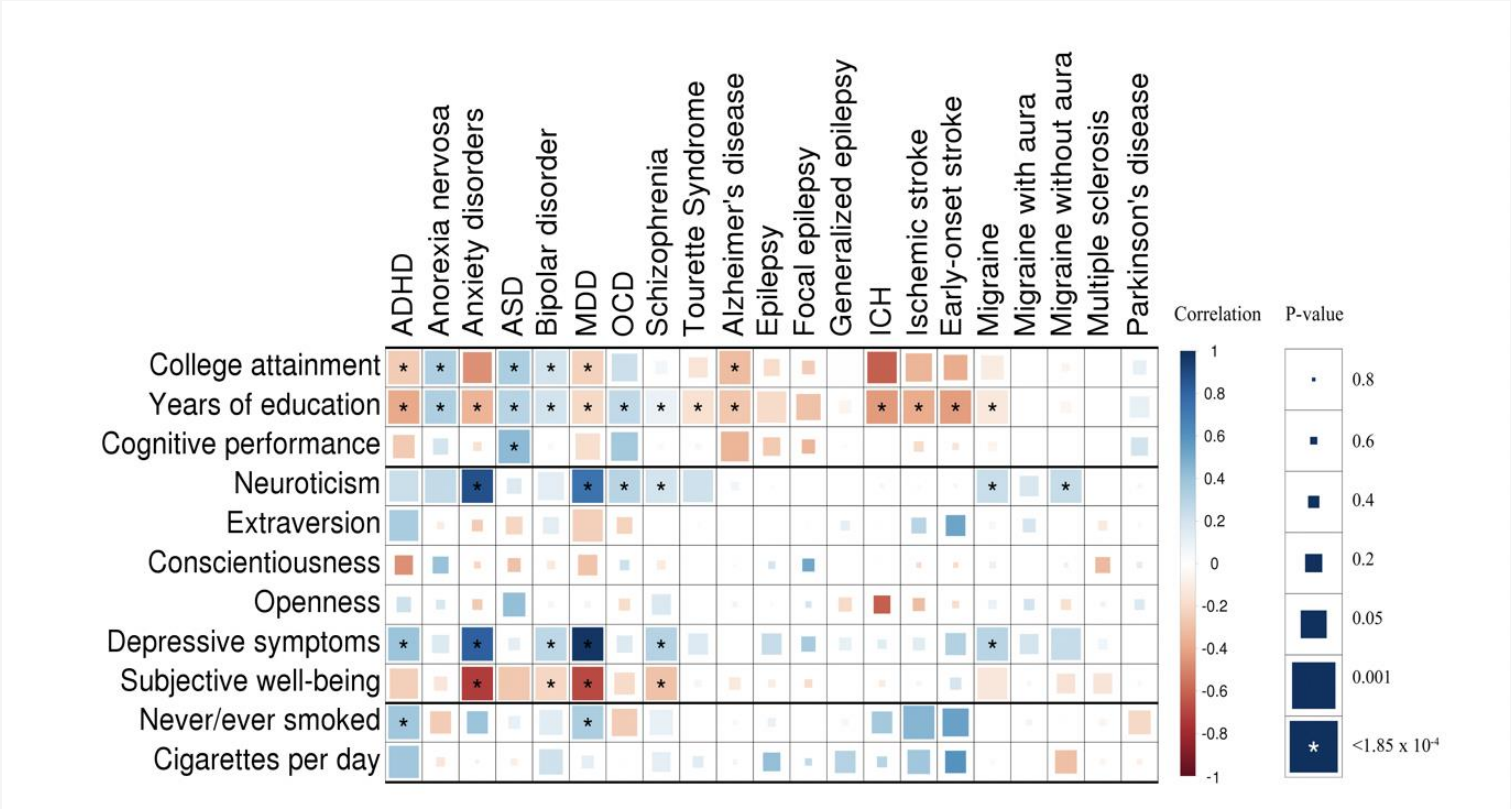
Between Psychiatric & Neurological Disorders



The Brainstorm Consortium et al. (2018). DOI:10.1126/science.aap8757

Example

Correlations of Psychiatric & Neurological Disorders with Behavioral-Cognitive Traits



The Brainstorm Consortium et al. (2018). DOI:10.1126/science.aap8757

Key Points

- LDSC allows us to estimate the genetic correlation between traits using only the GWAS sum stats.
- Traits need *not* be assessed in the same sample.
- Estimates are robust to bias from (shared) population stratification and sample overlap.
- Both GWASs need to be performed in samples with similar genetic ancestry

Key Points

We can estimate genetic correlations

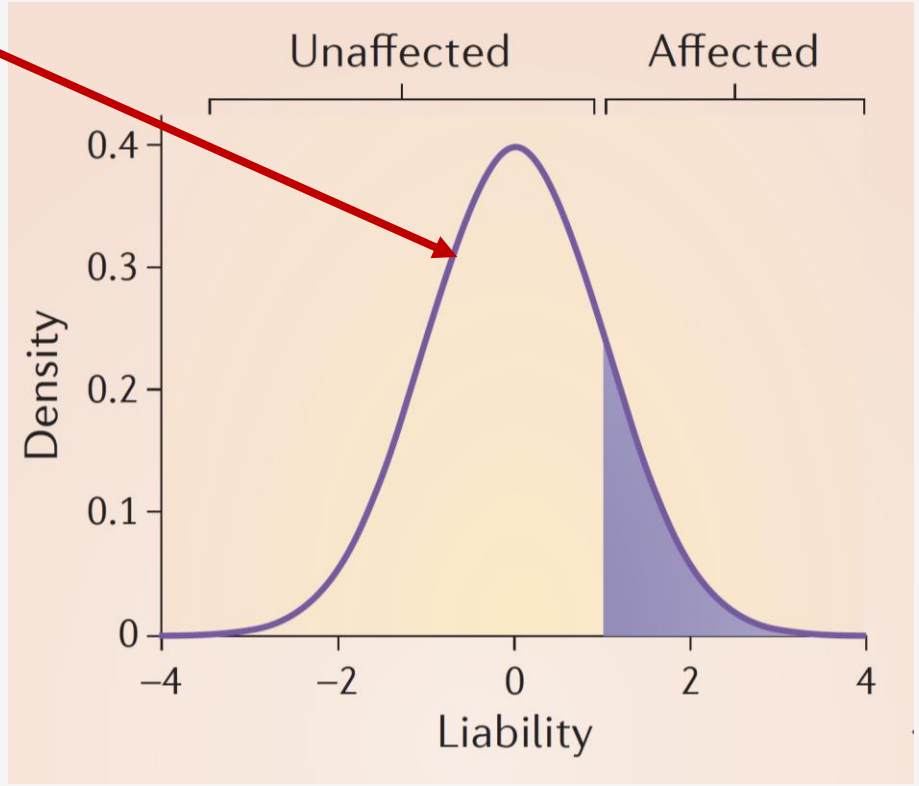
- Between quantitative traits
- Between binary traits
- Between quantitative and binary traits

For **binary traits**, the heritability may be estimated on

- Observed scale
- Liability scale

Liability Threshold Model

- Individuals have a **latent continuous liability** underlying complex binary traits.
- The liability is assumed to have a **normal distribution** in the population, with a mean of 0 and a variance of 1.

