Statistical Methods for Genome-wide Association Studies

Yongjin Park, UBC Path + Stat, BC Cancer

28 February, 2024

Announcement

- We have a DELIVERABLE for Seminar 8 (probably very useful for your project)
- The analysis assignment will be due soon.

Themes for the forthcoming lectures from a method's perspective

- Supervised Learning – lectures 13, 14, 15
- Unsupervised Learning – lectures 17, 18, 19

- Other types:
  - Semi-supervised learning
  - Active learning
  - Self-taught learning

Themes for the forthcoming lectures from a data's perspective

- Statistical genetics – lectures 13, 14
- Regulatory genomics – lectures 15
- Single-cell genomics – lectures 17, 18, 19

The goal of today's lecture

- Some background knowledge of statistical genetics
- Biological, statistical intuitions

Today's lecture

1. Mapping disease-specific locations in Genome
2. How GWAS can be interpreted wrongfully
3. Combining evidence from multiple studies
4. Appendix

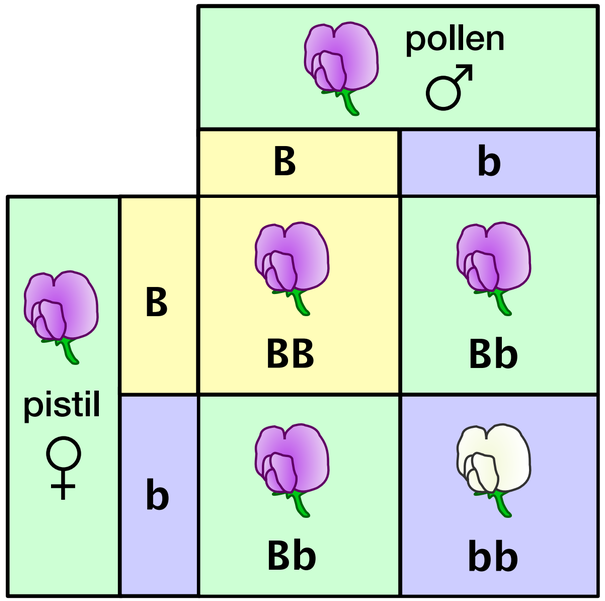
Goal: mapping disease-specific locations in Genome

GWAS

- Input:
  - A genotype matrix X $(n \times p)$, where $X_{ij} \in \{0, 1, 2\}$
  - A phenotype vector $\mathbf{y}$ $(n \times 1)$
- Output:
  - Estimate a coefficient (effect size) β_j $(j \in 1, \ldots, p)$

- Testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.
- Report p-values
- It seems straightforward...

Every story has a beginning



Gregor Mendel
(1822-1884)

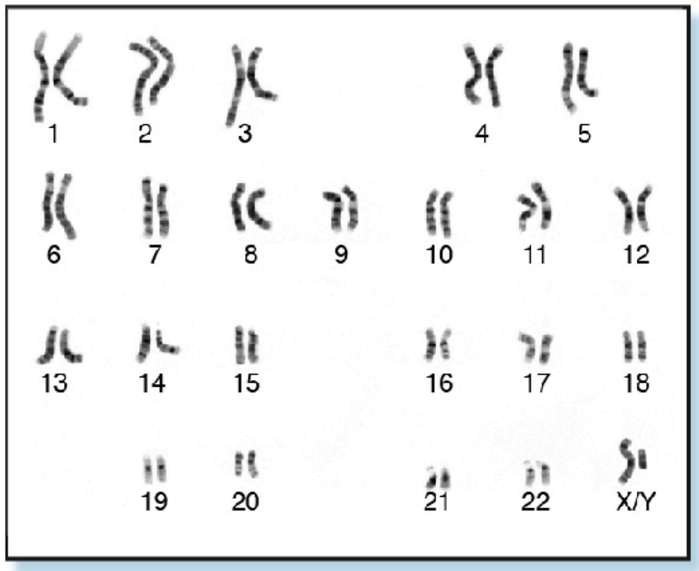
Key concepts emerged:

- Gene = a unit of heredity that transfers from a parent to offspring
- Allele = a different form of a gene [from a Greek word, αλληλο, αλλος, "allos", other]

Ploidy (diploid)

One copy from maternal genome another copy from paternal genome

When do we have polyploidy or aneuploidy?

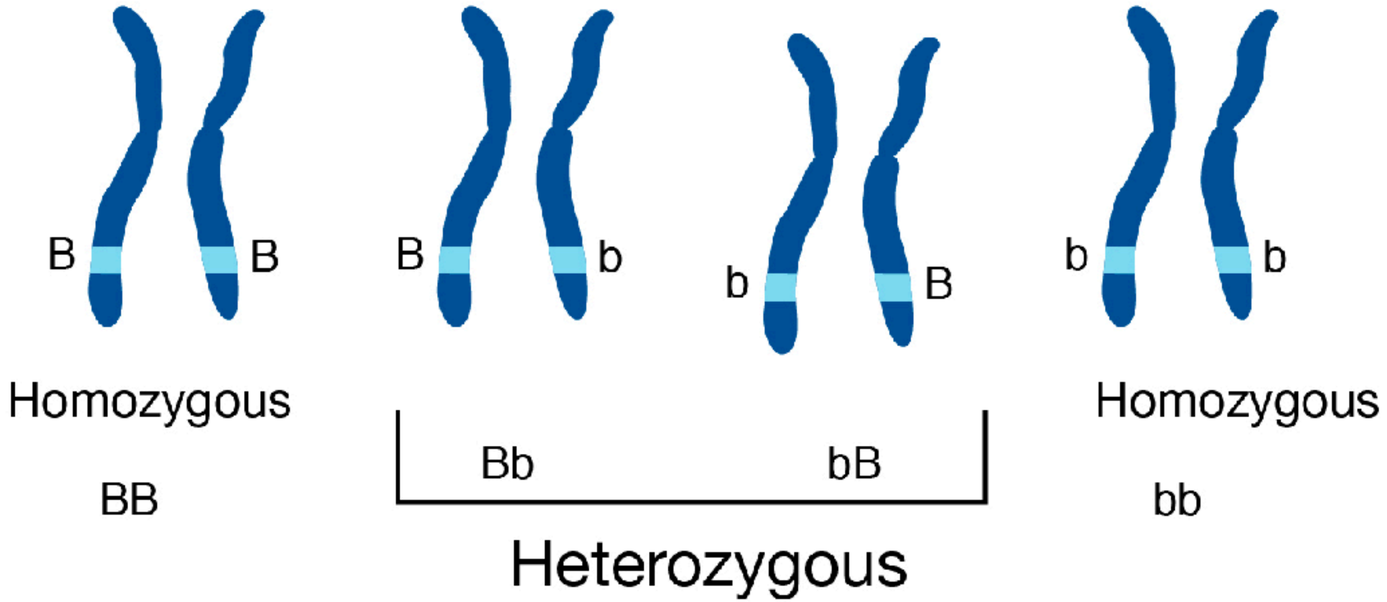


https://www.genome.gov/genetics-glossary/Diploid

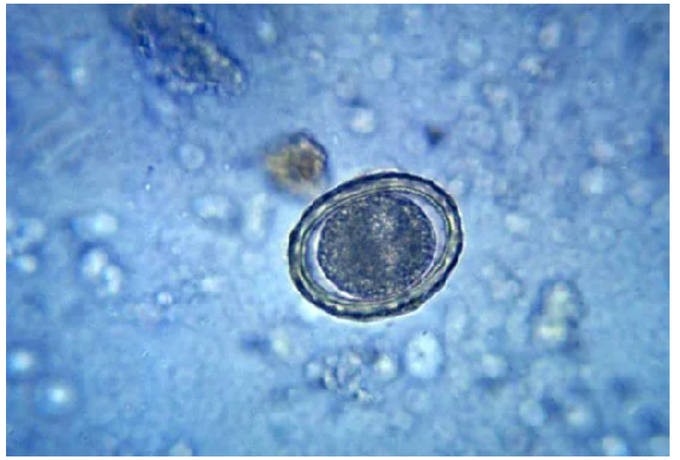
Homozygote vs. heterozygote

One copy from maternal genome
another copy from paternal genome

Homozygous: the maternal == paternal copy
Heterozygous: the maternal != paternal copy



- "Homo" vs. "hetero" ⟺ the same vs. different
- Zygote (fertilized egg cell)

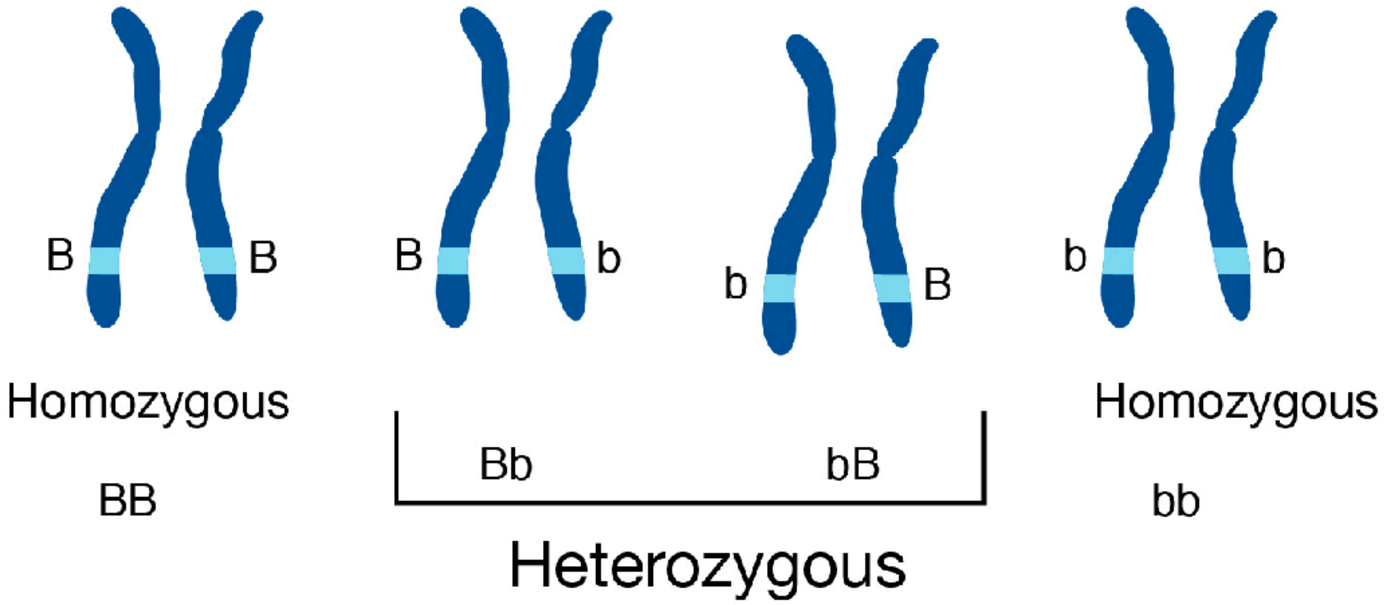


https://www.genome.gov/genetics-glossary/heterozygous

https://www.britannica.com/science/zygote

Biallelic vs. triallelic vs. multiallelic

"Bi"-allelic: two different forms exist for this heritable unit (most of the human variants)
"Tri"-allelic: three different forms exist (much rare than biallelic variants)



For biallelic variant:
- reference allele
- alternative allele

https://www.genome.gov/genetics-glossary/heterozygous

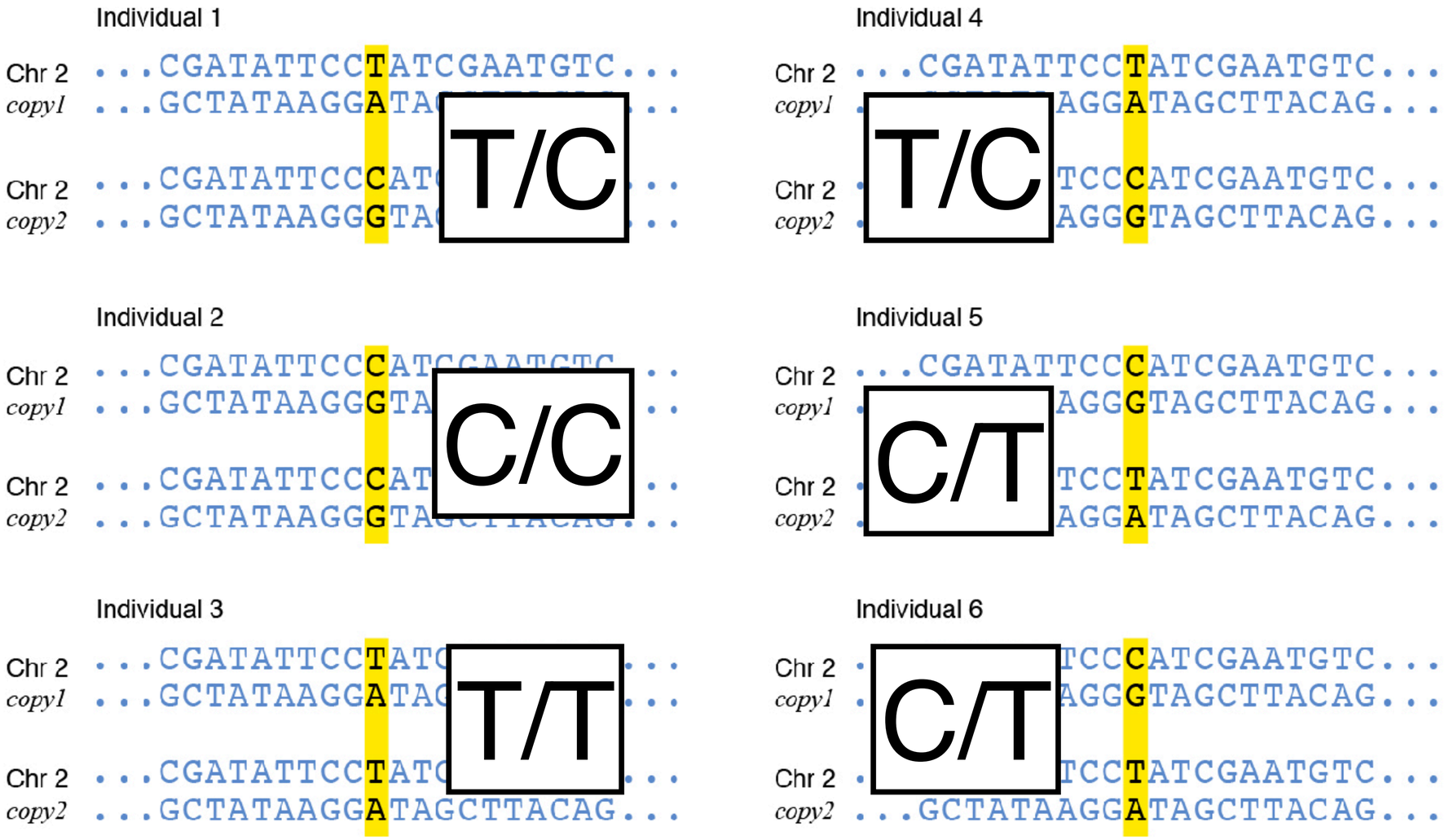
Single Nucleotide Polymorphism

~1 in every 1000bp

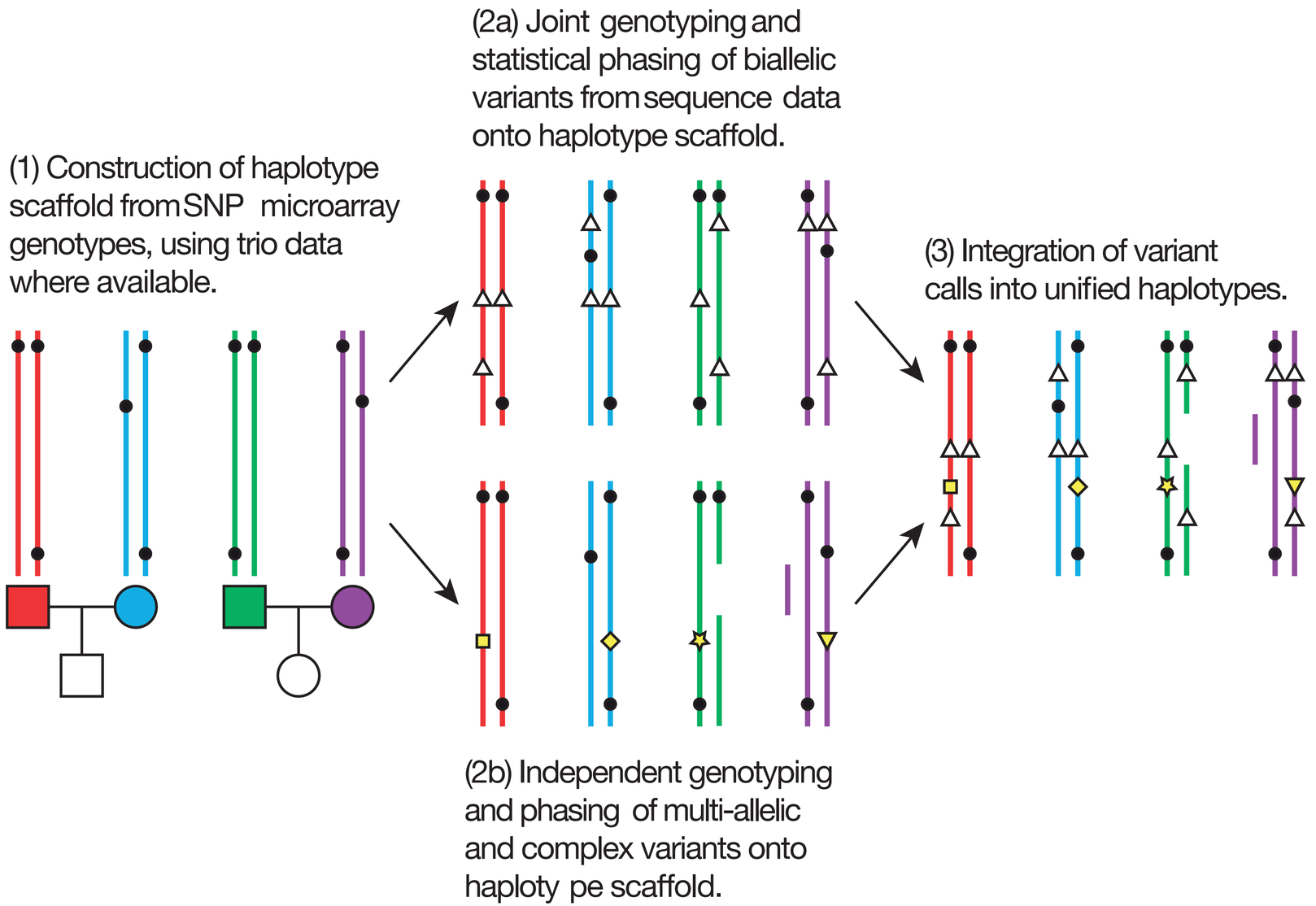
Single nucleotide polymorphism is way too many syllables, so you can understand why we just say "snip". And this is really a simple concept.

These are the places in the genome where people differ.

Francis Collins



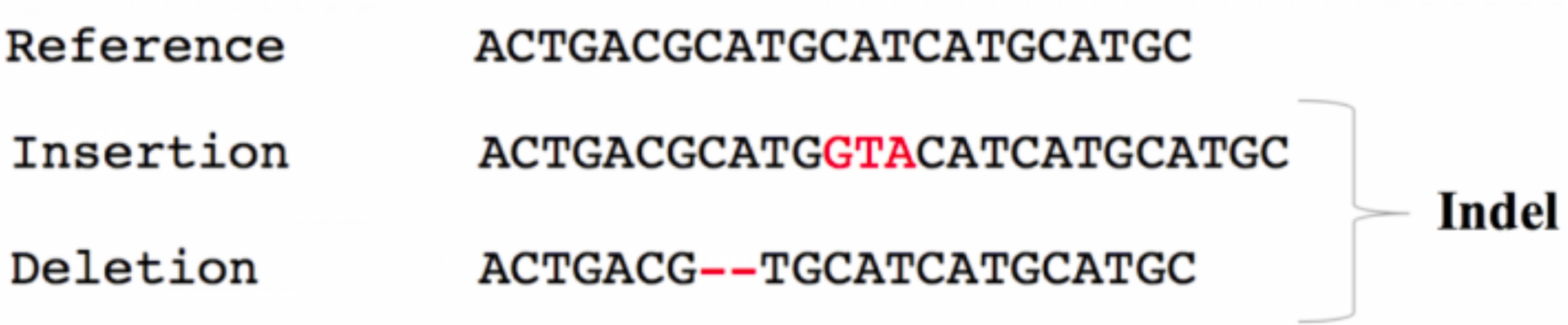
Haplotype


(2a) Joint genotyping and statistical phasing of biallelic variants from sequence data onto haplotype scaffold.
(1) Construction of haplotype scaffold from SNP microarray genotypes, using trio data where available.
(3) Integration of variant calls into unified haplotypes.
(2b) Independent genotyping and phasing of multi-allelic and complex variants onto haploty pe scaffold.

