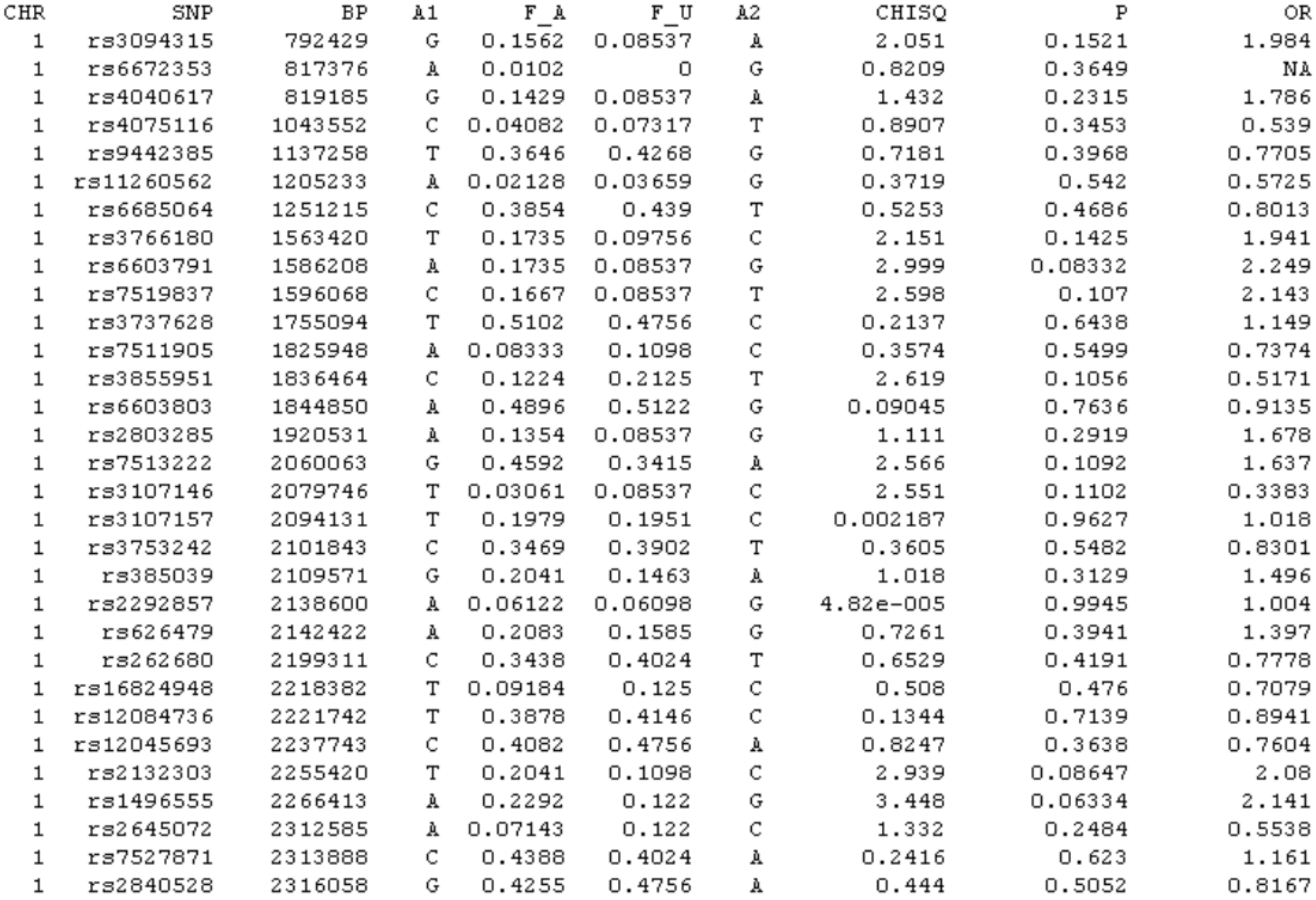
Genotek

Post-GWAS

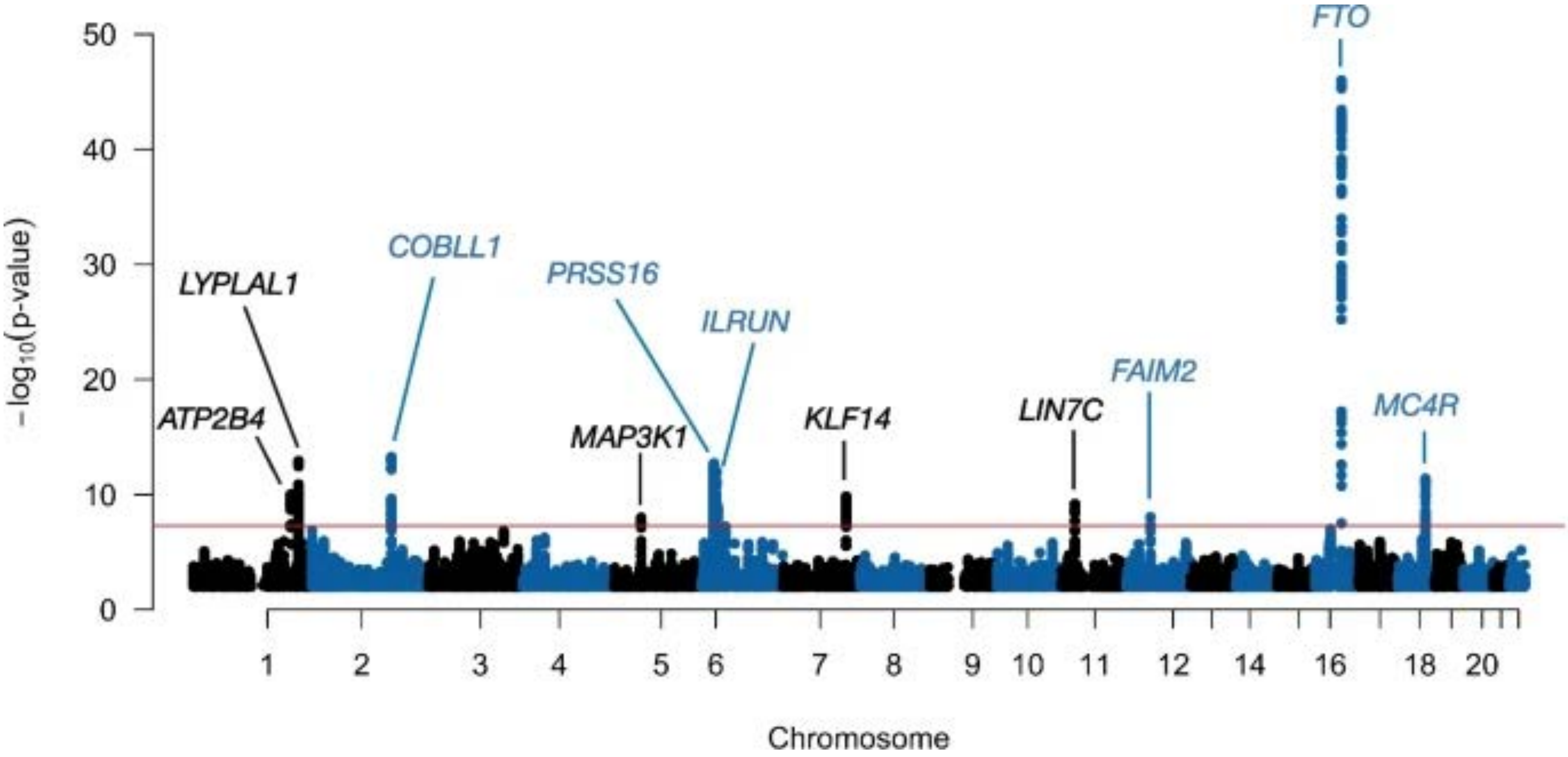
Семинар наставника 30.09.2024

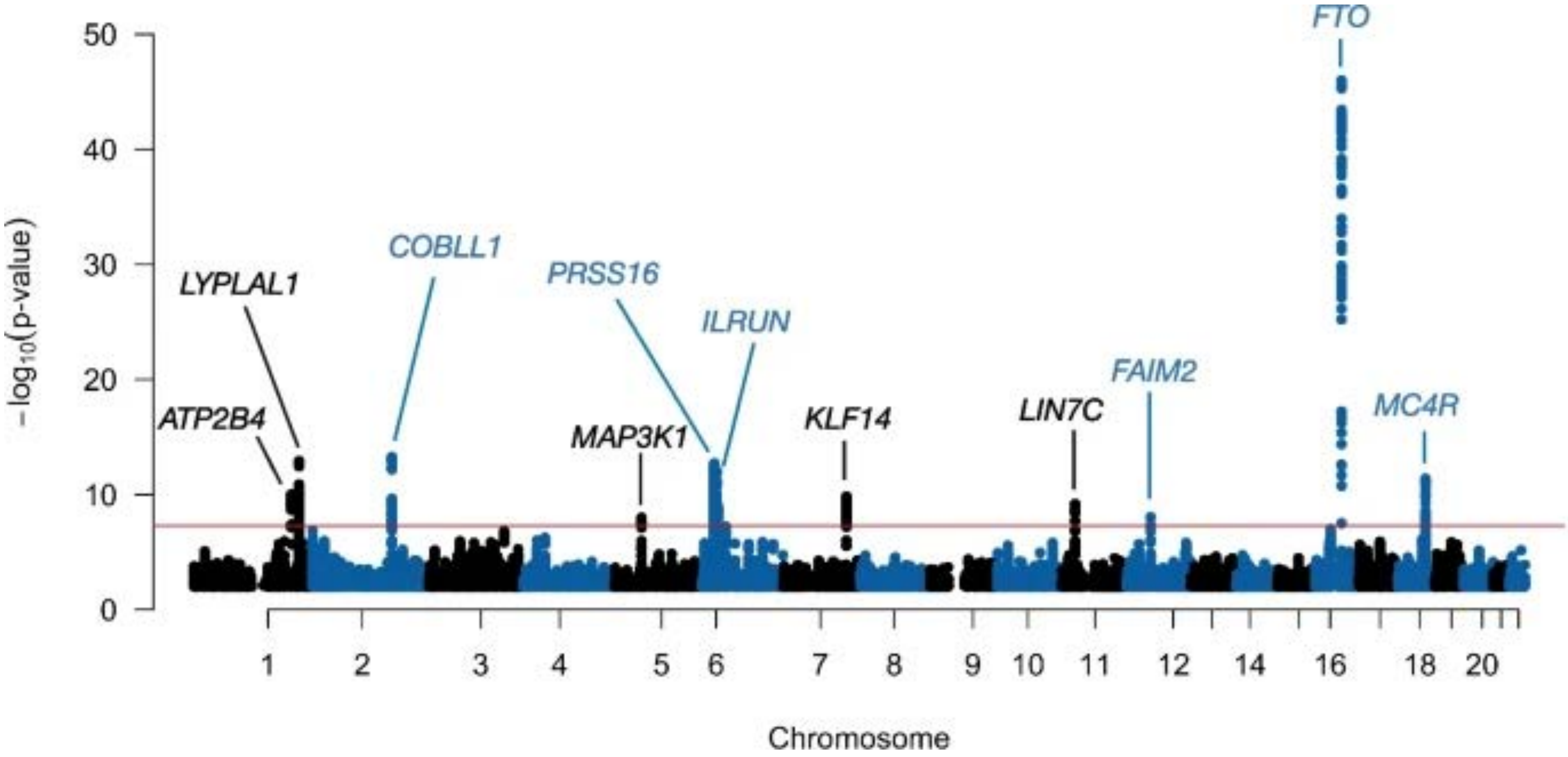
Александр Ракитько



```
CHR         SNP         BP   A1      F_A      F_U   A2       CHISQ          P        OR
  1   rs3094315     792429    G   0.1562  0.08537    A       2.051     0.1521     1.984
  1   rs6672353     817376    A   0.0102        0    G      0.8209     0.3649        NA
  1   rs4040617     819185    G   0.1429  0.08537    A       1.432     0.2315     1.786
  1   rs4075116    1043552    C  0.04082  0.07317    T      0.8907     0.3453     0.539
  1   rs9442385    1137258    T   0.3646   0.4268    G      0.7181     0.3968    0.7705
  1  rs11260562    1205233    A  0.02128  0.03659    G      0.3719      0.542    0.5725
  1   rs6685064    1251215    C   0.3854    0.439    T      0.5253     0.4686    0.8013
  1   rs3766180    1563420    T   0.1735  0.09756    C       2.151     0.1425     1.941
  1   rs6603791    1586208    A   0.1735  0.08537    G       2.999    0.08332     2.249
  1   rs7519837    1596068    C   0.1667  0.08537    T       2.598      0.107     2.143
  1   rs3737628    1755094    T   0.5102   0.4756    C      0.2137     0.6438     1.149
  1   rs7511905    1825948    A  0.08333   0.1098    C      0.3574     0.5499    0.7374
  1   rs3855951    1836464    C   0.1224   0.2125    T       2.619     0.1056    0.5171
  1   rs6603803    1844850    A   0.4896   0.5122    G     0.09045     0.7636    0.9135
  1   rs2803285    1920531    A   0.1354  0.08537    G       1.111     0.2919     1.678
  1   rs7513222    2060063    G   0.4592   0.3415    A       2.566     0.1092     1.637
  1   rs3107146    2079746    T  0.03061  0.08537    C       2.551     0.1102    0.3383
  1   rs3107157    2094131    T   0.1979   0.1951    C    0.002187     0.9627     1.018
  1   rs3753242    2101843    C   0.3469   0.3902    T      0.3605     0.5482    0.8301
  1    rs385039    2109571    G   0.2041   0.1463    A       1.018     0.3129     1.496
  1   rs2292857    2138600    A  0.06122  0.06098    G   4.82e-005     0.9945     1.004
  1    rs626479    2142422    A   0.2083   0.1585    G      0.7261     0.3941     1.397
  1    rs262680    2199311    C   0.3438   0.4024    T      0.6529     0.4191    0.7778
  1  rs16824948    2218382    T  0.09184    0.125    C       0.508      0.476    0.7079
  1  rs12084736    2221742    T   0.3878   0.4146    C      0.1344     0.7139    0.8941
  1  rs12045693    2237743    C   0.4082   0.4756    A      0.8247     0.3638    0.7604
  1   rs2132303    2255420    T   0.2041   0.1098    C       2.939    0.08647      2.08
  1   rs1496555    2266413    A   0.2292    0.122    G       3.448    0.06334     2.141
  1   rs2645072    2312585    A  0.07143    0.122    C       1.332     0.2484    0.5538
  1   rs7527871    2313888    C   0.4388   0.4024    A      0.2416      0.623     1.161
  1   rs2840528    2316058    G   0.4255   0.4756    A       0.444     0.5052    0.8167
```

