University of Colorado Boulder
Institute for Behavioral Genetics

LD-Score Heritability + Genetic Correlation

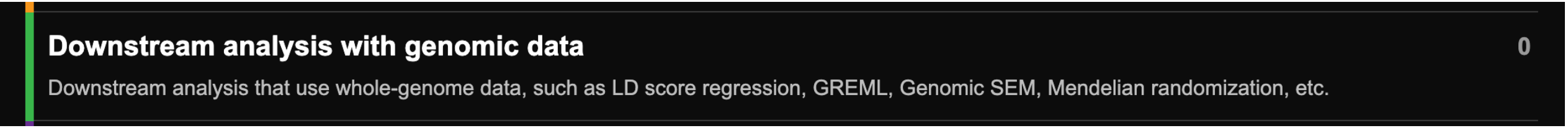
Andrew Grotzinger

2023 International Statistical Genetics Workshop

March 8th, 2023

Where to put questions for this practical on the forum

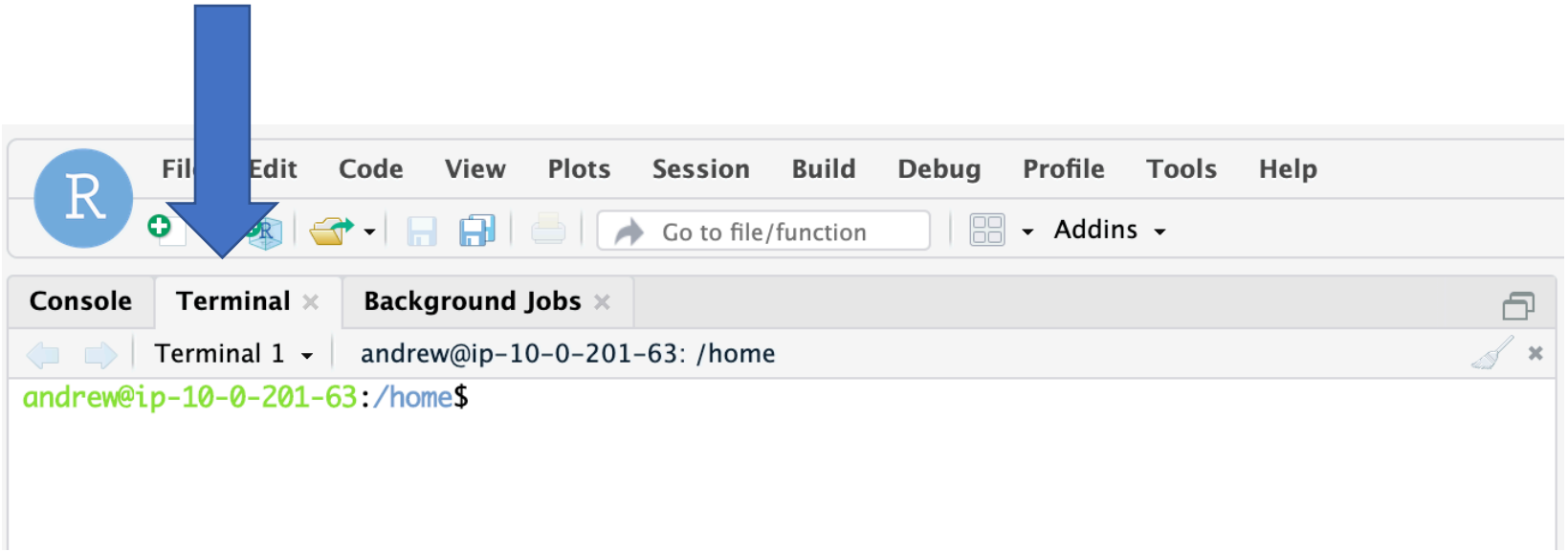
https://isgw-forum.colorado.edu/



Let’s start by going to:

https://workshop.colorado.edu/rstudio/

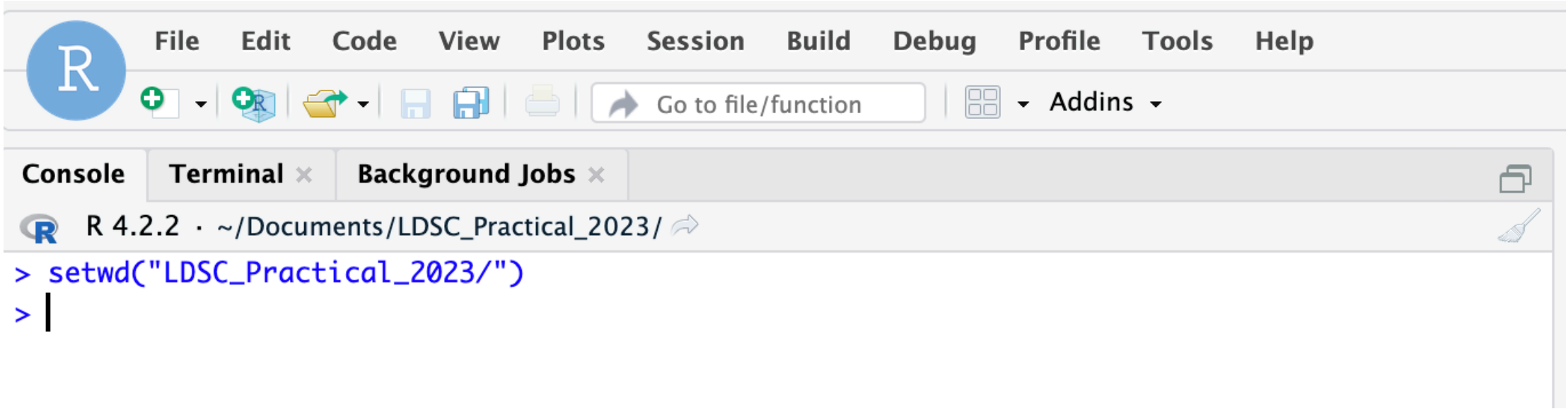
And clicking on the terminal tab.



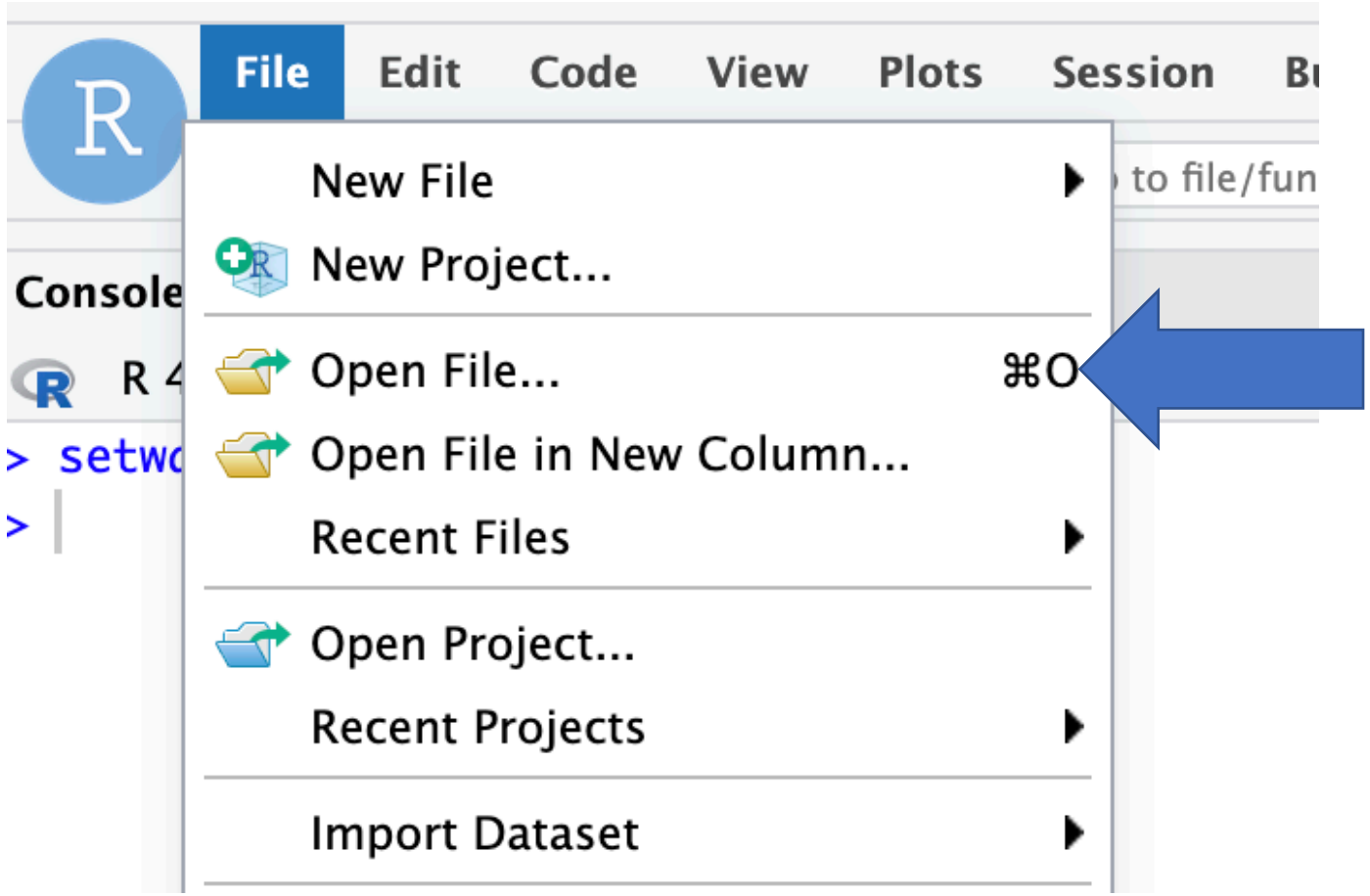
Copy over the practical files from that terminal tab

```
cp -r /faculty/andrew/LDSC_Practical_2023  .
```