Many types of genetic variants

ACTCGTGACCGCATGCATCTTCATTCATGC **Reference**

ACTCGTGACCGCATGCATCGTCATTCATGC **Alternative**

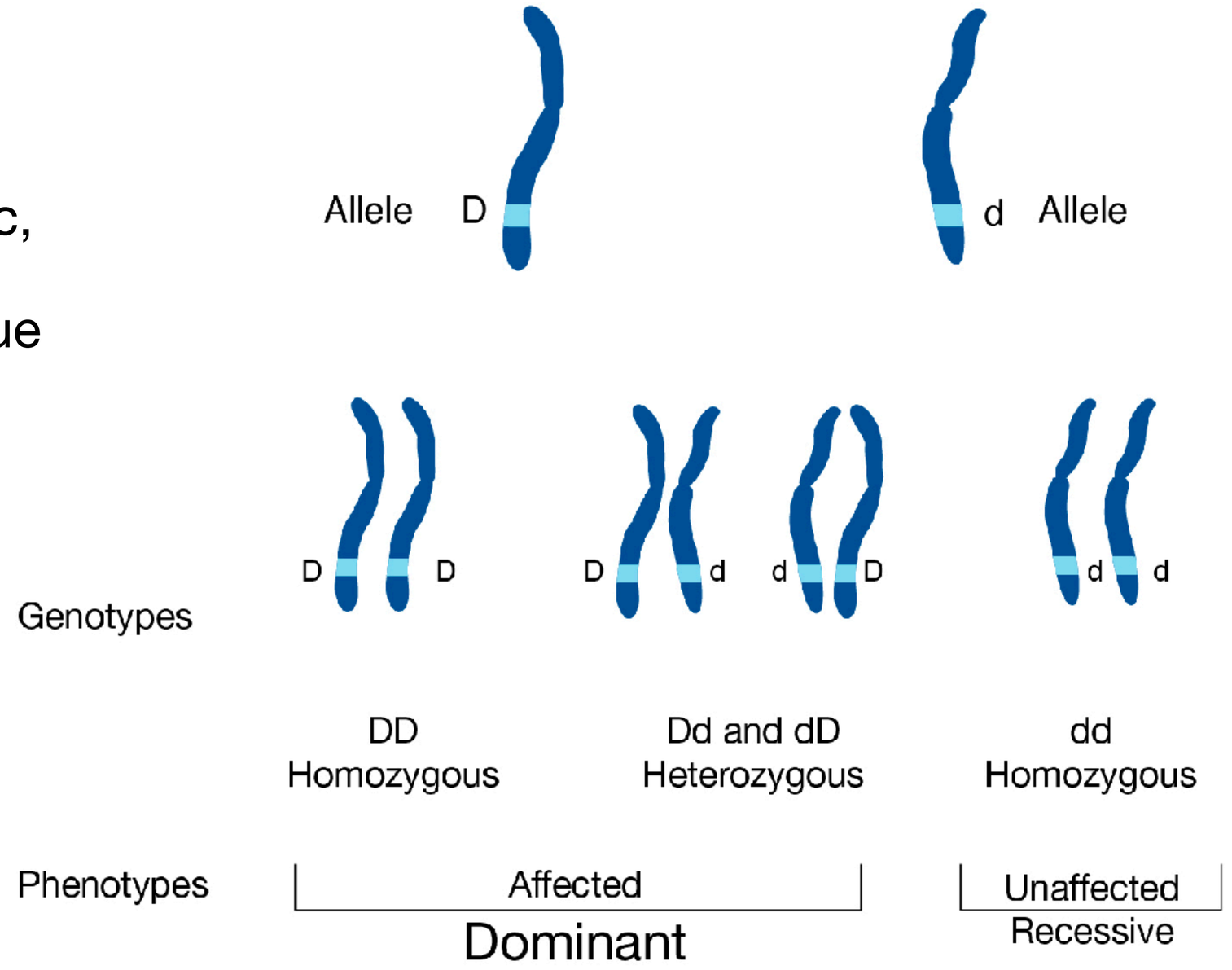


https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

Huntington's Disease

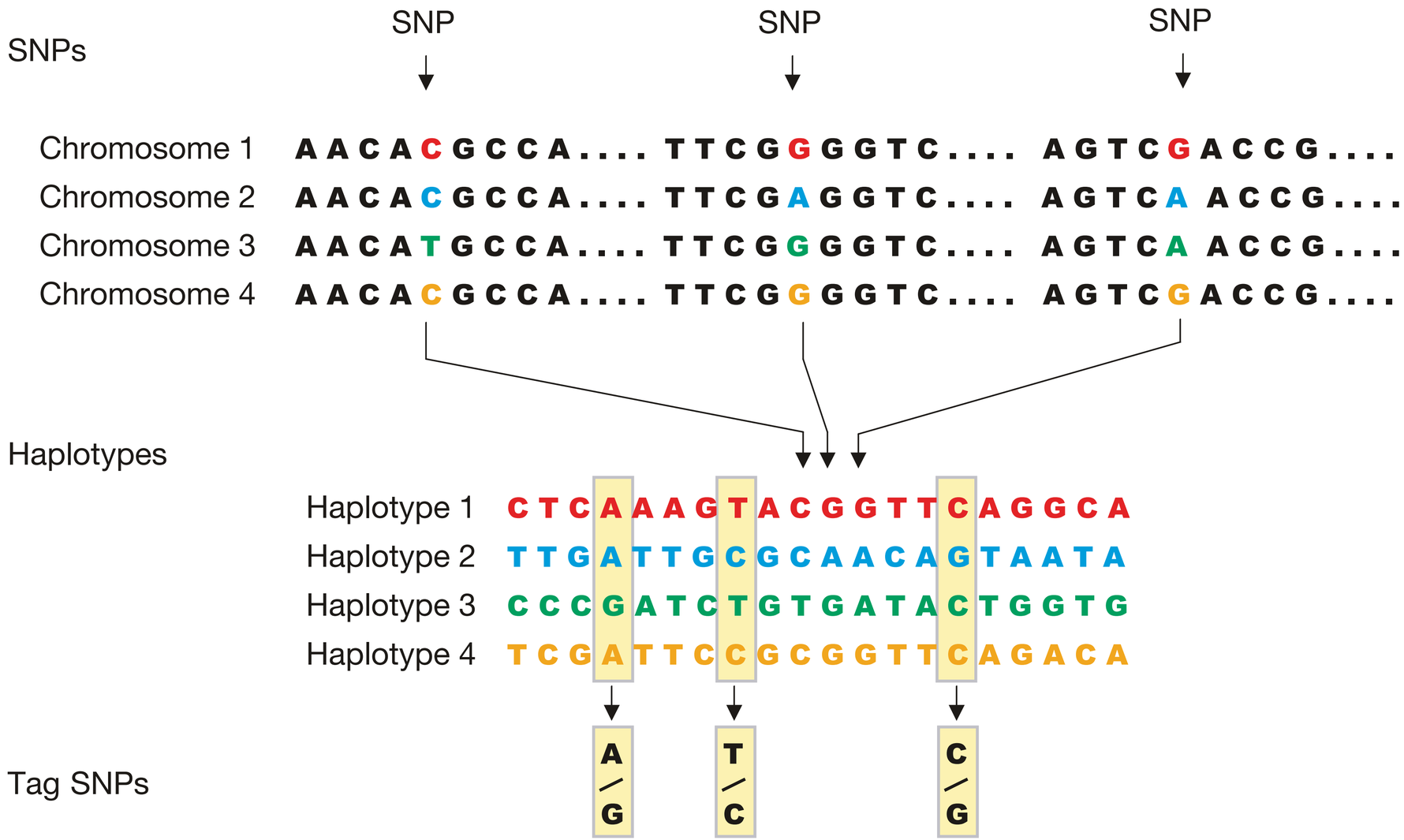
Dominance

If one allele of a gene is toxic, the other allele of the same, wildtype gene may not rescue



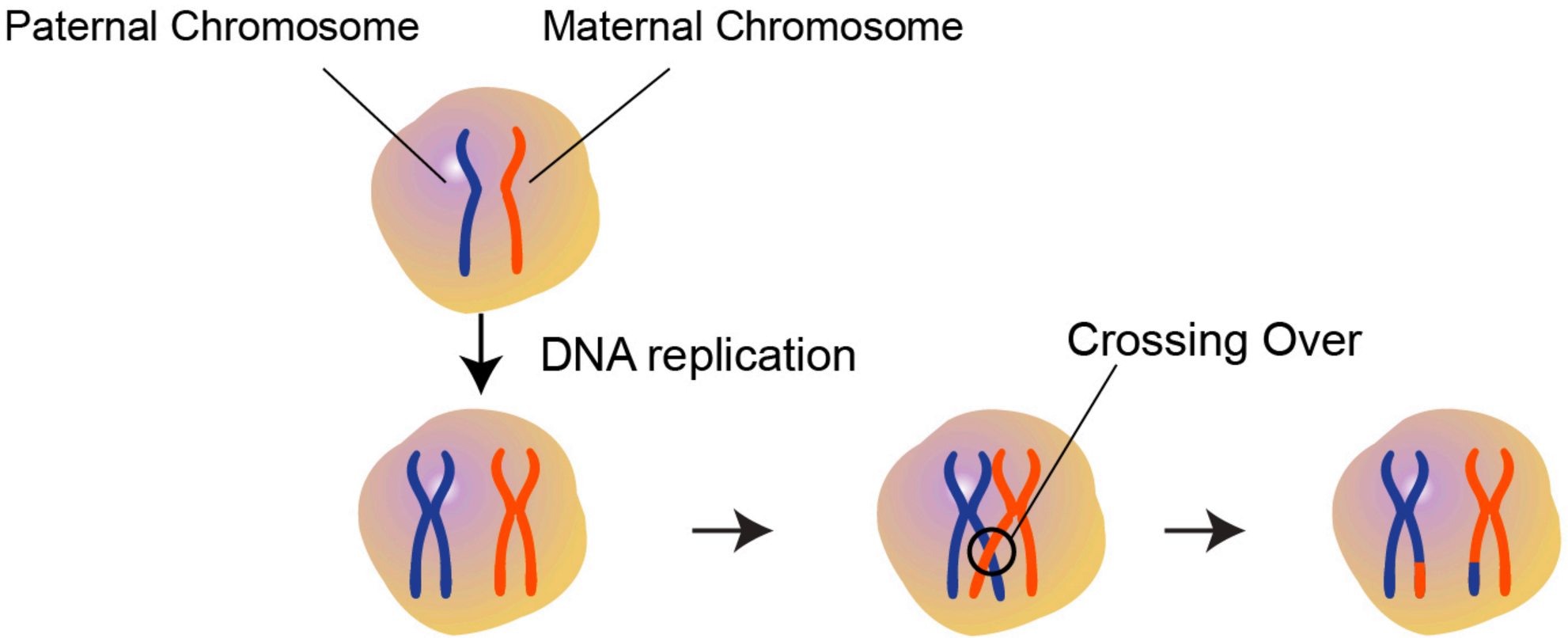
https://www.genome.gov/genetics-glossary/Dominant

~ 1M common SNPs



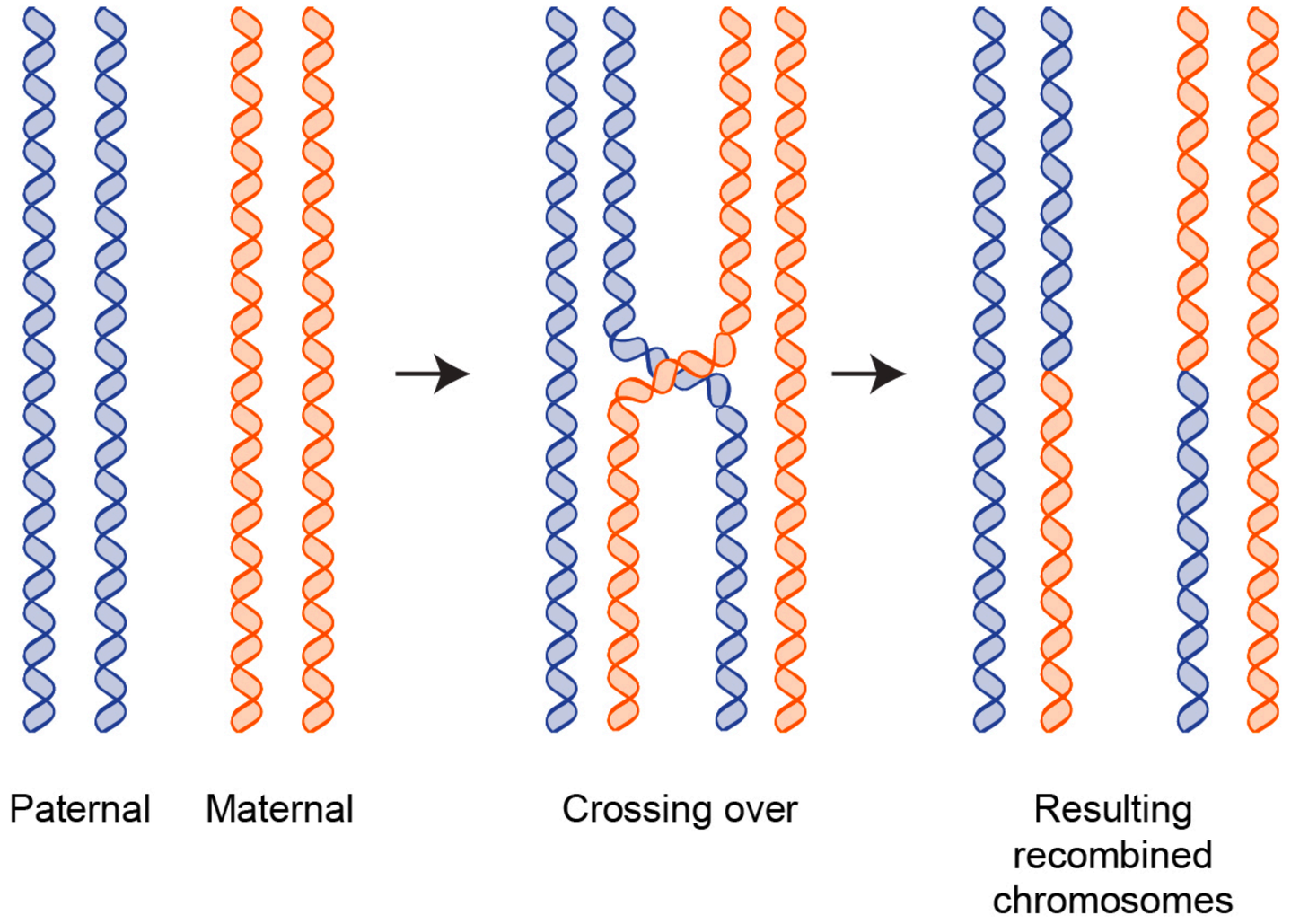
The International HapMap Project (2003)

Recombination: Mixing the maternal and paternal copies



https://www.genome.gov/genetics-glossary

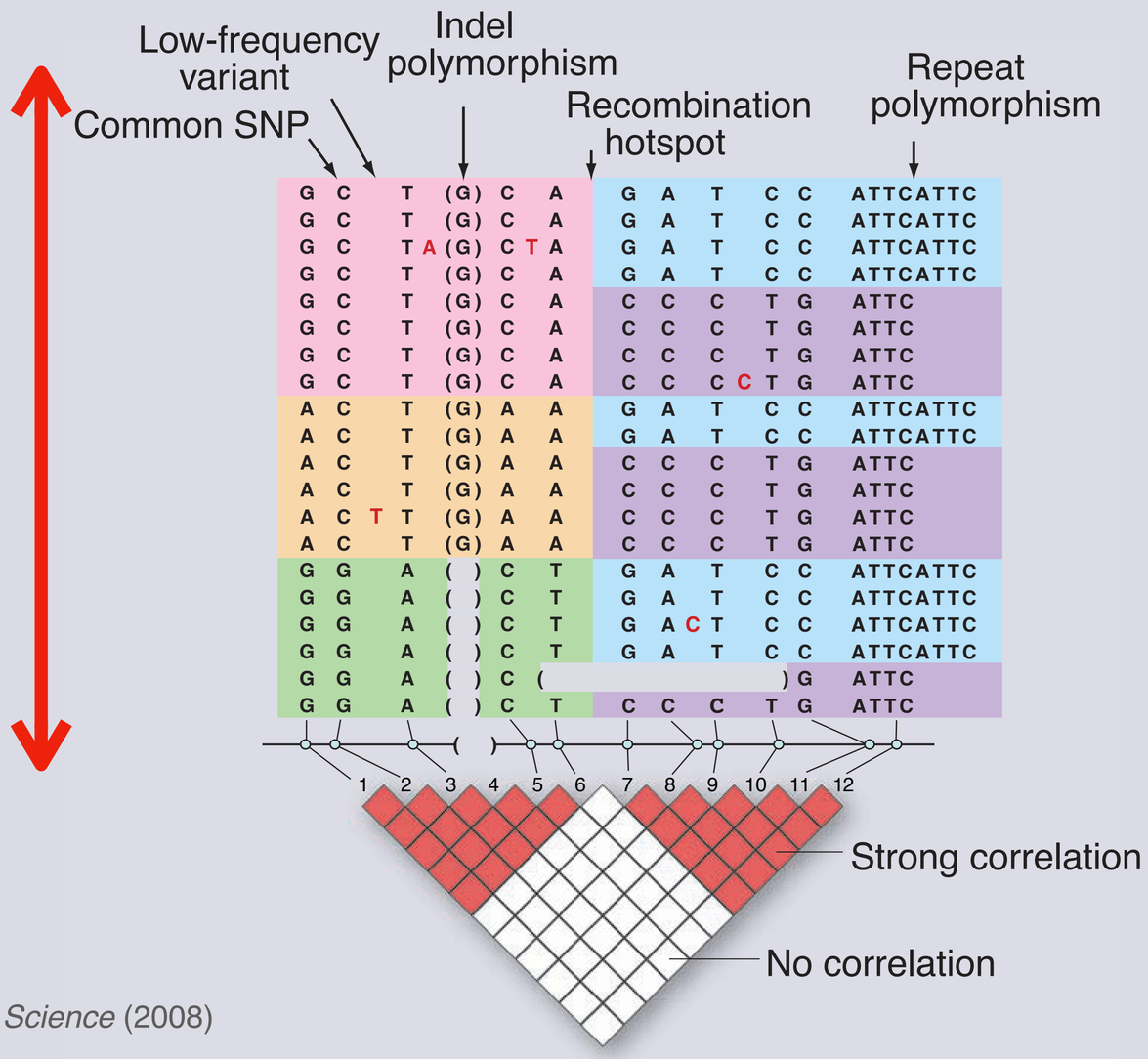
Recombination: Mixing the maternal and paternal copies



https://www.genome.gov/genetics-glossary

Genetic variation in one figure

across many individuals (diploid genomes)



Common SNPs
Insertion/deletion
Other low-freq. variants
Other structural variants
Recombination hotspot

Altshuler, Daly, Lander, *Science* (2008)

Definitions

Allele

- A different form of a gene
- A Greek word "allos," $\alpha\lambda\lambda\eta\lambda o$, meaning "other"

Variant and locus

- A specific region of the genome differs across two or more genomes
- A result of mutation
- A locus: a location where many variants lie (plural: loci).

Ploidy

- The number of copies of chromosomes within a cell/organism
- Haploid: one copy
- Diploid: two copies

"We found *A* allele for this genetic variant as an effect allele."

"We found a hundred variants within the locus of *APOE*."

"We identified ten loci associated with cancer."

More definitions

Biallelic variant

- bi + allelic
- Two forms for a variant
- Reference (more frequently observed) vs. alternative allele

Polymorphism

- Poly + morph
- Occurrence of different forms

SNP

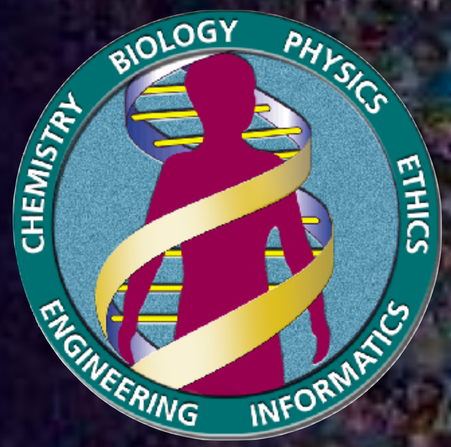
- Single Nucleotide Polymorphism
- A place in the genome where people differ by a single base pair

We pronounce SNP “snip” in North America.

Genetics is described in its own terminologies.

Don't be illiterate.