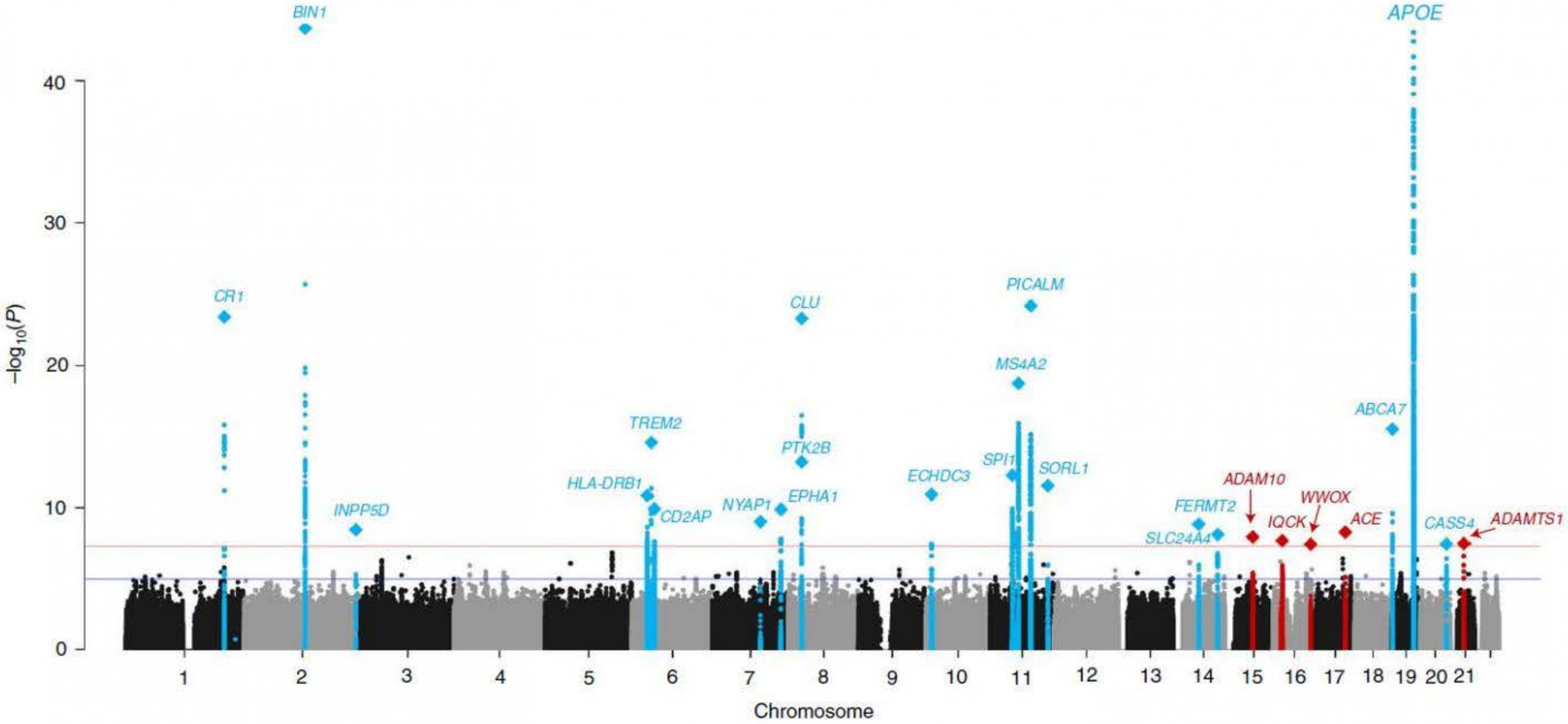
GWAS Summary Statistics

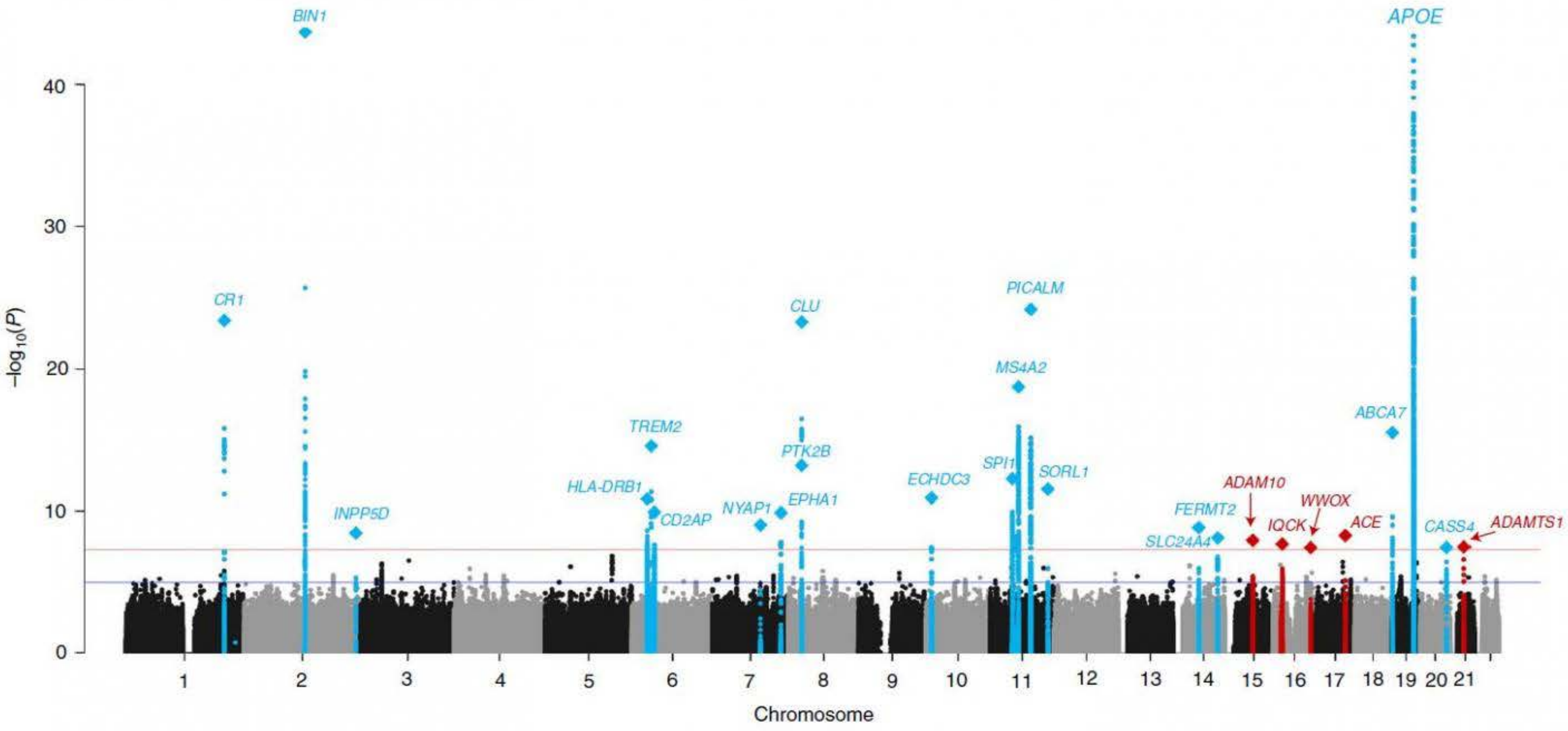

FTO
COBLL1
LYPLAL1
PRSS16
ILRUN
FAIM2
MC4R
ATP2B4
KLF14
LIN7C
MAP3K1
50
40
30
20
10
0
$-\log_{10}$(p-value)
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
20
Chromosome



Ожирение

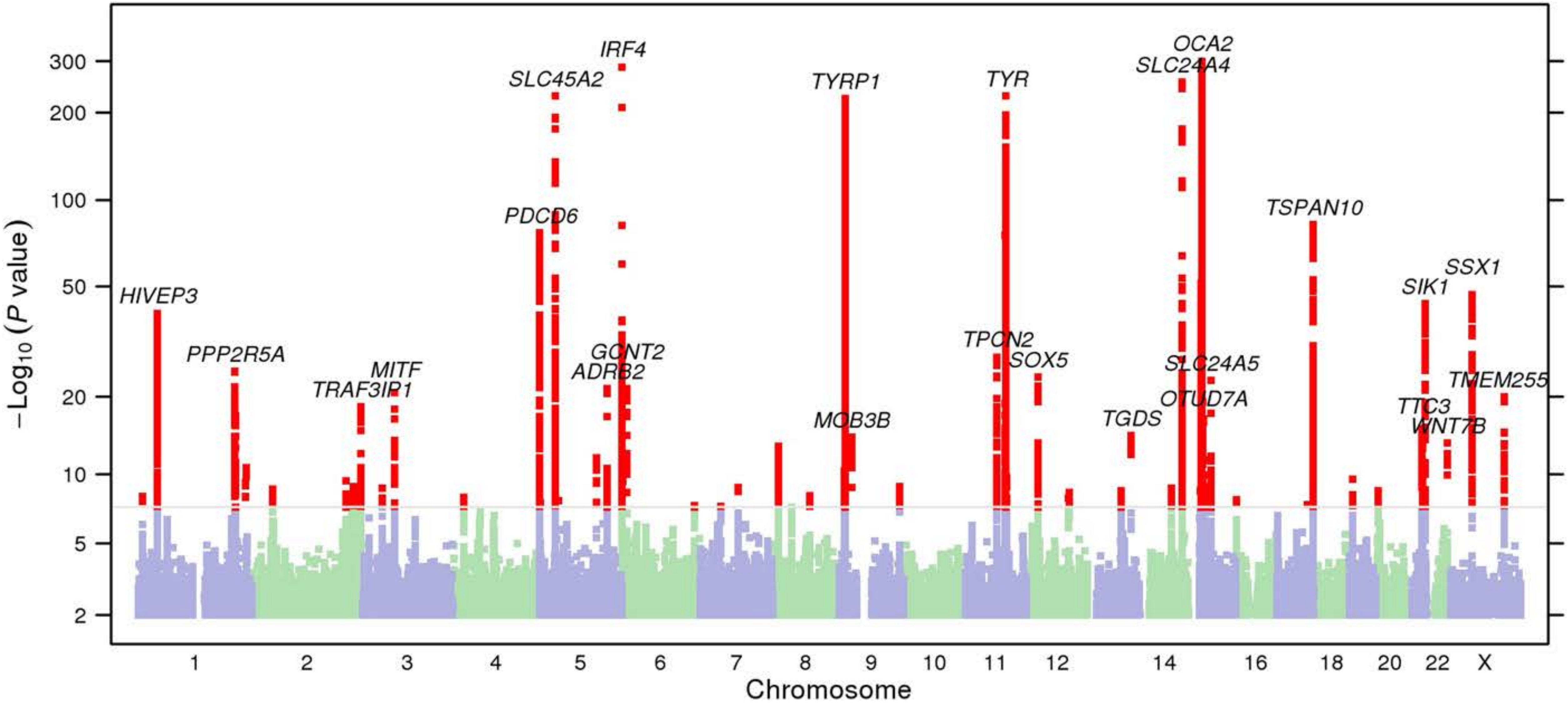
https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-024-01329-0

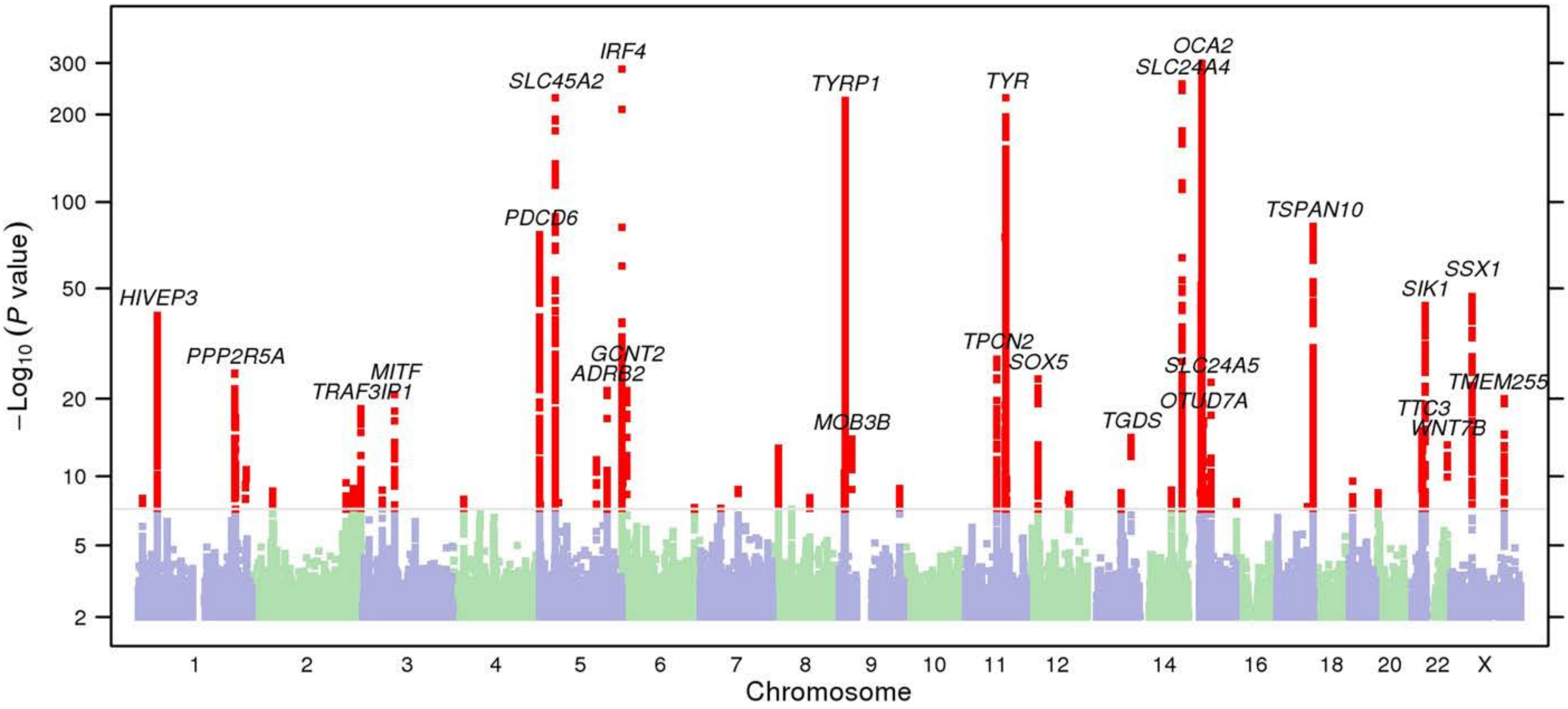

BIN1
APOE
CR1
CLU
PICALM
MS4A2
ABCA7
TREM2
PTK2B
SPI1
SORL1
ECHDC3
HLA-DRB1
CD2AP
NYAP1
EPHA1
INPP5D
FERMT2
SLC24A4
ADAM10
IQCK
WWOX
ACE
CASS4
ADAMTS1
40
30
20
10
0
−log10(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Chromosome