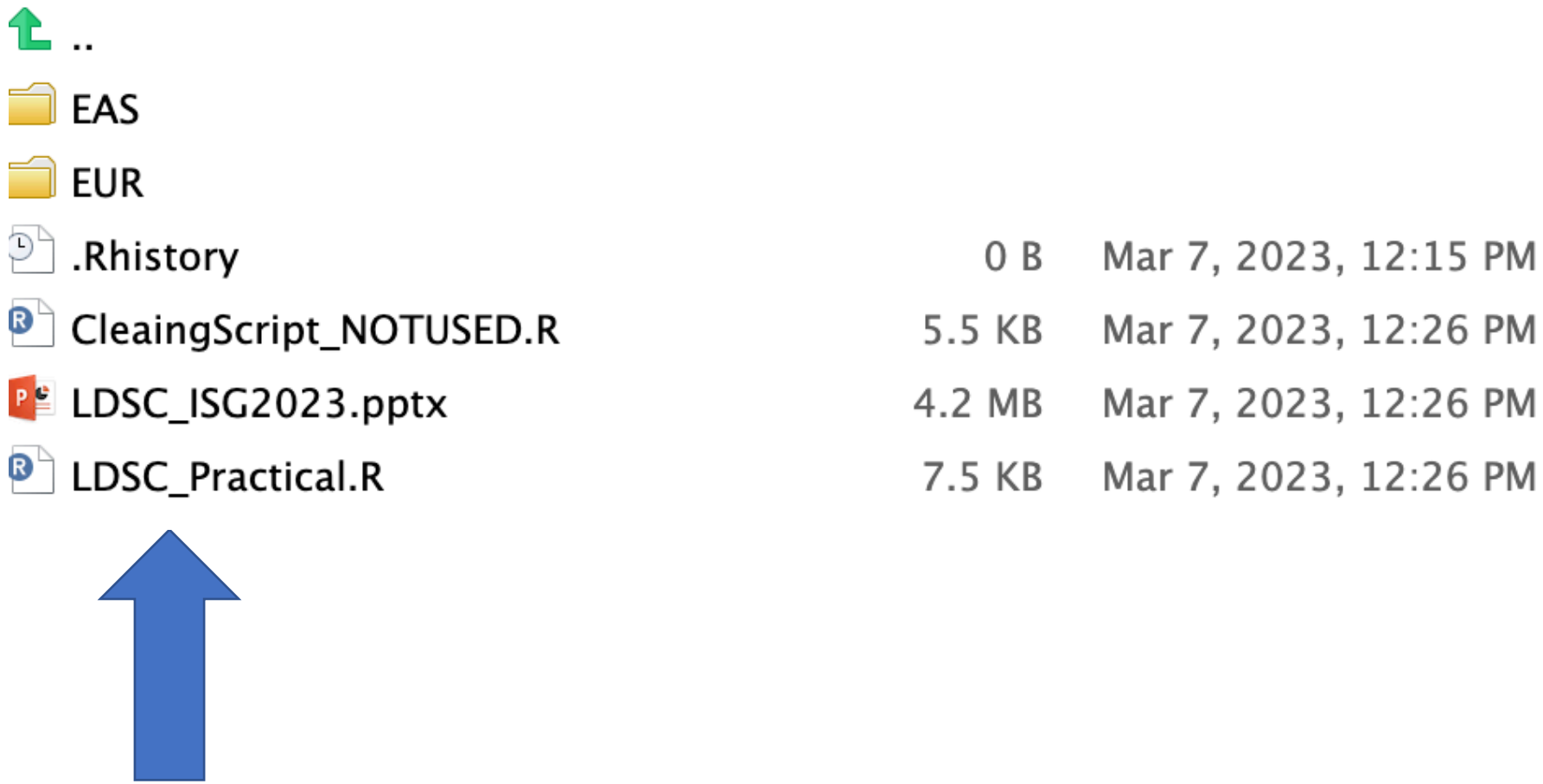
Now let's go over to the console and *setwd* for this new folder you just copied



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



Finally let’s open the R script “LDSC_Practical.R”

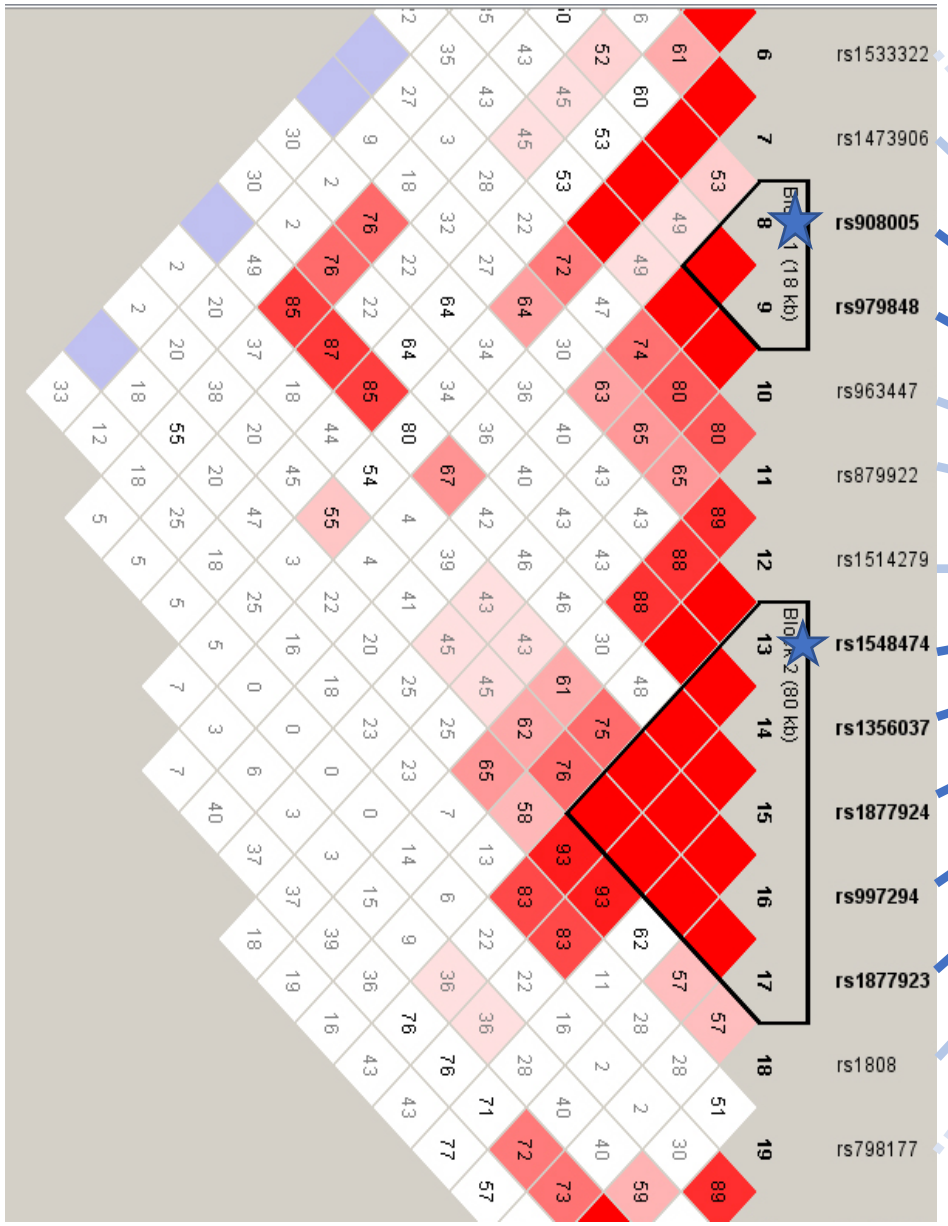


Outline

I. Background
II. Univariate LDSC
III. Bivariate LDSC
IV. Practical for Continuous Traits
V. Liability Scale Heritability for Binary Traits
VI. Practical for Binary Traits