Human Genome Project

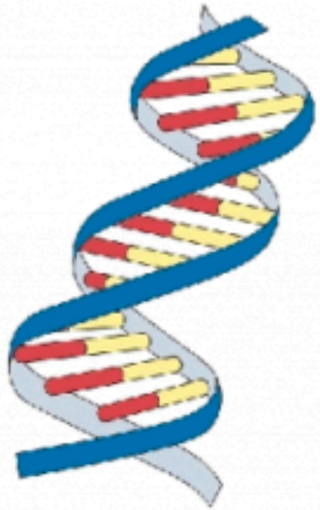


A reference DNA sequence for
3.2B basepairs x diploid for each individual

https://www.genome.gov/human-genome-project

Given reference, we can look up locations

Genetic association studies: genotype vs. phenotype



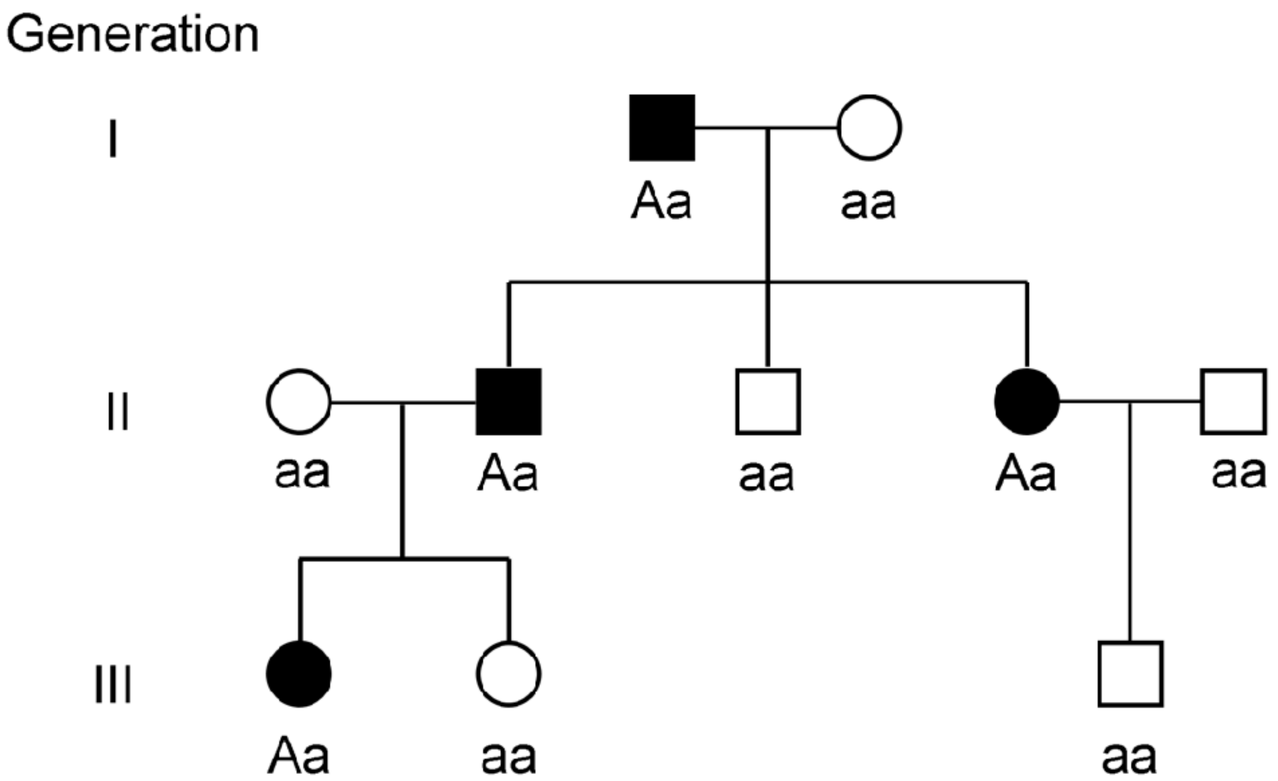
How do we test this?

Genotypes are the genetic make-up of an individual.

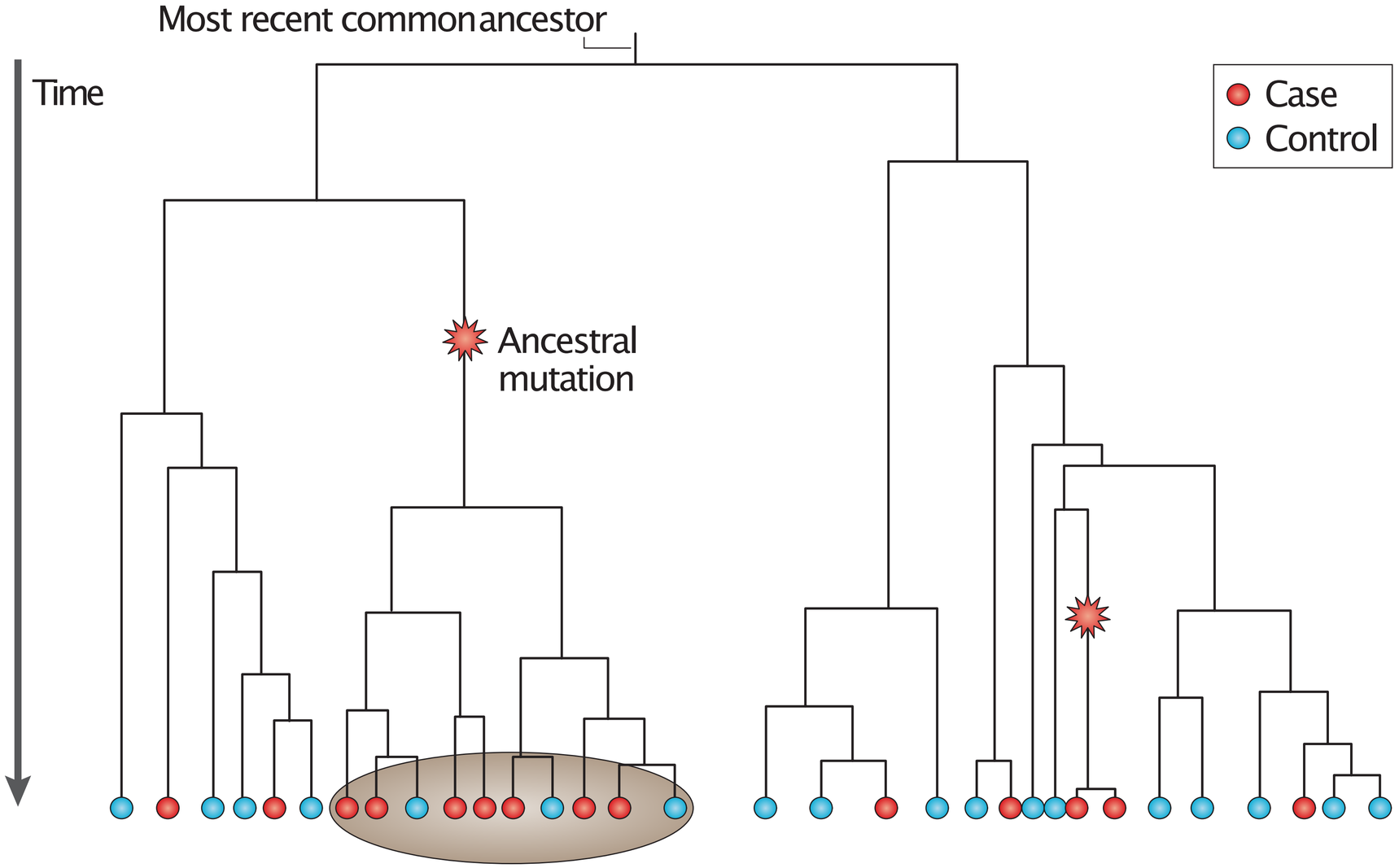
Phenotypes are the physical traits and characteristics of an individual and are influenced by their genotype and the environment.

https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction

How do we associate genetic variants to traits?

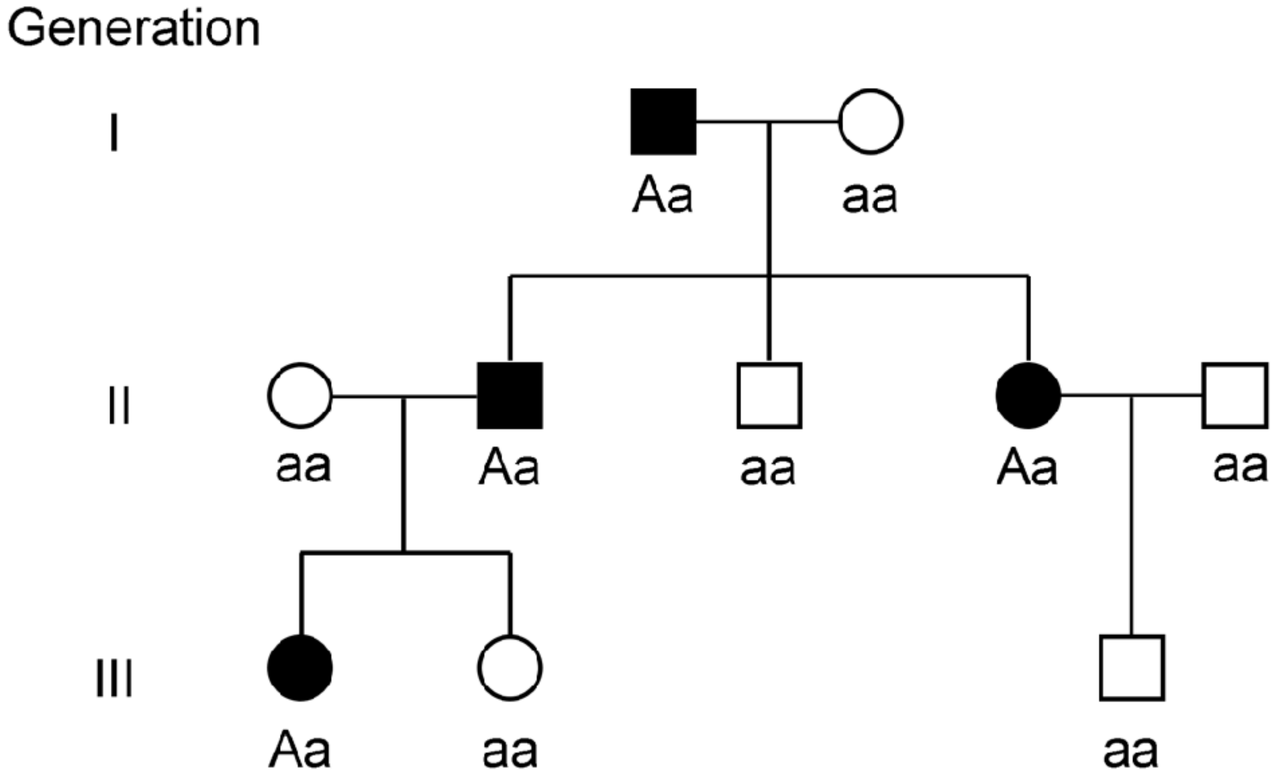


How can a genetic disease occur?



Balding, Nature Genetics Review, 2006

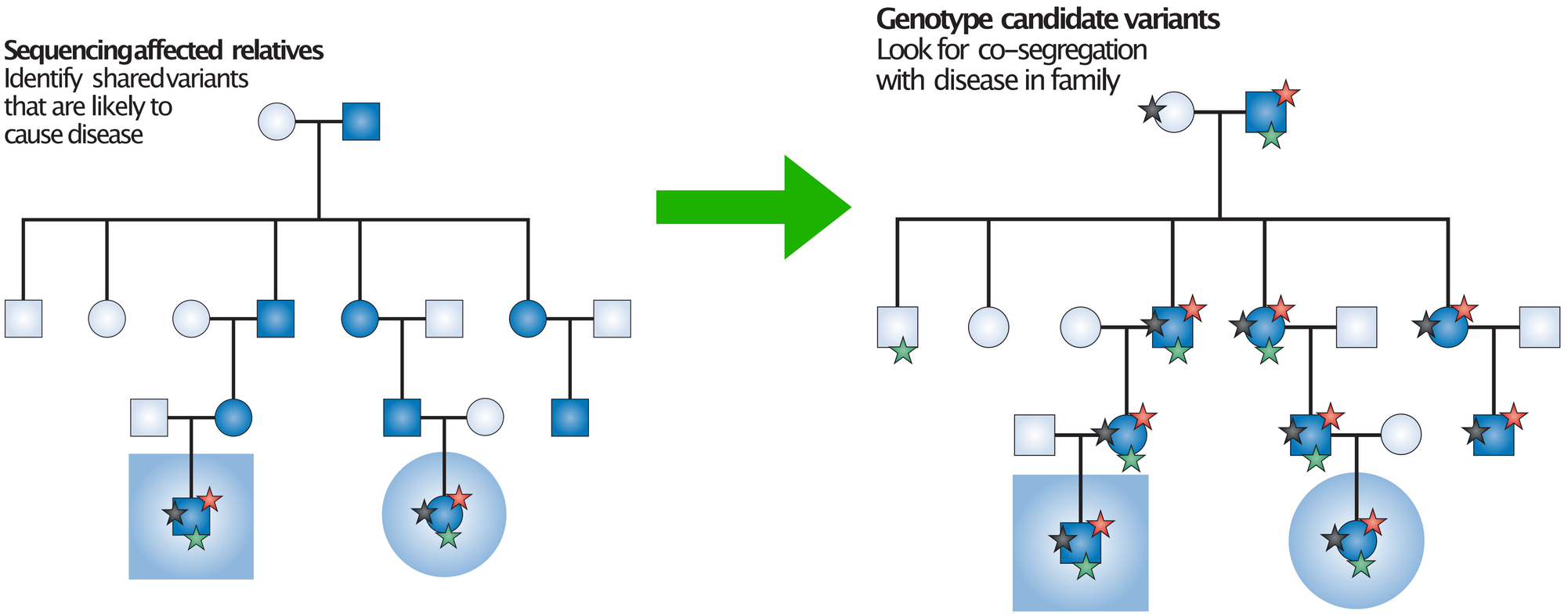
Mendelian disorders ≈ a single disease gene



- Cystic Fibrosis
- Sickle-cell anemia
- Phenylketonuria
- Huntington's disease
- ...

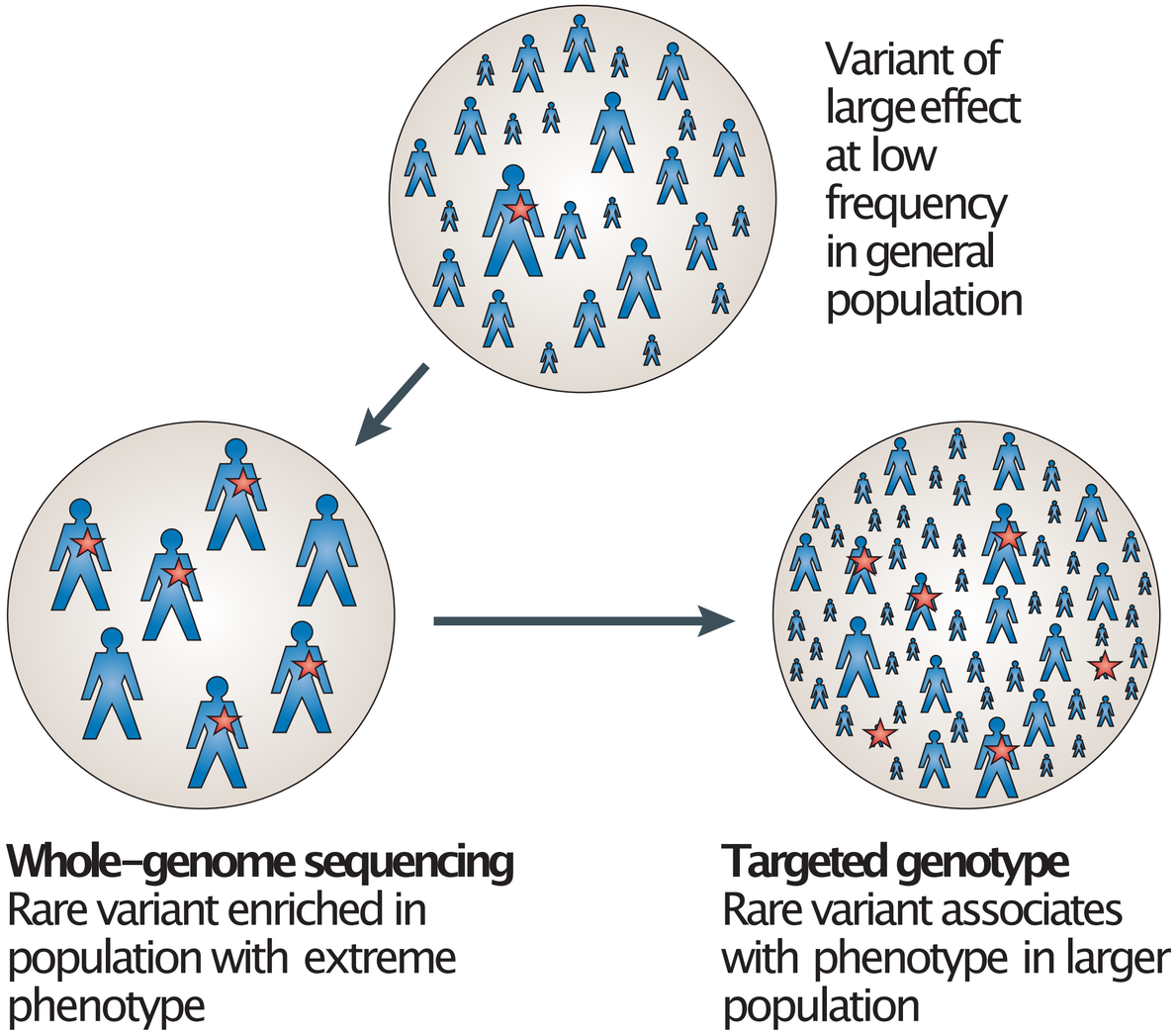
Very high penetrance, monogenic

Targetted genotyping and sequencing



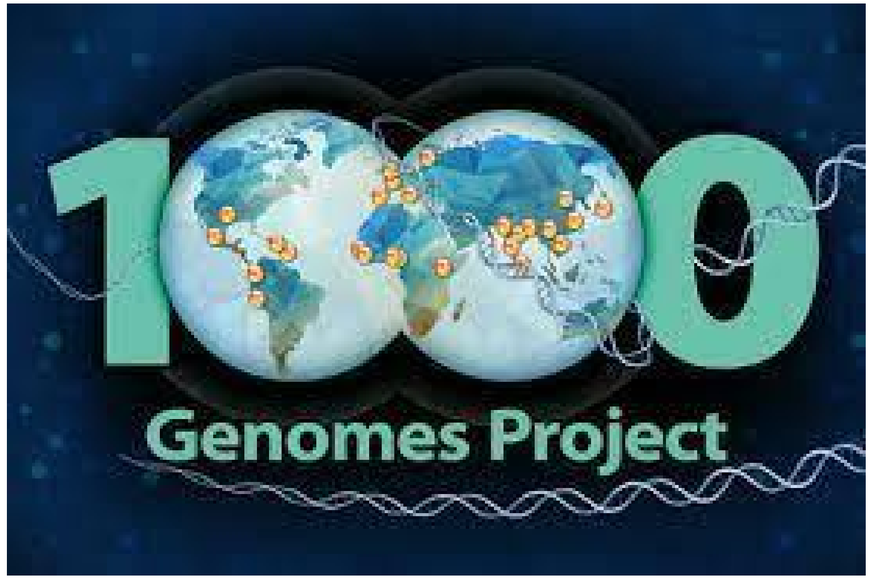
Cirulli and Goldstein, 2010

Population-level enrichment?



Cirulli and Goldstein, 2010

Human genetics data – The 1000 Genomes Project



Let's download the 1KG data using `bigsnpr` library.

```
library(bigsnpr)
download_1000G("../data/genotype/")
```

https://www.internationalgenome.org/

- International consortium
- Goal: find common genetic variants with frequencies of at least 1%
- The project planned to sequence each sample to 4x genomic coverage.