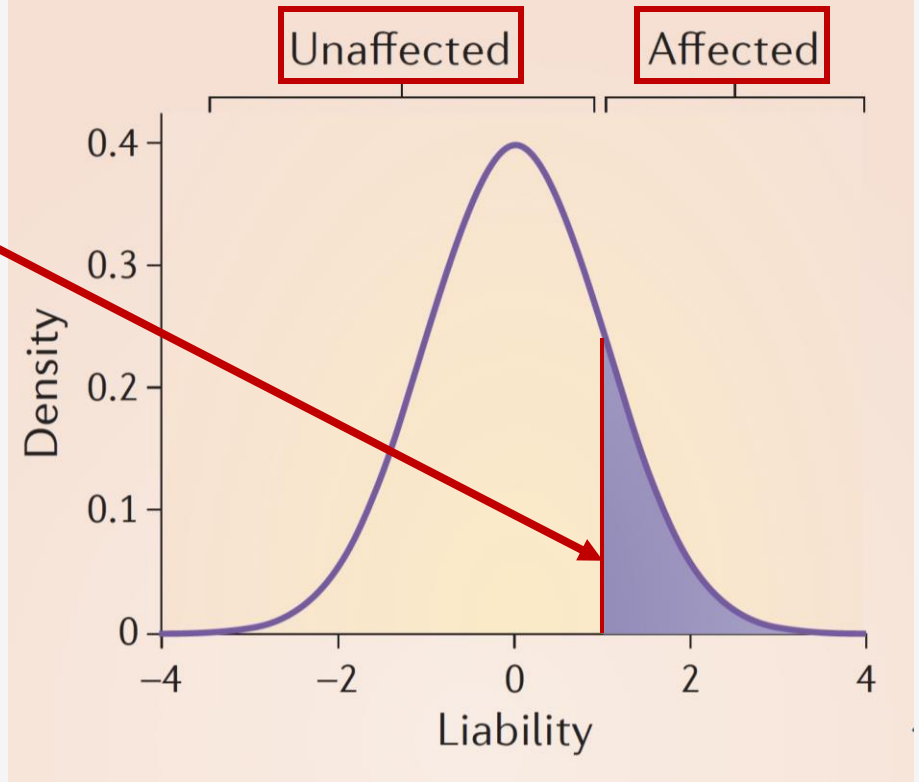


Witte, J., Visscher, P. & Wray, N. *Nat Rev Genet* **15**, 765–776 (2014).
https://doi.org/10.1038/nrg3786

Liability Threshold Model

There exists a certain **threshold, *t***, on the liability scale.

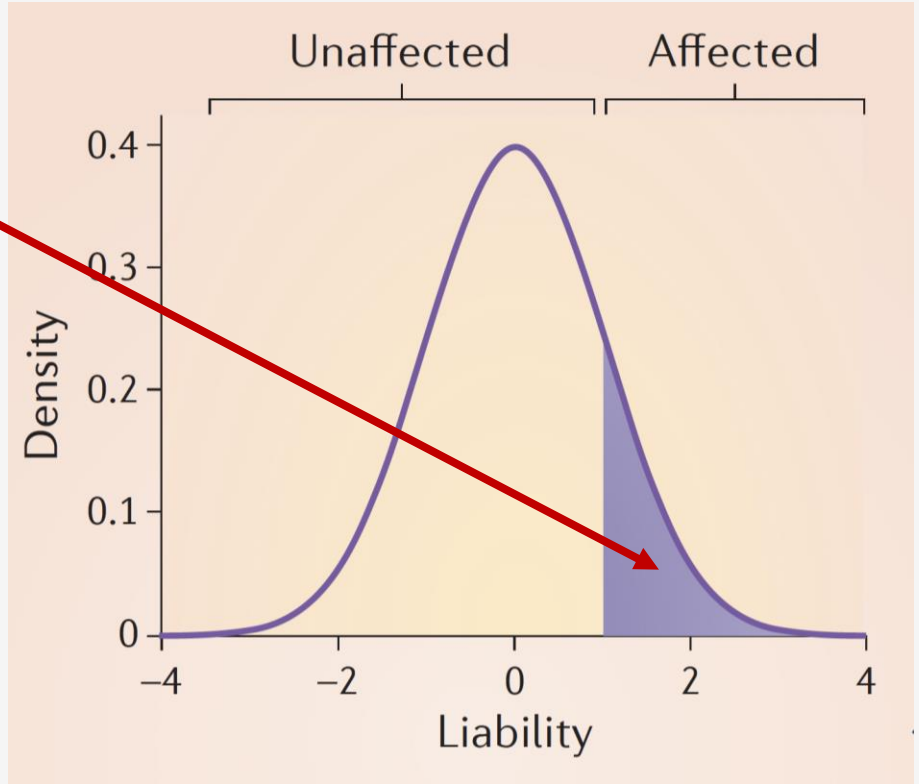
- All individuals above this threshold exhibit the trait. [“affected/cases”]
- All individuals below this threshold do not exhibit the trait. [“unaffected/controls”]



Witte, J., Visscher, P. & Wray, N. *Nat Rev Genet* **15**, 765–776 (2014).
https://doi.org/10.1038/nrg3786

Liability Threshold Model

Population prevalence (K) = Area under the distribution curve to the right of the threshold, *t*.



If we know the population prevalence, we may estimate *t*.

- By using the probit function.

$$probit(p) = \Phi^{-1}(p)$$

NB: probit is the inverse of the cumulative distribution function of the standard normal distribution (ϕ).

Witte, J., Visscher, P. & Wray, N. *Nat Rev Genet* **15**, 765–776 (2014). https://doi.org/10.1038/nrg3786

From population prevalence of a binary disease status
To threshold on the latent continuous liability