Болезнь Альцгеймера

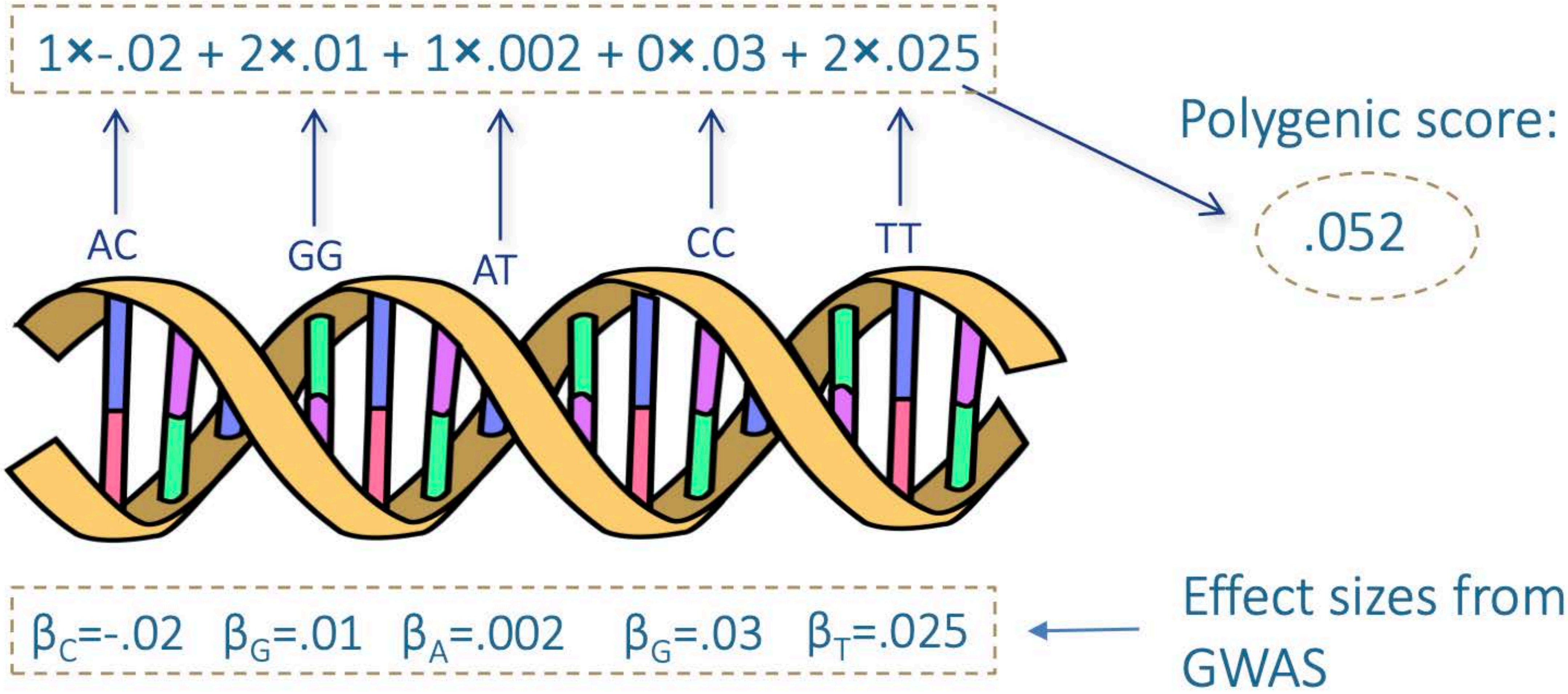
https://www.eurekalert.org/multimedia/863837


−Log10 (P value)
300
200
100
50
20
10
5
2
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
20
22
X
HIVEP3
PPP2R5A
TRAF3IP1
MITF
PDCD6
SLC45A2
IRF4
ADRB2
GCNT2
TYRP1
MOB3B
TPCN2
TYR
SOX5
TGDS
OCA2
SLC24A4
SLC24A5
OTUD7A
TSPAN10
SIK1
TTC3
WNT7B
SSX1
TMEM255



Цвет глаз

https://www.science.org/doi/10.1126/sciadv.abd1239



$1 \times -.02 + 2 \times .01 + 1 \times .002 + 0 \times .03 + 2 \times .025$

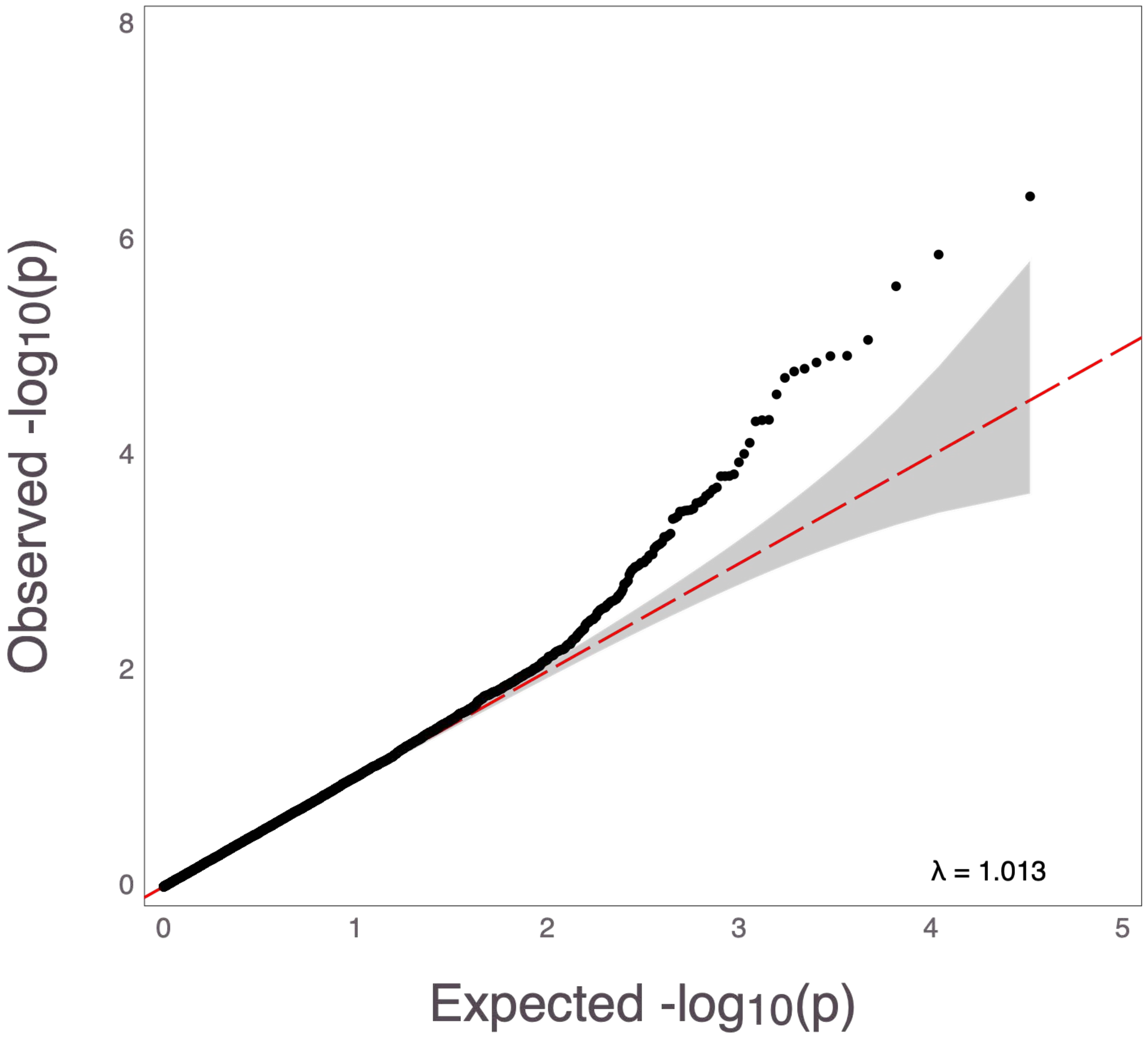
Polygenic score:

.052

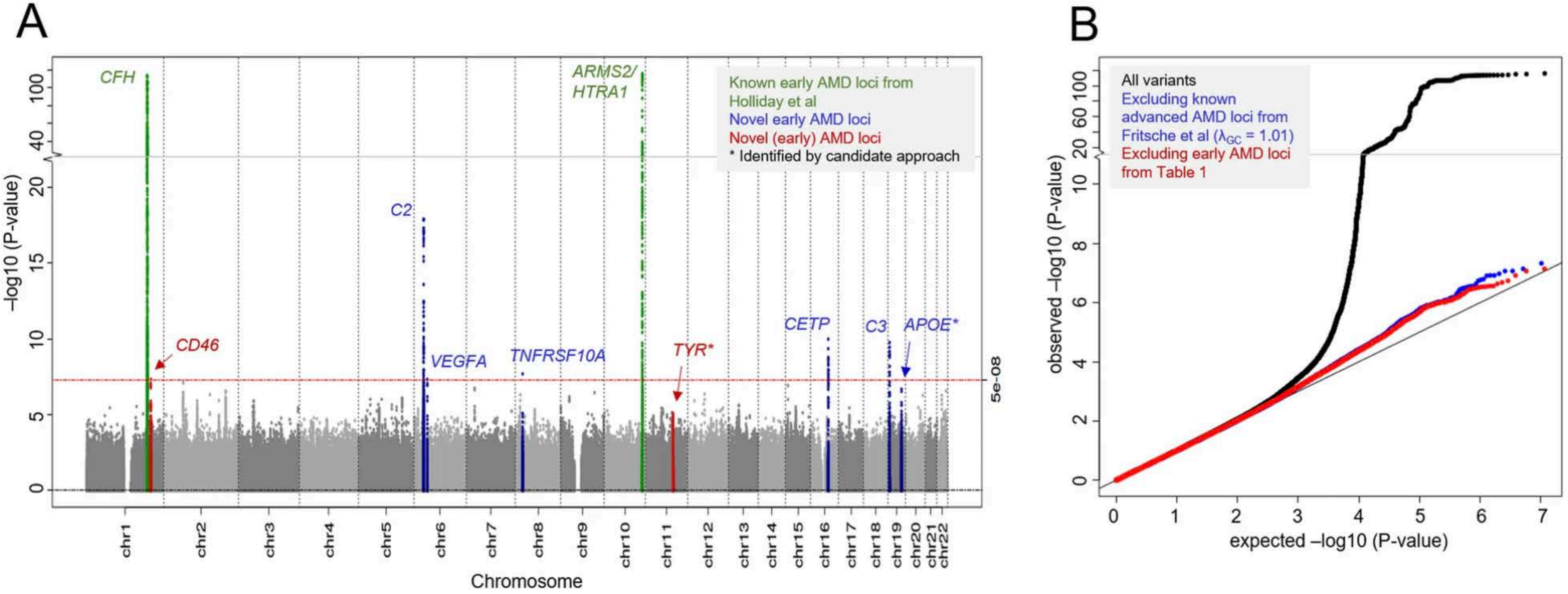
$\beta_C = -.02$ $\beta_G = .01$ $\beta_A = .002$ $\beta_G = .03$ $\beta_T = .025$

Effect sizes from GWAS

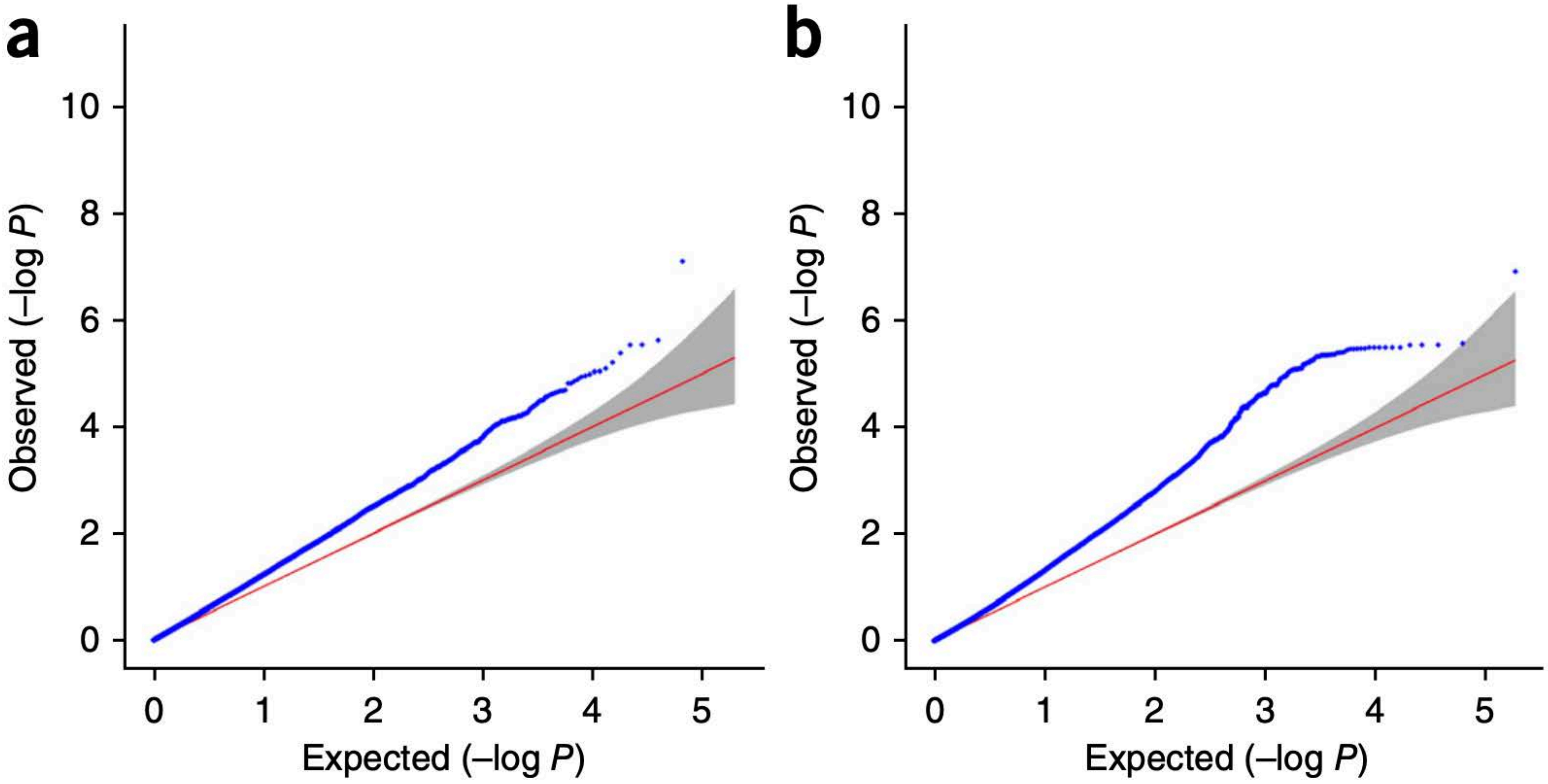
QQ Plot of GWAS p-values



From: Genome-wide association meta-analysis for early age-related macular degeneration highlights novel loci and insights for advanced disease



Early AMD meta-analysis. Shown are the association results of the meta-analysis for early AMD: **a** by their position on the genome (Manhattan plot) with color indicating whether the locus was previously identified by Holliday et al. [12] (blue), novel for early AMD (red), or among the other advanced AMD loci identified by Fritsche et al. [9] (green); and **b** their distribution (QQ plot)