I. Background

Toy Example: 2 True Causal Variants ★



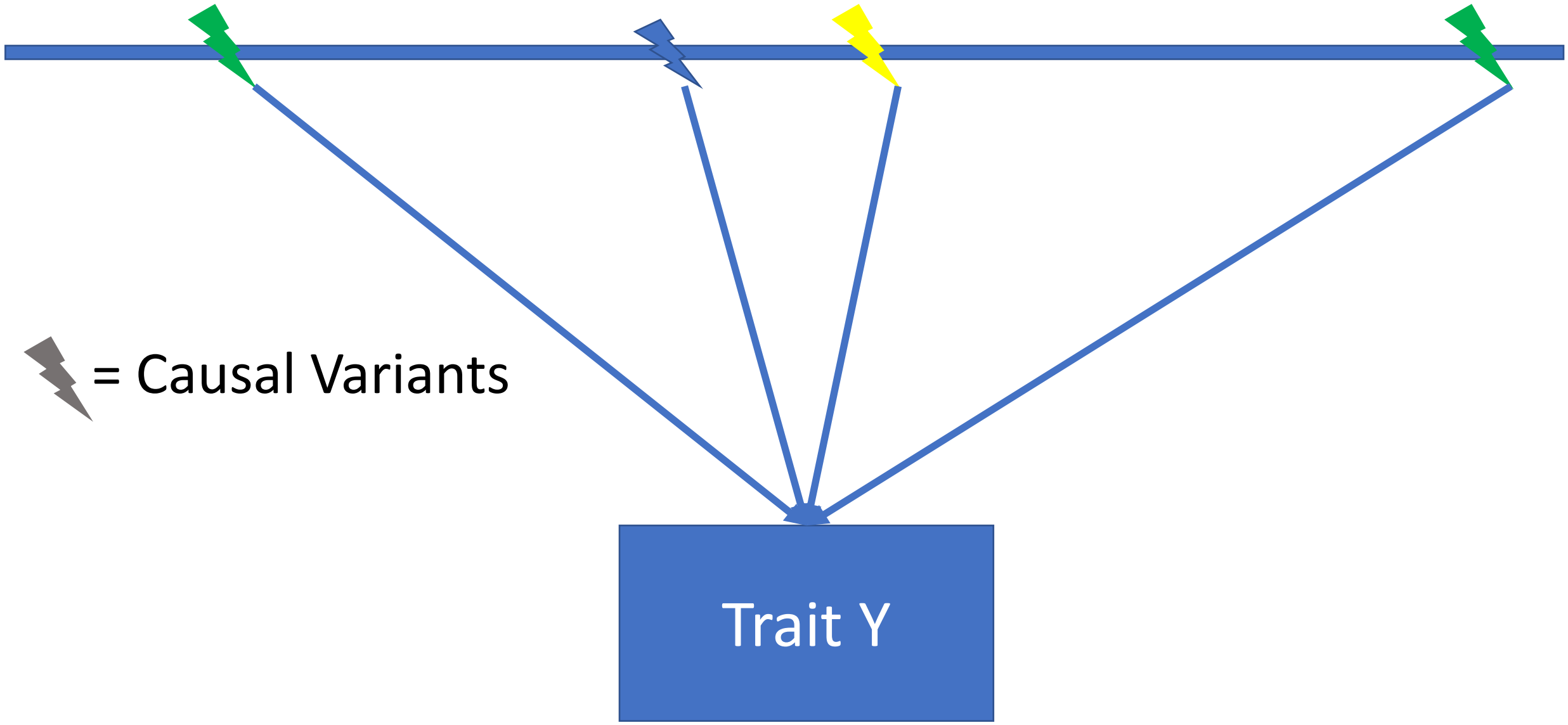
Phenotype Y

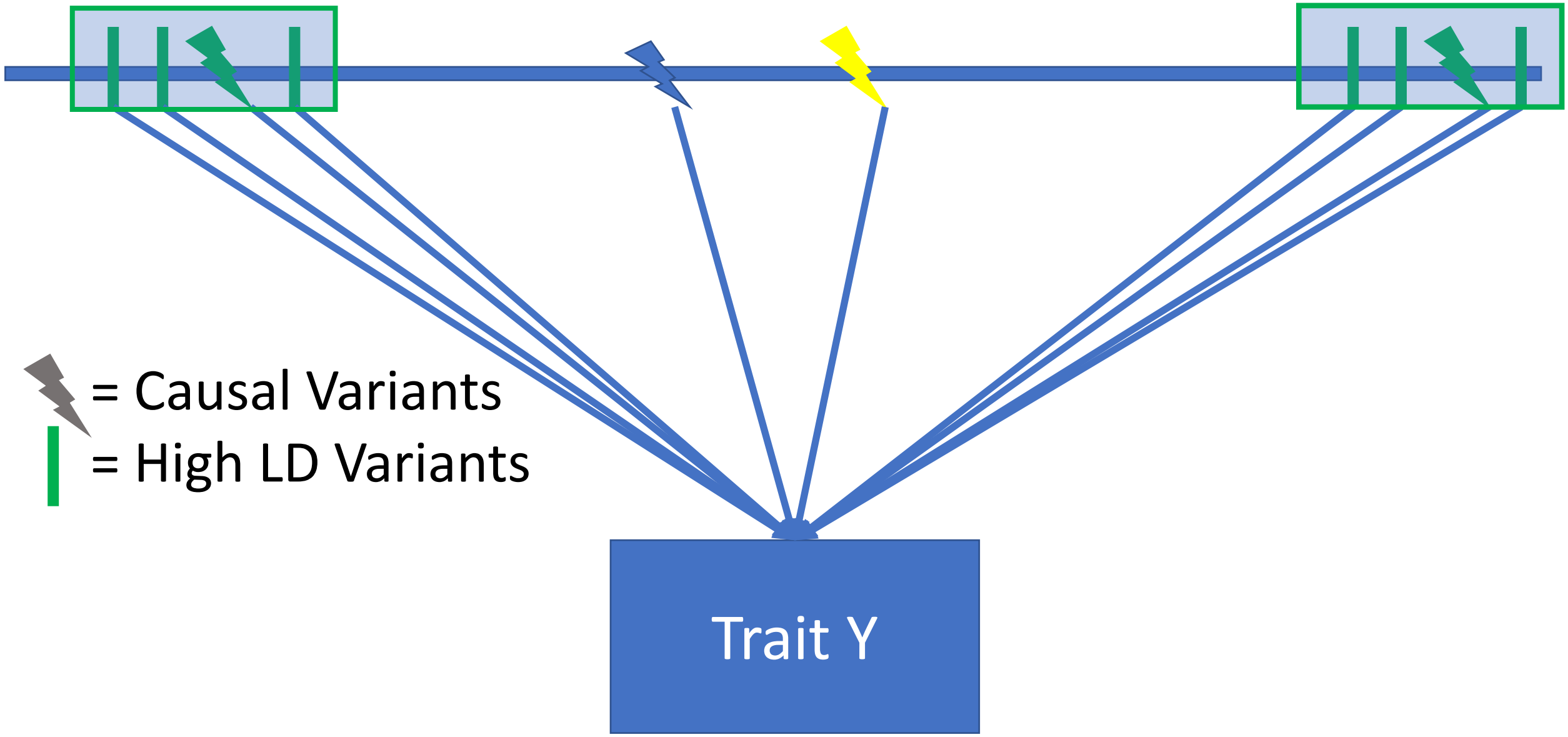
$$\ell_j = \sum_i r_{ij}^2$$

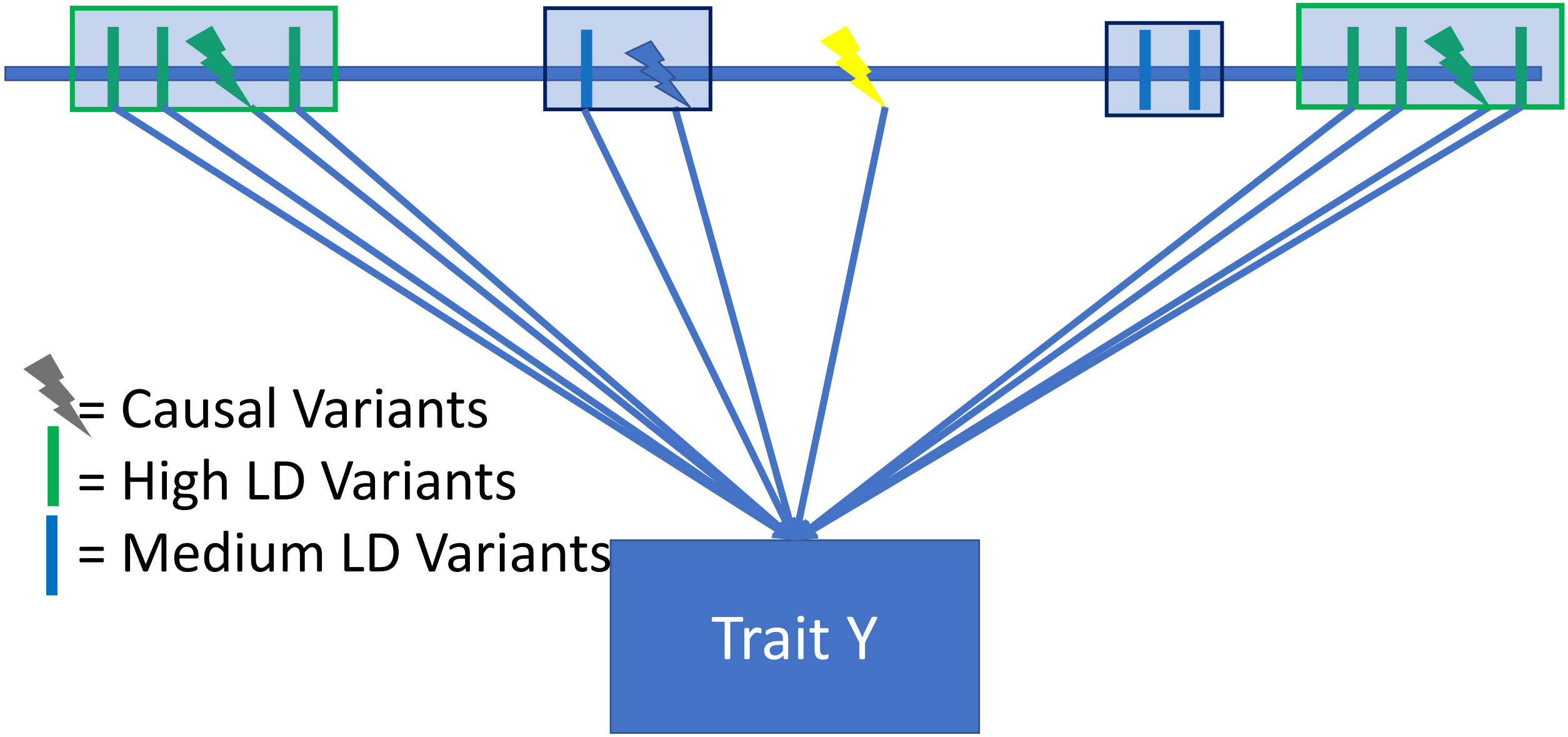
Imagine that there are two causal variants in the population.