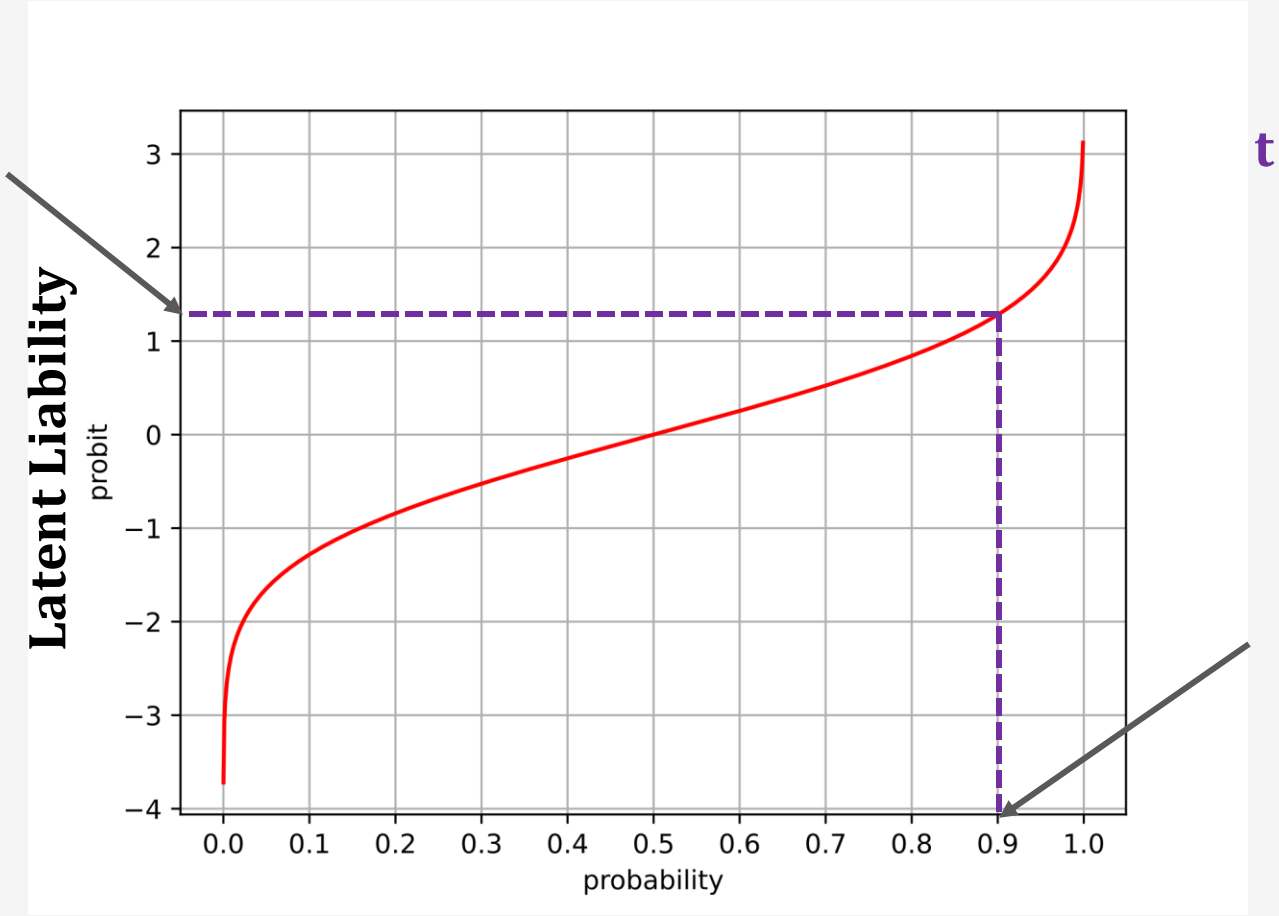
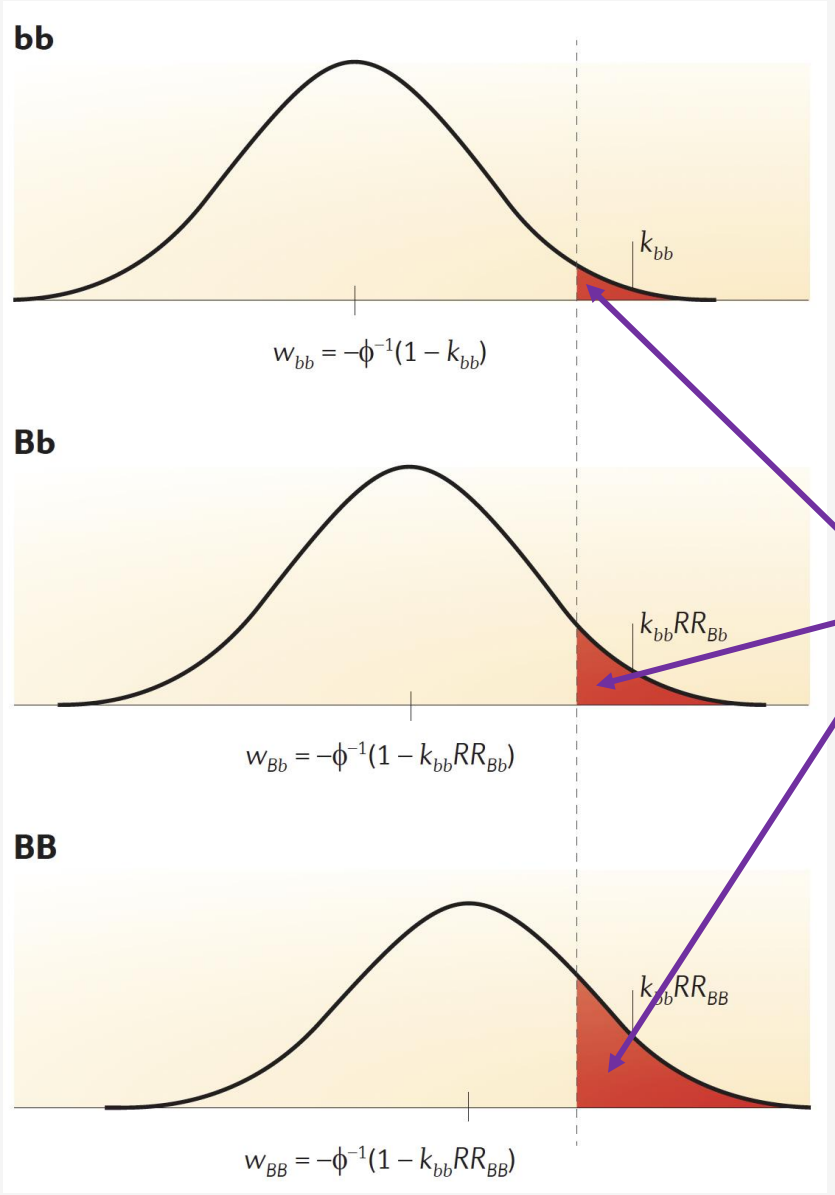


Image credit: https://en.wikipedia.org/wiki/Probit#/media/File:Probit_plot.svg

Observed scale (visualized under the liability threshold model)



Observed Scale

- **Multiplicative**
- **Non-Linear**

The observed disease risk (prevalence among individuals with that genotype) increases multiplicatively.

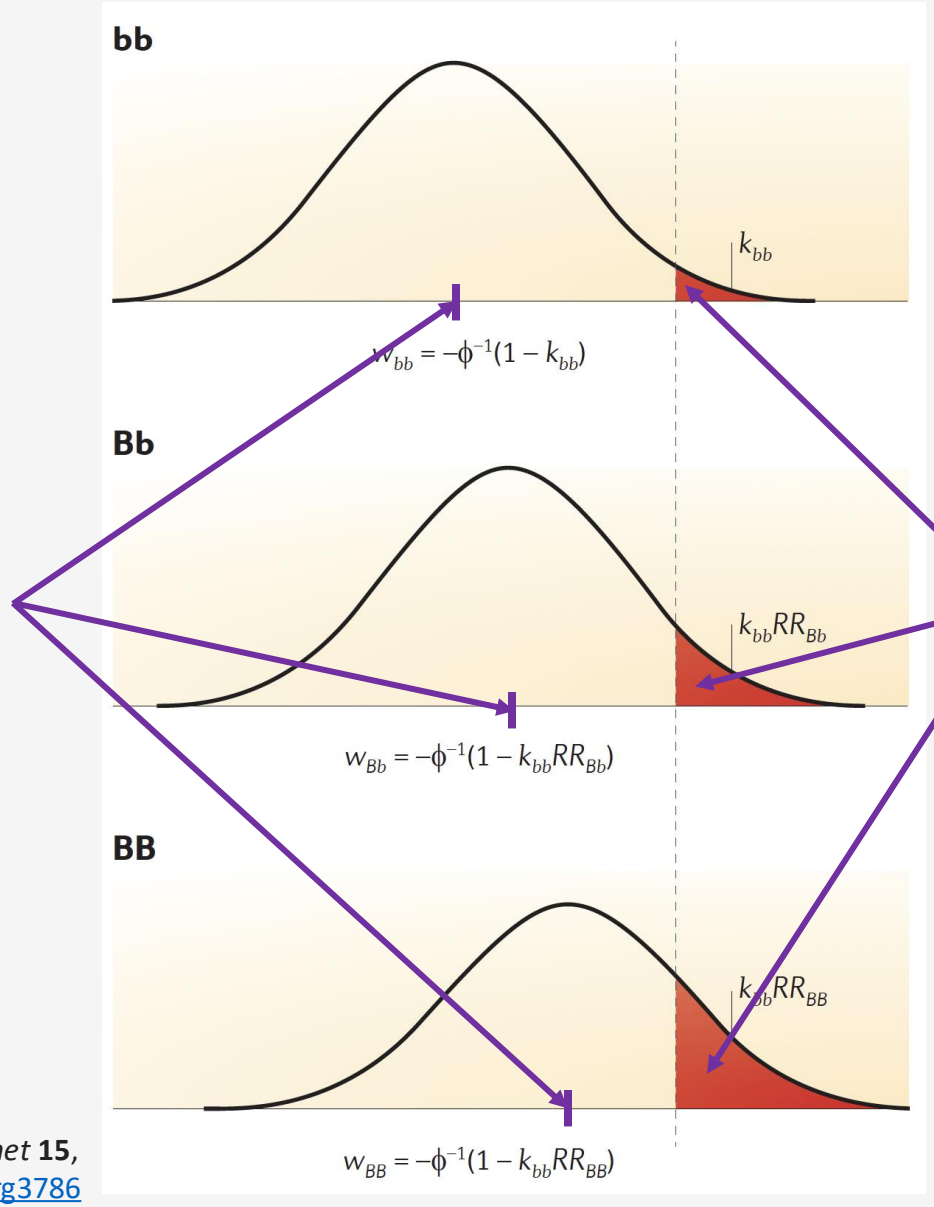
Witte, J., Visscher, P. & Wray, N. *Nat Rev Genet* **15**, 765–776 (2014). https://doi.org/10.1038/nrg3786