На одном из QQ-графиков симуляция с популяционной стратификацией, а на другом с полигонной наследуемость. Коэффициент инфляции в обоих случаях равен 1.32

nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6], Schizophrenia Working Group of the Psychiatric Genomics Consortium[7], Nick Patterson[1], Mark J Daly[1–3], Alkes L Price[1,4,8] & Benjamin M Neale[1–3]

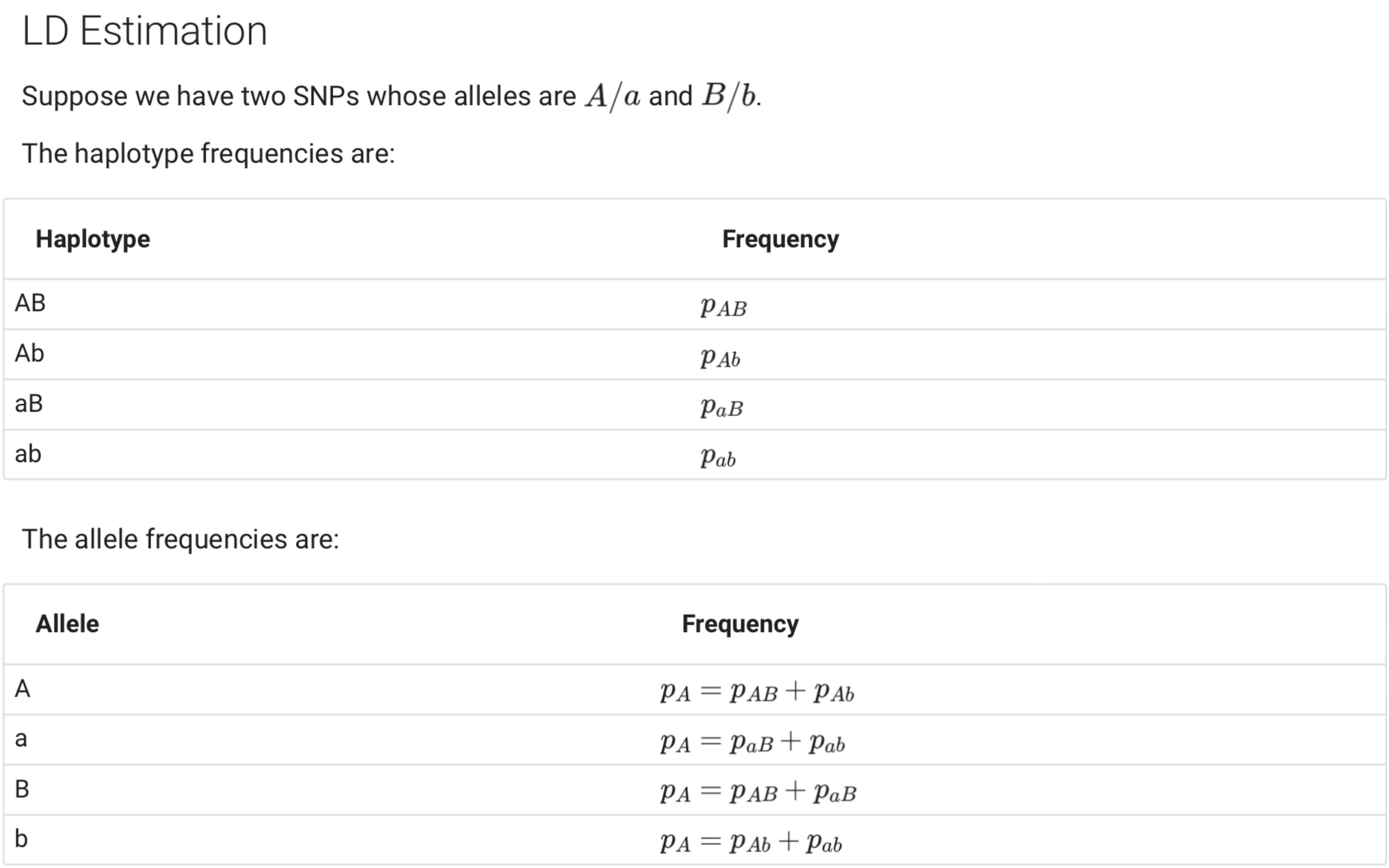
Both polygenicity (many small genetic effects) and confounding biases, such as cryptic relatedness and population stratification, can yield an inflated distribution of test statistics in genome-wide association studies (GWAS). However, current methods cannot distinguish between inflation from a true polygenic signal and bias. We have developed an approach, LD Score regression, that quantifies the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD). The LD Score regression intercept can be used to estimate a more powerful and accurate correction factor than genomic control. We find strong evidence that polygenicity accounts for the majority of the inflation in test statistics in many GWAS of large sample size.

of this equation is provided in the **Supplementary Note**).This relationship holds for meta-analyses and also for ascertained studies of binary phenotypes, in which case h^2 is on the observed scale. Consequently, if we regress the χ^2 statistics from GWAS against LD Score (LD Score regression), the intercept minus one is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.

RESULTS

Overview of methods

We estimated LD Scores from the European-ancestry samples in the 1000 Genomes Project[7] (EUR) using an unbiased estimator[8] of r^2 with 1-cM windows, singletons excluded (MAF > 0.13%) and no r^2 cutoff. Standard errors were estimated by jackknifing over blocks of



LD Estimation

Suppose we have two SNPs whose alleles are A/a and B/b.

The haplotype frequencies are:

Haplotype	Frequency
AB	p_{AB}
Ab	p_{Ab}
aB	p_{aB}
ab	p_{ab}

The allele frequencies are:

Allele	Frequency
A	$p_A = p_{AB} + p_{Ab}$
a	$p_A = p_{aB} + p_{ab}$
B	$p_A = p_{AB} + p_{aB}$
b	$p_A = p_{Ab} + p_{ab}$

https://cloufield.github.io/GWASTutorial/19_ld/#ld-estimation

D : the level of LD between A and B can be estimated using **coefficient of linkage disequilibrium (D)**, which is defined as:

$$D_{AB} = p_{AB} - p_A p_B$$

If A and B are in **linkage equilibrium**, we can get

$$D_{AB} = p_{AB} - p_A p_B = 0$$

which means the coefficient of linkage disequilibrium is 0 in this case.

D can be calculated for each pair of alleles and their relationships can be expressed as:

$$D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$$

So we can simply denote $D = D_{AB}$, and the relationship between haplotype frequencies and allele frequencies can be summarized in the following table.

Allele	A	a	Total
B	$p_{AB} = p_A p_B + D$	$p_{aB} = p_a p_B - D$	p_B
b	$p_{AB} = p_A p_b - D$	$p_{AB} = p_a p_b + D$	p_b
Total	p_A	p_a	1

⚠ **The range of possible values of D depends on the allele frequencies, which is not suitable for comparison between different pairs of alleles.**

Lewontin suggested a method for the normalization of D :

$$D_{normalized} = \frac{D}{D_{max}}$$

where

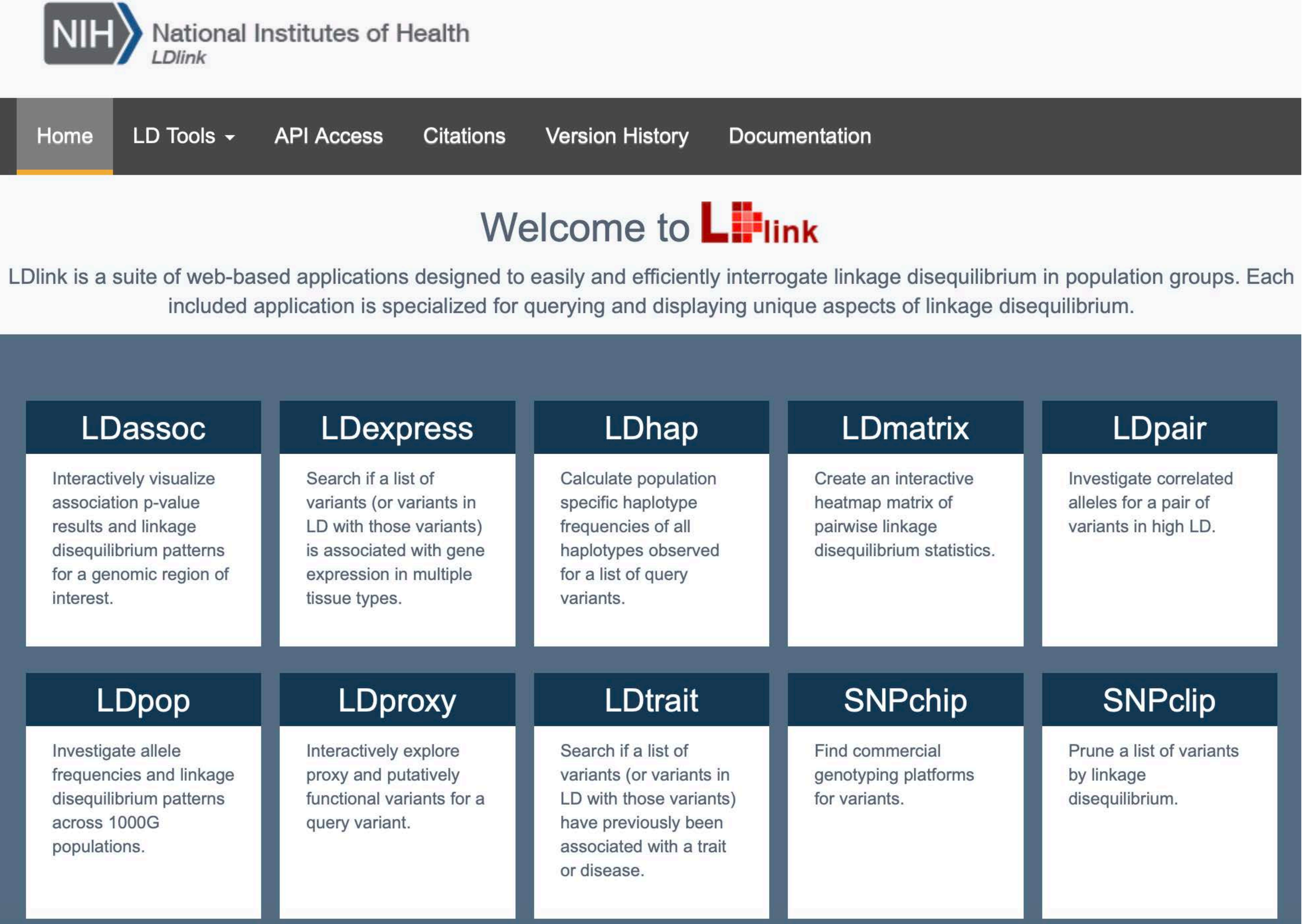
$$D_{max} = \begin{cases} max\{-p_A p_B, -(1-p_A)(1-p_B)\} & \text{when } D < 0 \\ min\{p_A(1-p_B), p_B(1-p_A)\} & \text{when } D > 0 \end{cases}$$

It measures how much proportion of the haplotypes had undergone recombination.

In practice, the most commonly used alternative metric to $D_{normalized}$ is r^2, the correlation coefficient, which can be obtained by:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

Reference: Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics, 9(6), 477-485.



Welcome to LDlink

LDlink is a suite of web-based applications designed to easily and efficiently interrogate linkage disequilibrium in population groups. Each included application is specialized for querying and displaying unique aspects of linkage disequilibrium.

LDassoc

Interactively visualize association p-value results and linkage disequilibrium patterns for a genomic region of interest.

LDexpress

Search if a list of variants (or variants in LD with those variants) is associated with gene expression in multiple tissue types.

LDhap

Calculate population specific haplotype frequencies of all haplotypes observed for a list of query variants.

LDmatrix

Create an interactive heatmap matrix of pairwise linkage disequilibrium statistics.

LDpair

Investigate correlated alleles for a pair of variants in high LD.

LDpop

Investigate allele frequencies and linkage disequilibrium patterns across 1000G populations.

LDproxy

Interactively explore proxy and putatively functional variants for a query variant.

LDtrait

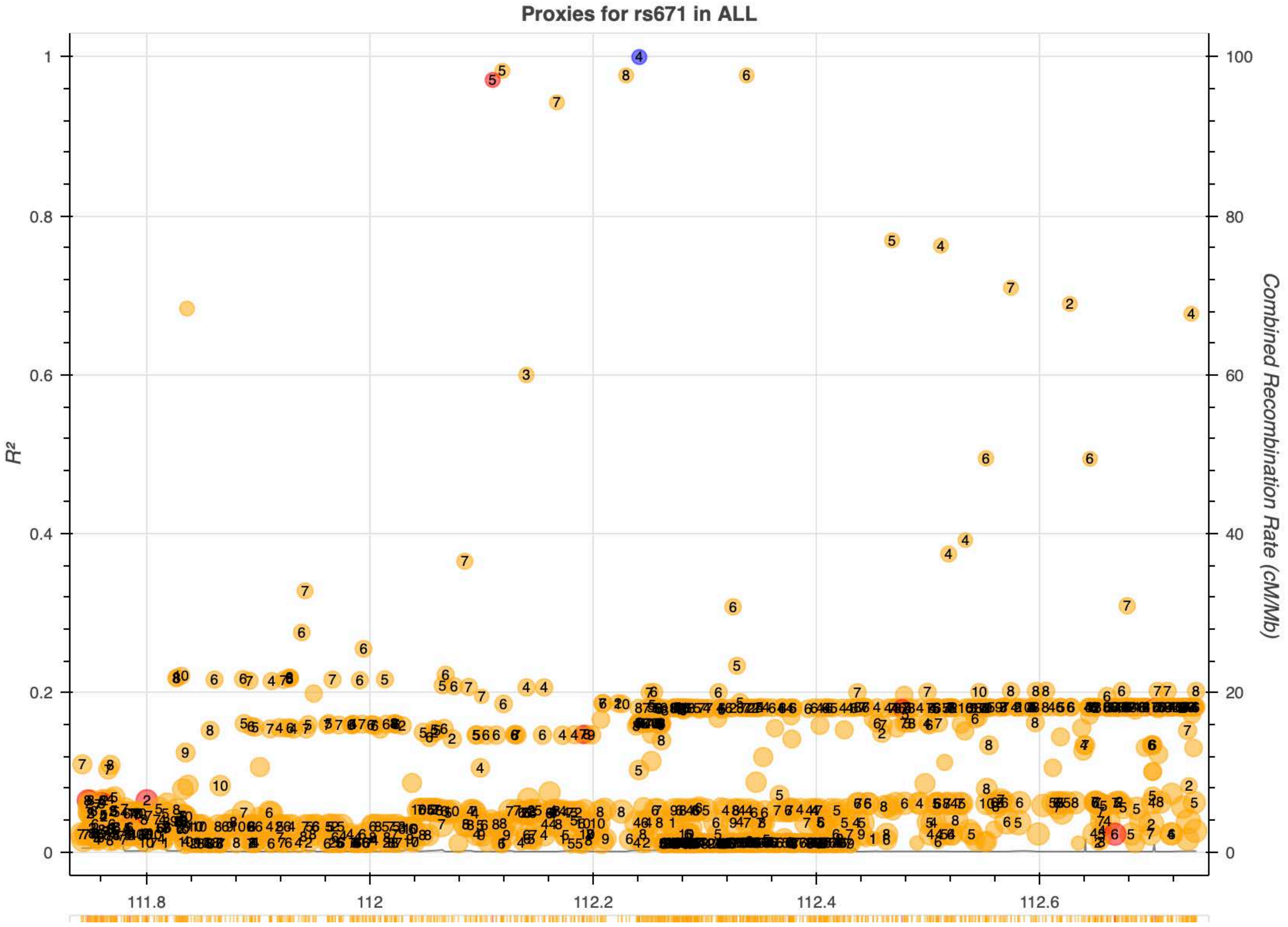
Search if a list of variants (or variants in LD with those variants) have previously been associated with a trait or disease.