A big matrix data set with row and column data

Normally, we have a set of three files (PLINK format):

1. `../data/genotype/1000G_phase3_common_norel.bed`
2. `../data/genotype/1000G_phase3_common_norel.bim`
3. `../data/genotype/1000G_phase3_common_norel.fam`

What are the rows? samples/individuals/subjects

```
fread(.fam.file, nrows=5)
```

```
##       V1       V2    V3    V4    V5    V6
##    <int>   <char> <int> <int> <int> <int>
## 1:     0 HG00096     0     0     1    -9
## 2:     0 HG00097     0     0     2    -9
## 3:     0 HG00099     0     0     2    -9
## 4:     0 HG00100     0     0     2    -9
## 5:     0 HG00101     0     0     1    -9
```

1. Family ID ('FID')
2. Within-family ID (**'IID'**; cannot be '0')
3. Within-family ID of father ('0' if father isn't in dataset)
4. Within-family ID of mother ('0' if mother isn't in dataset)
5. Sex code (1: male, 2: female, 0: unknown)
6. Phenotype value[a] (1: control, 2: case, '-9': missing)

[a] not much used

https://www.cog-genomics.org/plink/1.9/formats

What are the columns? genetic variants/SNPs

```
fread(.bim.file, nrow=5)
```

```
##       V1          V2    V3     V4      V5      V6
##    <int>      <char> <int>  <int> <char> <char>
## 1:     1  rs2185539     0 566875      T      C
## 2:     1  rs3131972     0 752721      G      A
## 3:     1 rs12184325     0 754105      T      C
## 4:     1  rs3131969     0 754182      G      A
## 5:     1  rs3131967     0 754334      C      T
```

https://www.cog-genomics.org/plink/1.9/formats

1. Chromosome code
2. Variant identifier
3. Position in morgans or centimorgans (safe to use dummy value of '0')
4. Base-pair coordinate (1-based; limited to 2^{31}-2)
5. Allele 1 (corresponding to clear bits in .bed; usually minor)
6. Allele 2 (corresponding to set bits in .bed; usually major)

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325  0 754105  T  C
```

1. Chromosome
2. Variant
3. Morgan (ignore)
4. Base-pair coordinate
5. Allele 1
6. Allele 2

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325   0 754105   T   C
```

1. Chromosome: **1**

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325  0 754105  T  C
```

1. Chromosome: **1**
2. Variant identifier: **rs12184325**, a unique ID used in dbSNP; can be anything unique; it can change depending on the dbSNP human genome build (don't rely on it).

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325  0 754105  T  C
```

1. Chromosome: **1**
2. Variant identifier: **rs12184325**, a unique ID used in `dbSNP`; can be anything unique; it can change depending on the `dbSNP` human genome build (don't rely on it).
3. A unit for measuring genetic linkage (just put the dummy value "0")
4. Genomic position (bp) within each chromosome

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325  0 754105  T  C
```

1. Chromosome: **1**
2. Variant identifier: **rs12184325**, a unique ID used in `dbSNP`; can be anything unique; it can change depending on the `dbSNP` human genome build (don't rely on it).
3. A unit for measuring genetic linkage (just put the dummy value "0")
4. Genomic position (bp) within each chromosome
5. **A1**: If this allele (haplotype) is T, we use **0**
6. **A2**: If this allele (haplotype) is C, we use **1**

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325   0 754105   T   C
```

- How many haplotypes? A conventional way to code "genotype" is to sum them up (additive effect); there are several other options (dominance, heterozygous).

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325   0 754105   T   C
```

- How many haplotypes? A conventional way to code "genotype" is to sum them up (additive effect); there are several other options (dominance, heterozygous).
- How do you code "TT"?
- How do you code "CC"?
- How do you code "CT"?

Let's understand how we "name" and "code/type" variants

A variant $\approx$ SNP (throughout the course)

example

```
1 rs12184325   0 754105   T   C
```

- How many haplotypes? A conventional way to code "genotype" is to sum them up (additive effect); there are several other options (dominance, heterozygous).
- How do you code "TT"? **0**
- How do you code "CC"? **2**
- How do you code "CT"? **1**

How these genotypes instantiated in 1KG data

By calling `snp_readBed(.bed.file)`, we can convert the "BED"-formatted data to a "RDS" file for faster access. Later, we need to "attach" that RDS.

```
data <- snp_attach(.bk.file)
str(data, max.level=1, strict.width = "cut")
```

```
## List of 3
##  $ genotypes:Reference class 'FBM.code256' [package "bigstatsr"] with
##    ..and 26 methods, of which 12 are  possibly relevant
##  $ fam      :'data.frame':   2490 obs. of  6 variables:
##  $ map      :'data.frame':   1664852 obs. of  6 variables:
##  - attr(*, "class")= chr "bigSNP"
```

How these genotypes instantiated in 1KG data

```
dim(data$genotype)
```

```
## [1]    2490 1664852
```

```
dim(data$fam)
```

```
## [1] 2490    6
```

```
dim(data$map)
```

```
## [1] 1664852       6
```

- How many rows and columns?
- What are the rows and columns?

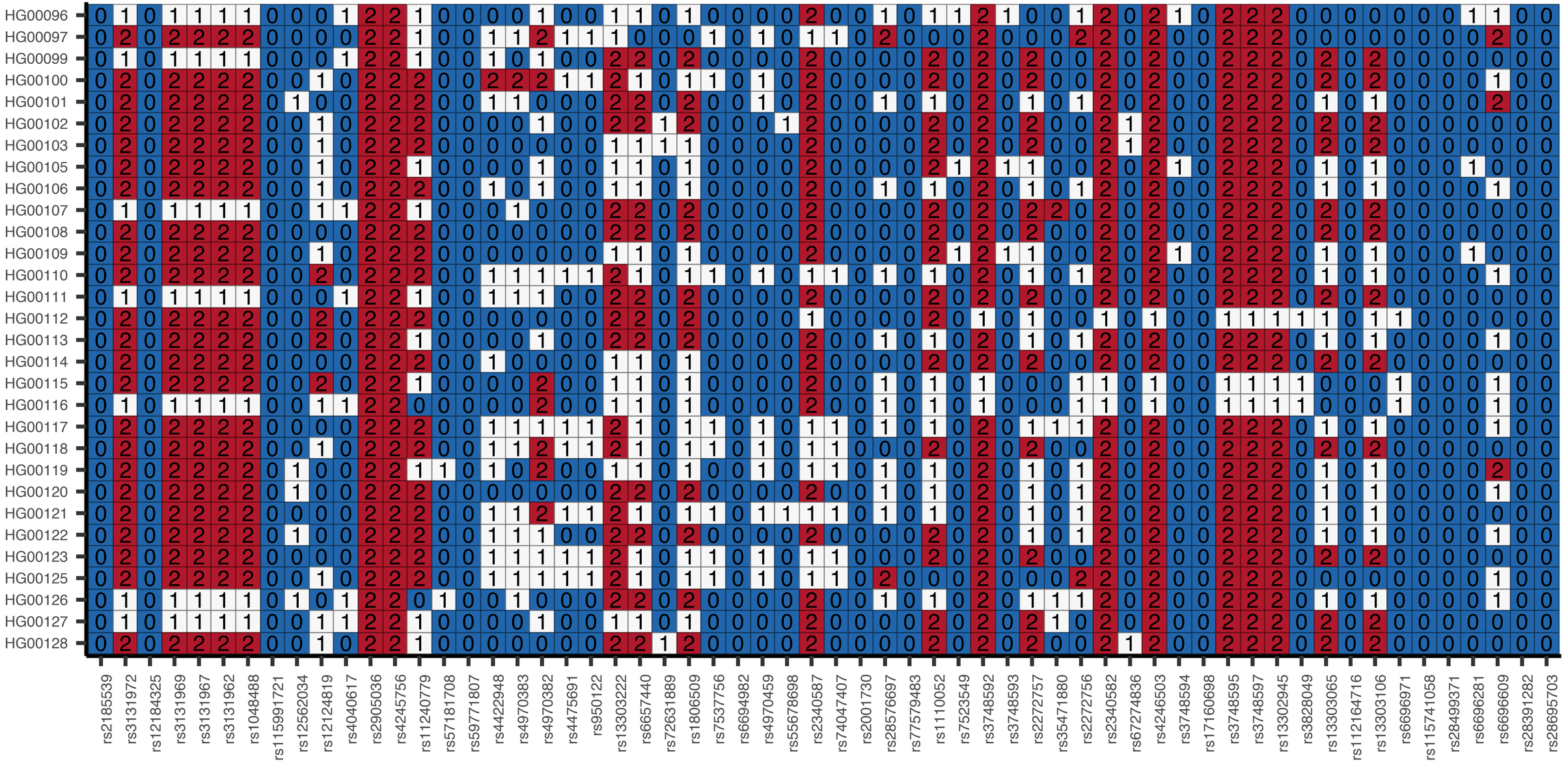
How these genotypes instantiated in 1KG data

A genotype/dosage matrix:

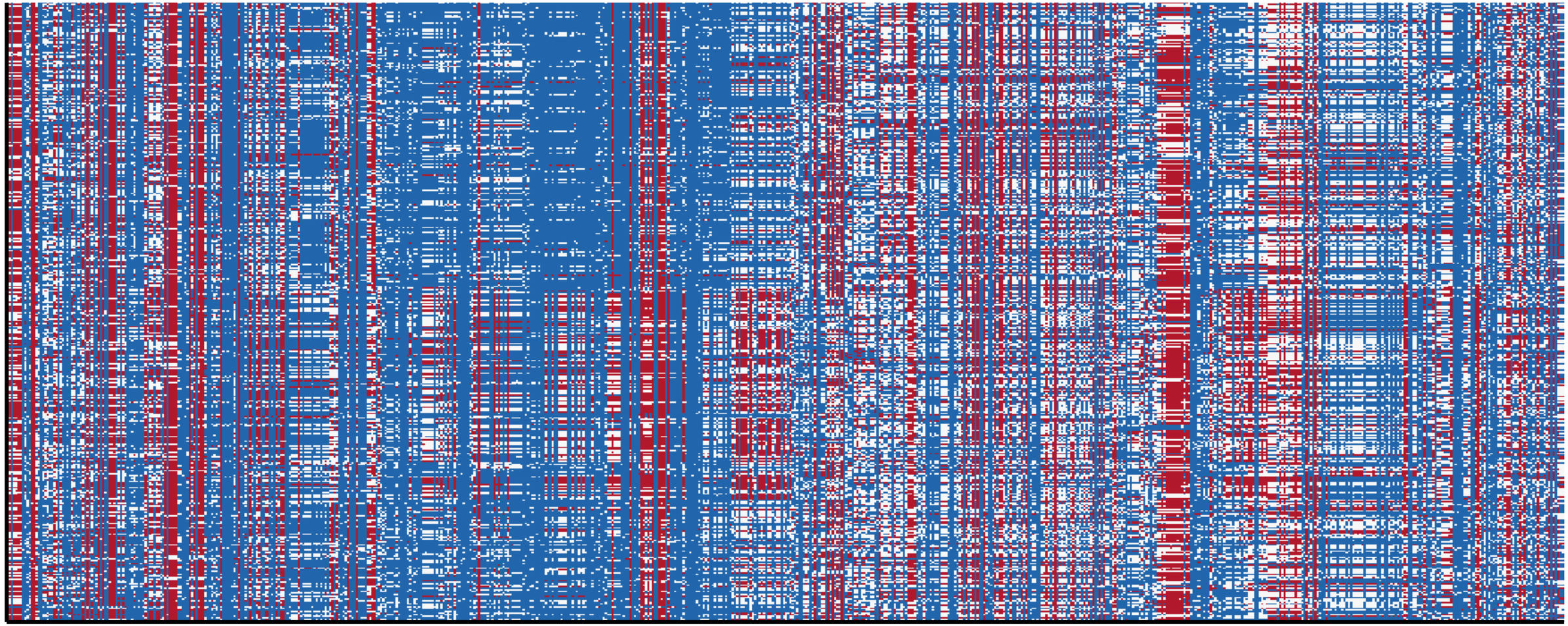
```
data$genotypes[1:5, 1:10]
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    0    1    0    1    1    1    1    0    0     0
## [2,]    0    2    0    2    2    2    2    0    0     0
## [3,]    0    1    0    1    1    1    1    0    0     0
## [4,]    0    2    0    2    2    2    2    0    0     1
## [5,]    0    2    0    2    2    2    2    0    1     0
```

A tiny fraction of the full genotype matrix



Just a bit more: Do you see any patterns?



We will get back to the structure.

Goal: mapping disease-specific locations in Genome

GWAS

- Input:
  - A genotype matrix X $(n \times p)$, where $X_{ij} \in \{0, 1, 2\}$
  - A phenotype vector $\mathbf{y}$ $(n \times 1)$; it can be anything...
- Output:
  - Estimate a coefficient (effect size) β_j $(j \in 1, \ldots, p)$

- Testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.
- Report p-values
- Is it still straightforward after seeing the X matrix–so direly structured?

What are the differences between DEG and GWAS?

What are the differences between DEG and GWAS?

So many variants, much fewer samples/individuals

What are the differences between DEG and GWAS?

So many variants, much fewer samples/individuals

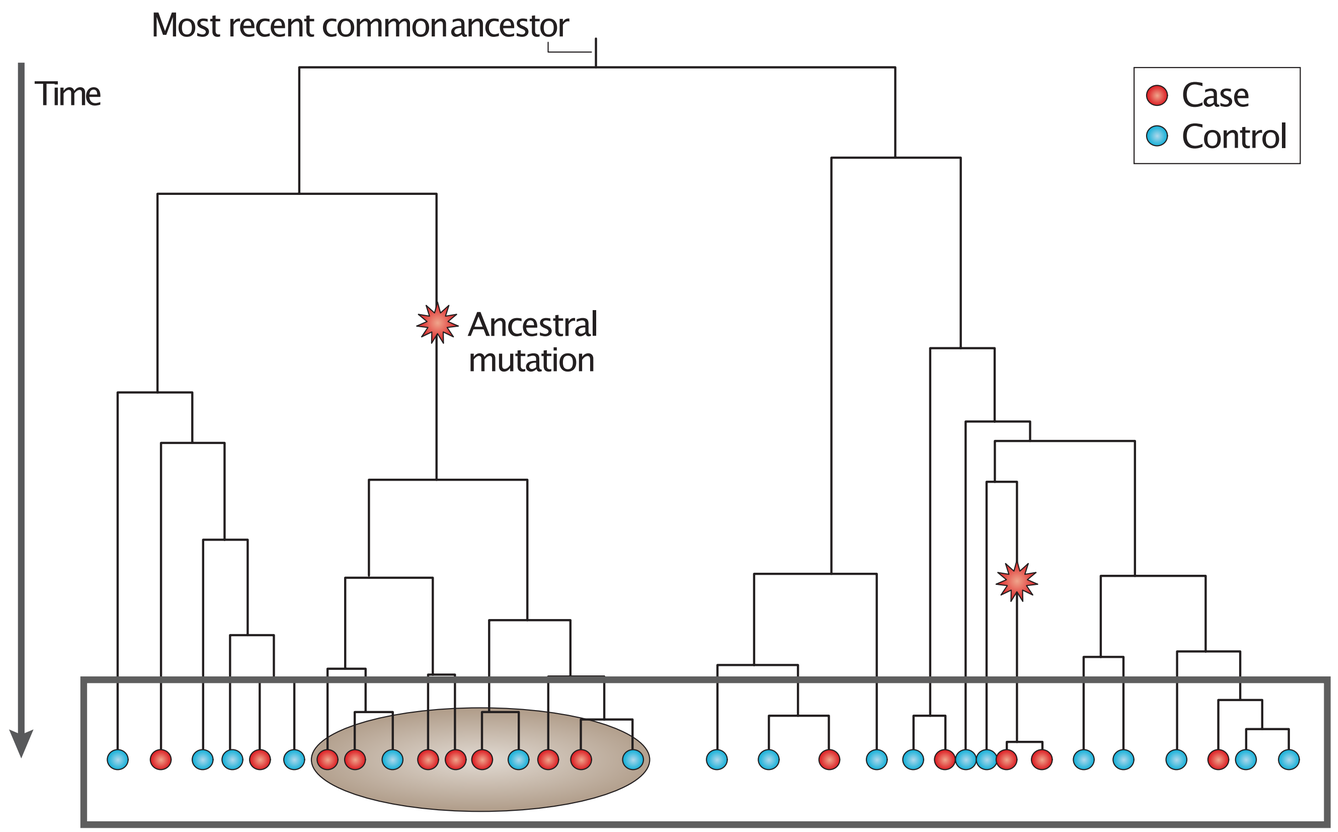
Expression $\sim$ phenotype vs. phenotype $\sim$ genetic variants

What are the differences between DEG and GWAS?

So many variants, much fewer samples/individuals

Goal: mapping a genetic variant $\rightarrow$ a phenotype

Genetic association tests compare genotype vs. phenotype variation across individuals



Don't care about the history

Observed complex phenotype

vs.

Genotype at a certain locus

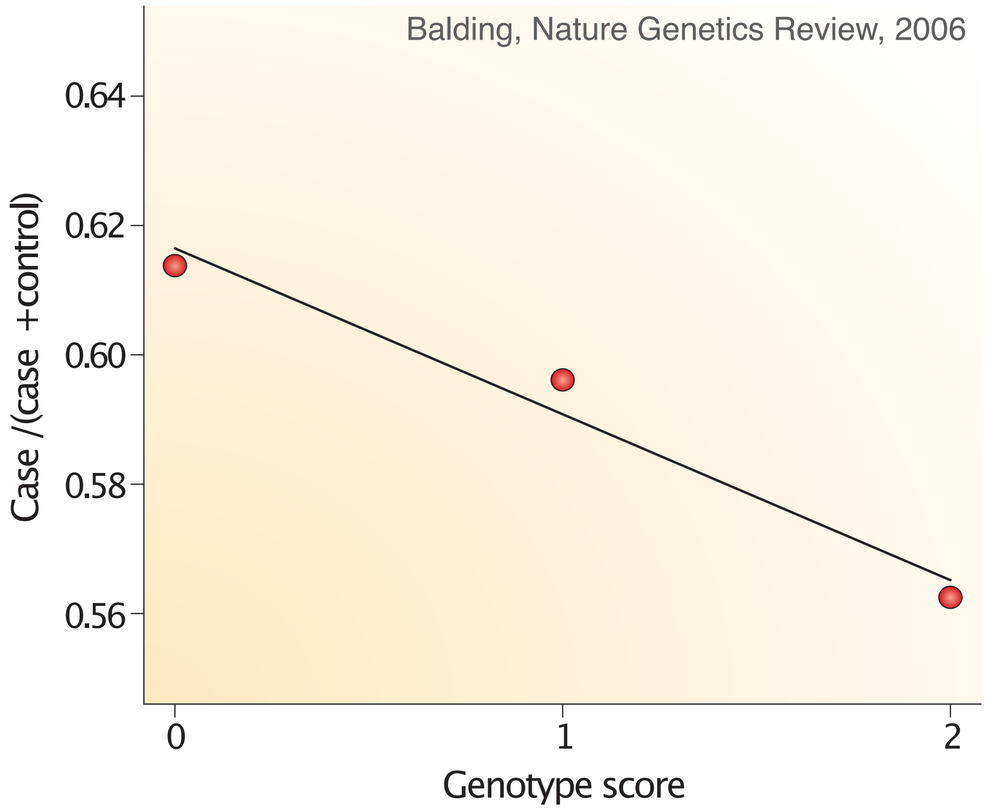
Balding, Nature Genetics Review, 2006

A typical genetics study design

Resistant
to a disease

More
susceptible
to a disease
(e.g., COVID)

T/T	0
C/C	**2**
T/T	0
T/**C**	**1**
T/T	0
T/T	0
T/T	0
T/T	0

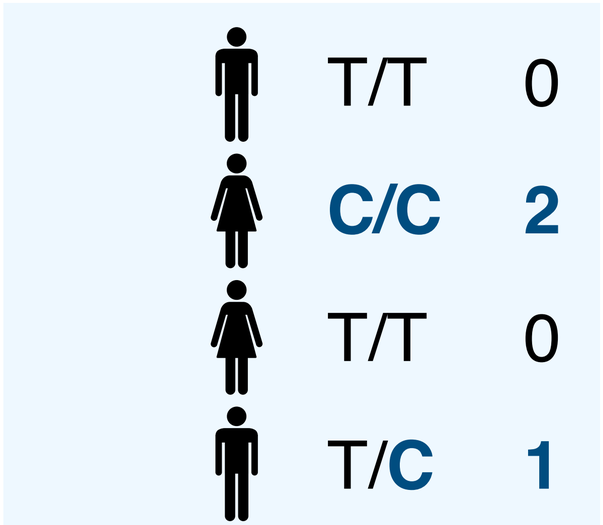


T = a major allele

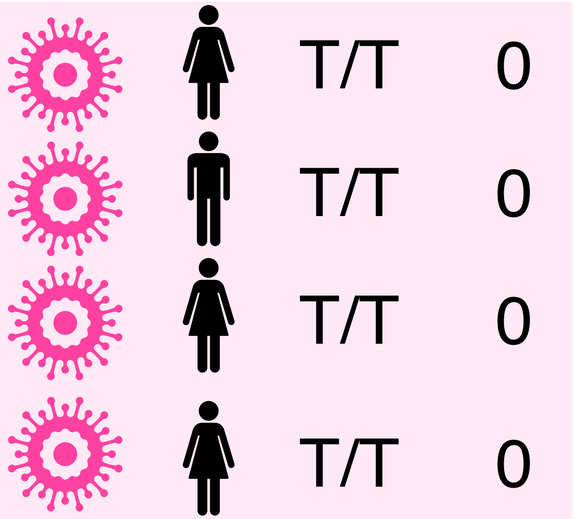
C = a minor allele

A typical genetics study design

Resistant
to a disease



More
susceptible
to a disease
(e.g., COVID)



Genotype (dosage) of a variant *j*

$$X_{ij} \in \{0,1,2\}$$

$$Y_i \in \{0,1\}$$

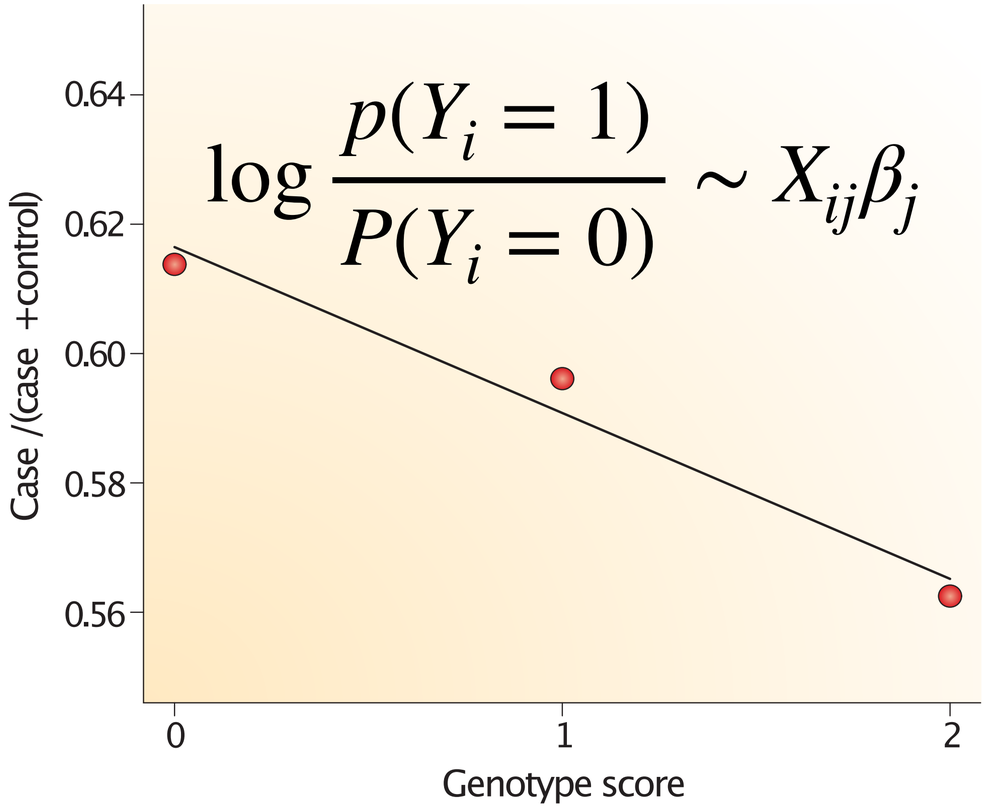
Genotype (dosage) of a variant *j*

$$\log \frac{p(Y_i = 1)}{P(Y_i = 0)} \sim X_{ij}\beta_j$$

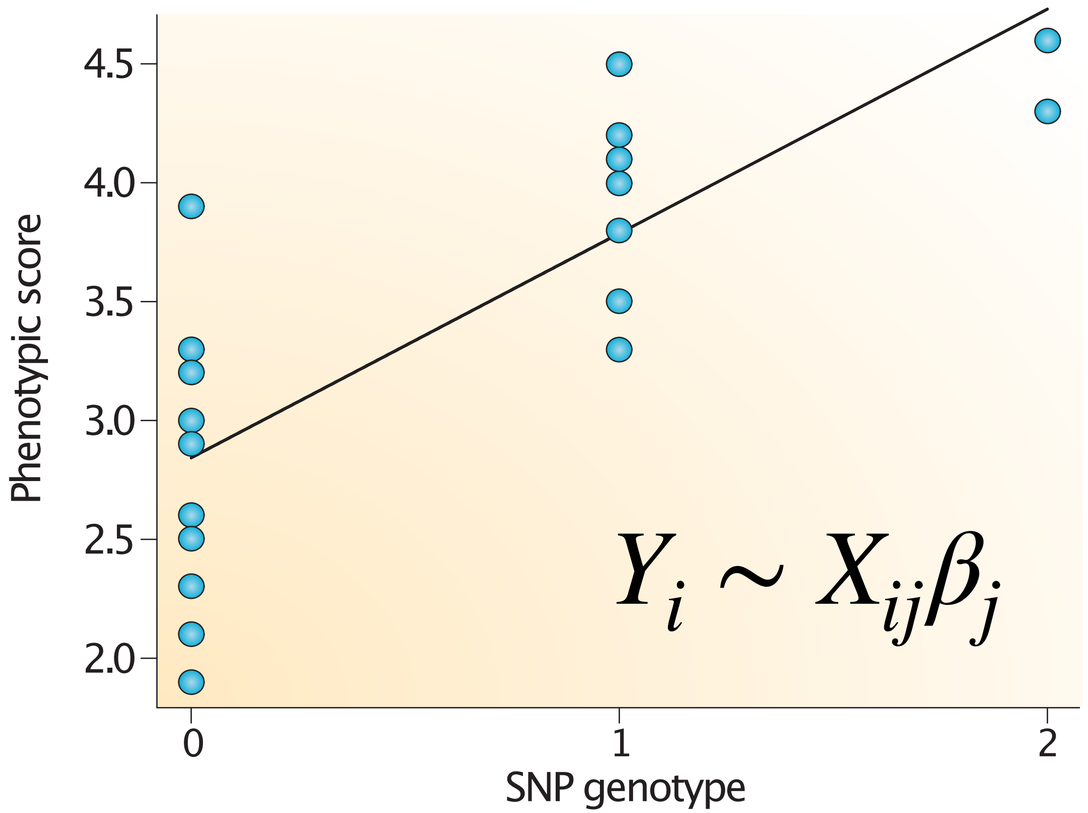
T = a major allele C = a minor allele

A traditional genetics study design

Case-control GWAS



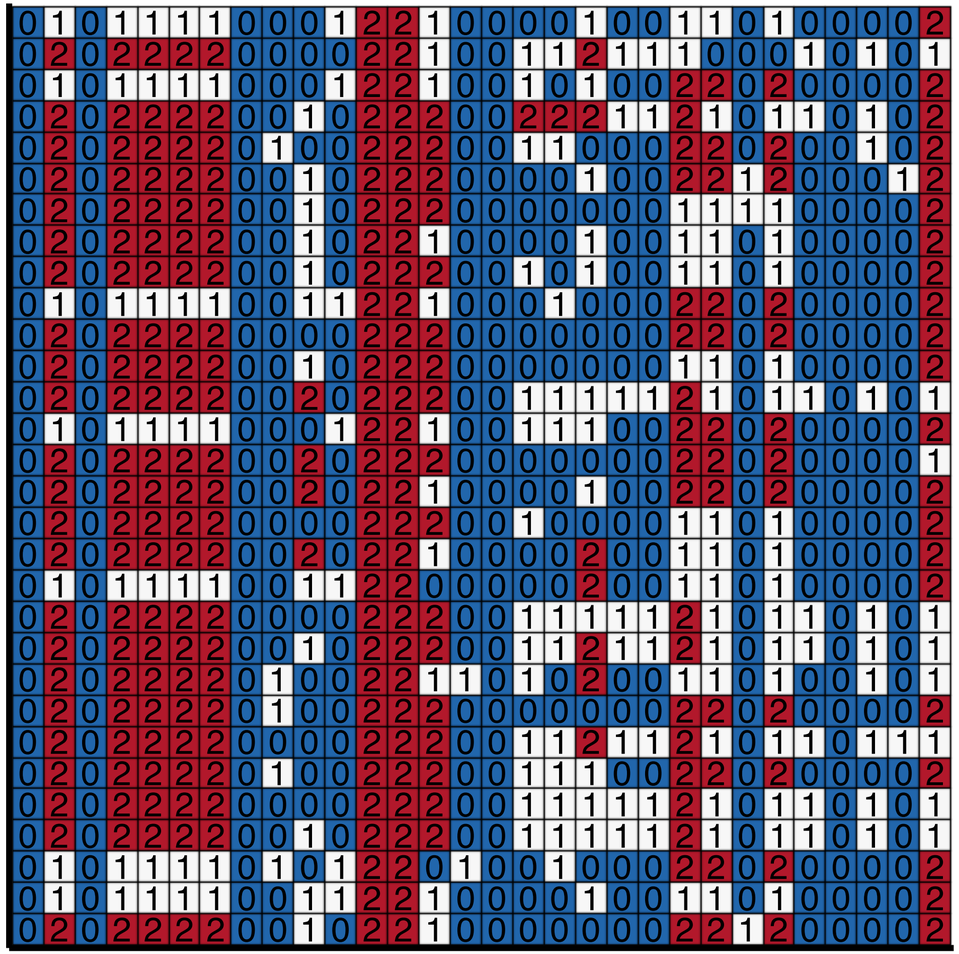
Quantitative trait GWAS



Balding, Nature Genetics Review, 2006

A toy example: GWAS for "Obsessive ggplot Disorder"[1]

A genotype matrix X ($X_{ij} \in \{0, 1, 2\}$):



A vector of phenotypes Y ($Y_i = 1$ if case vs. $Y_i = 0$ if control)