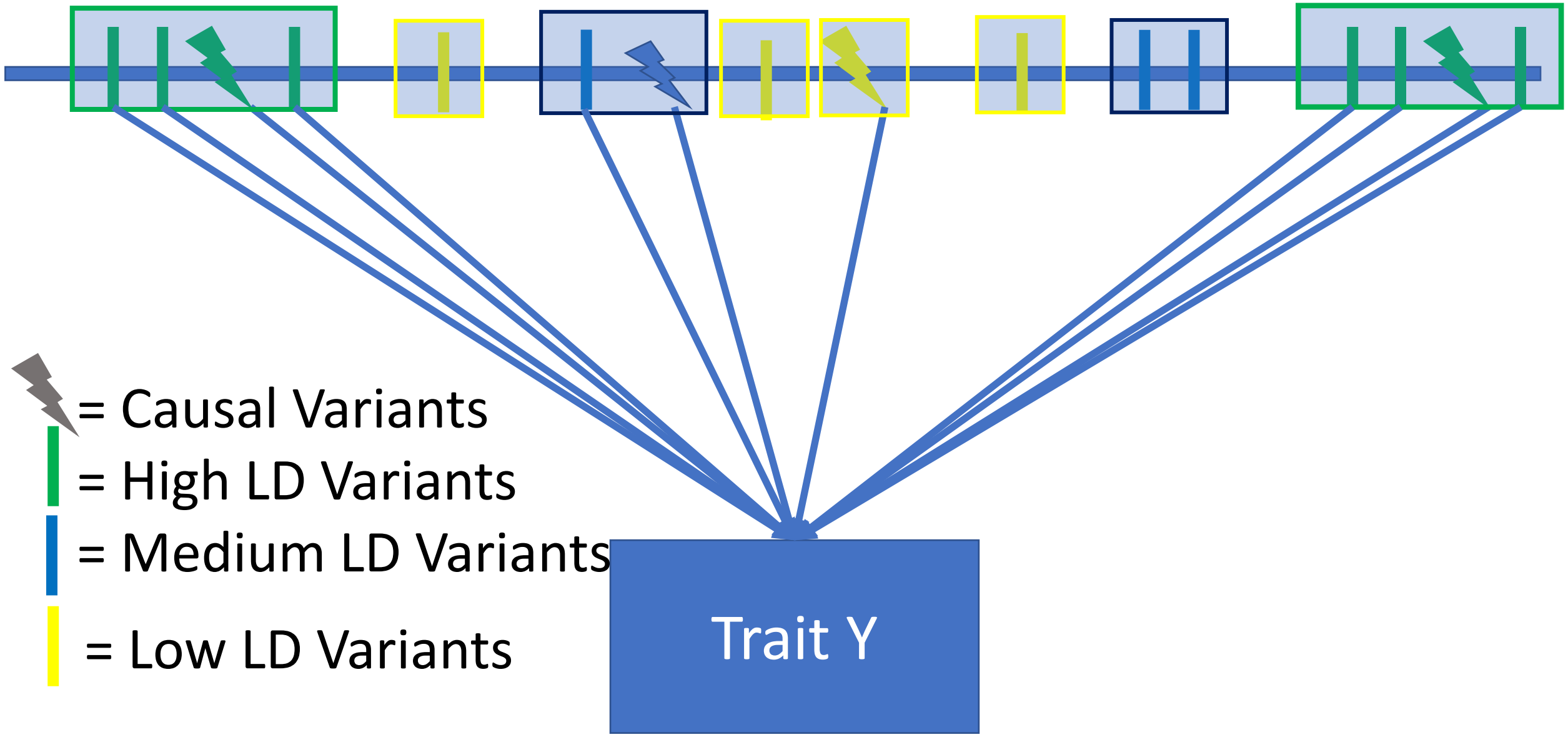
All variants in LD with those variants are going to be estimated as having an affect on the phenotype, Y

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


= Causal Variants
Trait Y


= Causal Variants
= High LD Variants
Trait Y


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


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

Because traits are *highly* polygenic:

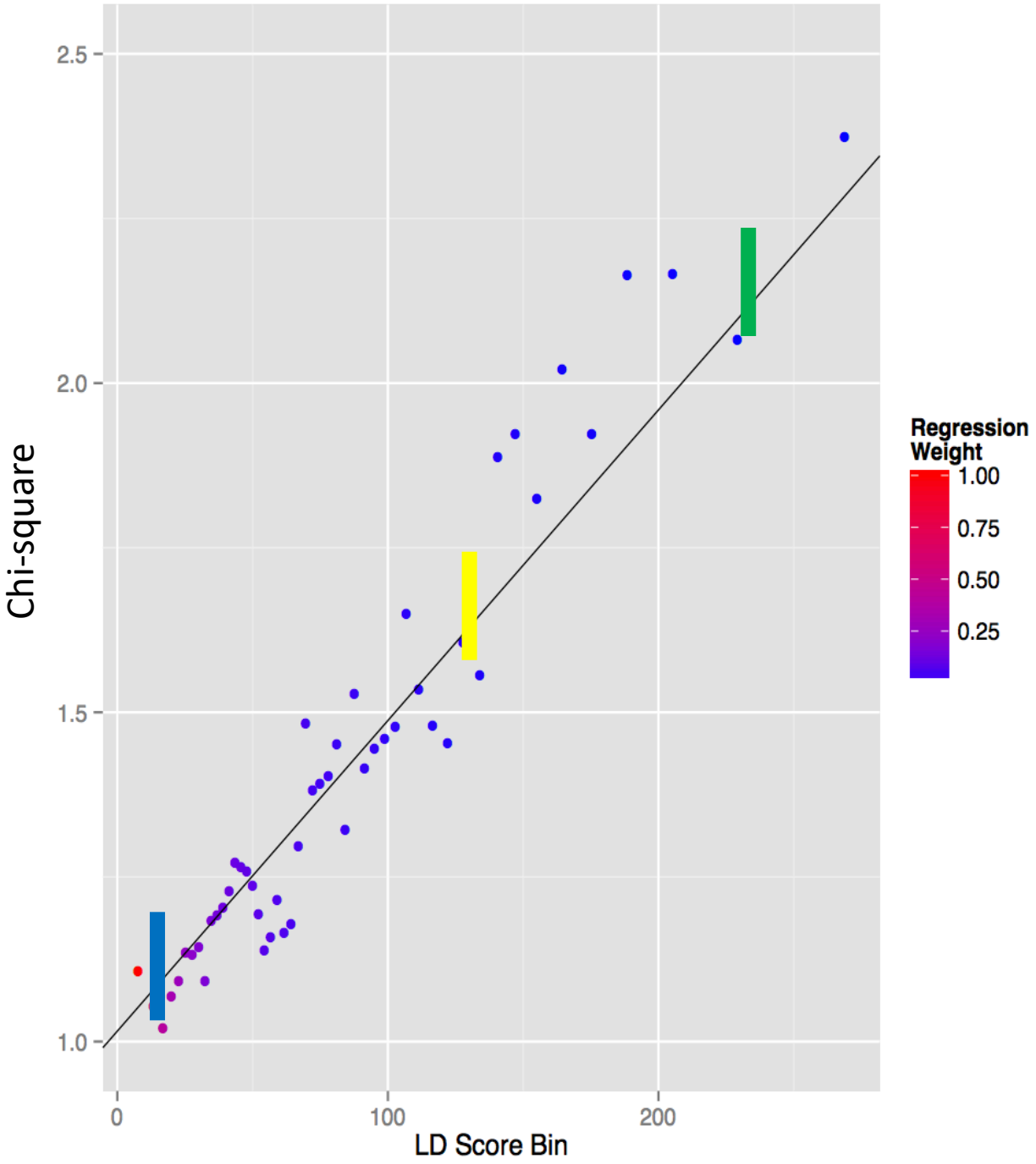
Variants with higher LD are more likely to pick up effect of causal variant

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

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

Bulik-Sullivan et al., 2015

II. Univariate LDSC



- 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
  - Only effect the average test statistic (the LDSC intercept) but not the relationship between test statistics and LD Scores (the slope)

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

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

The expectation for the GWAS chi-square given the LD-score for the SNP *j*

This is the "Y" variable in the linear regression

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

N **= Sample size**

M **= number of SNPs used to estimate the LD-scores**

h^2 **= SNP-based heritability**

This whole piece can be thought as the estimated "beta" in the linear regression

If you have two traits that have the exact same GWAS estimates (i.e., same p-value)

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

Trait 1 must be more heritable given equivalent results at 1/2 the sample size

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

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

The LD-score of SNP j

This is the ”X” variable in the linear regression

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

N is sample size again

a **is confounding biases**

This whole piece is the “intercept” in the linear regression

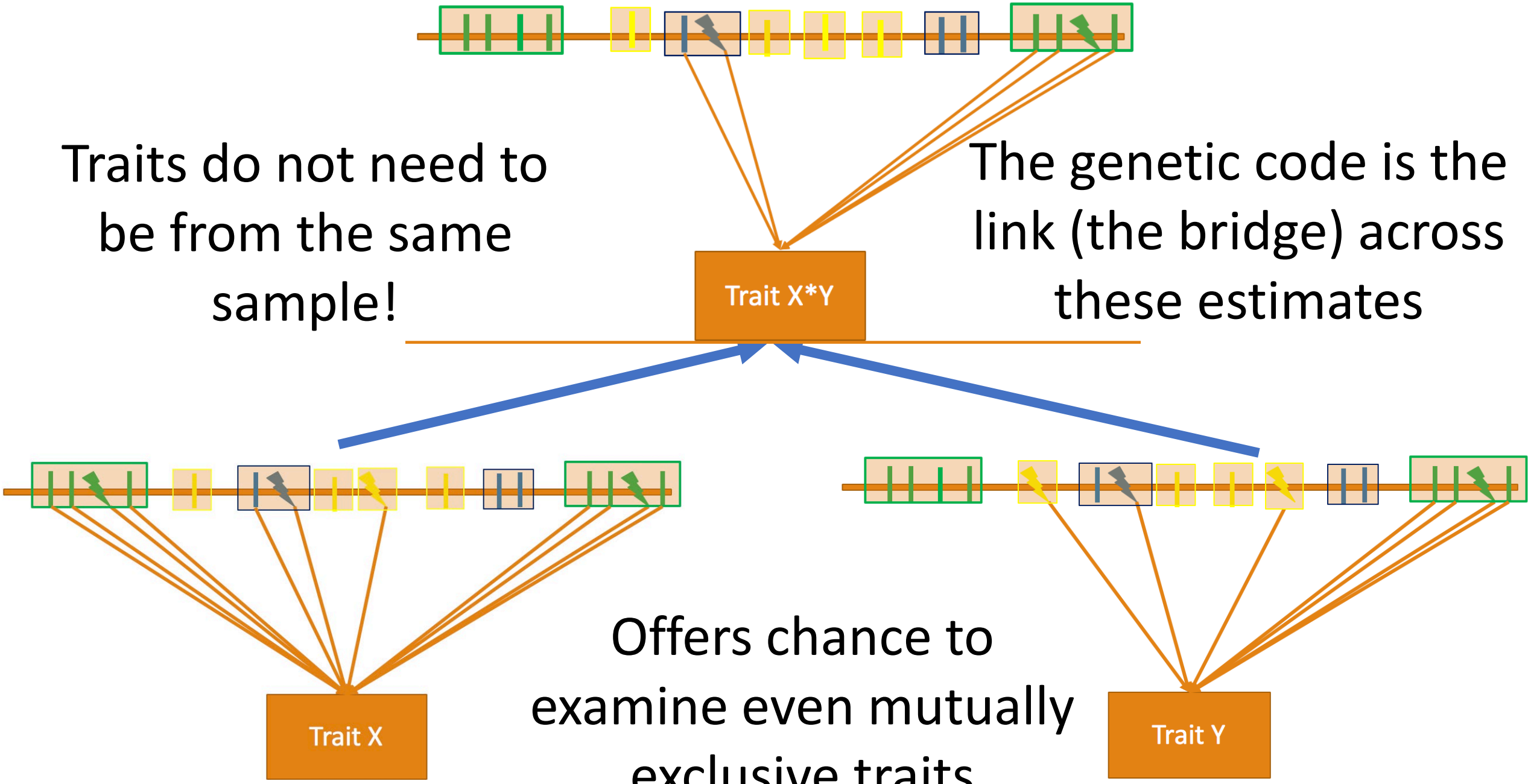
Putting it all together, this is ultimately a simple linear regression equation to back out the implied heritability estimates