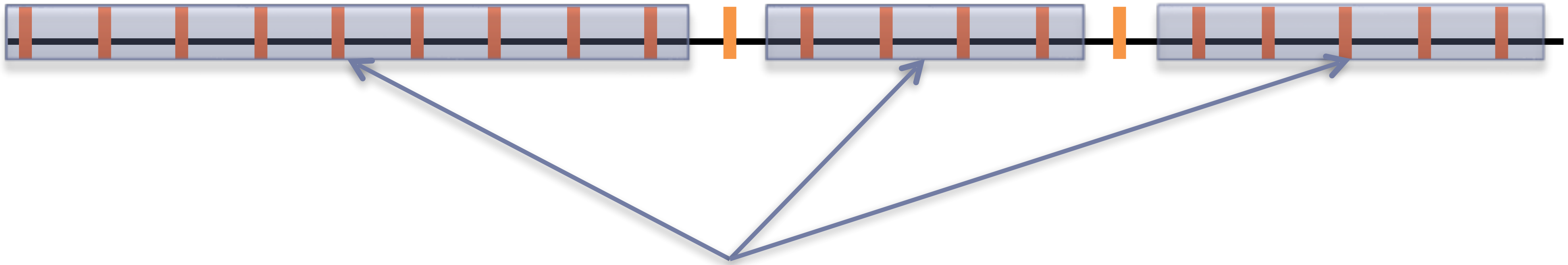
SNPchip

Find commercial genotyping platforms for variants.

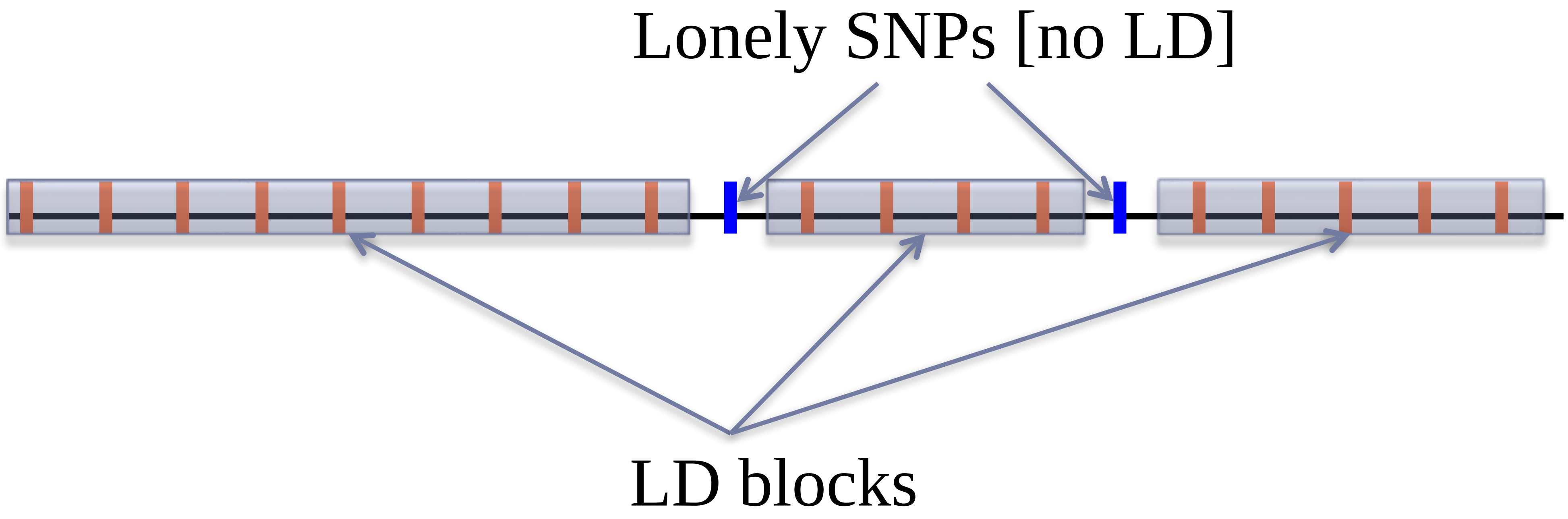
SNPclip

Prune a list of variants by linkage disequilibrium.


Proxies for rs671 in ALL
R²
Combined Recombination Rate (cM/Mb)
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
111.8
112
112.2
112.4
112.6

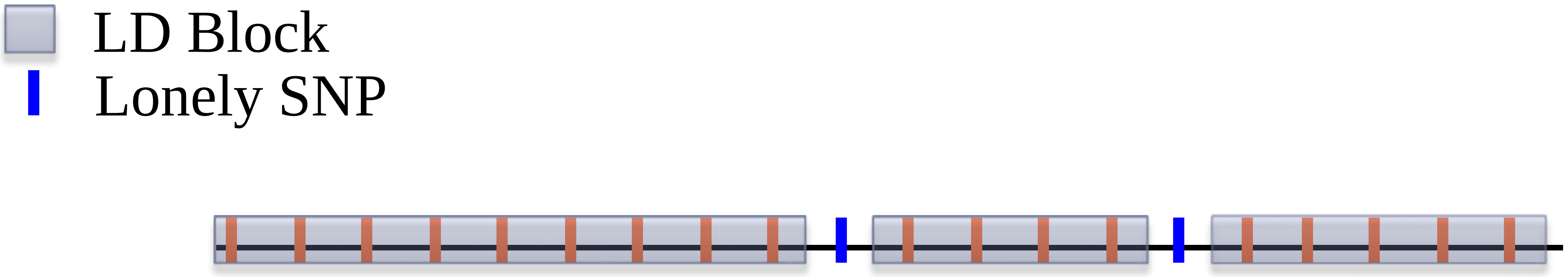


LD blocks



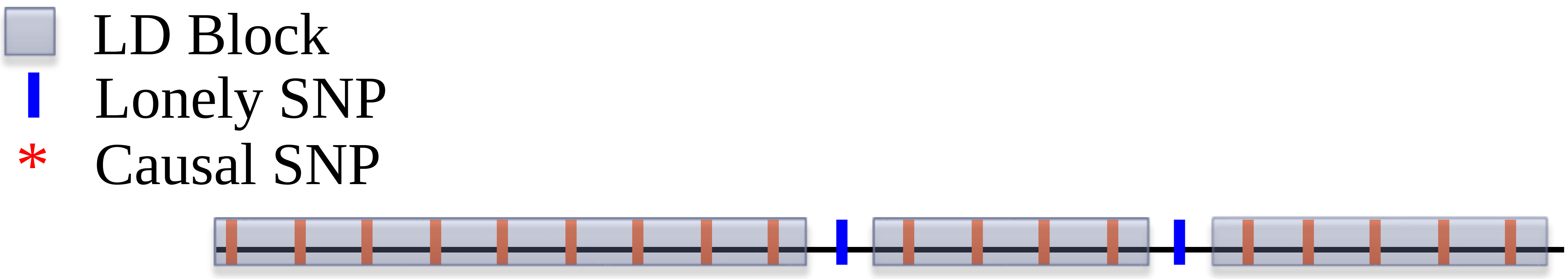
Lonely SNPs [no LD]