```
table(Y)
```

```
## Y
##    0    1
## 1272 1218
```

Goal: Find locations (columns of X) where a genotype $\mathbf{x}_j$ is associated with a phenotype $\mathbf{y}$

[1] an illustration purpose

Let's do GWAS to find the disease-associated variants

A variant-by-variant association t-test for a variant j:

$$H_0 : \mathbb{E}\left[X_{ij}|Y_i = 1\right] = \mathbb{E}\left[X_{ij}|Y_i = 0\right] \quad \text{vs.} \quad \mathbb{E}\left[X_{ij}|Y_i = 1\right] \neq \mathbb{E}\left[X_{ij}|Y_i = 0\right]$$

- The average genotype is the same between the case and control under the null.

```
j <- 1
t.test(X[Y == 0, j], X[Y == 1, j])
```

```
##
##  Welch Two Sample t-test
##
## data:  X[Y == 0, j] and X[Y == 1, j]
## t = 0.16273, df = 2468.6, p-value = 0.8707
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02015587  0.02380397
## sample estimates:
##  mean of x  mean of y
## 0.06996855 0.06814450
```

Can we calculate GWAS statistics more efficiently?

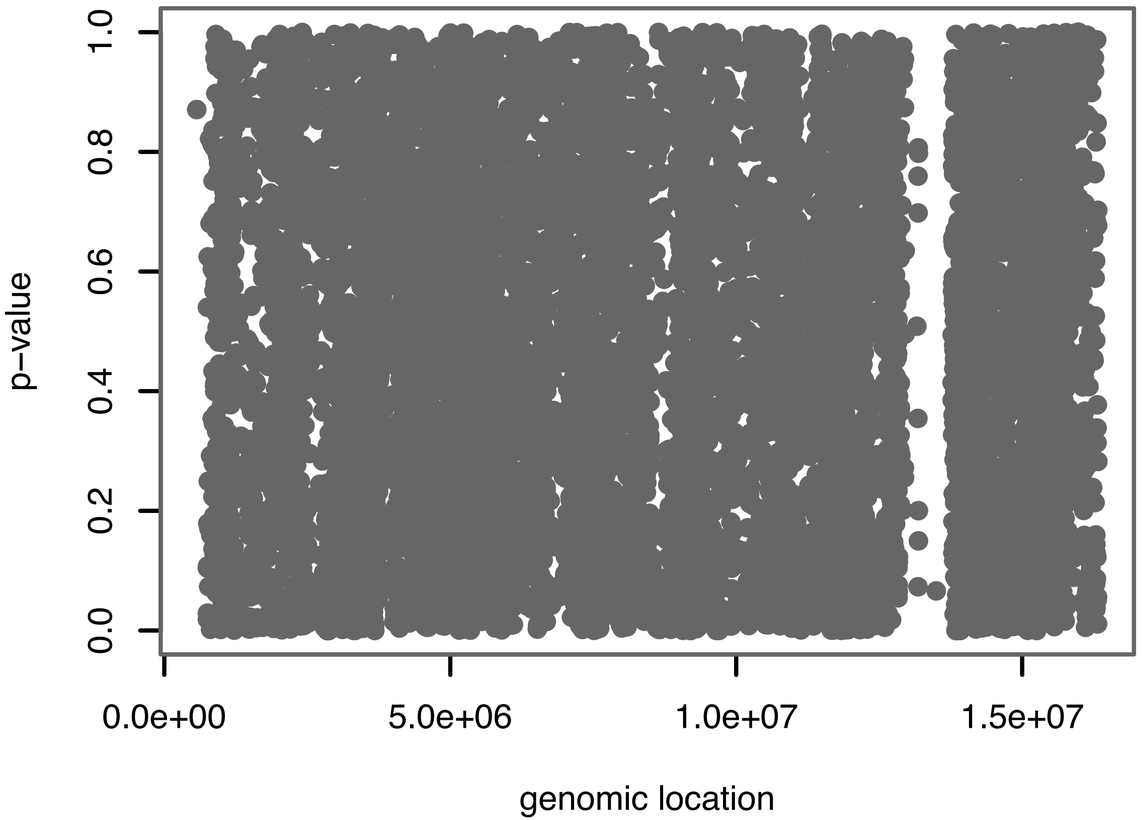
```
## library(matrixTests)
.gwas <- col_t_welch(X[Y == 0, ], X[Y == 1, ])
```

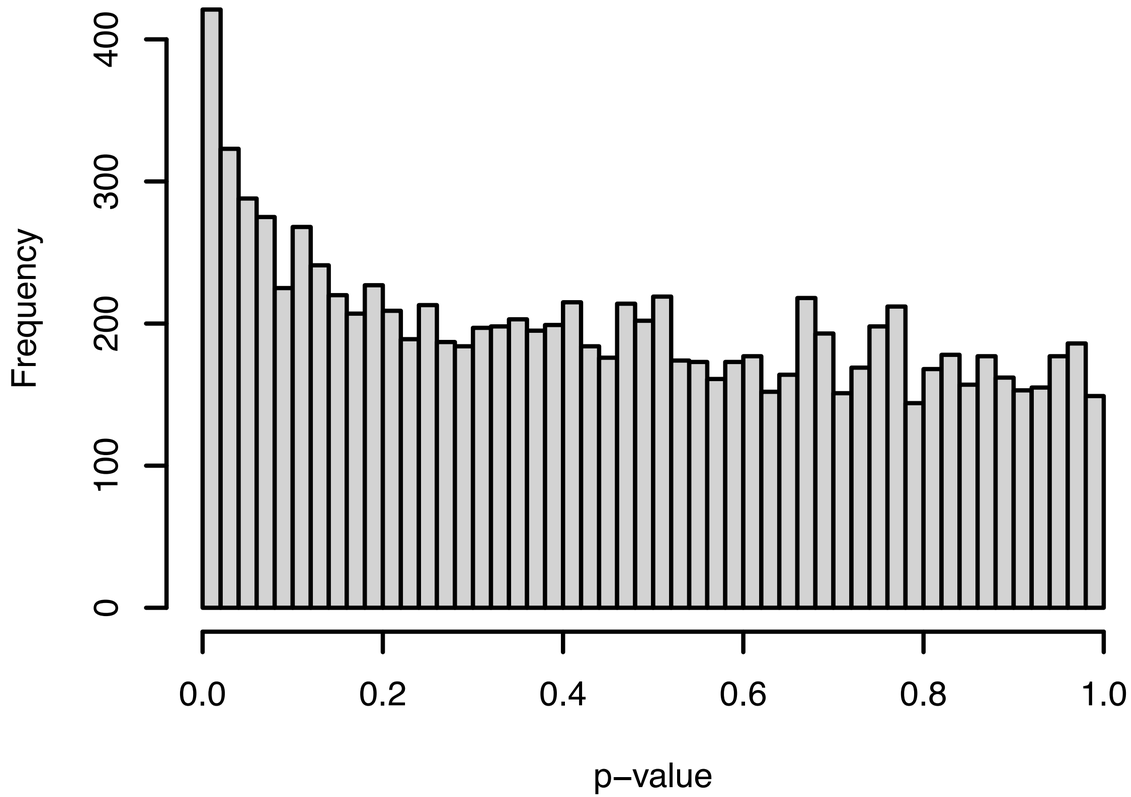
- For each pair of columns in the two matrices, the function performs t-test and summarize all the results into a list of vectors.