Observed scale to Liability scale

Liability Scale

- **Additive**
- **Linear**

The mean latent liability (among individuals with that genotype) increases additively.

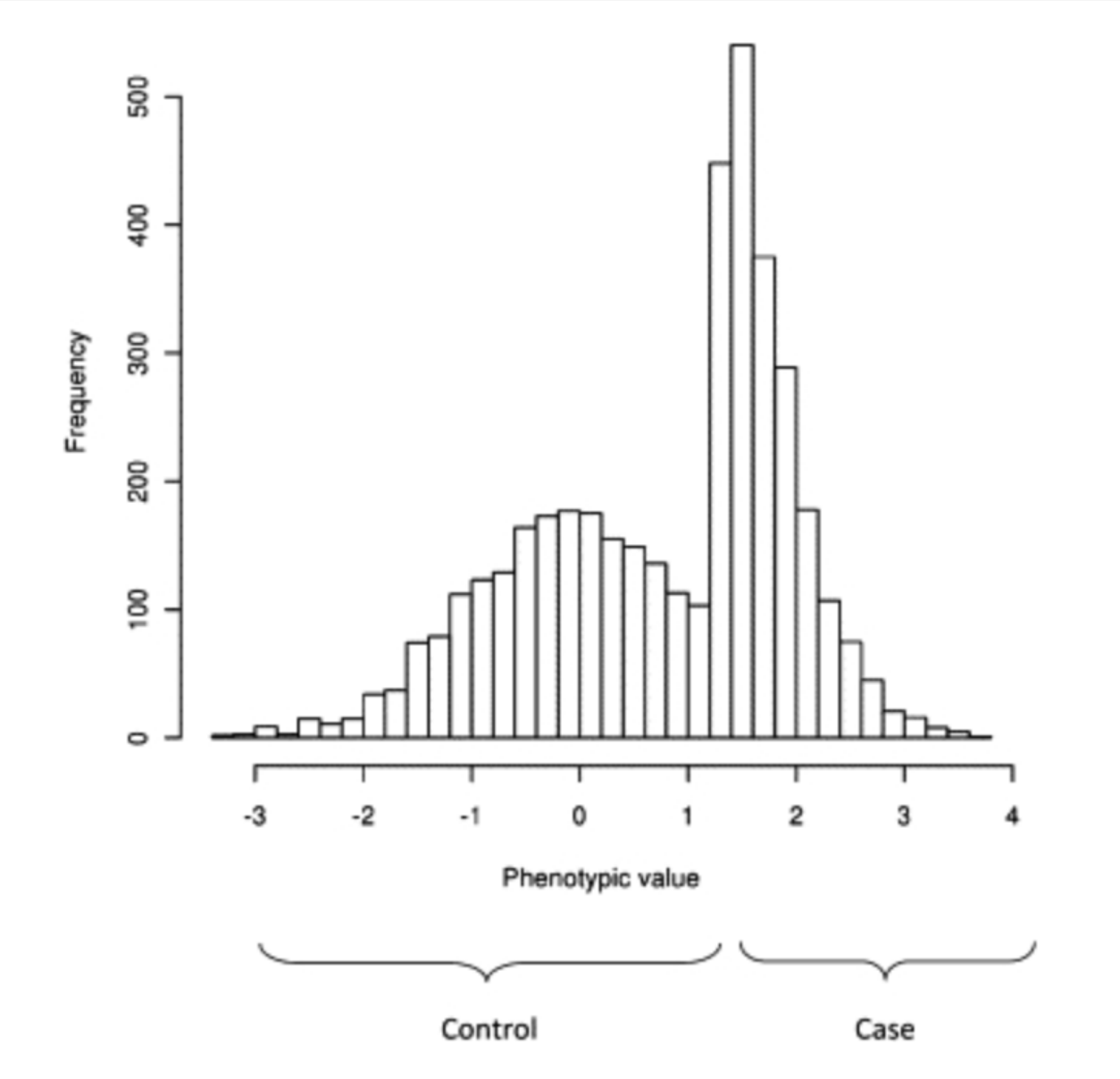


Observed Scale

- **Multiplicative**
- **Non-Linear**

The observed disease risk (prevalence among individuals with that genotype) *still* increases multiplicatively.

Witte, J., Visscher, P. & Wray, N. *Nat Rev Genet* **15**, 765–776 (2014). https://doi.org/10.1038/nrg3786



"In case-control studies the proportion of cases is usually (much) larger than the prevalence in the population yet **estimates of genetic variation are most interpretable if they are not biased by this ascertainment.**"

Lee et al. (2011)

From observed-scale heritability
To liability-scale heritability

$$h_l^2 = \hat{h}_o^2 \frac{K\,(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

We need to specify two parameters:

- **Sample prevalence (*P*)**
- **Population prevalence (*K*)**
  - Note that *z* is computed from *K*, so need not be specified.

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). *Am J Hum Genet*, *88*(3), 294–305. https://doi.org/10.1016/j.ajhg.2011.02.002

From observed-scale heritability To liability-scale heritability

$$h_l^2 = \hat{h}_o^2 \frac{K\,(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

The first part of this equation converts the heritability estimate to the liability scale.

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). *Am J Hum Genet*, *88*(3), 294–305. https://doi.org/10.1016/j.ajhg.2011.02.002

From observed-scale heritability
To liability-scale heritability

$$h_l^2 = \hat{h}_o^2 \frac{K\,(1-K)}{z^2} \boxed{\frac{K(1-K)}{P(1-P)}}$$

The second part of this equation performs the correction for ascertainment.

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). *Am J Hum Genet*, *88*(3), 294–305. https://doi.org/10.1016/j.ajhg.2011.02.002

NOTE!