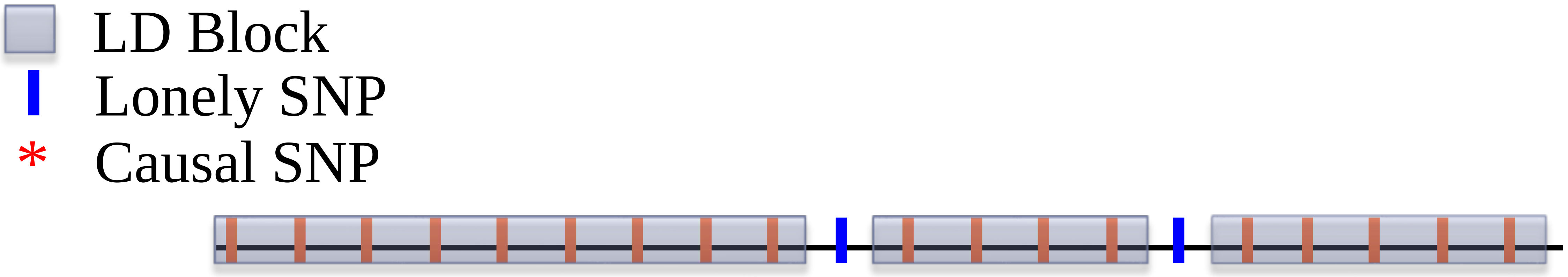
LD blocks



LD Block

Lonely SNP


LD Block
Lonely SNP
Causal SNP

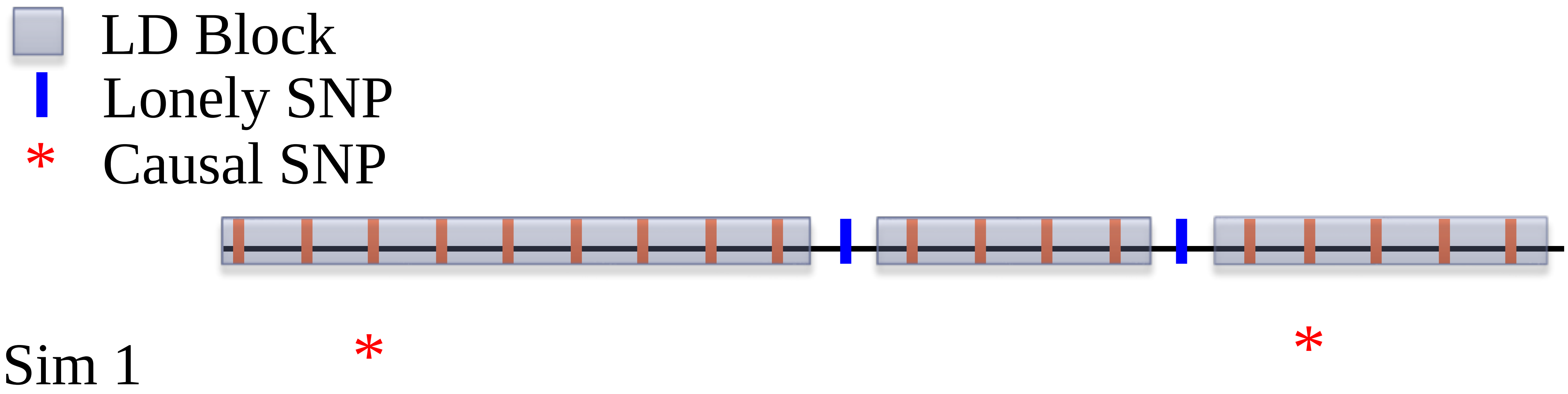


▭ LD Block

| Lonely SNP

* Causal SNP

Sim 1

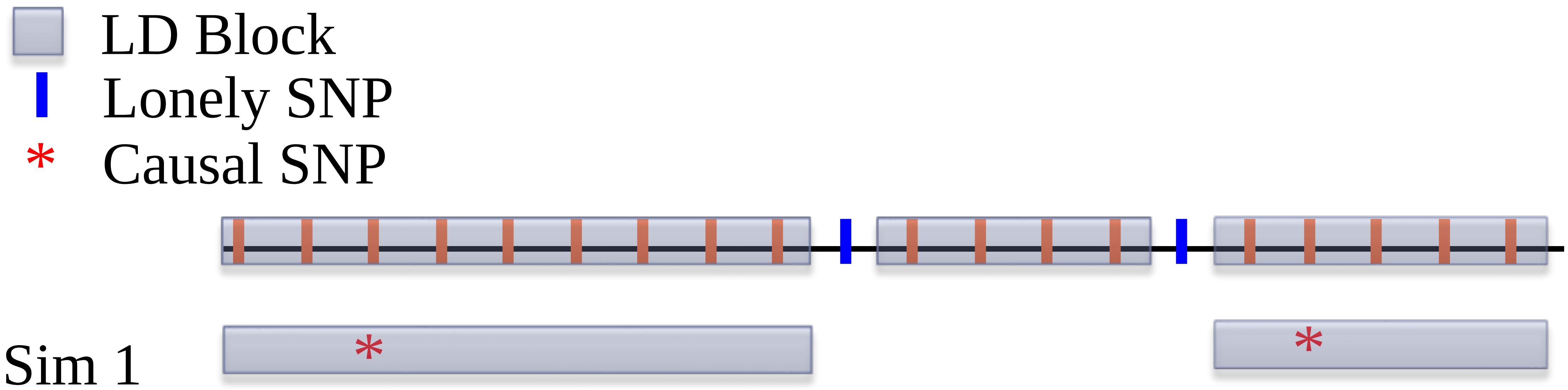


▢ LD Block

| Lonely SNP

* Causal SNP

Sim 1

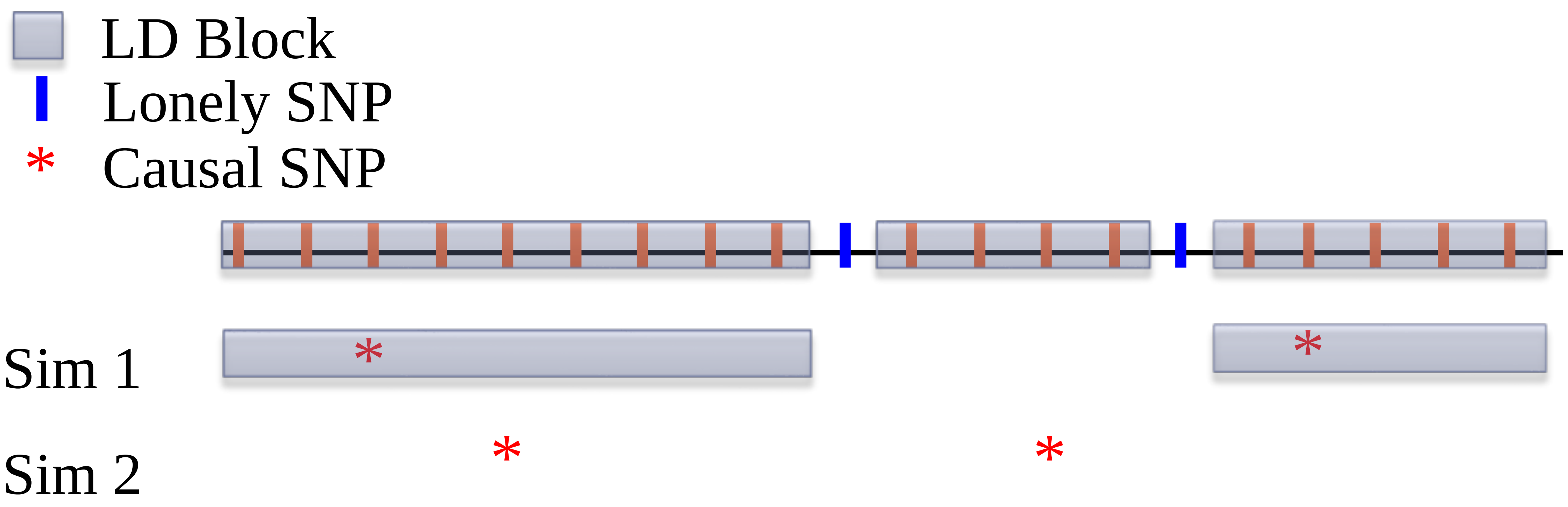


▨ LD Block

| Lonely SNP

* Causal SNP

Sim 1



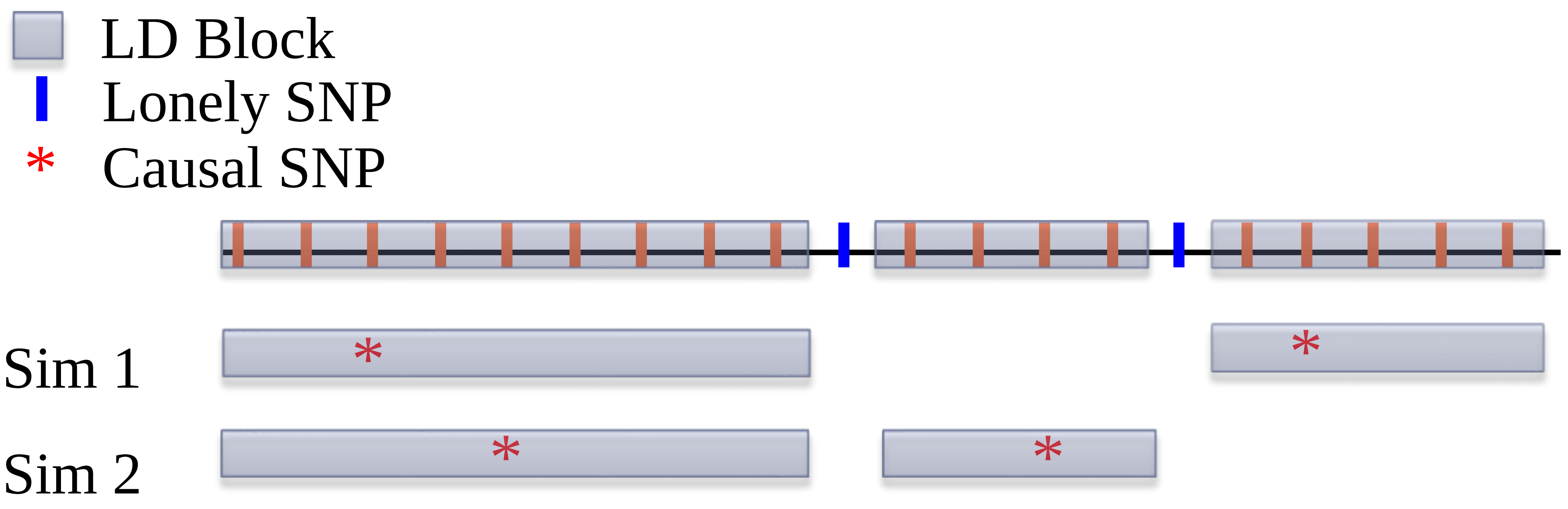
▭ LD Block

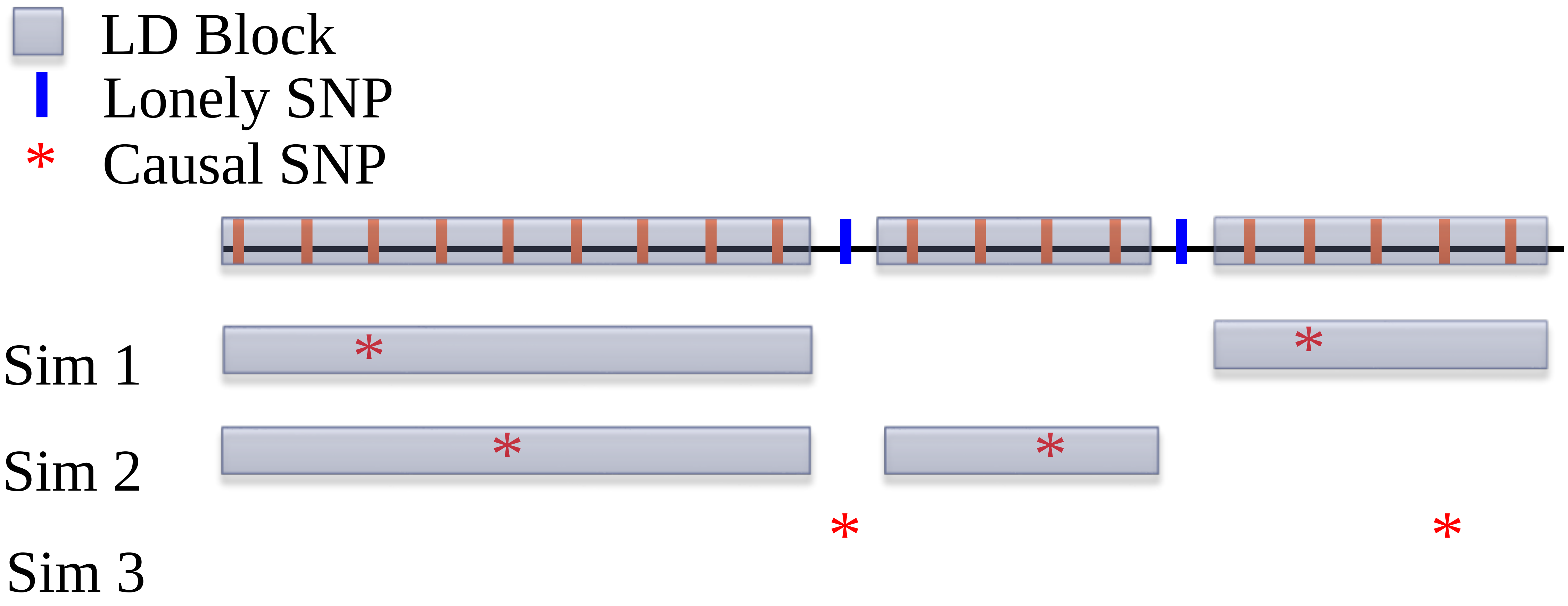
| Lonely SNP

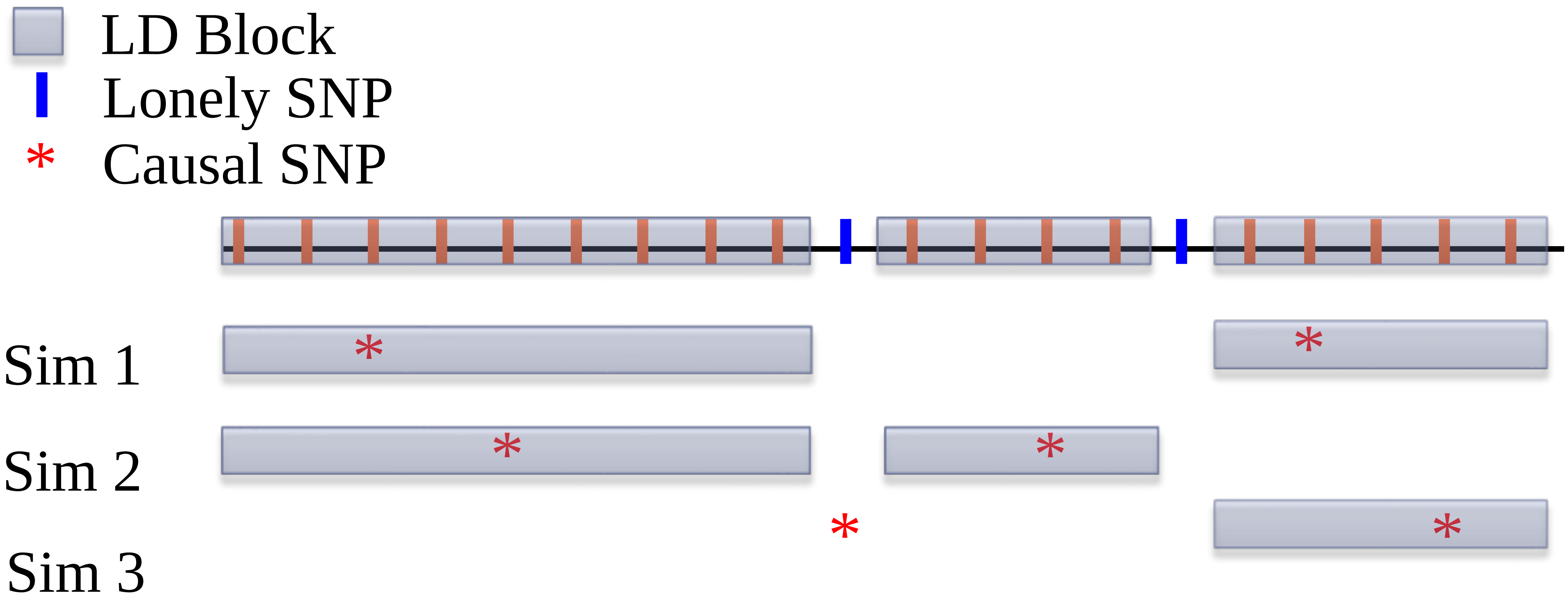
* Causal SNP

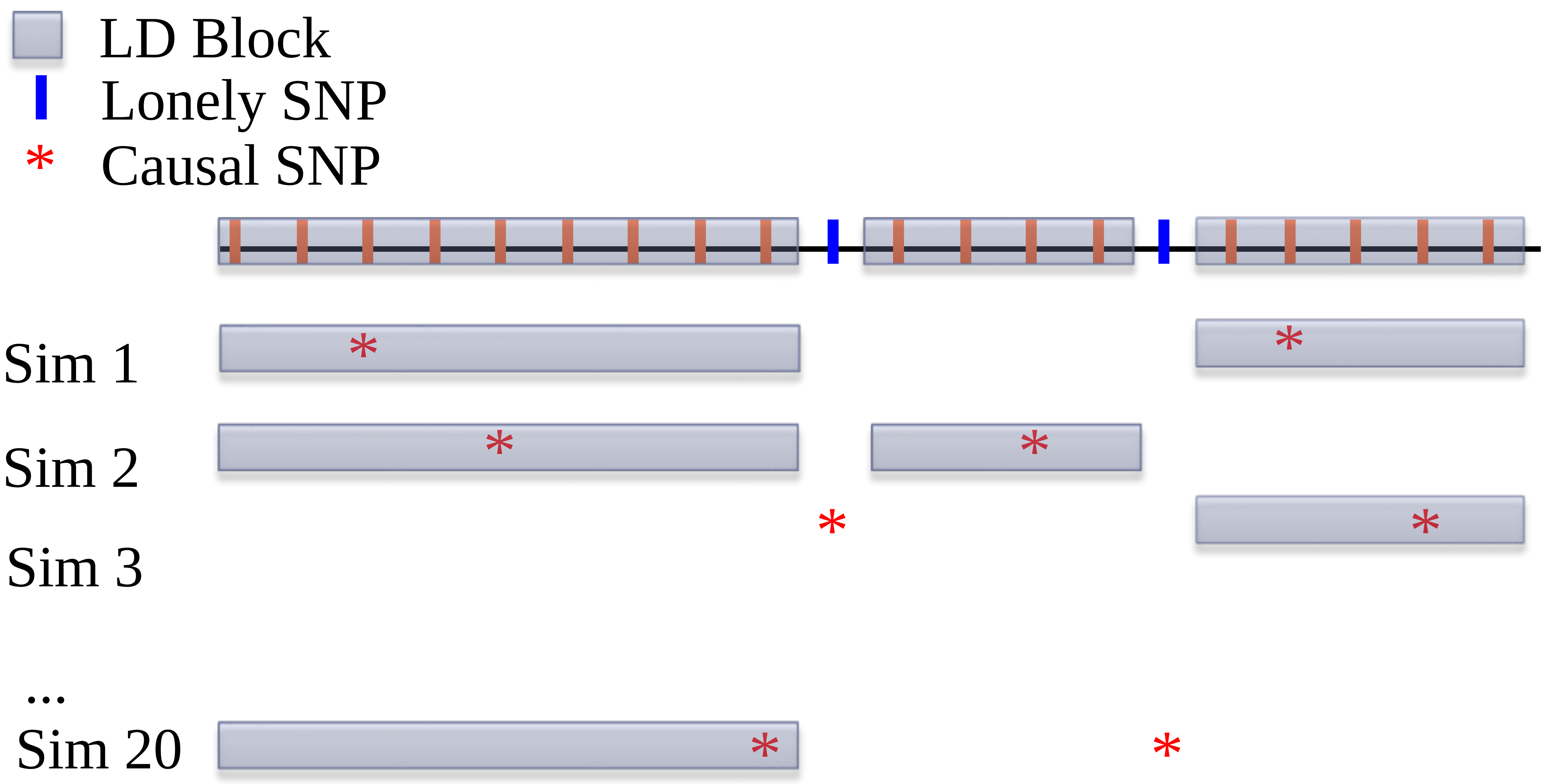
Sim 1

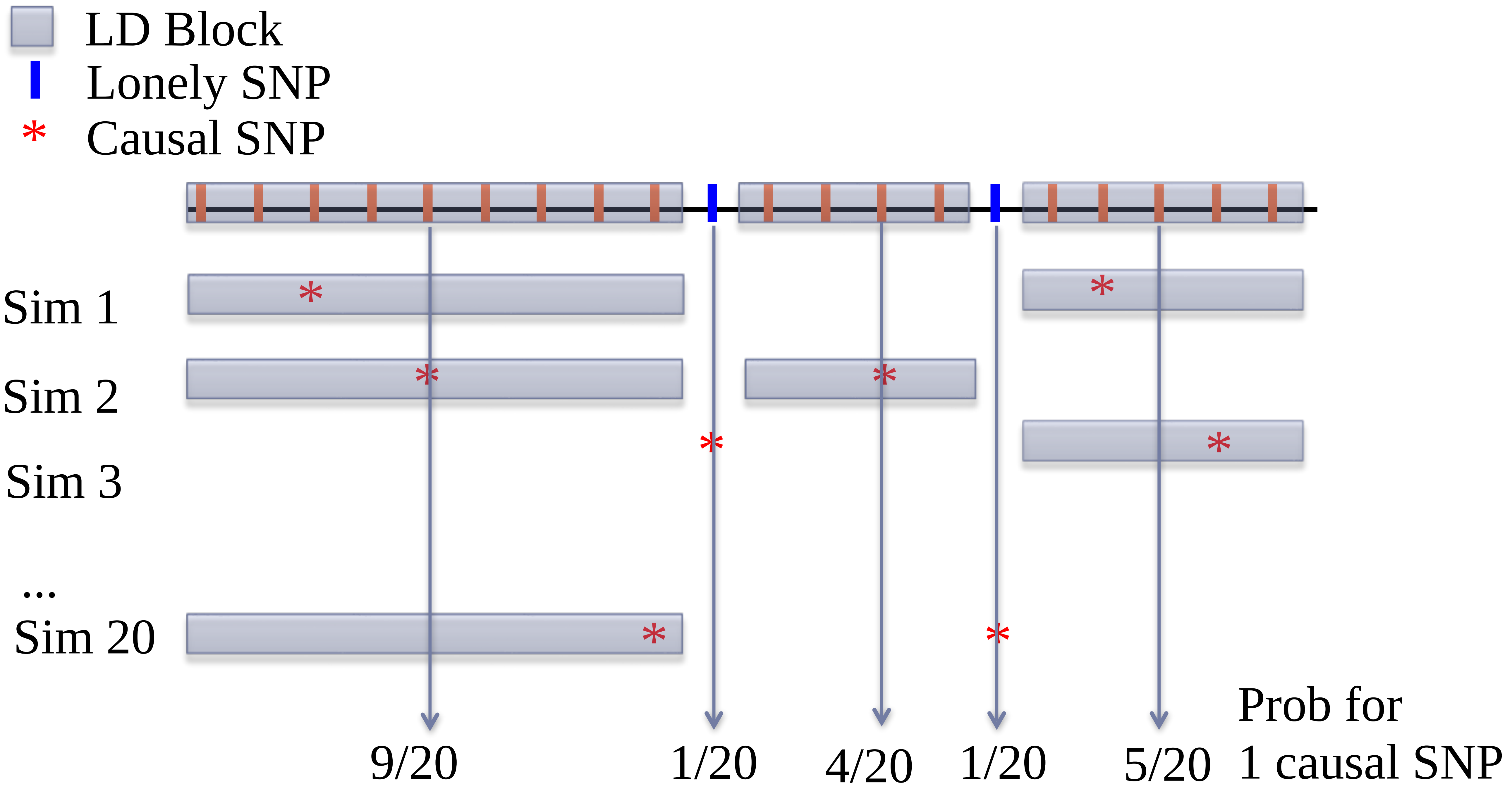
Sim 2


LD Block
Lonely SNP
Causal SNP
Sim 1
Sim 2


LD Block
Lonely SNP
Causal SNP
Sim 1
Sim 2
Sim 3


LD Block
Lonely SNP
Causal SNP
Sim 1
Sim 2
Sim 3


LD Block
Lonely SNP
Causal SNP
Sim 1
Sim 2
Sim 3
...
Sim 20



▢ LD Block

| Lonely SNP

* Causal SNP

Sim 1

Sim 2

Sim 3

...

Sim 20

9/20 1/20 4/20 1/20 5/20

Prob for
1 causal SNP

Heritability

Heritability is a term used in genetics to describe how much phenotypic variation can be explained by genetic variation.

For any phenotype, its variation $Var(P)$ can be modeled as the combination of **genetic effects** $Var(G)$ and **environmental effects** $Var(E)$.

$$Var(P) = Var(G) + Var(E)$$

Broad-sense Heritability

The **broad-sense heritability** $H^2_{broad-sense}$ is mathmatically defined as :

$$H^2_{broad-sense} = \frac{Var(G)}{Var(P)}$$

LD: Linkage disequilibrium

Linkage disequilibrium (LD) : non-random association of alleles at different loci in a given population. (Wiki)

LD score

LD score l_j for a SNP j is defined as the sum of r^2 for the SNP and other SNPs in a region.