```
names(.gwas)
```

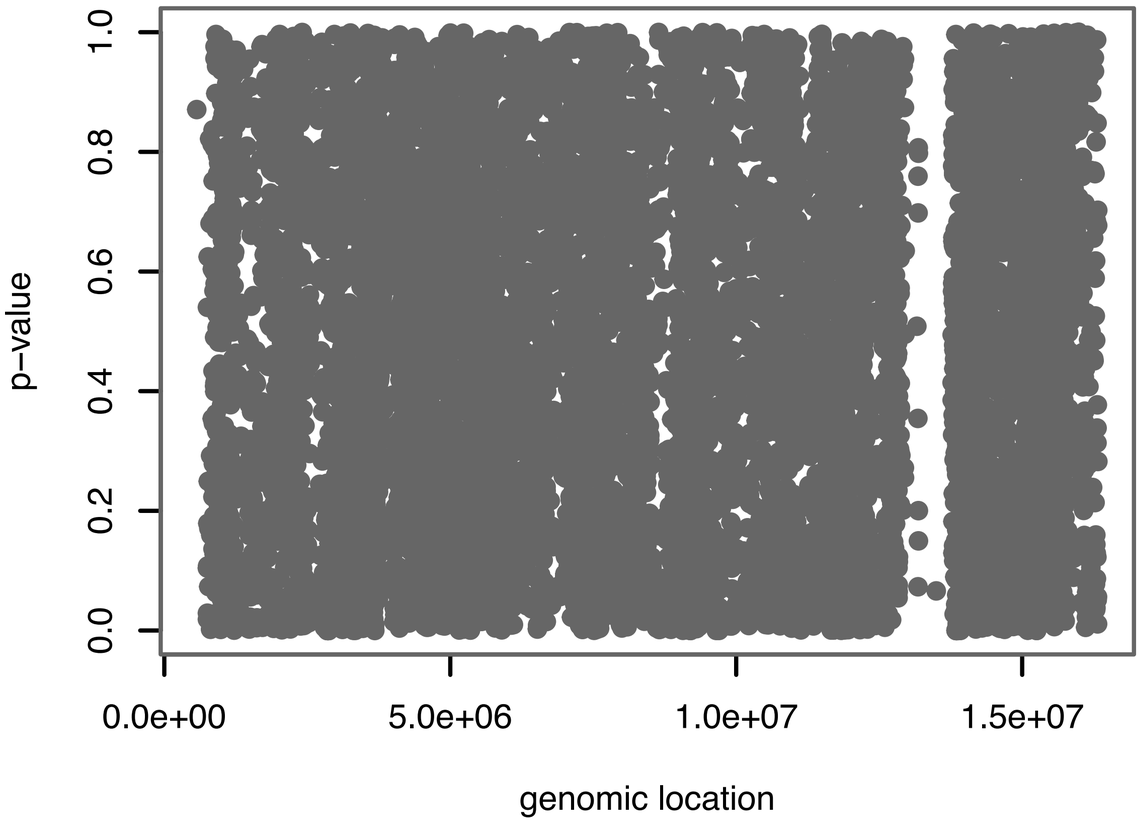
```
##  [1] "obs.x"       "obs.y"       "obs.tot"     "mean.x"      "mean.y"
##  [6] "mean.diff"   "var.x"       "var.y"       "stderr"      "df"
## [11] "statistic"   "pvalue"      "conf.low"    "conf.high"   "mean.null"
## [16] "alternative" "conf.level"
```

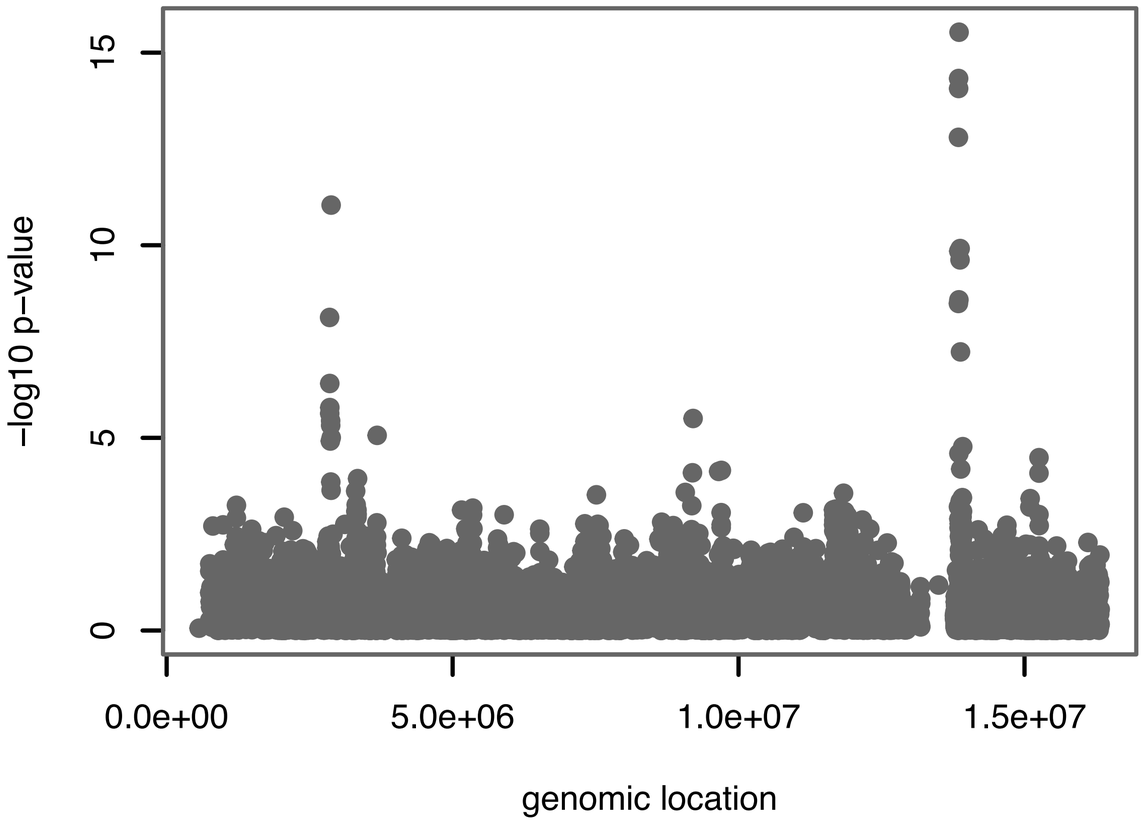
Manhattan plot: A quick summary of all the GWAS p-values





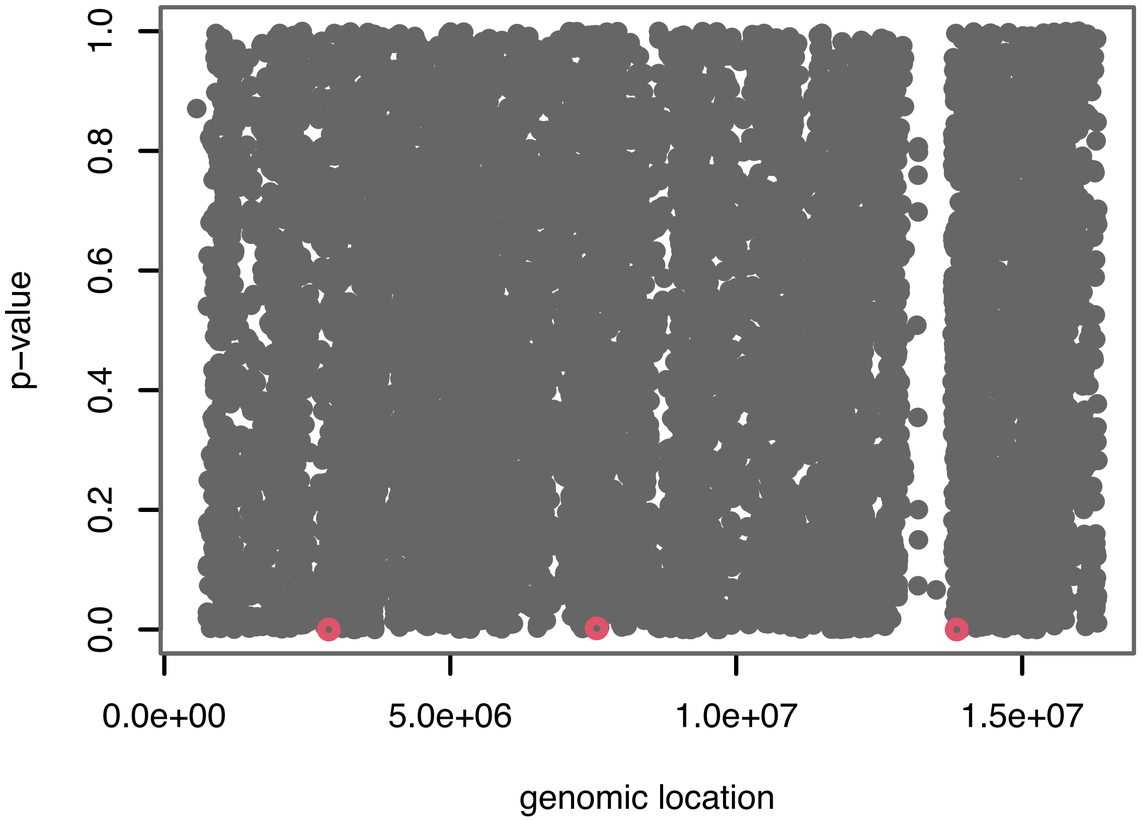
Manhattan plot: A quick summary of all the GWAS p-values



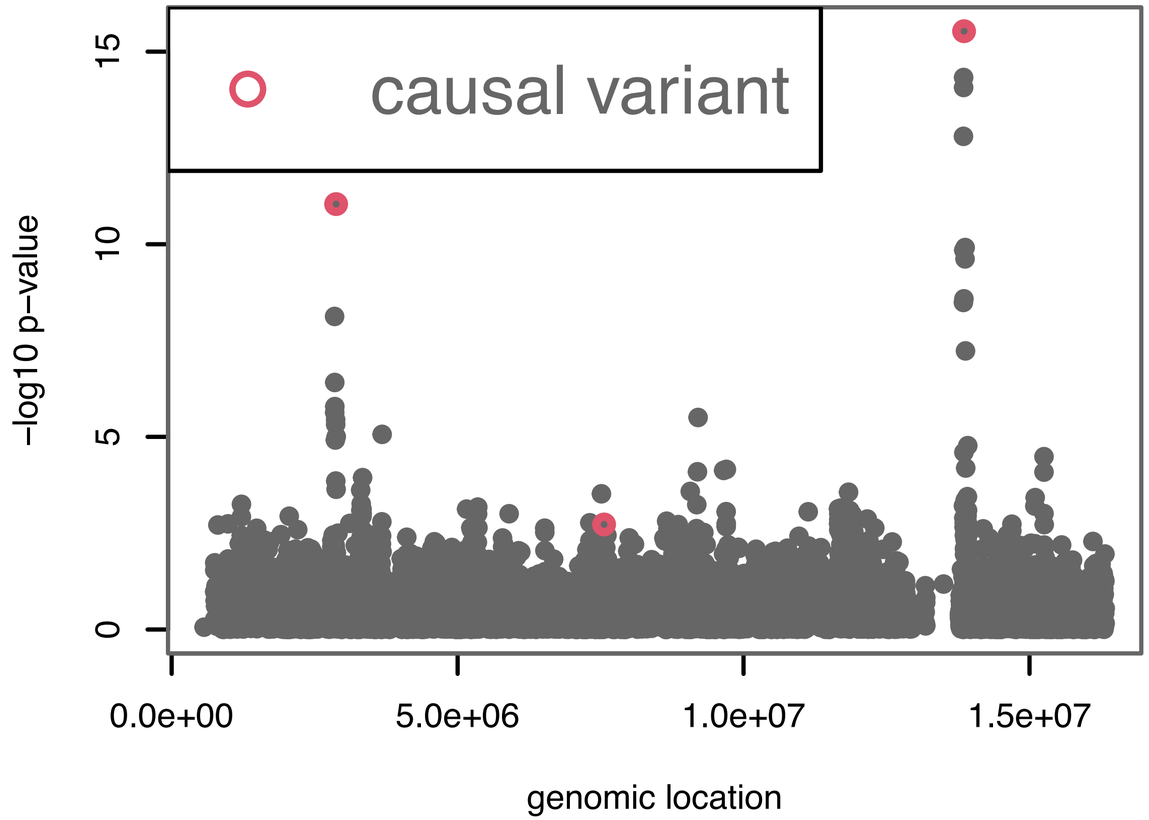


Which one is more informative?

Manhattan plot: A quick summary of all the GWAS p-values



Where are the variants significantly associated with OGD?



Which one is more informative?

Multiple hypothesis adjustment

We have many p-values, namely, $p_1, p_2, \dots, p_{10000}$.

1. In GWAS, we usually take a very conservative stance. So, we adjust the nominal p-values by Bonferroni adjustment.

```
padj <- p.adjust(.gwas$pvalue)
sum(padj < .05)
```

```
## [1] 18
```

Multiple hypothesis adjustment

We have many p-values, namely, $p_1, p_2, \dots, p_{10000}$.

1. In GWAS, we usually take a very conservative stance. So, we adjust the nominal p-values by Bonferroni adjustment.
2. What will be the adjusted p-value for p_j?

```
padj <- p.adjust(.gwas$pvalue)
sum(padj < .05)
```

```
## [1] 18
```

Multiple hypothesis adjustment

We have many p-values, namely, $p_1, p_2, \dots, p_{10000}$.

1. In GWAS, we usually take a very conservative stance. So, we adjust the nominal p-values by Bonferroni adjustment.
2. What will be the adjusted p-value for p_j?
3. $p_j^{\text{adj}} \leftarrow \min\{1, p_j \times 10000\}$

```
padj <- p.adjust(.gwas$pvalue)
sum(padj < .05)
```

```
## [1] 18
```

Multiple hypothesis adjustment

We have many p-values, namely, $p_1, p_2, \dots, p_{10000}$.

1. In GWAS, we usually take a very conservative stance. So, we adjust the nominal p-values by Bonferroni adjustment.
2. What will be the adjusted p-value for p_j?
3. $p_j^{\text{adj}} \leftarrow \min\{1, p_j \times 10000\}$
4. $p_j < \alpha/10000$ where 10000 is the total number of hypothesis tests

```
padj <- p.adjust(.gwas$pvalue)
sum(padj < .05)
```

```
## [1] 18
```

Genome-wide significance level

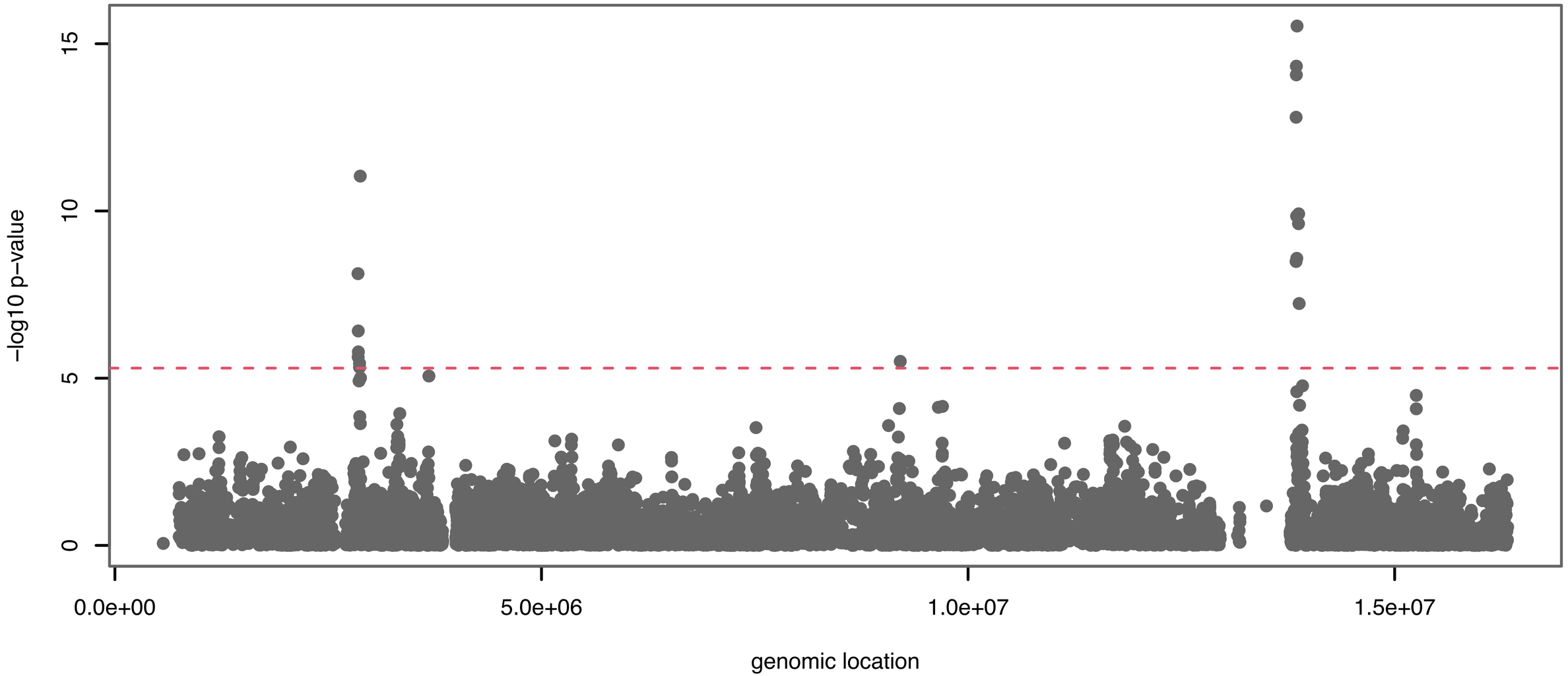
$$p_j < \frac{0.05}{\text{number of (independent) tests}} = \frac{0.05}{10^6} = 5 \times 10^{-8}$$

$$\text{FWER} = P\left(\cup_{j=1}^{10^6} \{p_j < 0.05/10^6\} \,|\, H_0 \right)$$

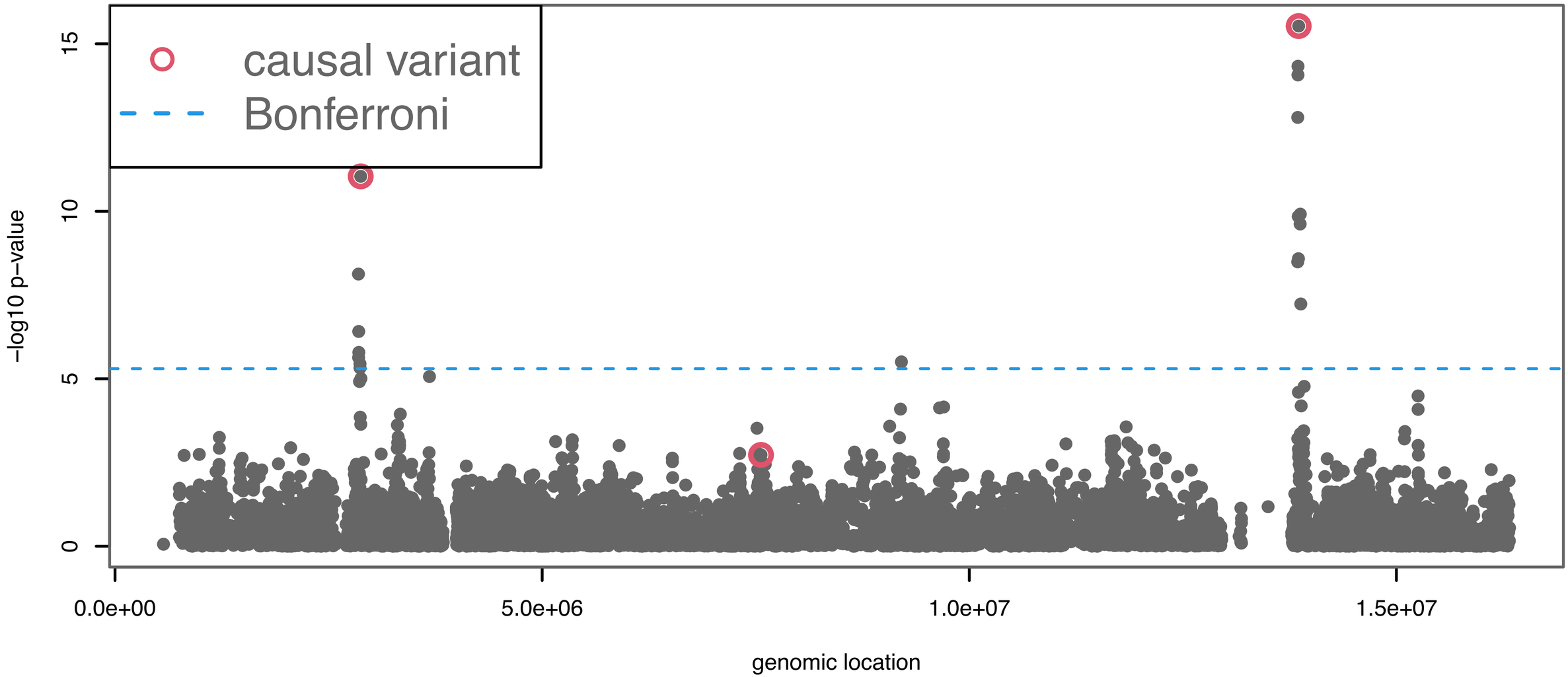
Union bound

$$< \sum_{j=1}^{10^6} P(p_j < 0.05/10^6 \,|\, H_0) = \sum_{j=1}^{10^6} \frac{0.05}{10^6} = 0.05$$

How many variants vs. how many “independent” variants



How many variants vs. how many "independent" variants



Adjusted p-value should control FWER or FDR only if...

```
padj <- p.adjust(gwas.p, method="bonferroni")
false.discoveries <- setdiff(which(padj < .05), causal.variants)
empirical.FDR <- length(false.discoveries) / sum(padj < .05)
print(empirical.FDR)
```

```
## [1] 0.8888889
```

```
padj <- p.adjust(gwas.p, method="BH")
false.discoveries <- setdiff(which(padj < .05), causal.variants)
empirical.FDR <- length(false.discoveries) / sum(padj < .05)
print(empirical.FDR)
```