Conversion to the liability scale influences the estimates of

- SNP Heritability
- Genetic *Covariance*

However, it *DOES NOT* influence

- Genetic *Correlation*
  - Both the numerator and the denominator are on the same scale.
- LDSC Intercept

$$r_g = \frac{\rho_g}{\sqrt{h^2_{Y1} h^2_{Y2}}}$$

Cohort-specific ascertainment in Meta-Analysis

$$h_l^2 = \hat{h}_o^2 \frac{K\,(1-K)}{z^2} \frac{K(1-K)}{\sum_i P_i(1-P_i)}$$

We need to calculate the sum of the sample prevalence of each contributing cohort.

The ascertainment calculated using total cases and controls is not the same as ascertainment calculated within each cohort.

See GenomicSEM Wiki page for more details
https://github.com/GenomicSEM/GenomicSEM/wiki/2.1-Calculating-Sum-of-Effective-Sample-Size-and-Preparing-GWAS-Summary-Statistics

Effective Sample Size (N_{Eff})

$$N_{Eff} = 4\boldsymbol{P}(1 - \boldsymbol{P})\boldsymbol{N}$$

P = Sample prevalence

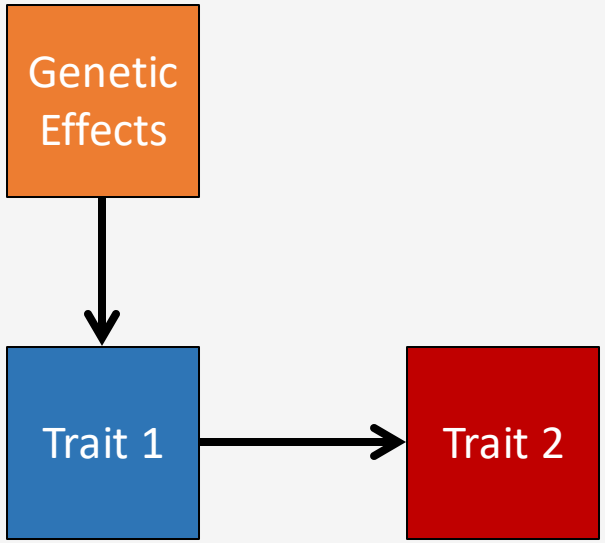
N = Total sample size (cases + controls)

It is the sample size we would have had if the study design was balanced (50% cases and 50% controls).

To account for cohort-specific ascertainment in GWAS meta-analysis, calculate the effective sample size of each cohort and then sum across cohorts.

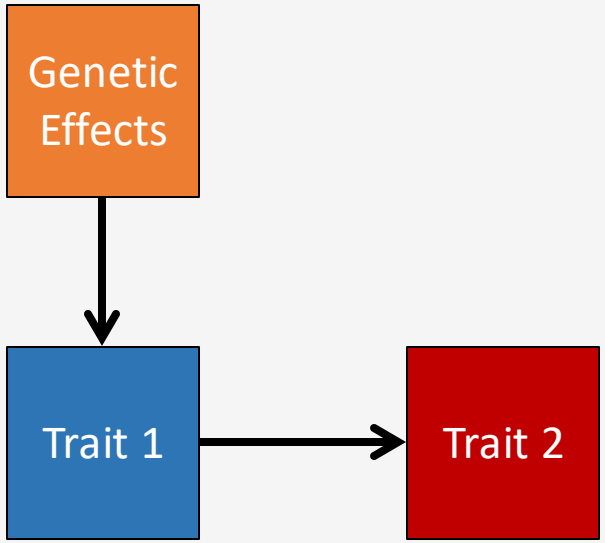
Why Are Two Traits Genetically Correlated?

Why Are Two Traits Genetically Correlated?

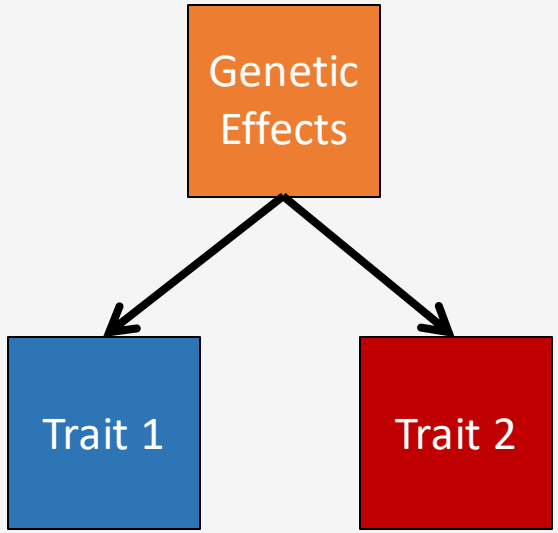


Trait 1 has a causal effect on Trait 2
(Vertical Pleiotropy)

Why Are Two Traits Genetically Correlated?

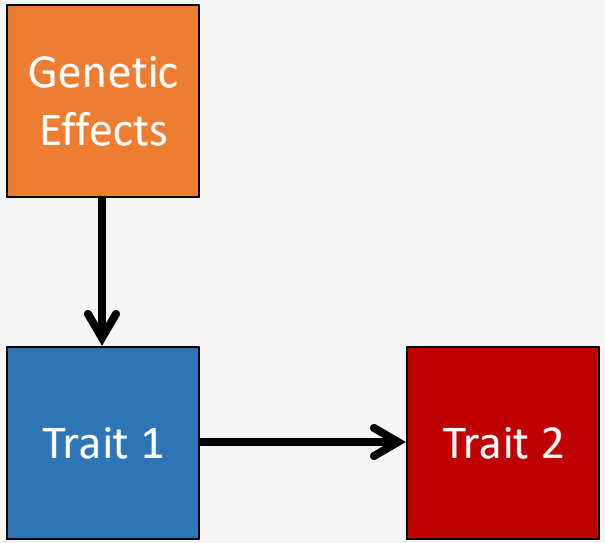


Trait 1 has a causal effect on Trait 2
(Vertical Pleiotropy)

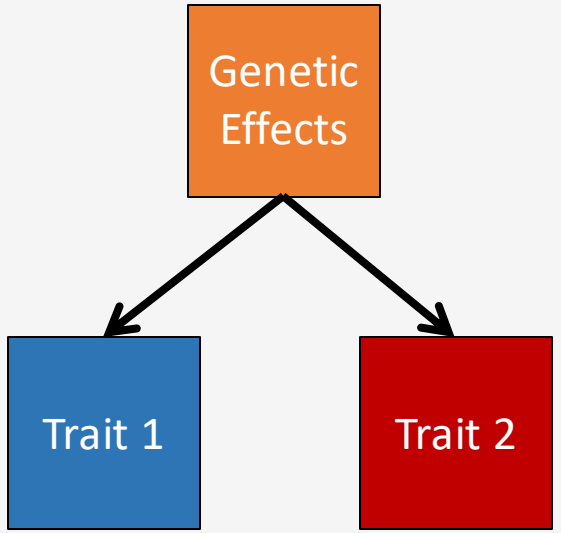


Genetic effects influence Trait 1 and Trait 2
(Horizontal Pleiotropy)

Why Are Two Traits Genetically Correlated?