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

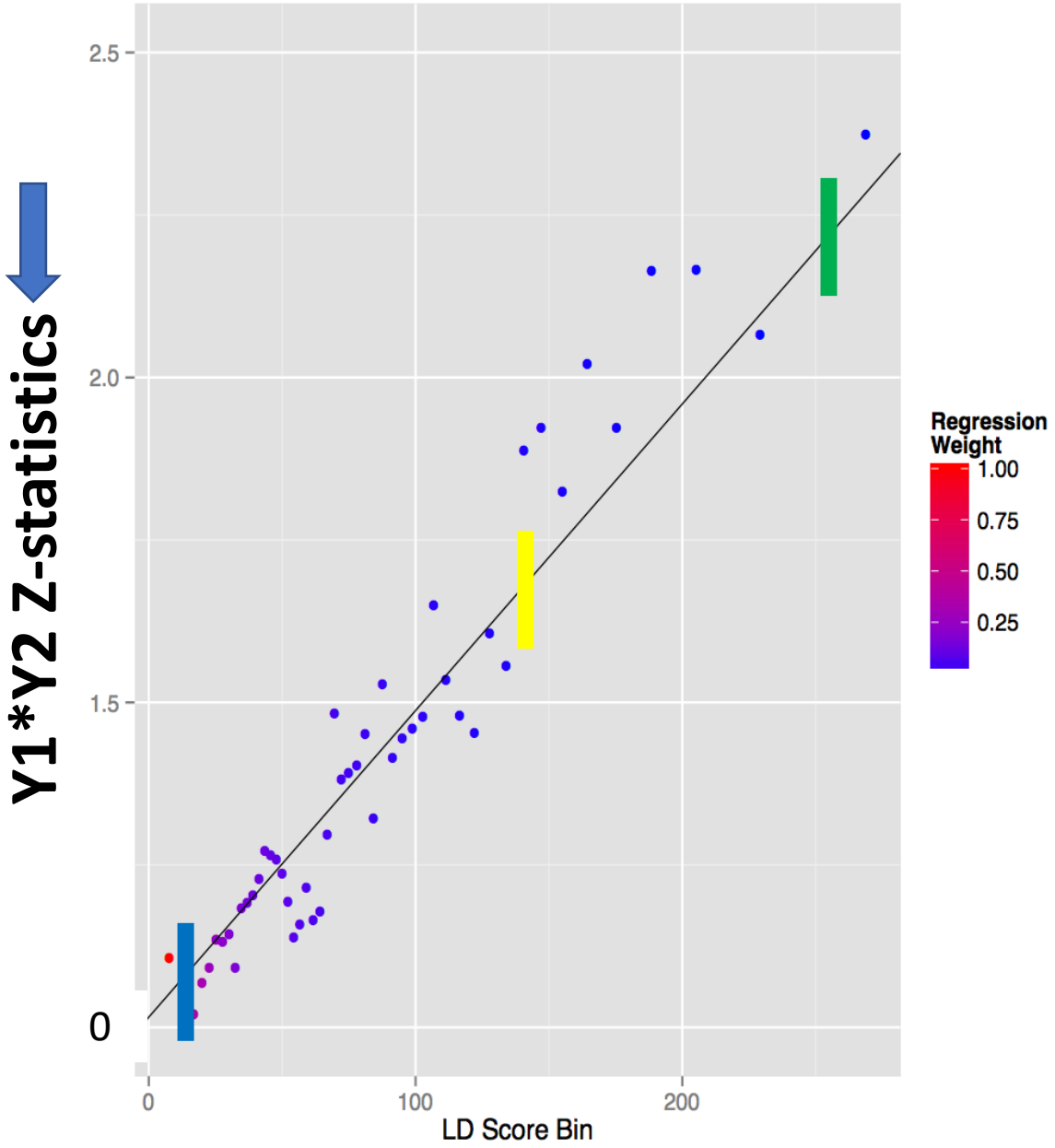
III. Bivariate LDSC

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)

Bulik-Sullivan et al., 2015

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

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

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

This again is the "Y" variable in the linear regression

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

$\sqrt{N_1N_2}$ reflects the square root of the sample size for trait 1 and trait 2

M is the number of SNPs

ρ_g is the genetic covariance

This is the "beta" where again we back out the estimate we care about, which in this case is the genetic covariance

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

Again, this is the LD-score for a given SNP j, which reflects our “X” variable

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

Here we have the bivariate LDSC intercept.

$\sqrt{N_1N_2}a$ reflects shared sources of population stratification across the two samples.

$\frac{\rho N_s}{\sqrt{N_1N_2}}$ reflects the phenotypic correlation among overlapping participant samples weighted by proportional sample overlap

Genetic Correlation

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

- In many instances, we are interested in the amount of genetic overlap on the standardized scale (i.e., genetic correlation).
  - What are some important things to keep in mind when interpreting this estimate?

- Why is the bivariate LDSC equation only going to work within ancestry?

Allows us to produce genetic “heat maps” of genetic correlations across traits

Analysis of shared heritability in common disorders of the brain

The Brainstorm Consortium*†

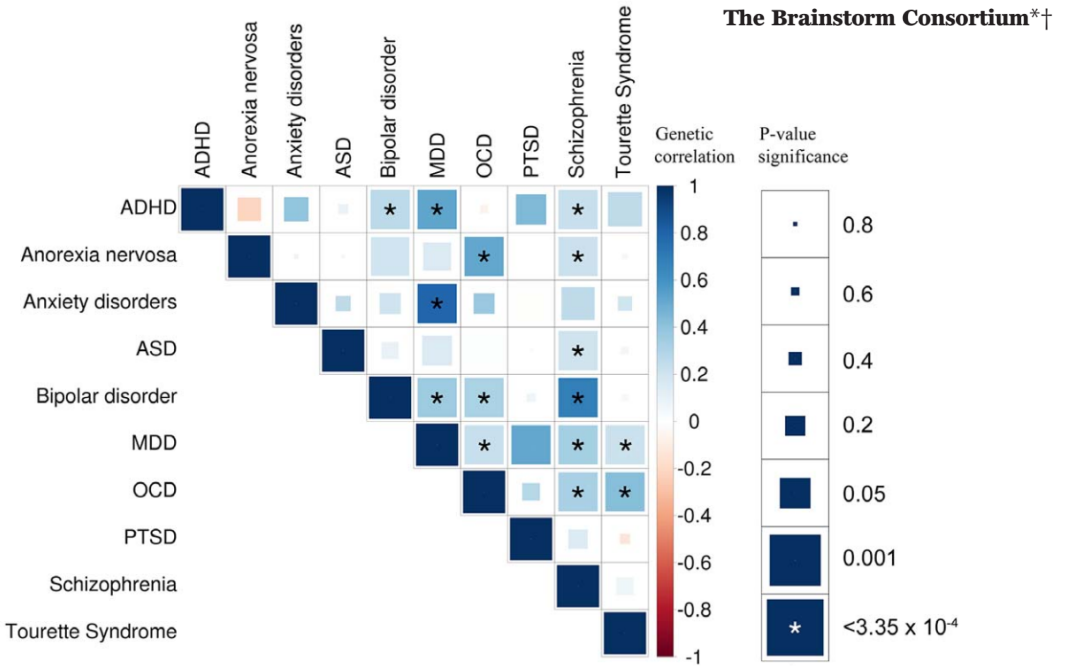


Fig. 1. Genetic correlations across psychiatric phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic correlations that are significantly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

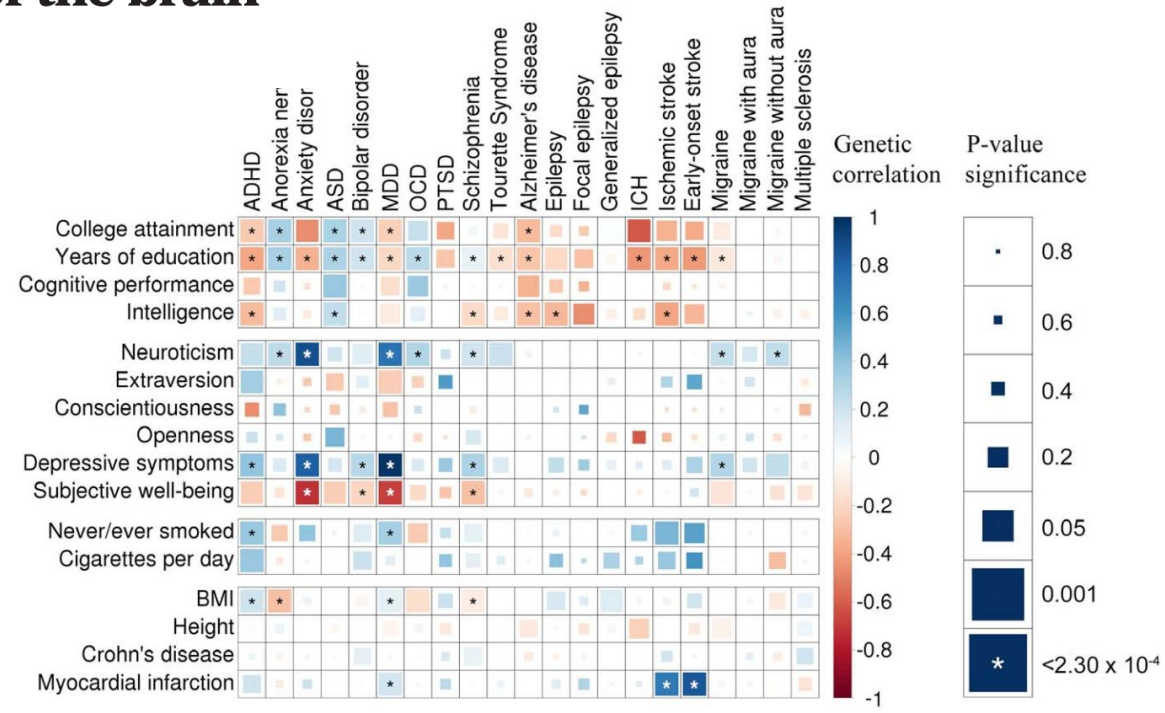
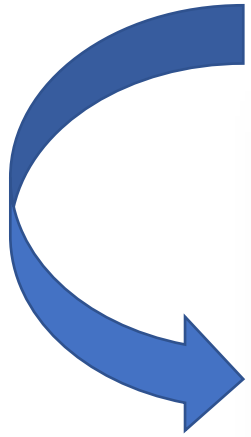