```
## [1] 0.9354839
```

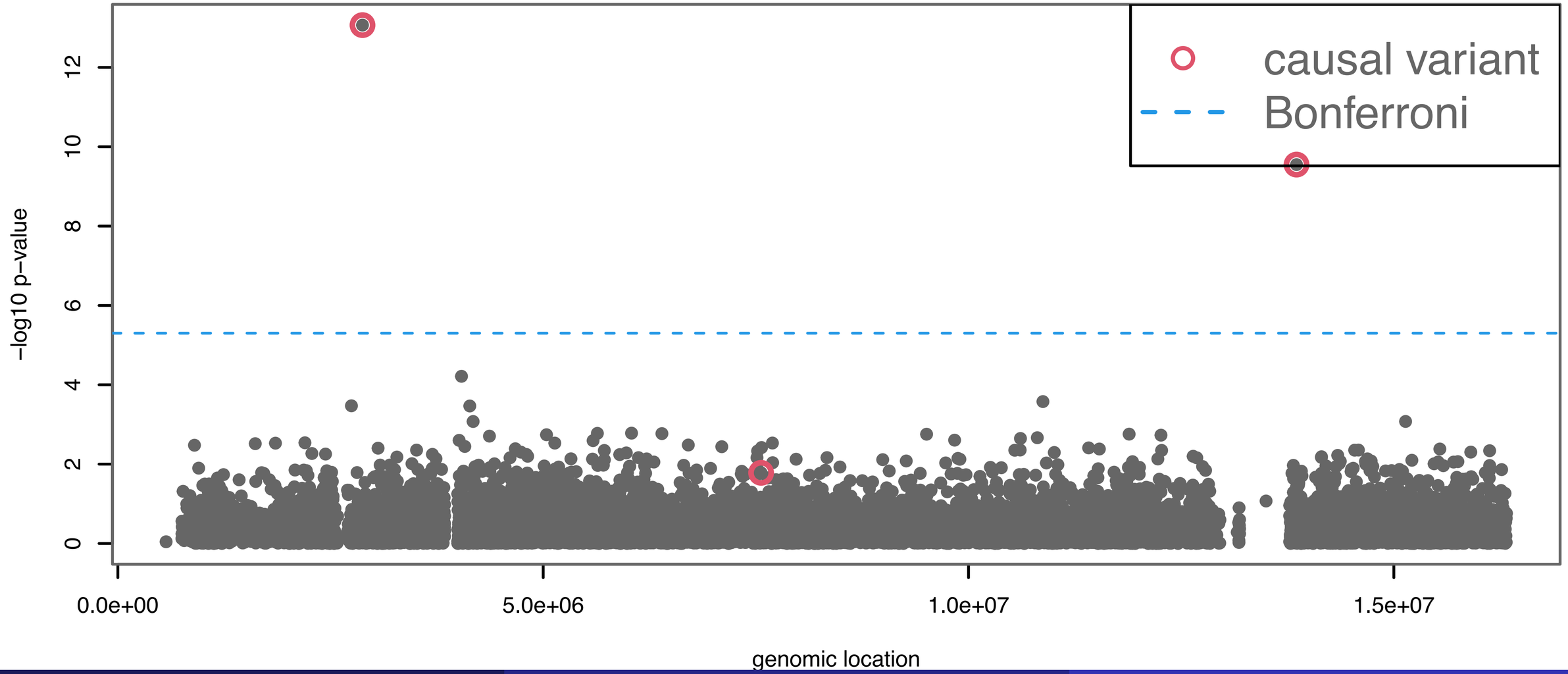
What went wrong with GWAS?

What went wrong with GWAS?

Actually, everything is perfectly expected.

What if we do IID variant association studies?

- Instead of using a real-world genotype matrix, we could do the same exercise with IID 0,1,2 matrix (sample permutation).



Today's lecture

1. Mapping disease-specific locations in Genome
2. How GWAS can be interpreted wrongfully
3. Combining evidence from multiple studies
4. Appendix

Exploratory Data Analysis of the 1KG data

In the 1KG data, we have 2490 rows/individuals and 1664852 columns/variants.

- Consistently, we will use n to refer to the number of samples and p to the number of variants.
- Let's assume that we are mostly interested in biallelic (two alleles) variants/SNPs.

Exploratory Data Analysis of the 1KG data

In the 1KG data, we have 2490 rows/individuals and 1664852 columns/variants.

- Consistently, we will use n to refer to the number of samples and p to the number of variants.
- Let's assume that we are mostly interested in biallelic (two alleles) variants/SNPs.
- *Recall*: Each SNP/genotype comprises two numbers (haplotypes)–one from the maternal and the other from the paternal genome.

Exploratory Data Analysis of the 1KG data

In the 1KG data, we have 2490 rows/individuals and 1664852 columns/variants.

- Consistently, we will use n to refer to the number of samples and p to the number of variants.
- Let's assume that we are mostly interested in biallelic (two alleles) variants/SNPs.
- *Recall*: Each SNP/genotype comprises two numbers (haplotypes)–one from the maternal and the other from the paternal genome.
- There are major (usually reference) and minor (usually effect) alleles. To save bits (in the early 2000s), geneticists used 0 for the major and 1 for the minor allele.

Exploratory Data Analysis of the 1KG data

In the 1KG data, we have 2490 rows/individuals and 1664852 columns/variants.

- Consistently, we will use n to refer to the number of samples and p to the number of variants.
- Let's assume that we are mostly interested in biallelic (two alleles) variants/SNPs.
- *Recall*: Each SNP/genotype comprises two numbers (haplotypes)–one from the maternal and the other from the paternal genome.
- There are major (usually reference) and minor (usually effect) alleles. To save bits (in the early 2000s), geneticists used 0 for the major and 1 for the minor allele.
- So, each variant, we have a minor allele frequency (MAF), namely, f_j, for a variant j.

If each variant follows Binomial distribution

In Binomial distribution, with a MAF f_j, we have

- What will be a useful summary statistic for a variant j?

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

If each variant follows Binomial distribution

In Binomial distribution, with a MAF f_j, we have

- What will be a useful summary statistic for a variant j?

$$\widehat{\mathbb{E}}[X_j] = 2f_j$$

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

If each variant follows Binomial distribution

In Binomial distribution, with a MAF f_j, we have

- What will be a useful summary statistic for a variant j?

$$\widehat{\mathbb{E}}[X_j] = 2f_j$$

- What is the variance of this variant?

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

If each variant follows Binomial distribution

In Binomial distribution, with a MAF f_j, we have

- What will be a useful summary statistic for a variant j?

$$\hat{\mathbb{E}}\left[X_j\right] = 2f_j$$

- What is the variance of this variant?

$$\hat{\mathbb{V}}\left[X_j\right] = 2f_j(1 - f_j)$$

Remark: Technically, the dosage (0,1,2) does not follow binomial distribution. Why? The underlying data generation process involves haplotypes (separating the maternal and paternal 0/1 counts) and dependency along the genomic axis.

Variant-level MAF across individuals

We can easily calculate MAF using `bigsnpr`:

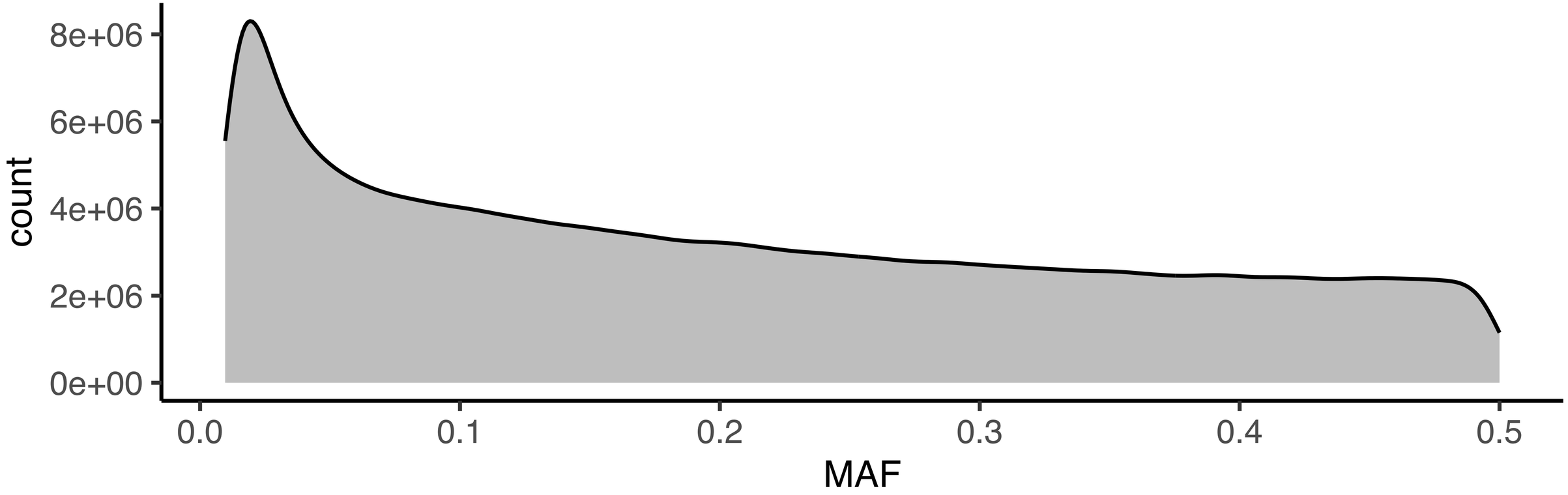
```
maf <- snp_MAF(data$genotypes, ncores=16)
```

- What are the x- and y-axis? Where is the peak?

Variant-level MAF across individuals

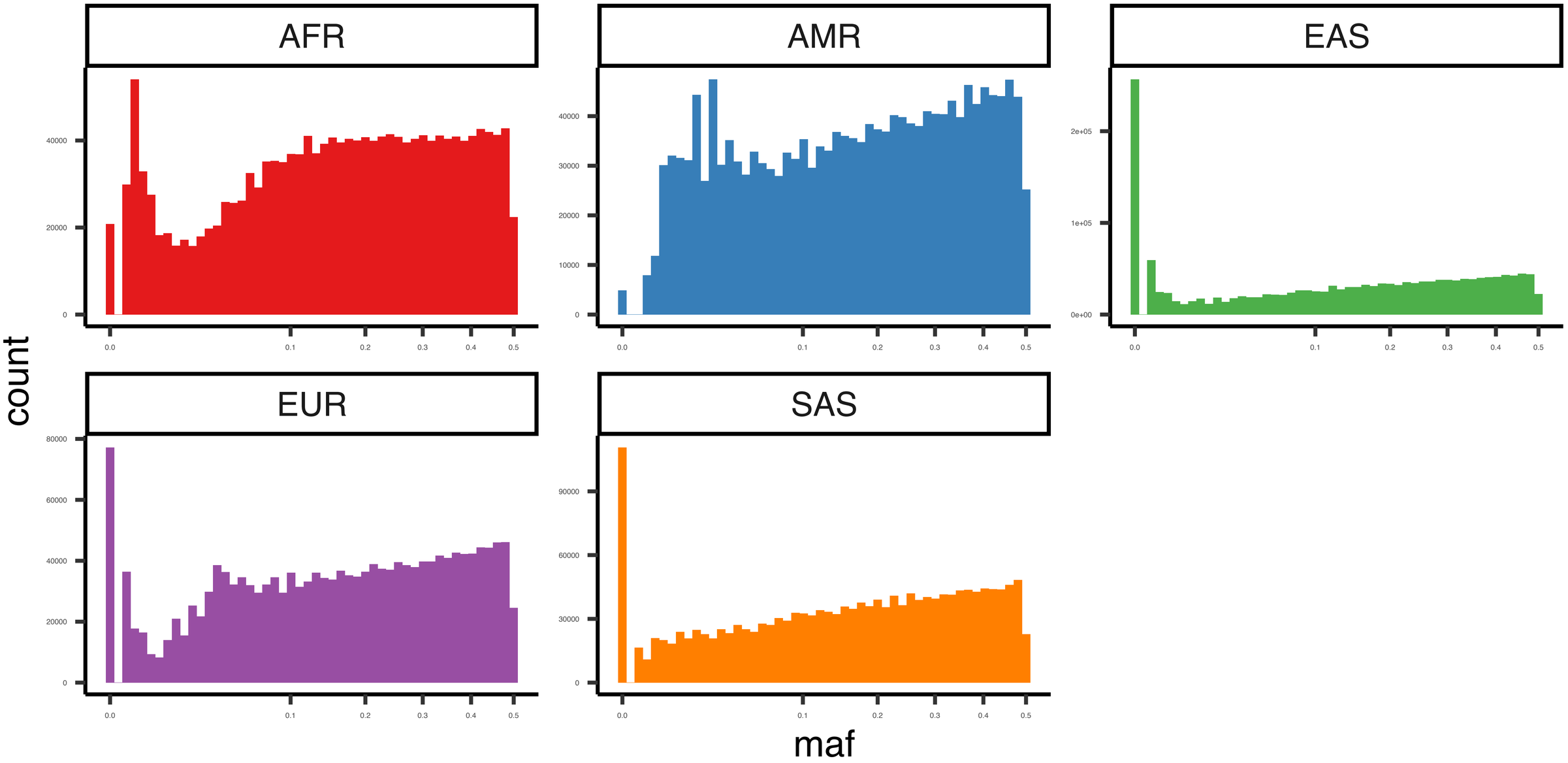
We can easily calculate MAF using `bigsnpr`:

```
maf <- snp_MAF(data$genotypes, ncores=16)
```



- What are the x- and y-axis? Where is the peak?

MAF distributions generally differ across ancestry groups (AG)

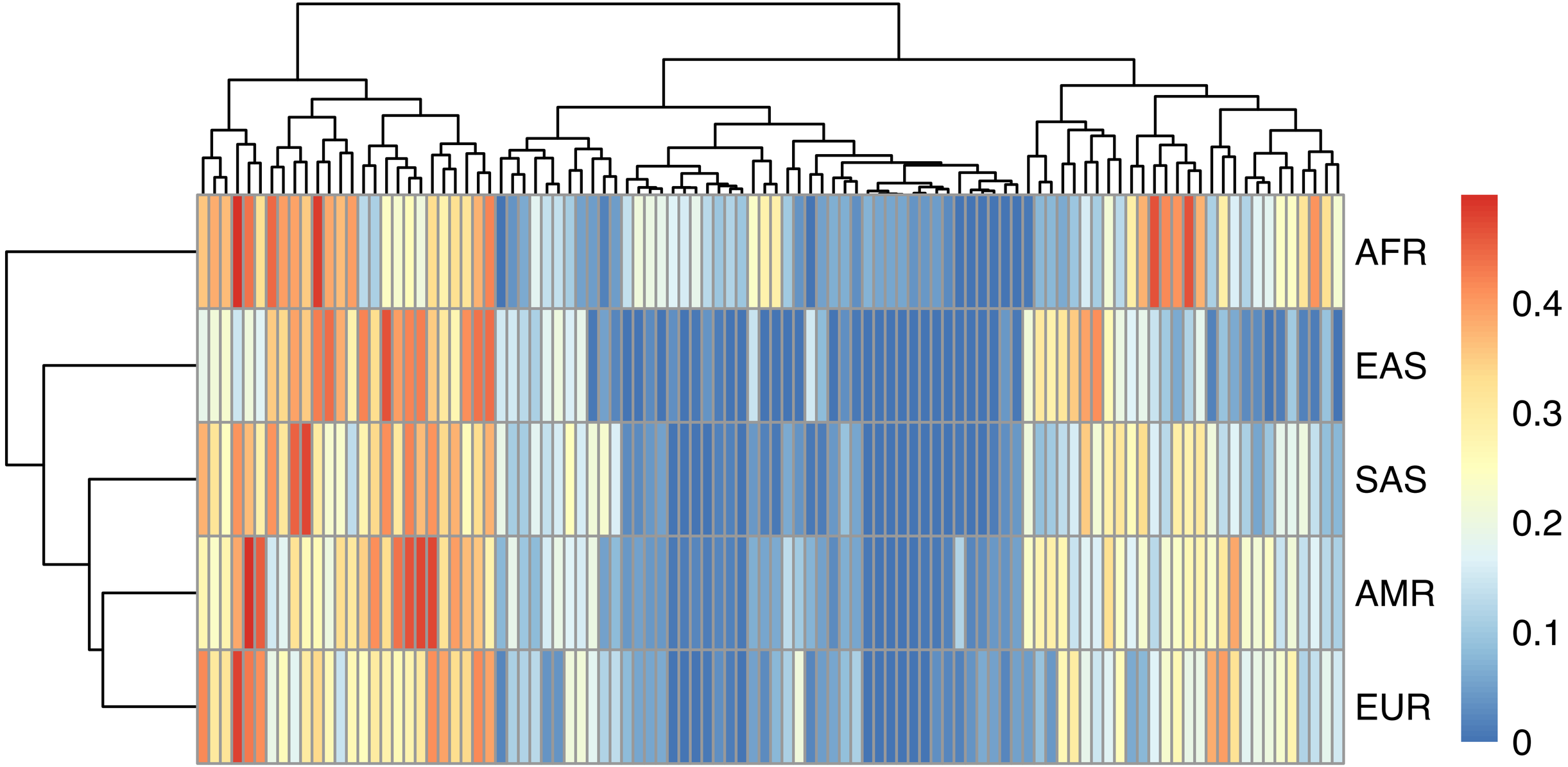


The same variant could have vastly different MAF values across AG

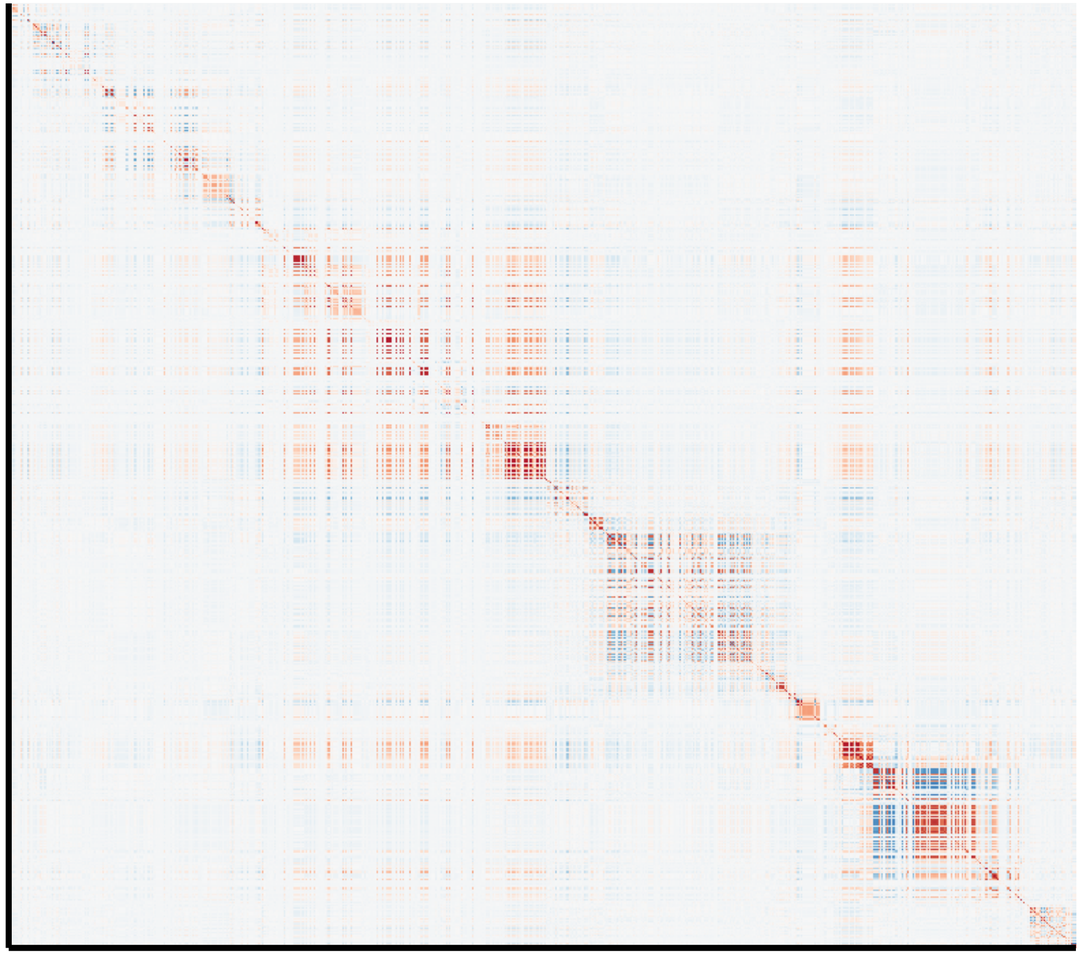
```
head(maf.mat)
```