Trait 1 has a causal effect on Trait 2
(Vertical Pleiotropy)

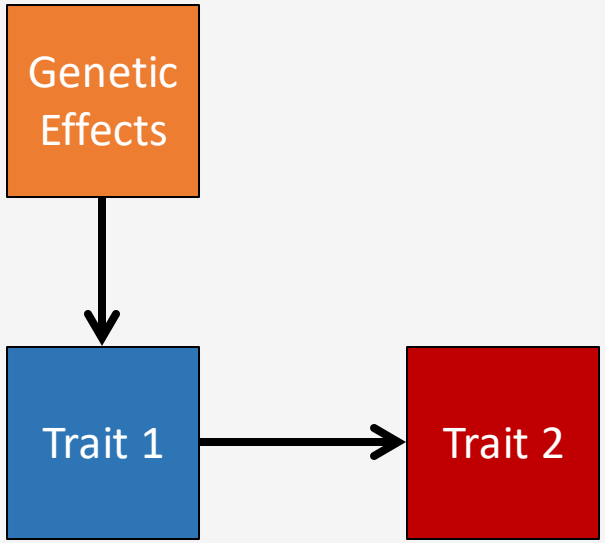


Genetic effects influence Trait 1 and Trait 2
(Horizontal Pleiotropy)

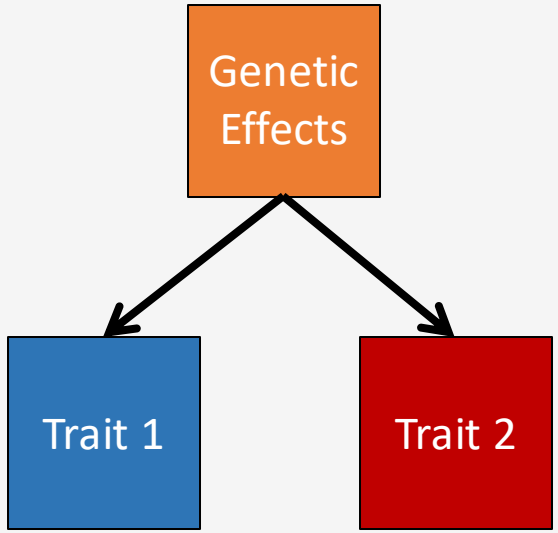
May be due to an unmeasured (or yet unknown) intermediate trait

→ Causal effects on both Trait 1 and Trait 2

Why Are Two Traits Genetically Correlated?



Trait 1 has a causal effect on Trait 2
(Vertical Pleiotropy)



Genetic effects influence Trait 1 and Trait 2
(Horizontal Pleiotropy)

Further research to test these hypotheses, e.g.,

- Mendelian Randomization
- Genomic SEM

Practical for Continuous Traits

Only TWO Primary Steps to Run LDSC

We are using **`{GenomicSEM}`** library in R to run LDSC

1. Munge the summary statistics: **`munge()`**

   munge = convert raw data from one form to another

2. Run LD-Score Regression: **`ldsc()`**

We will be running LDSC for both European and East Asian Samples

Using European GWAS sumstats for:

Height (Yengo et al., 2022)

BMI from GIANT + UKB

Using East Asian GWAS sumstats for:

Height (Yengo et al., 2022)

BMI from Biobank Japan

BIOBANK JAPAN BioBank Japan PheWeb (PheWeb.jp)

munge .log file

The two sum stats files are munged separately

BMI

```
Munging file: GIANT_UKB_BMI_EUR_chr1.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 BETA column as the effect column.
Interpreting the P column as the P column.
Interpreting the N column as the N column.
Interpreting the MAF column as the MAF column.
Interpreting the SE column as the SE column.
Merging file:GIANT_UKB_BMI_EUR_chr1.txt with the reference file:eur_w_ld_chr/w_hm3.snplist
175116 rows present in the full GIANT_UKB_BMI_EUR_chr1.txt summary statistics file.
94386 rows were removed from the GIANT_UKB_BMI_EUR_chr1.txt summary statistics file as the rs-ids for these rows were not present in the reference file.
No INFO column, cannot filter on INFO, which may influence results
4 rows were removed from the GIANT_UKB_BMI_EUR_chr1.txt summary statistics file due to missing MAF information or MAFs below the designated threshold of0.01
80726SNPs are left in the summary statistics file GIANT_UKB_BMI_EUR_chr1.txt after QC.
I am done munging file: GIANT_UKB_BMI_EUR_chr1.txt
The file is saved as BMI.sumstats.gz in the current working directory.
```

Height

```
Munging file: Yengo_Height_EUR_chr1.txt
Interpreting the RSID column as the SNP column.
Interpreting the EFFECT_ALLELE column as the A1 column.
Interpreting the OTHER_ALLELE column as the A2 column.
Interpreting the BETA column as the effect column.
Interpreting the P column as the P column.
Interpreting the N column as the N column.
Interpreting the MAF column as the MAF column.
Interpreting the SE column as the SE column.
Merging file:Yengo_Height_EUR_chr1.txt with the reference file:eur_w_ld_chr/w_hm3.snplist
96851 rows present in the full Yengo_Height_EUR_chr1.txt summary statistics file.
7910 rows were removed from the Yengo_Height_EUR_chr1.txt summary statistics file as the rs-ids for these rows were not present in the reference file.
No INFO column, cannot filter on INFO, which may influence results
0 rows were removed from the Yengo_Height_EUR_chr1.txt summary statistics file due to missing MAF information or MAFs below the designated threshold of0.01
88941SNPs are left in the summary statistics file Yengo_Height_EUR_chr1.txt after QC.
I am done munging file: Yengo_Height_EUR_chr1.txt
The file is saved as Height.sumstats.gz in the current working directory.
```

`ldsc` .log file

Three parts (for two traits)

[1/3]

```
Estimating heritability [1/3] for: BMI.sumstats.gz
Heritability Results for trait: BMI.sumstats.gz
Mean Chi^2 across remaining SNPs: 4.0559
Lambda GC: 2.7889
Intercept: 0.9355 (0.0803)
Ratio: -0.0211 (0.0263)
Total Observed Scale h2: 0.2092 (0.021)
h2 Z: 9.96
```

`ldsc` .log file

Three parts (for two traits)

[2/3]

```
Calculating genetic covariance [2/3] for traits: BMI.sumstats.gz and Height.sumstats.gz
73669 SNPs remain after merging BMI.sumstats.gz and Height.sumstats.gz summary statistics
Results for genetic covariance between: BMI.sumstats.gz and Height.sumstats.gz
Mean Z*Z: -0.3982
Cross trait Intercept: -0.0688 (0.0772)
Total Observed Scale Genetic Covariance (g_cov): -0.008 (0.0129)
g_cov Z: -0.623
g_cov P-value: 0.53315
```

`ldsc` .log file

Three parts (for two traits)