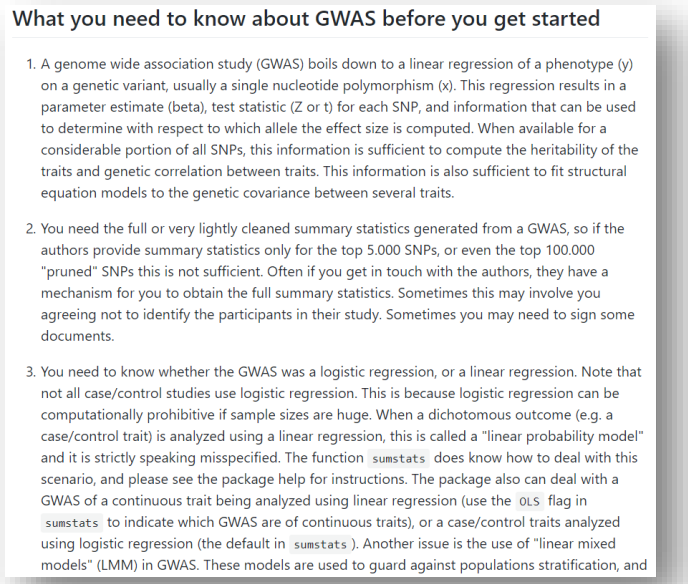
Fig. 4. Genetic correlations across brain disorders and behavioral-cognitive phenotypes. The color of each box indicates the magnitude of the correlation, and the size of the box indicates its significance (LDSC), with significant correlations filling each square completely. Asterisks indicate genetic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

IV. Practical for Continuous Traits

Where to get summary statistics

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



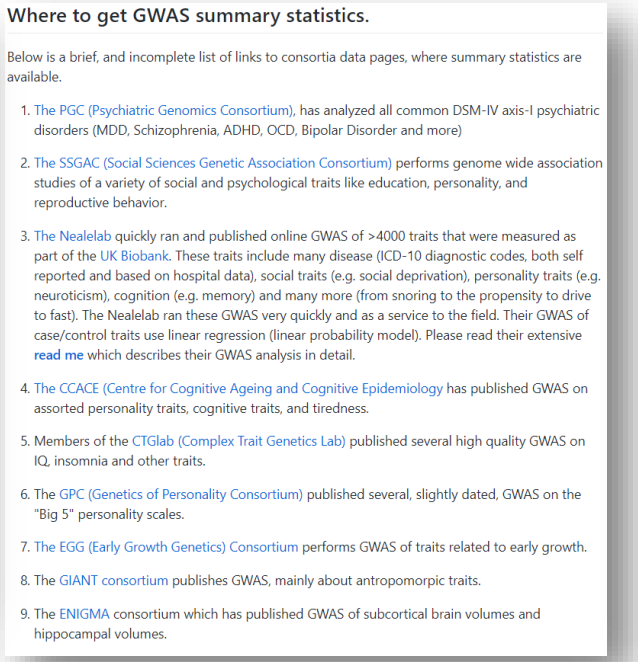


What you need to know about GWAS before you get started

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.

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

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

BioBank Japan PheWeb (PheWeb.jp)

https://pheweb.jp/

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

Pan-UK Biobank

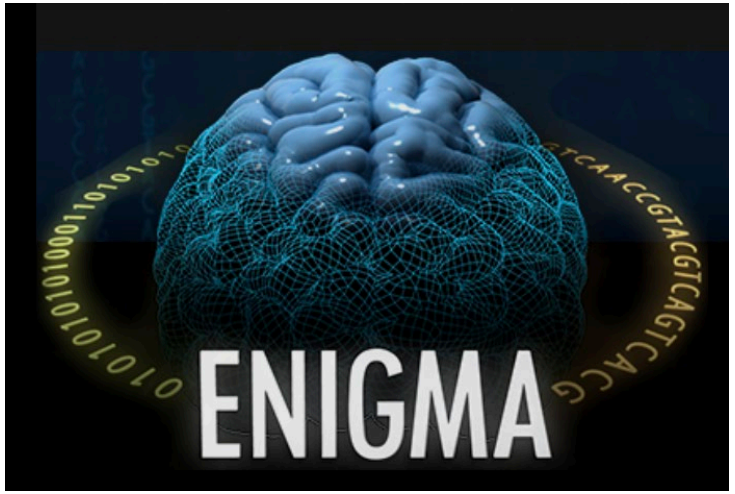
Pan-ancestry genetic analysis of the UK Biobank

https://docs.google.com/spreadsheets/d/1AeeADtT0U1AuklliiNyiVzVRdLYPkTbruQSk38DeutU8/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/

Only TWO Primary Steps to estimate ldsc

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

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

Munge: convert raw data from one form to another

The summary statistics files input to the *munge* function at a minimum need to contain five pieces of information:

1. The rsID of the SNP.
2. An A1 allele column, indicating the effect allele.
3. An A2 allele column, indicating the non-effect allele.
4. A signed (+/-) effect column.
5. The *p*-value associated with this effect.

The *munge* function takes 6 arguments:

1. **files**: The name of the summary statistics files
2. **hm3**: The name of the reference file. Here we use Hapmap 3 SNPs.
3. **trait.names**: The trait names that will be used to name the saved files
4. **N**: The sample sizes associated with the traits.
5. **info.filter**: INFO filter. Package default is to retain SNPs with INFO > 0.9.
6. **maf.filter**: MAF filter. Package default is to retain SNPs with MAF > 0.01.

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. If the trait is continuous, the values should equal NA.
3. **population.prev**: A vector of population prevalences. If the trait is continuous the values should equal NA.
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.

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

Using European GWAS sumstats for:

Height (Yengo et al., 2022)

BMI from Pan UKB

Using East Asian GWAS sumstats for:

Height (Yengo et al., 2022)

BMI from Biobank Japan

LET'S GO TO THE CODE

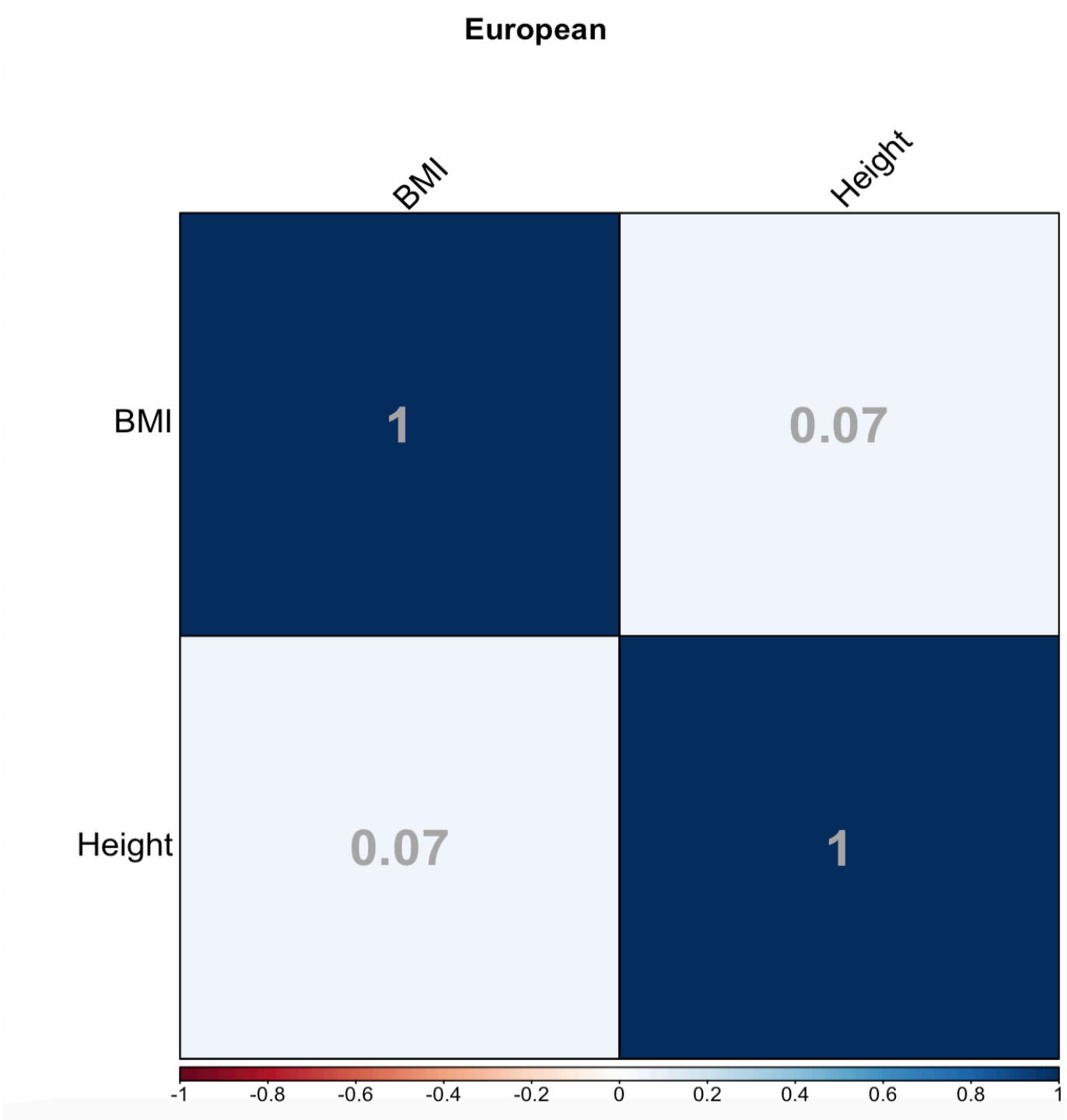
In this first practical you will get practice running the *munge* and *ldsc* functions and plotting the output as a heatmap.