$$l_j = \Sigma_k r^2_{j,k}$$

LD score regression

Key idea: A variant will have higher test statistics if it is in LD with causal variant, and the elevation is proportional to the correlation (r^2) with the causal variant.

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

- N: sample size.
- M : number of SNPs.
- h^2 : observed-scale heritability
- a : the effect of confounding factors, including crytic relatedness and populatiuon stratification.

For more details of LD score regression, please refer to : - Bulik-Sullivan, Brendan K., et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." Nature genetics 47.3 (2015): 291-295.

In words:

Test statistic = average causal effect per SNP * LD score + inflation due to population stratification + 1

$$E(\chi^2 \mid l_j) = \frac{Nh^2}{M} l_j + Na + 1$$

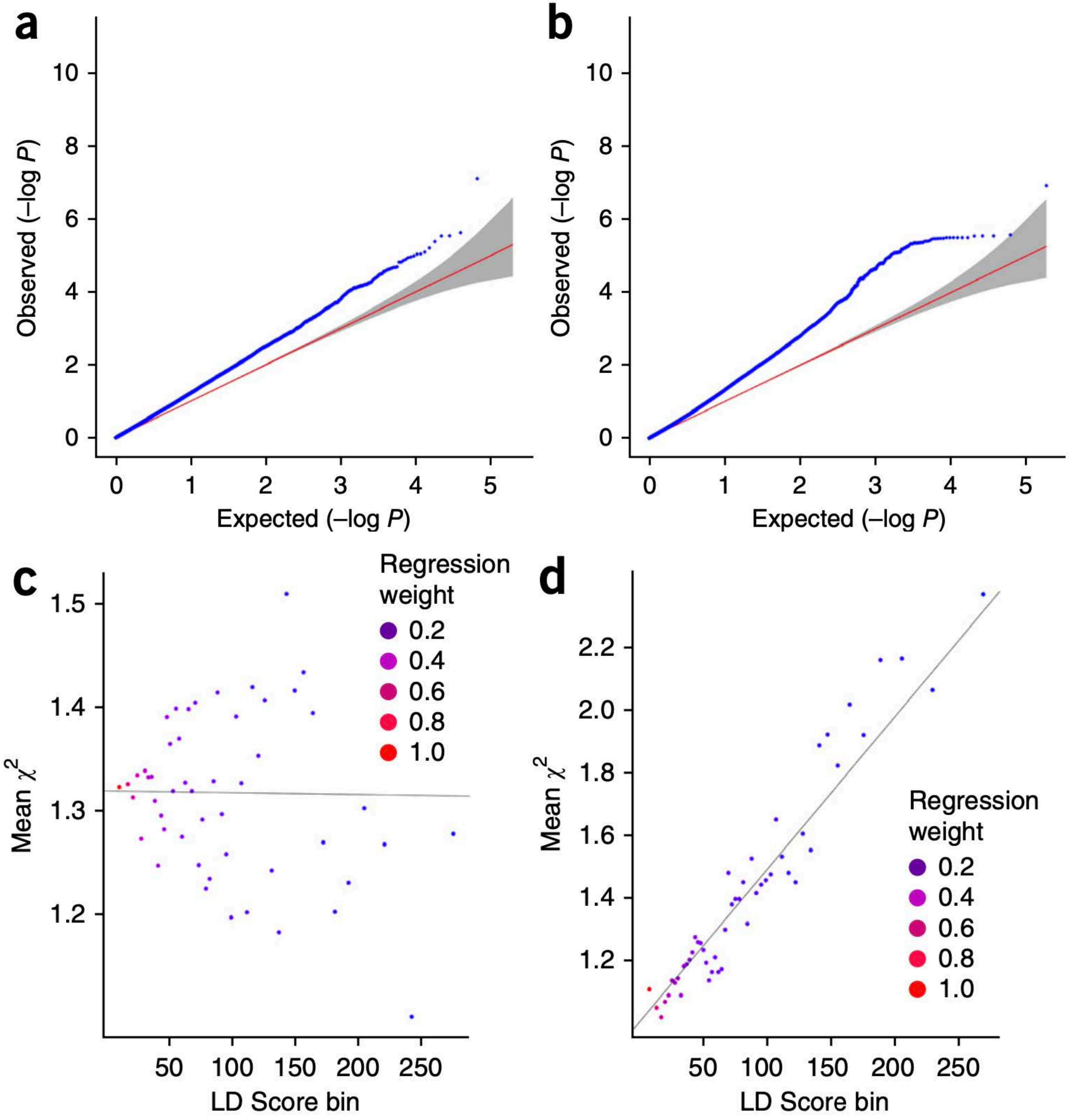
$$l_j = \sum_k r_{jk}^2$$ (LD Score)

N = sample size

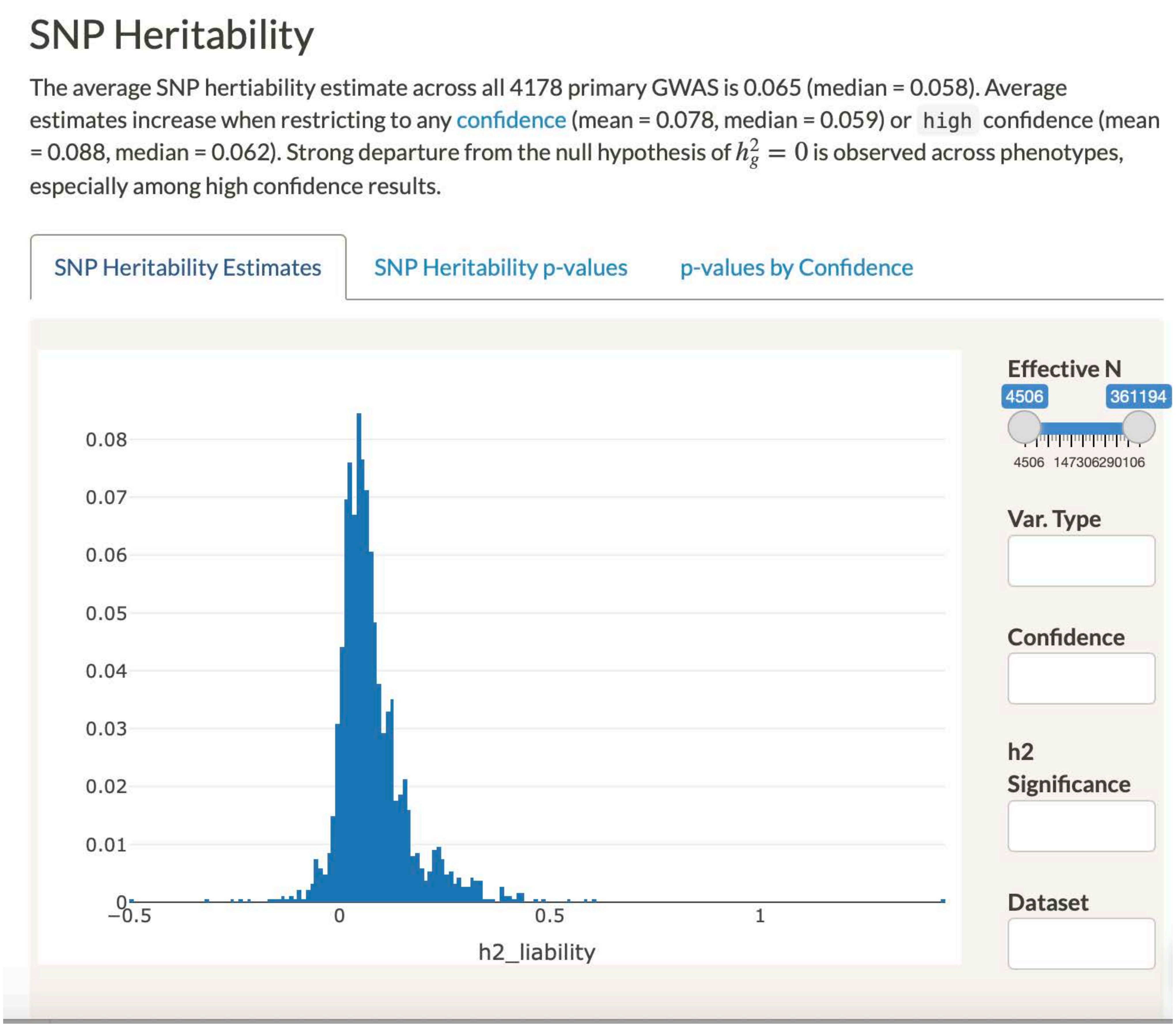
M = number of SNPs

h^2 / M = average heritability per SNP

a = Population structure /
cryptic relatedness

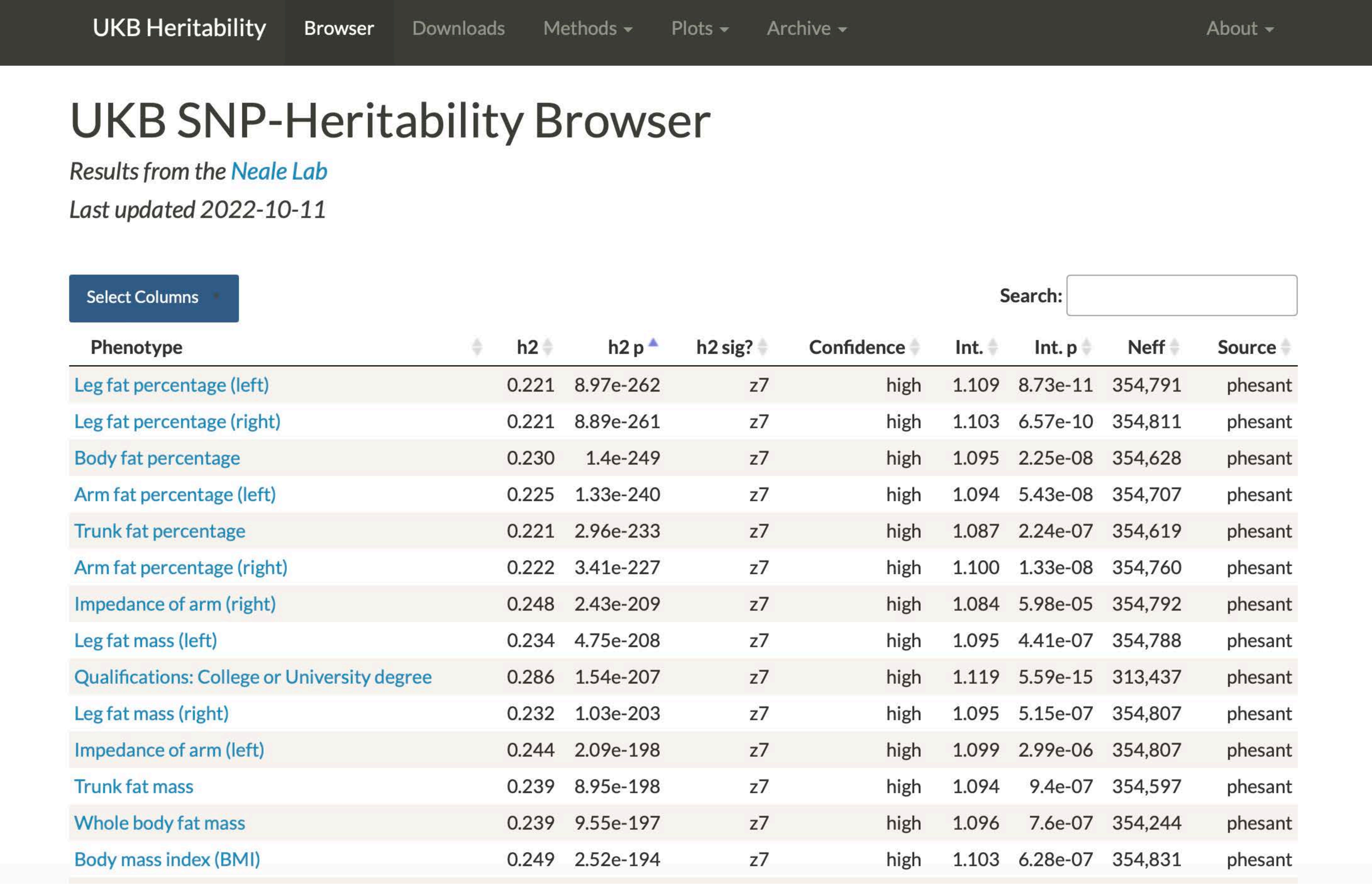


a
b
Observed (−log P)
Expected (−log P)
0
2
4
6
8
10
0
1
2
3
4
5
c
d
Mean χ^2
LD Score bin
1.2
1.3
1.4
1.5
1.2
1.4
1.6
1.8
2.0
2.2
50
100
150
200
250
Regression weight
● 0.2
● 0.4
● 0.6
● 0.8
● 1.0



SNP Heritability

The average SNP hertiability estimate across all 4178 primary GWAS is 0.065 (median = 0.058). Average estimates increase when restricting to any confidence (mean = 0.078, median = 0.059) or `high` confidence (mean = 0.088, median = 0.062). Strong departure from the null hypothesis of $h_g^2 = 0$ is observed across phenotypes, especially among high confidence results.