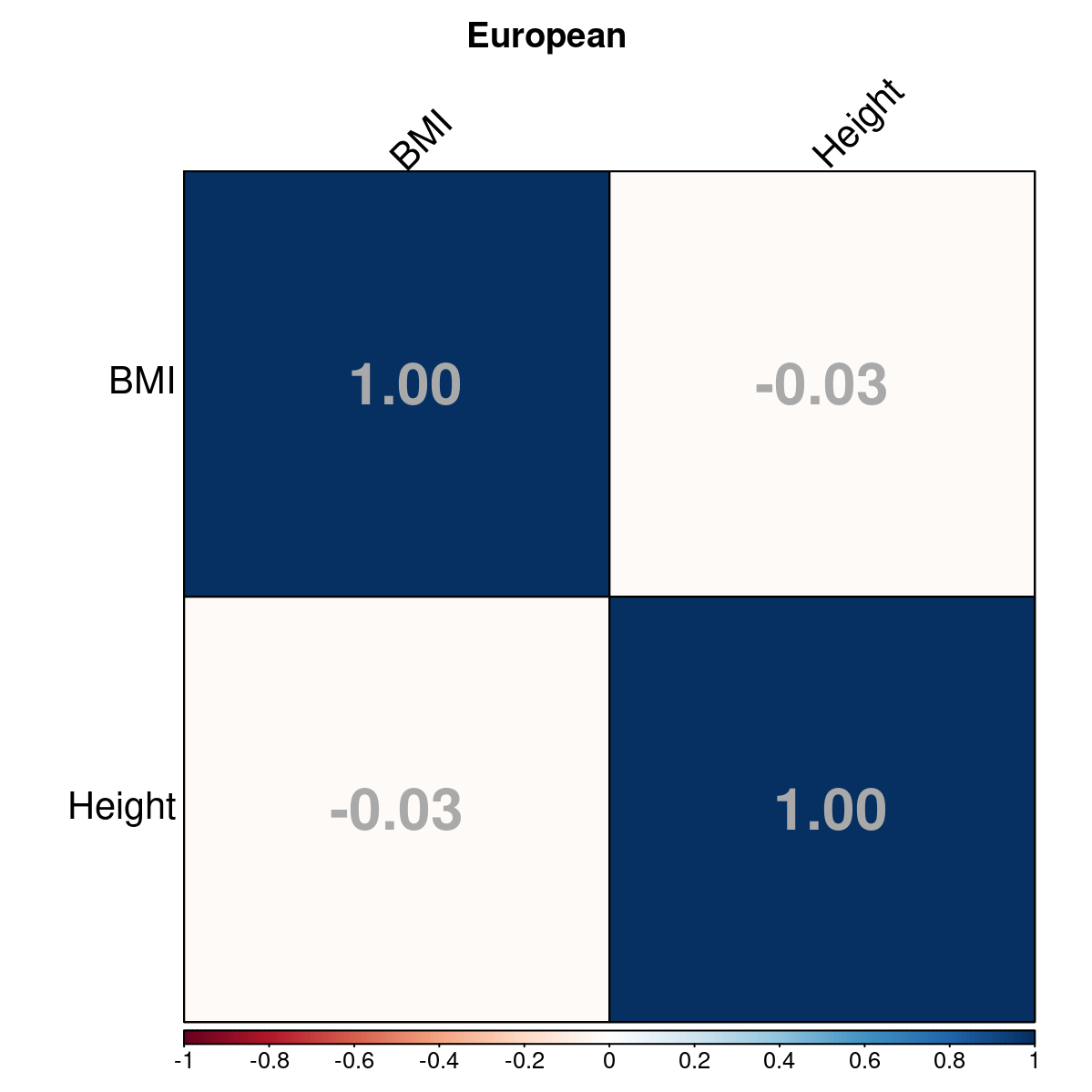
[3/3]

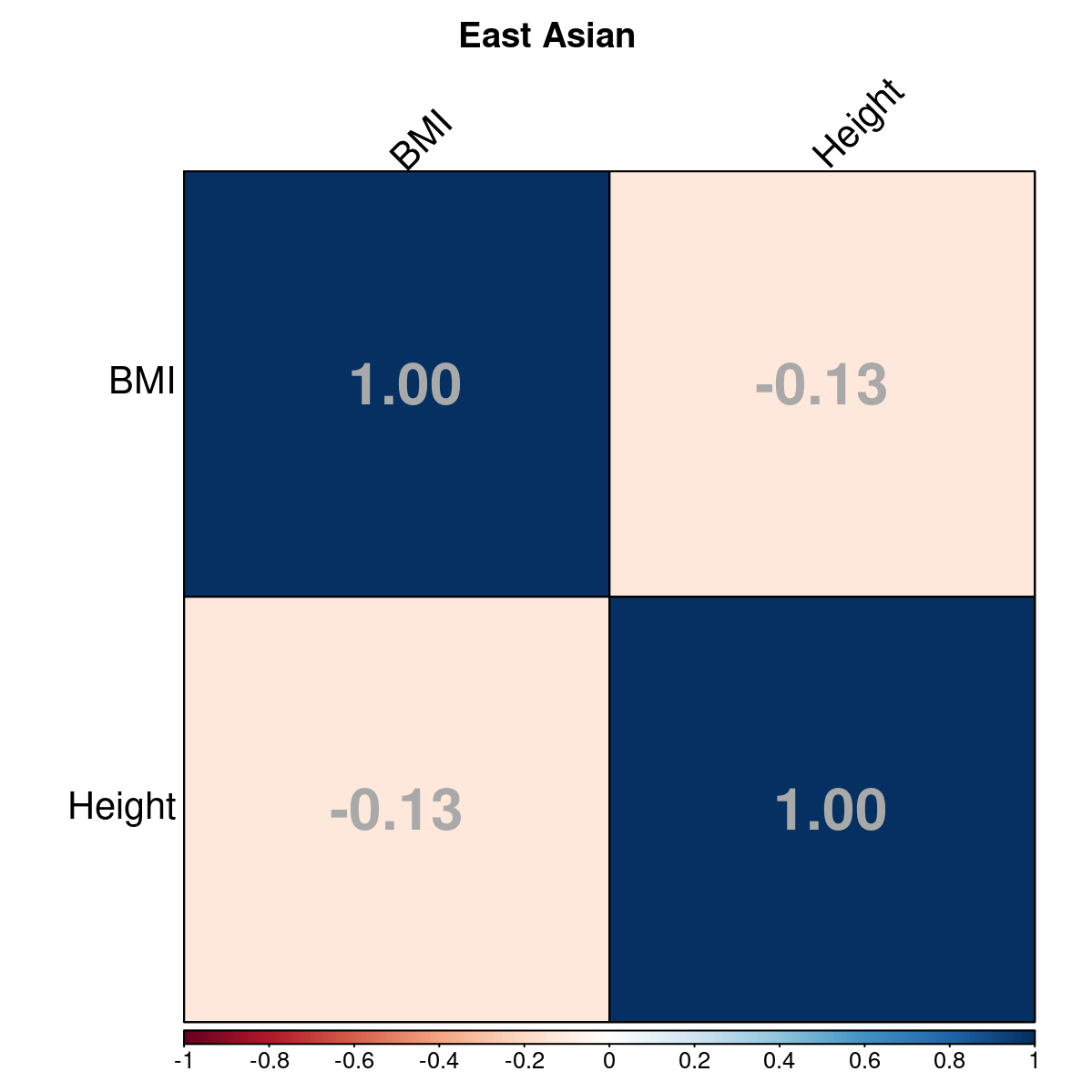
```
Estimating heritability [3/3] for: Height.sumstats.gz
Heritability Results for trait: Height.sumstats.gz
Mean Chi^2 across remaining SNPs: 15.303
Lambda GC: 5.6174
Intercept: 1.3365 (0.3003)
Ratio: 0.0235 (0.021)
Total Observed Scale h2: 0.4158 (0.0448)
h2 Z: 9.29
Genetic Correlation Results
Genetic Correlation between BMI and Height: -0.0272 (0.0436)
```

`ldsc` .log file

(Standardized) Genetic Correlation printed at the end of the log file.

```
Genetic Correlation Results
Genetic Correlation between BMI and Height: -0.0272 (0.0436)
```


European
BMI
Height
BMI
Height
1.00
-0.03
-0.03
1.00
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


East Asian
BMI
Height
BMI
Height
1.00
-0.13
-0.13
1.00
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Practical for Binary Traits

We will be estimating LDSC for European Ancestry Samples for Schizophrenia and Bipolar