Make sure to look at the .log files produced by each function to get a sense of what they are each doing "behind the scenes"

In practice, you will always want to inspect these to ensure that columns are being interpreted correctly

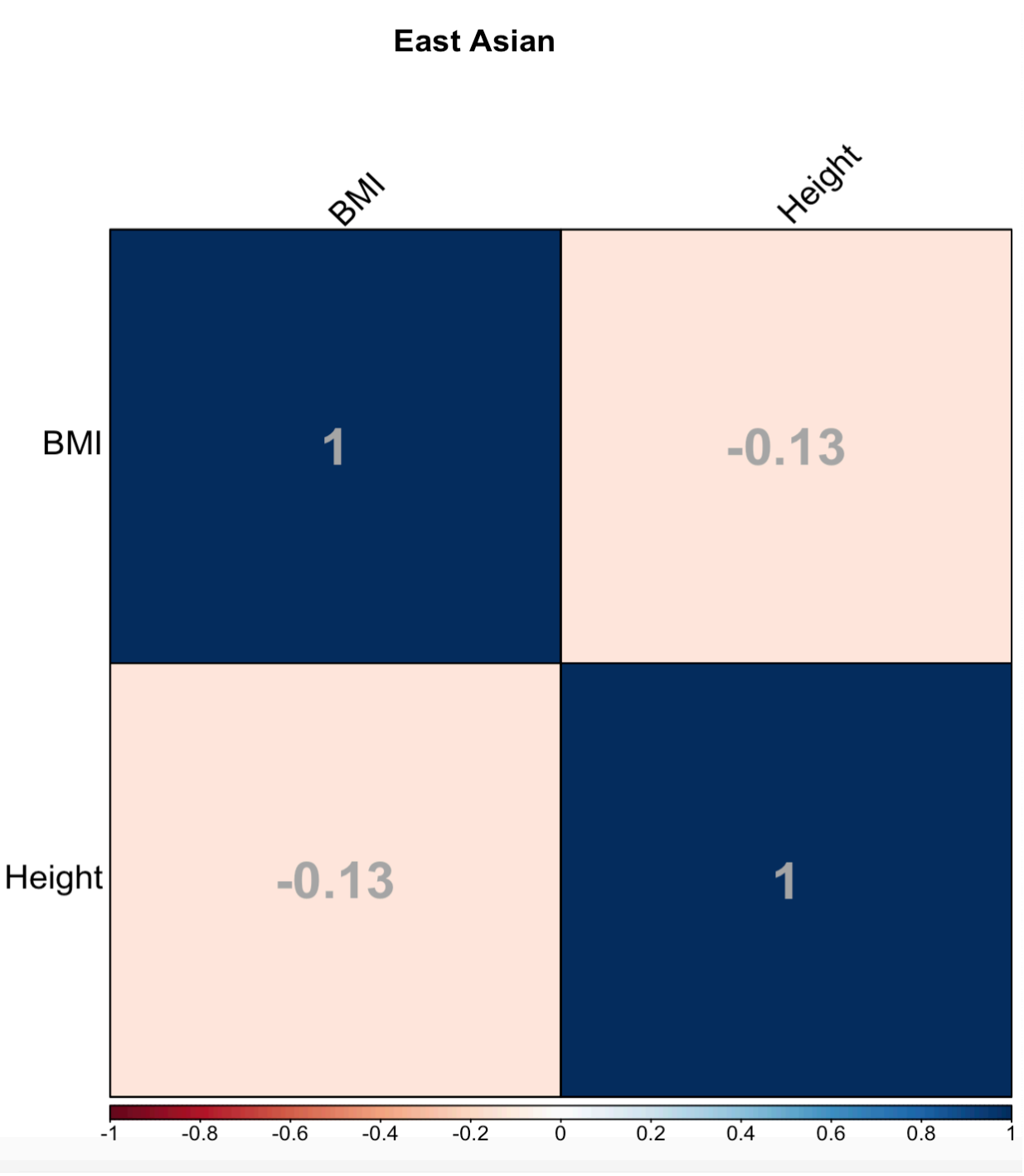
munge .log file

```
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.
```

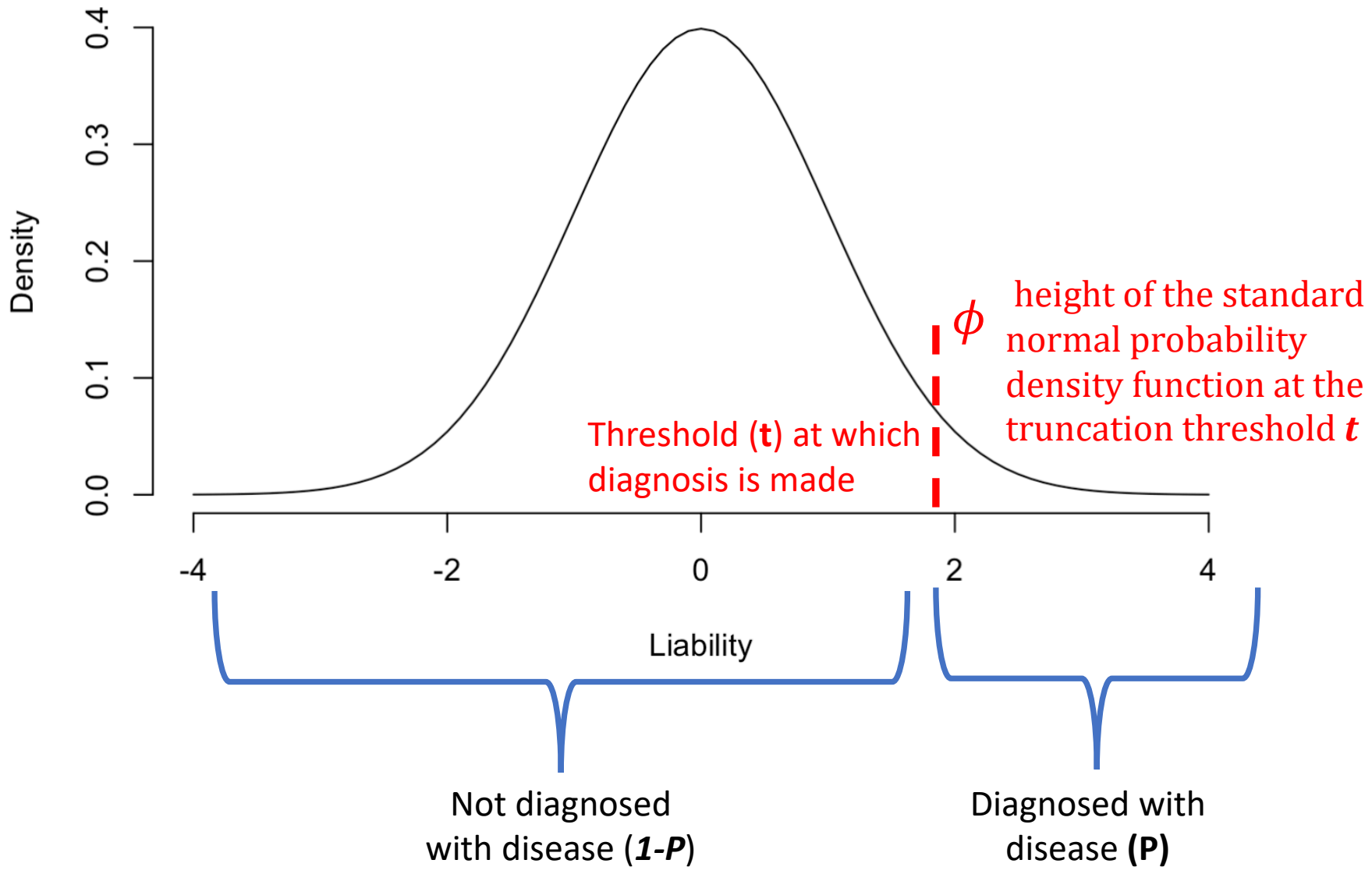

European
BMI
Height
BMI
Height
1
0.07
0.07
1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

ldsc .log file

```
Calculating genetic covariance [2/3] for traits: BMI.sumstats.gz and Height.sumstats.gz
71781 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: -0.1603
Cross trait Intercept: -0.028 (0.0305)
Total Observed Scale Genetic Covariance (g_cov): -0.0239 (0.0143)
g_cov Z: -1.67
g_cov P-value: 0.094448

Estimating heritability [3/3] for: Height.sumstats.gz
Heritability Results for trait: Height.sumstats.gz
Mean Chi^2 across remaining SNPs: 2.4191
Lambda GC: 1.6247
Intercept: 1.0755 (0.0685)
Ratio: 0.0532 (0.0483)
Total Observed Scale h2: 0.2414 (0.0325)
h2 Z: 7.42


Genetic Correlation Results
Genetic Correlation between BMI and Height: -0.1288 (0.077)


LDSC finished running at 2023-03-07 12:23:52
Running LDSC for all files took 0 minutes and 12 seconds
```


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

V. Liability Scale Heritability for Binary Traits


Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Liability
ϕ
height of the standard normal probability density function at the truncation threshold ***t***
Threshold (**t**) at which diagnosis is made
Not diagnosed with disease (***1-P***)
Diagnosed with disease **(P)**

$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \frac{P(1-P)}{v(1-v)}$$

This first part of this equation backs out the expected heritability estimate on that continuous distribution of risk (liability)

P is the population prevalence of the disorder.