UKB SNP-Heritability Browser

Results from the Neale Lab

Last updated 2022-10-11

Phenotype	h2	h2 p	h2 sig?	Confidence	Int.	Int. p	Neff	Source
Leg fat percentage (left)	0.221	8.97e-262	z7	high	1.109	8.73e-11	354,791	phesant
Leg fat percentage (right)	0.221	8.89e-261	z7	high	1.103	6.57e-10	354,811	phesant
Body fat percentage	0.230	1.4e-249	z7	high	1.095	2.25e-08	354,628	phesant
Arm fat percentage (left)	0.225	1.33e-240	z7	high	1.094	5.43e-08	354,707	phesant
Trunk fat percentage	0.221	2.96e-233	z7	high	1.087	2.24e-07	354,619	phesant
Arm fat percentage (right)	0.222	3.41e-227	z7	high	1.100	1.33e-08	354,760	phesant
Impedance of arm (right)	0.248	2.43e-209	z7	high	1.084	5.98e-05	354,792	phesant
Leg fat mass (left)	0.234	4.75e-208	z7	high	1.095	4.41e-07	354,788	phesant
Qualifications: College or University degree	0.286	1.54e-207	z7	high	1.119	5.59e-15	313,437	phesant
Leg fat mass (right)	0.232	1.03e-203	z7	high	1.095	5.15e-07	354,807	phesant
Impedance of arm (left)	0.244	2.09e-198	z7	high	1.099	2.99e-06	354,807	phesant
Trunk fat mass	0.239	8.95e-198	z7	high	1.094	9.4e-07	354,597	phesant
Whole body fat mass	0.239	9.55e-197	z7	high	1.096	7.6e-07	354,244	phesant
Body mass index (BMI)	0.249	2.52e-194	z7	high	1.103	6.28e-07	354,831	phesant

https://nealelab.github.io/UKBB_ldsc/h2_browser.html

nature genetics

An atlas of genetic correlations across human diseases and traits

Brendan Bulik-Sullivan[1–3,9], Hilary K Finucane[4,9], Verneri Anttila[1–3], Alexander Gusev[5,6], Felix R Day[7], Po-Ru Loh[1,5], ReproGen Consortium[8], Psychiatric Genomics Consortium[8], Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3[8], Laramie Duncan[1–3], John R B Perry[7], Nick Patterson[1], Elise B Robinson[1–3], Mark J Daly[1–3], Alkes L Price[1,5,6,10] & Benjamin M Neale[1–3,10]

Identifying genetic correlations between complex traits and diseases can provide useful etiological insights and help prioritize likely causal relationships. The major challenges preventing estimation of genetic correlation from genome-wide association study (GWAS) data with current methods are the lack of availability of individual-level genotype data and widespread sample overlap among meta-analyses. We circumvent these difficulties by introducing a technique—cross-trait LD Score regression—for estimating genetic correlation that requires only GWAS summary statistics and is not biased by sample overlap. We use this method to estimate 276 genetic correlations among 24 traits. The results include genetic correlations between anorexia nervosa and schizophrenia, anorexia and obesity, and educational attainment and several diseases. These results highlight the power of genome-wide analyses, as there currently are no significantly associated SNPs for anorexia nervosa and only three for educational attainment.

inferences from such studies can be challenging because of issues such as confounding and reverse causation, which can lead to spurious associations and mask the effects of real risk factors[1,2]. Genetics can help elucidate cause and effect, as inherited genetic risks cannot be subject to reverse causation and are correlated with a smaller list of confounders.

The first methods to test for genetic overlap were family studies[3–7]. To estimate the genetic overlap for many pairs of phenotypes, family study designs require the measurement of multiple traits for the same individuals. Consequently, it is challenging to scale these designs to a large number of traits, especially traits that are difficult or costly to measure (for example, low-prevalence diseases). More recently, GWAS have allowed effect size estimates to be obtained for specific genetic variants, so it is possible to test for shared genetics by looking for correlations in effect sizes across traits, which does not require measuring multiple traits per individual.

There exists a large class of methods for interrogating genetic overlap via GWAS that focus only on genome-wide significant SNPs. One of the most influential methods in this class is Mendelian randomiza-

Cross-trait LD score regression

Cross-trait LD score regression is employed to estimate the genetic correlation between a pair of traits.

Key idea: replace `\chi^2` in univariate LD score regression and the relationship (SNPs with high LD) still holds.

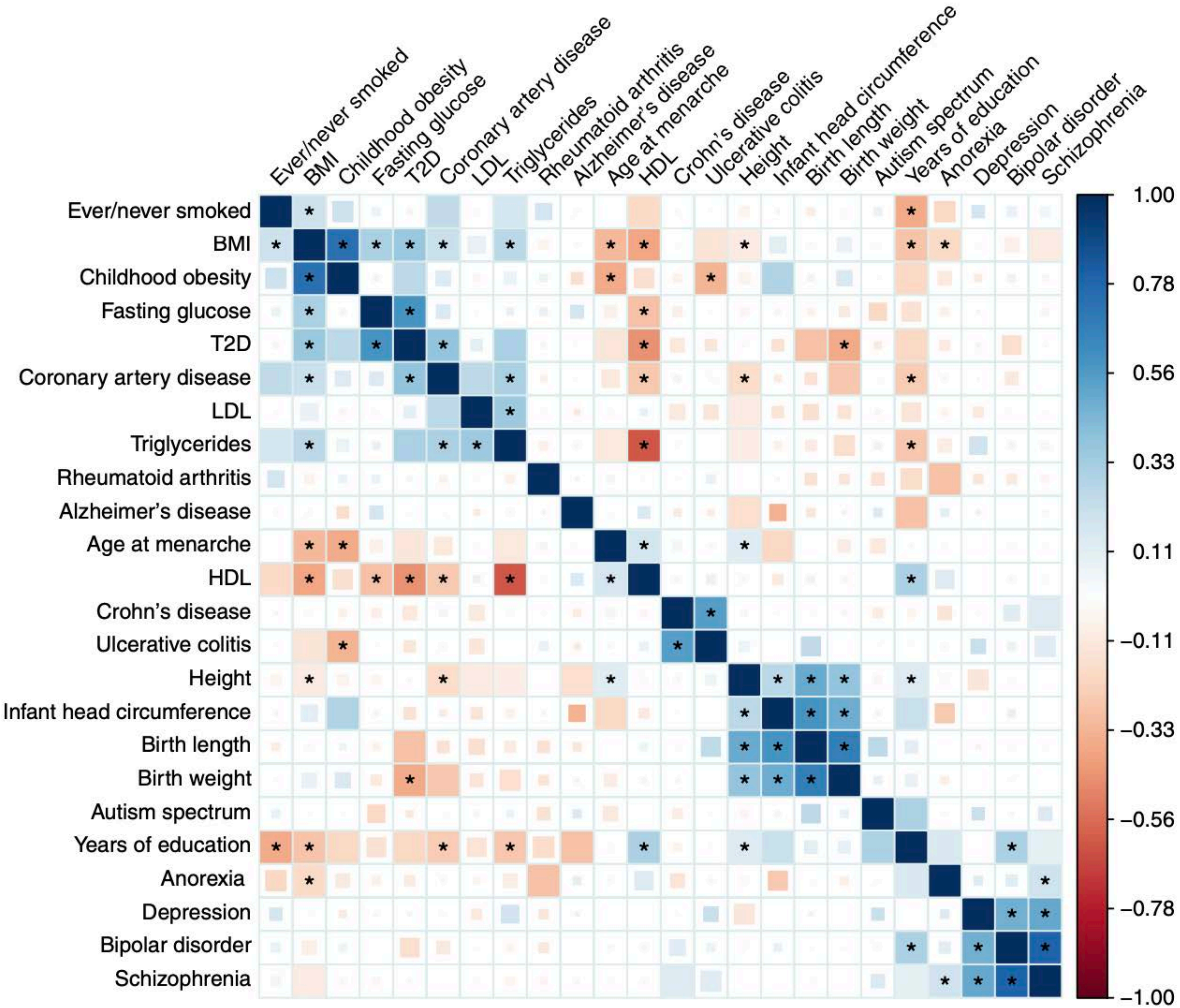
$$E[z_{1j}z_{2j}] = \frac{\sqrt{N_1N_2}\rho_g}{M}l_j + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

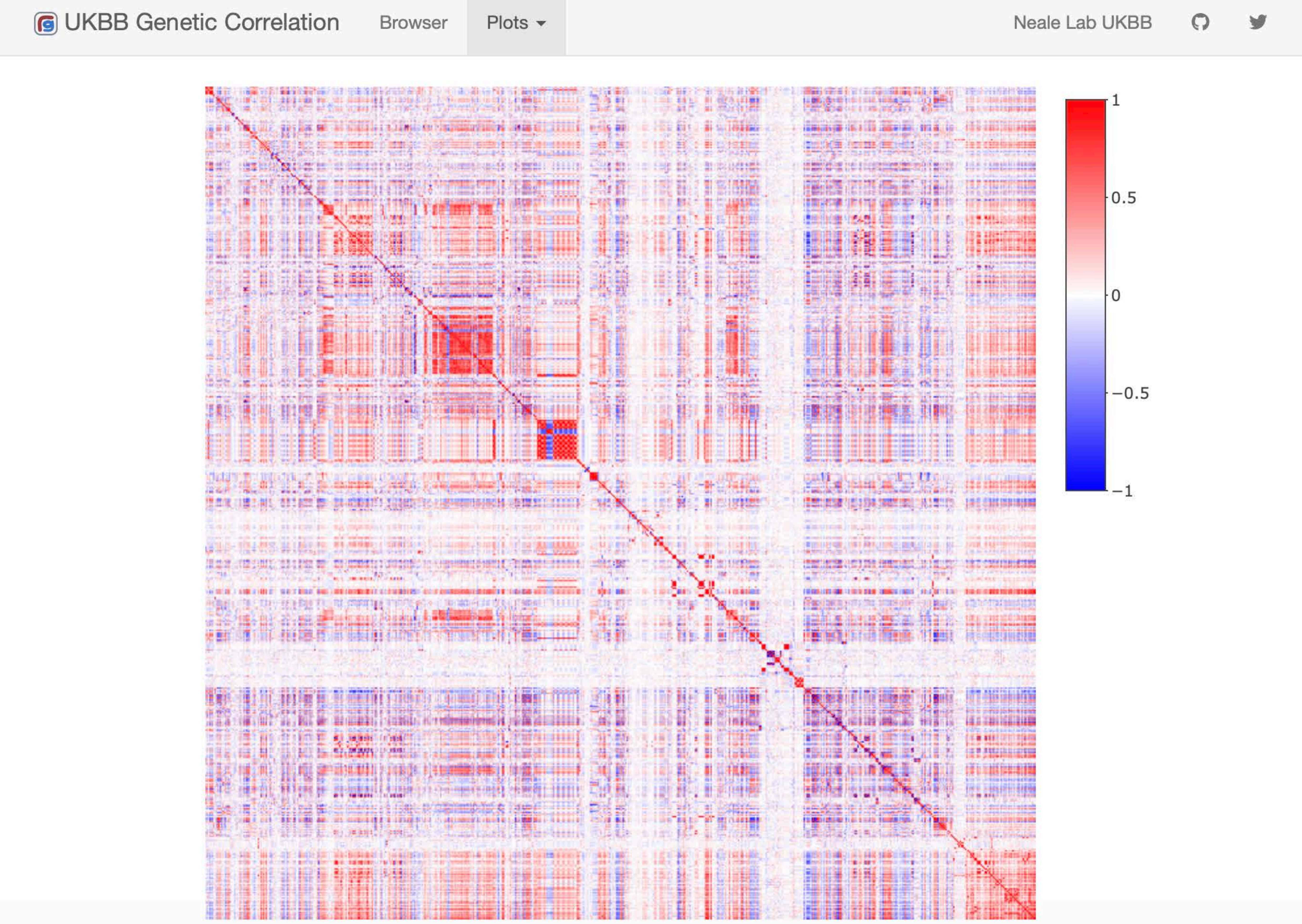
- $z_i j$: z score of trait i for SNP j
- N_i : sample size of trait i
- ρ : phenotypic correlation
- ρ_g : genetic covariance
- l_j : LD score
- M : number of SNPs

Then we can get the genetic correlation by :

$$r_g = \frac{\rho_g}{\sqrt{h_1^2h_2^2}}$$

- Reference: Bulik-Sullivan, Brendan, et al. "An atlas of genetic correlations across human diseases and traits." Nature genetics 47.11 (2015): 1236-1241.


Ever/never smoked
BMI
Childhood obesity
Fasting glucose
T2D
Coronary artery disease
LDL
Triglycerides
Rheumatoid arthritis
Alzheimer's disease
Age at menarche
HDL
Crohn's disease
Ulcerative colitis
Height
Infant head circumference
Birth length
Birth weight
Autism spectrum
Years of education
Anorexia
Depression
Bipolar disorder
Schizophrenia
1.00
0.78
0.56
0.33
0.11
−0.11
−0.33
−0.56
−0.78
−1.00


UKBB Genetic Correlation
Browser
Plots
Neale Lab UKBB
1
0.5
0
−0.5
−1