Article | Published: 08 April 2022

Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Vassily Trubetskoy, Antonio F. Pardiñas, Ting Qi, Georgia Panagiotaropoulou, Swapnil Awasthi, Tim B. Bigdeli, Julien Bryois, Chia-Yen Chen, Charlotte A. Dennison, Lynsey S. Hall, Max Lam, Kyoko Watanabe, Oleksandr Frei, Tian Ge, Janet C. Harwood, Frank Koopmans, Sigurdur Magnusson, Alexander L. Richards, Julia Sidorenko, Yang Wu, Jian Zeng, Jakob Grove, Minsoo Kim, Zhiqiang Li, Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, The SynGO Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium

+ Show authors

Nature **604**, 502–508 (2022) | Cite this article

48k Accesses | **229** Citations | **461** Altmetric | Metrics

Article | Published: 17 May 2021

Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology

Niamh Mullins ✉, Andreas J. Forstner, Kevin S. O'Connell, Brandon Coombes, Jonathan R. I. Coleman, Zhen Qiao, Thomas D. Als, Tim B. Bigdeli, Sigrid Børte, Julien Bryois, Alexander W. Charney, Ole Kristian Drange, Michael J. Gandal, Saskia P. Hagenaars, Masashi Ikeda, Nolan Kamitaki, Minsoo Kim, Kristi Krebs, Georgia Panagiotaropoulou, Brian M. Schilder, Laura G. Sloofman, Stacy Steinberg, Vassily Trubetskoy, Bendik S. Winsvold, HUNT All-In Psychiatry, … Ole A. Andreassen ✉

+ Show authors

Nature Genetics **53**, 817–829 (2021) | Cite this article

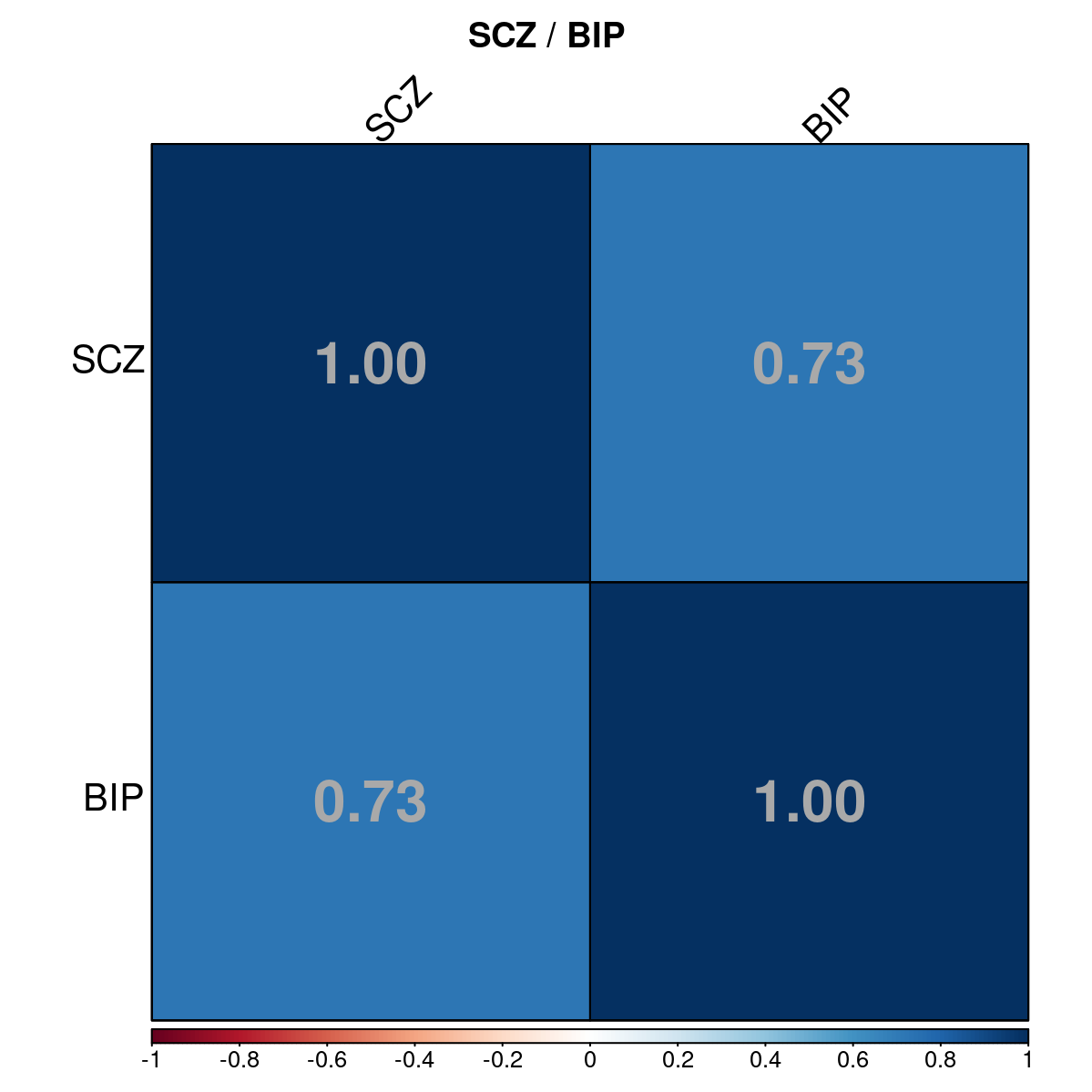
25k Accesses | **224** Citations | **321** Altmetric | Metrics

The *ldsc* function takes 6 arguments:

1. **traits**: a vector of file names/paths to files which point to the munged sumstats.
2. **sample.prev**: A vector of sample prevalences of length equal to the number of traits. Enter 0.5 if inputting [sum of] effective N.
3. **population.prev**: A vector of population prevalences.
4. **ld**: A folder of LD scores used as the independent variable in LDSC
5. **wld**: A folder of LDSC weights (Typically same folder as specified for the ld argument)
6. **trait.names**: The trait names.

The *ldsc* function takes 6 arguments:

1. **traits**: a vector of file names/paths to files which point to the munged sumstats.
2. **sample.prev**: A vector of sample prevalences of length equal to the number of traits. **Enter 0.5 if inputting [sum of] effective N.**
3. **population.prev**: A vector of population prevalences.
4. **ld**: A folder of LD scores used as the independent variable in LDSC
5. **wld**: A folder of LDSC weights (Typically same folder as specified for the ld argument)
6. **trait.names**: The trait names.


SCZ / BIP
SCZ
BIP
SCZ
BIP
1.00
0.73
0.73
1.00
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

Practical Working Directory

```
cd  ~/practicals/3.2.GeneticCorrelation_MadhurSingh/final
```

Qualtrics link

This link is also on top of the R script: ***LDSC_Practical2_GeneticCorrelation.R***

https://qimr.az1.qualtrics.com/jfe/form/SV_781CgwvIn2YqyqO

References

Bulik-Sullivan, B., Finucane, H., Anttila, V. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47, 1236–1241 (2015). https://doi.org/10.1038/ng.3406

Witte, J., Visscher, P. & Wray, N. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet* 15, 765–776 (2014). https://doi.org/10.1038/nrg3786

Grotzinger, A. D., Fuente, J. de la, Privé, F., Nivard, M. G. & Tucker-Drob, E. M. Pervasive Downward Bias in Estimates of Liability-Scale Heritability in Genome-wide Association Study Meta-analysis: A Simple Solution. *Biol. Psychiatry* 93, 29–36 (2023). https://doi.org/10.1016/j.biopsych.2022.05.029