ϕ is the height of the continuous distribution of liability at the threshold *t* at which a diagnosis of the disorder is made

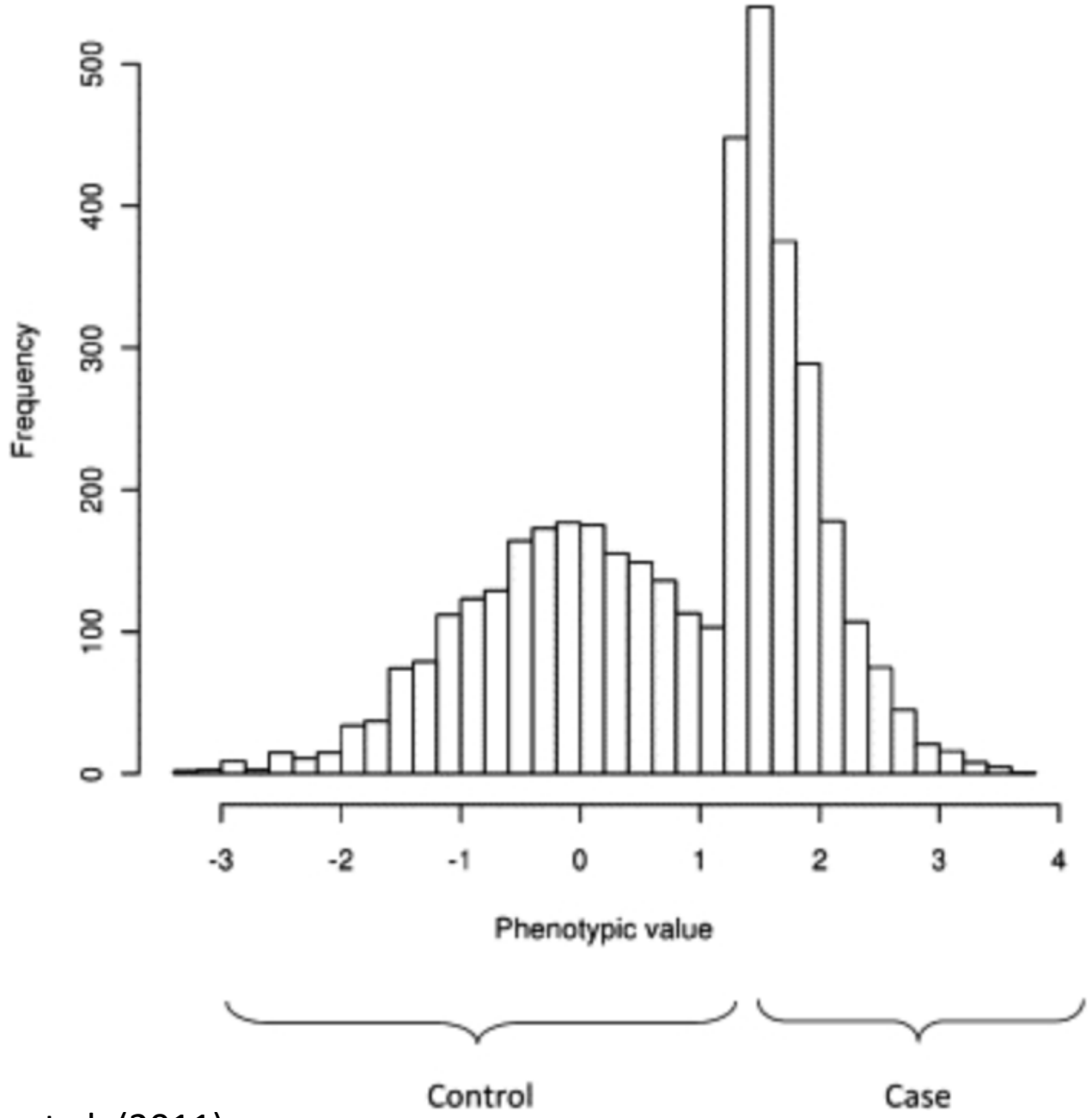
$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \boxed{\frac{P(1-P)}{v(1-v)}}$$

This second part of this equation performs the correction for participant ascertainment

P is again the population prevalence of the disorder.

v is the prevalence of the disorder in our participant sample.

The ratio then reflects the degree of ascertainment



"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)

Cohort-specific ascertainment

- As GWAS have continued to grow in sample size they often reflect meta-analyses of a series of contributing cohorts.

- For pragmatic reasons relating to data sharing with raw genotypes cohorts often share the GWAS summary stats to be meta-analyzed with other summary stats

- When this happens a correction for ascertainment within each cohort is required.
  - The reason: the ascertainment calculated using total cases and controls is not the same as ascertainment calculated within cohort
    - Sum of parts != sum of totals

$$h_l^2 = h_o^2 \frac{P(1-P)}{\phi^2} \frac{P(1-P)}{\sum_k v_k(1-v_k)}$$

In order to appropriately perform the ascertainment correction we need to calculate the sum of ascertainment across the contributing cohorts, *k*

$$Eff N_k = 4v_k(1 - v_k)n_k$$

In practice, we use what's call the effective sample size for this cohort-specific ascertainment correction.

The effective sample size is the sample size you would have had if the study design was balanced (50% cases and 50% controls)

Thus, it corrects the sample size for ascertainment and allows for summing sample size across cohorts.

We use sum of effective N as many GWAS pipelines (e.g., Ricopili) automatically output this for the GWAS

VI. Practical for Binary Traits

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

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:

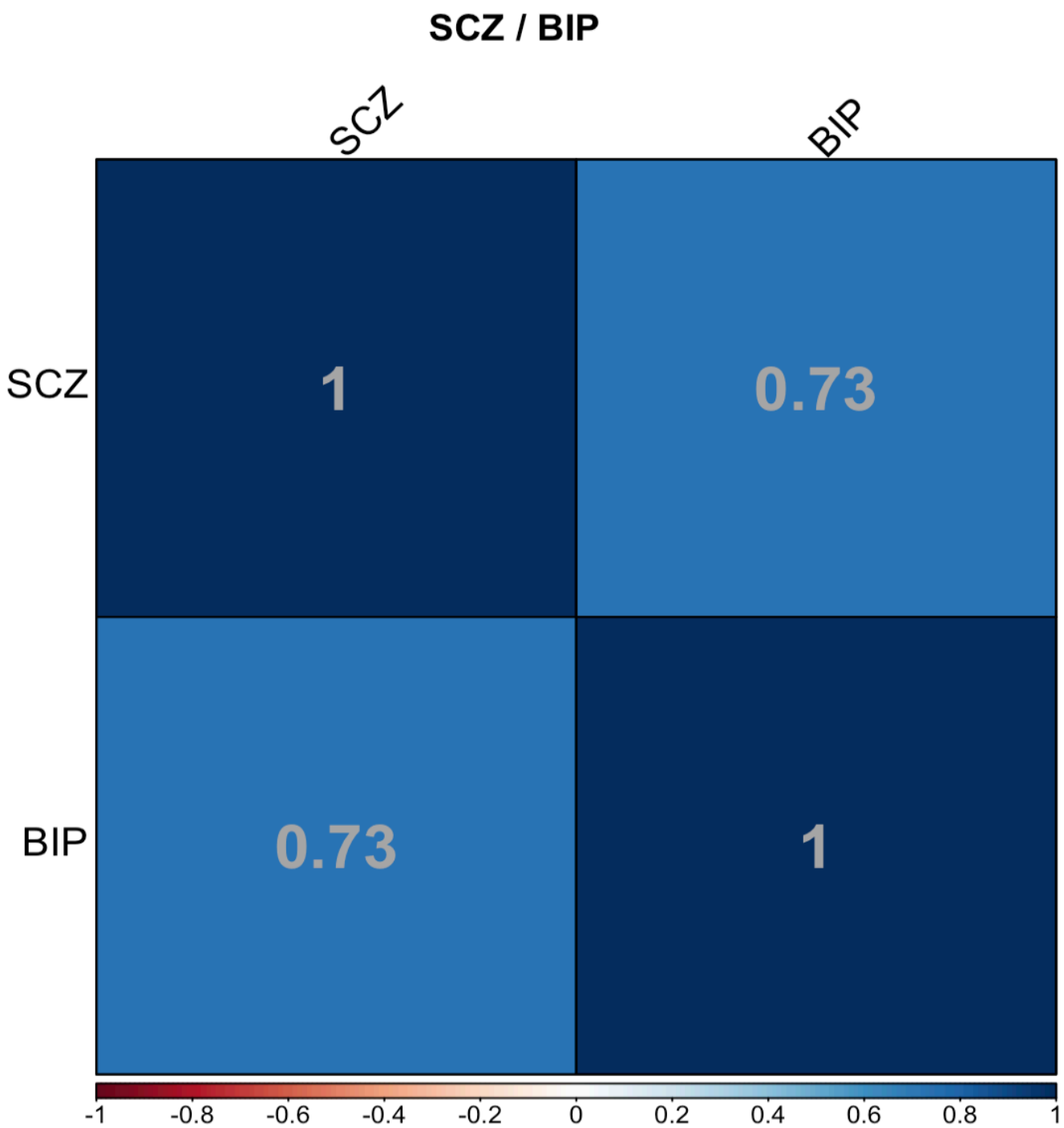
1. **traits**: a vector of file names/paths to files which point to the munged sumstats.
2. **sample.prev**: A vector of sample prevalences of length equal to the number of traits. Enter 0.5 if inputting sum of effective N.
3. **population.prev**: A vector of population prevalences.
4. **ld**: A folder 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.

LET'S GO TO THE CODE

In this second practical you will get practice running the *munge* and *ldsc* functions for binary traits

Note that bipolar disorder and schizophrenia are diagnostically exclusionary of one another.

What does a method like LDSC tell us about these two traits?


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