```
##             AFR         AMR       EAS         EUR       SAS
##           <num>       <num>     <num>       <num>     <num>
## 1: 0.12576687 0.008645533 0.0000000 0.001988072 0.00000000
## 2: 0.28987730 0.263688761 0.2341270 0.161033797 0.22004132
## 3: 0.02607362 0.028818444 0.0000000 0.040755467 0.04028926
## 4: 0.35199387 0.243515850 0.2668651 0.128230616 0.19008264
## 5: 0.35199387 0.244956772 0.2371032 0.128230616 0.19008264
## 6: 0.45092025 0.175792507 0.1170635 0.130218688 0.16632231
```

Random 100 variants



What are the covariance matrices of X? – between the variants



More discussions in the next lectures.

For a genotype matrix X $(n \times p)$,

Linkage disequilibrium matrix

A $p \times p$ matrix:

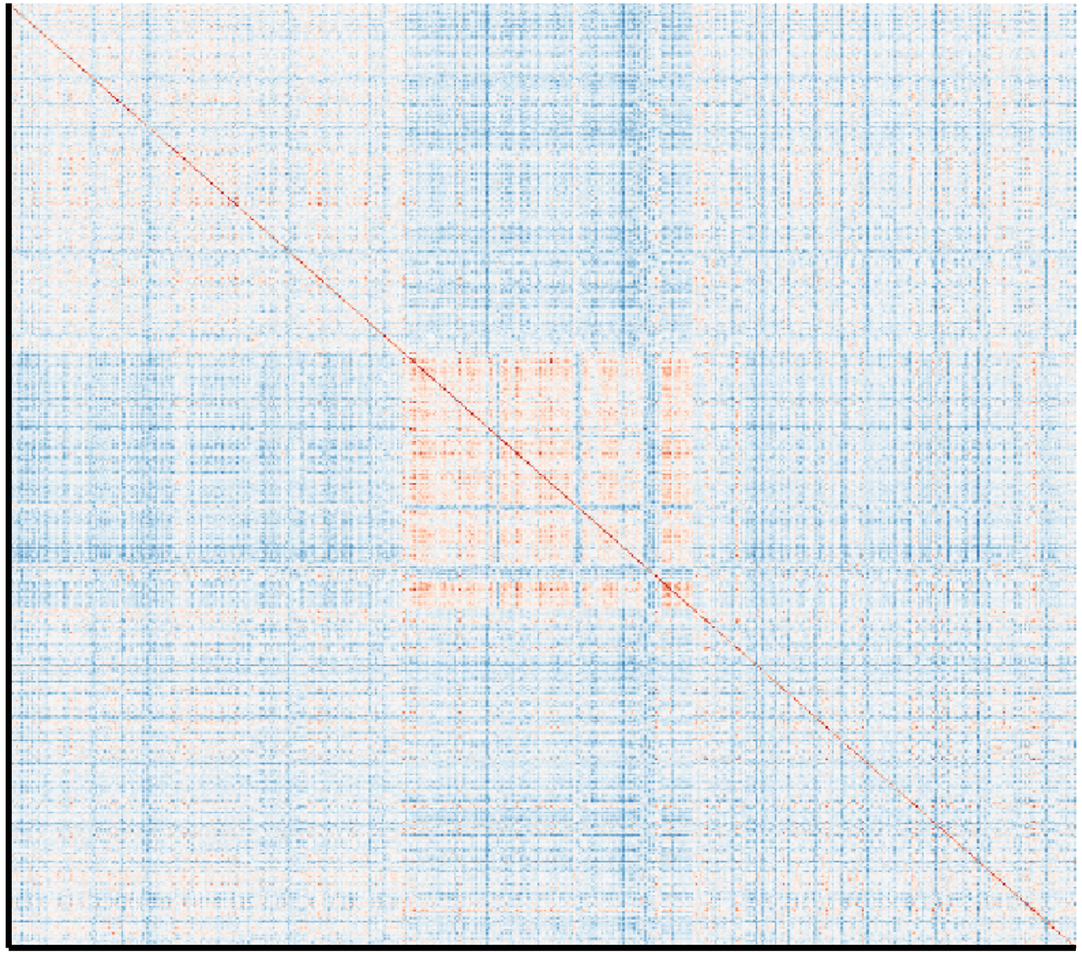
$$\hat{R} = \frac{1}{n} X^\top X$$

given each column $\mathbf{x}_j$ standardized (mean 0, SD 1).

- n: individuals
- p: variants/SNPs

$$R_{ij} = \frac{1}{n} \sum_{r=1}^{n} X_{ri} X_{rj}$$

What are the covariance matrices of X? – between the individuals



For a genotype matrix X $(n \times p)$

Kinship/genetic relatedness matrix

An $n \times n$ matrix:

$$\hat{K} = \frac{1}{p} X X^\top$$

given each row $\mathbf{x}_i$ standardized (mean 0, SD 1).

- n: individuals
- p: variants/SNPs

$$K_{ij} = \frac{1}{p} \sum_{g=1}^{p} X_{ig} X_{jg}$$

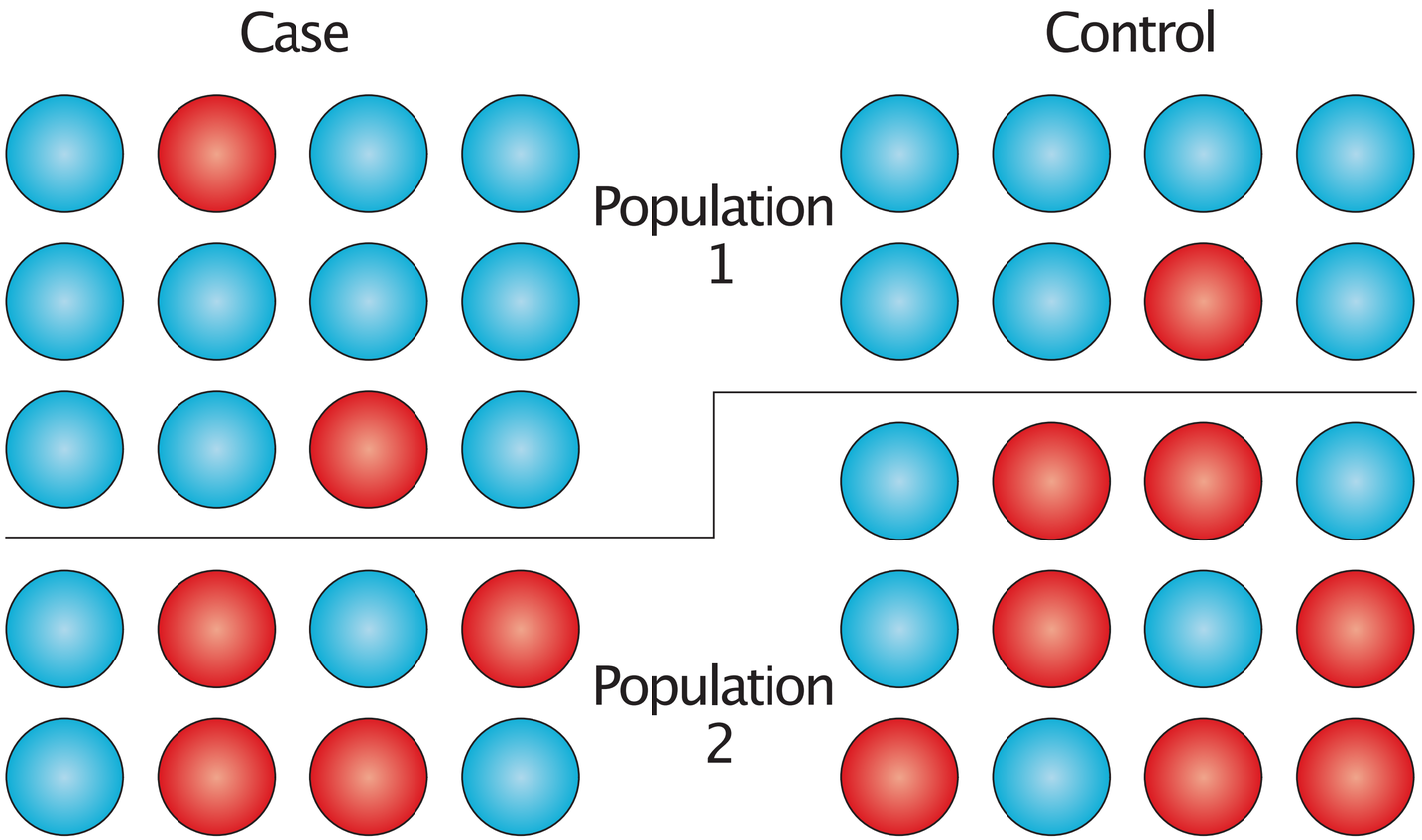
What would happen if there were AG-specific biases?

$$g\left(\underset{\text{phenotype}}{Y}\right) \sim \underbrace{\sum_{k \in \text{causal variants}} X_k \beta_k}_{\text{true genetic effect}} + \underbrace{\sum_{g} \mathbf{U}_g \gamma_g}_{\text{ancestry group factors}} + \epsilon$$

Do GWAS as usual by variant-by-variant t-test:

```
.gwas.biased <- col_t_welch(X[Y == 0, ], X[Y == 1, ])
```

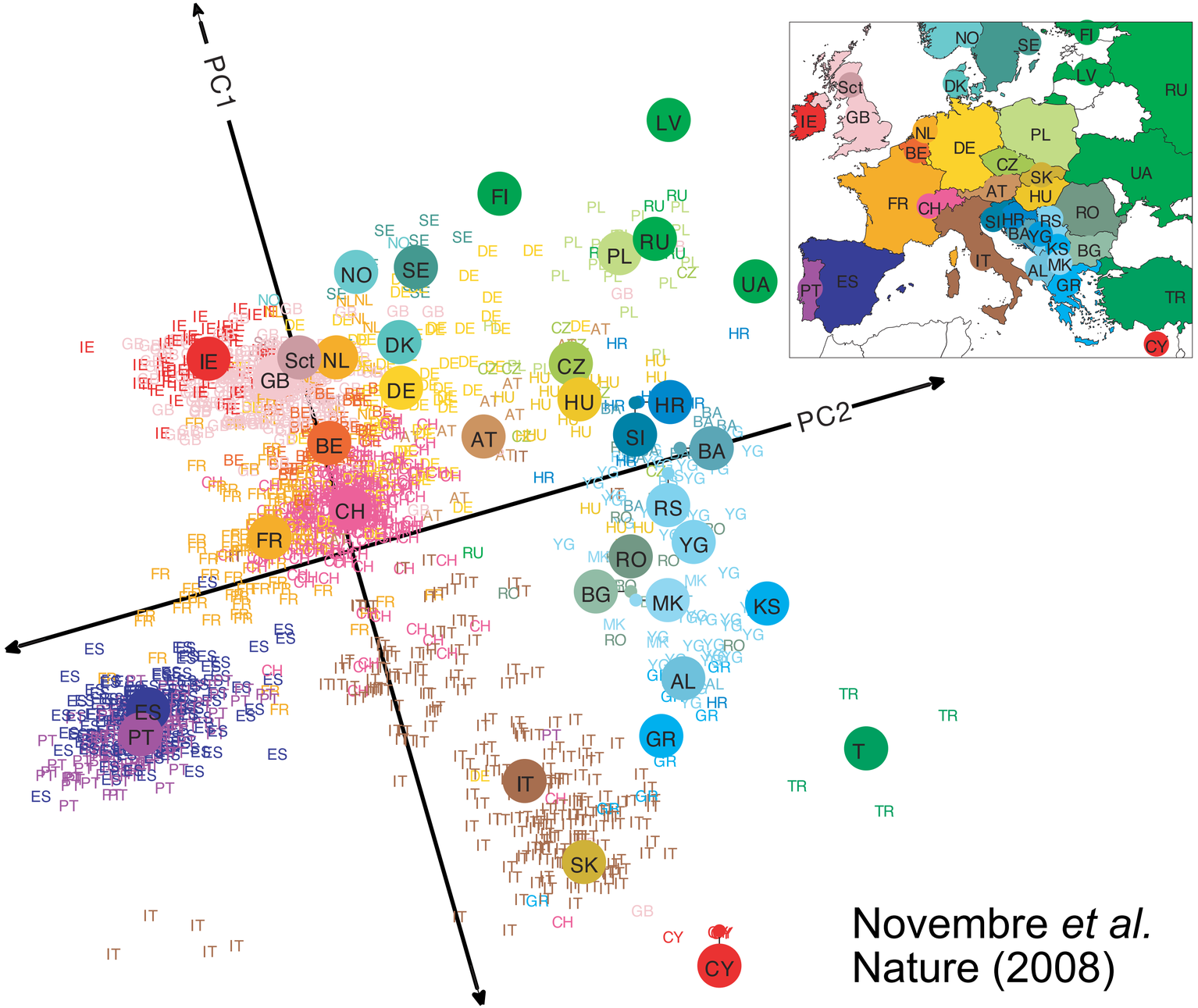
What if there were a hidden population structure?



Balding, Nature Genetics Review, 2006

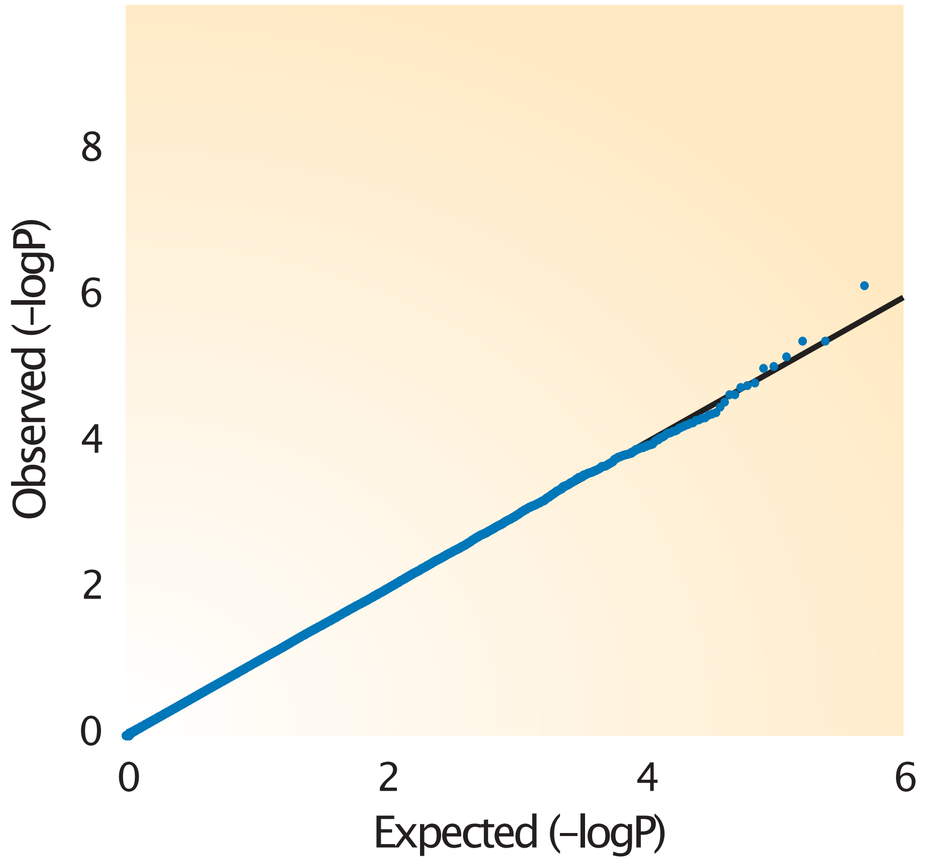
The human population as a result of the human migration history

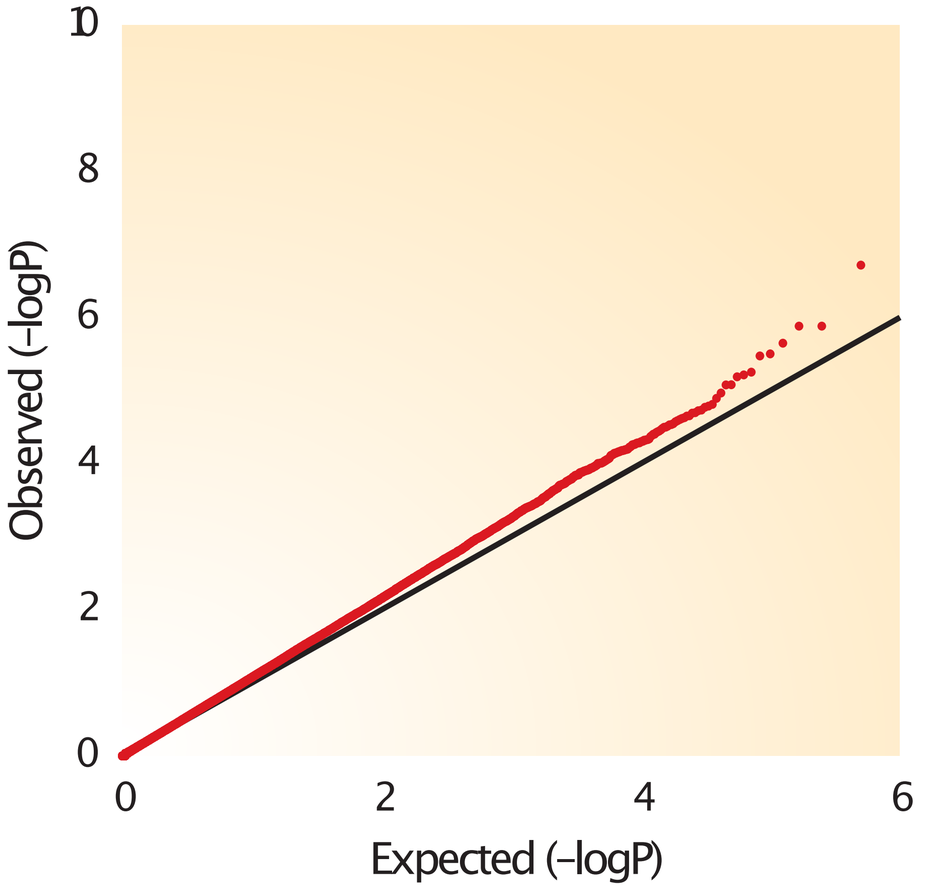
https://www.nytimes.com/2018/03/20/science/david-reich-human-migrations.html



Novembre *et al.*
Nature (2008)

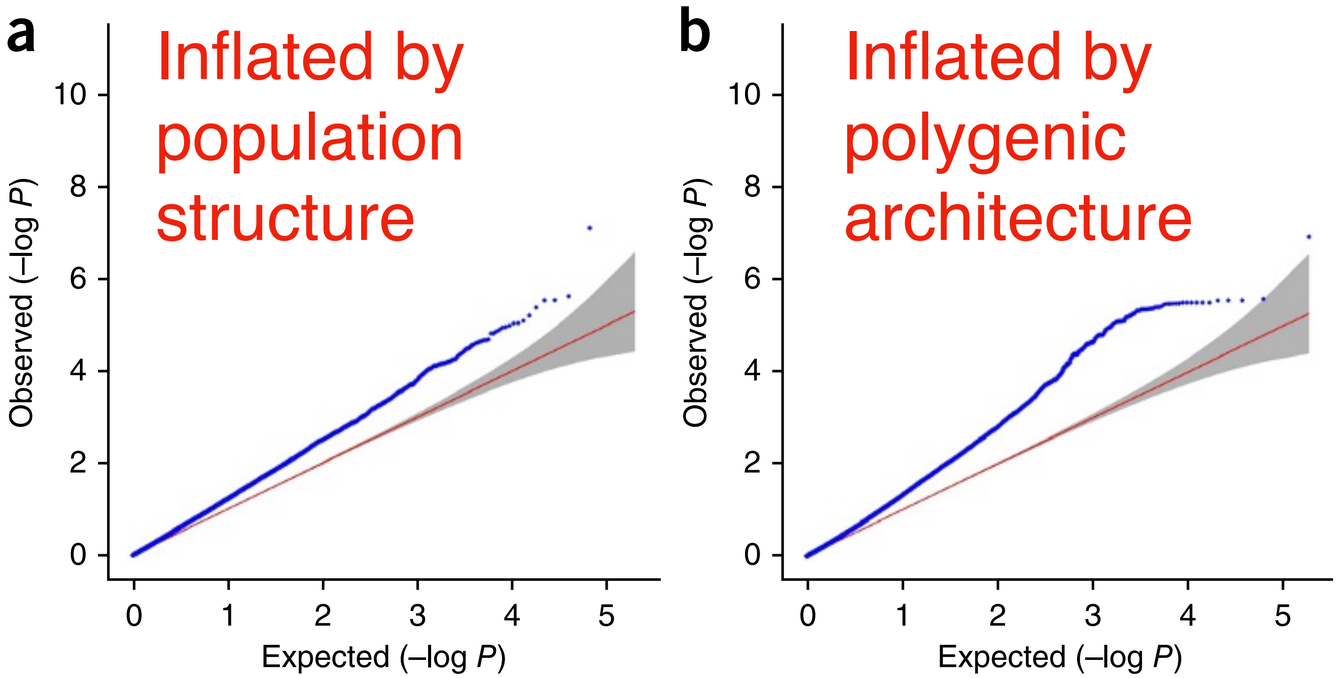
Population structures may inflate GWAS stats





Price, Zaitlen, Reich, Patterson

Population structures and polygenicity may inflate GWAS stats

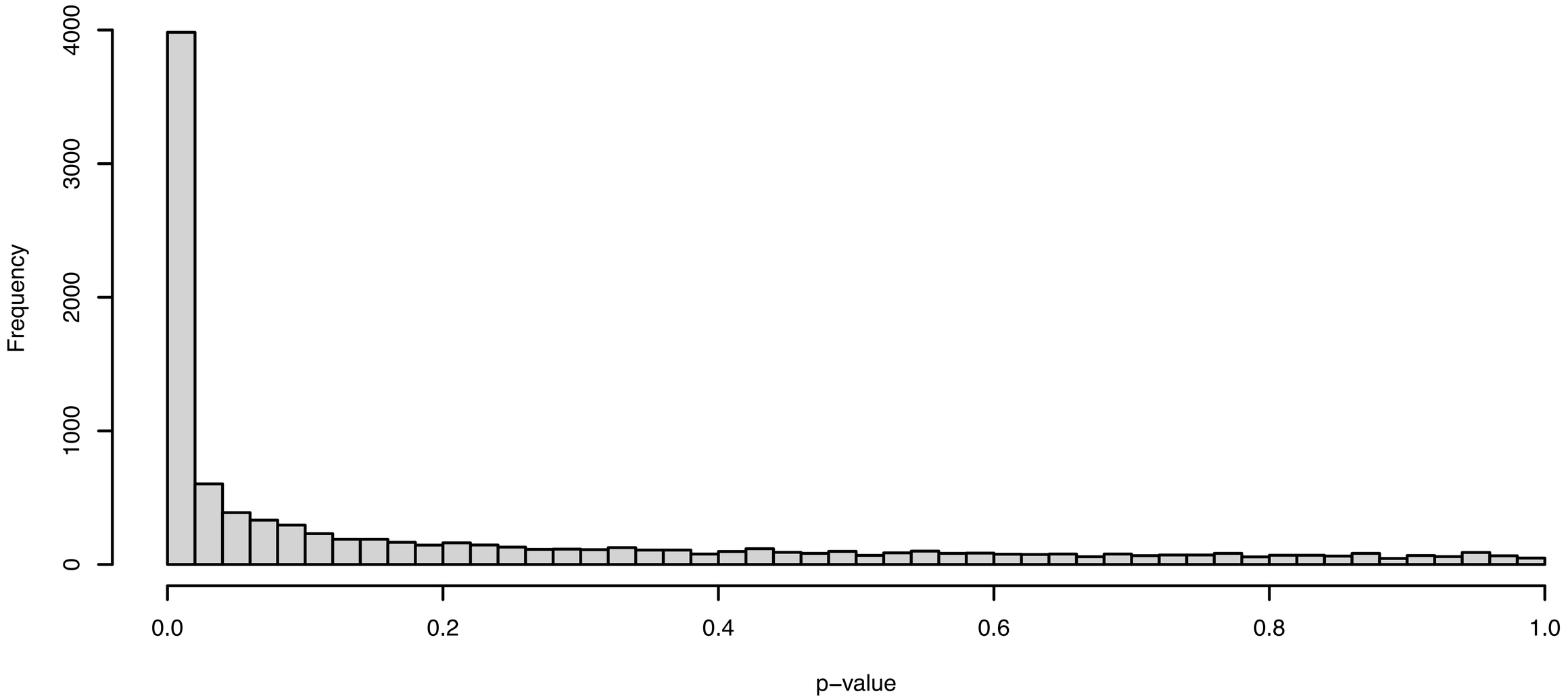


Unmeasured confounding

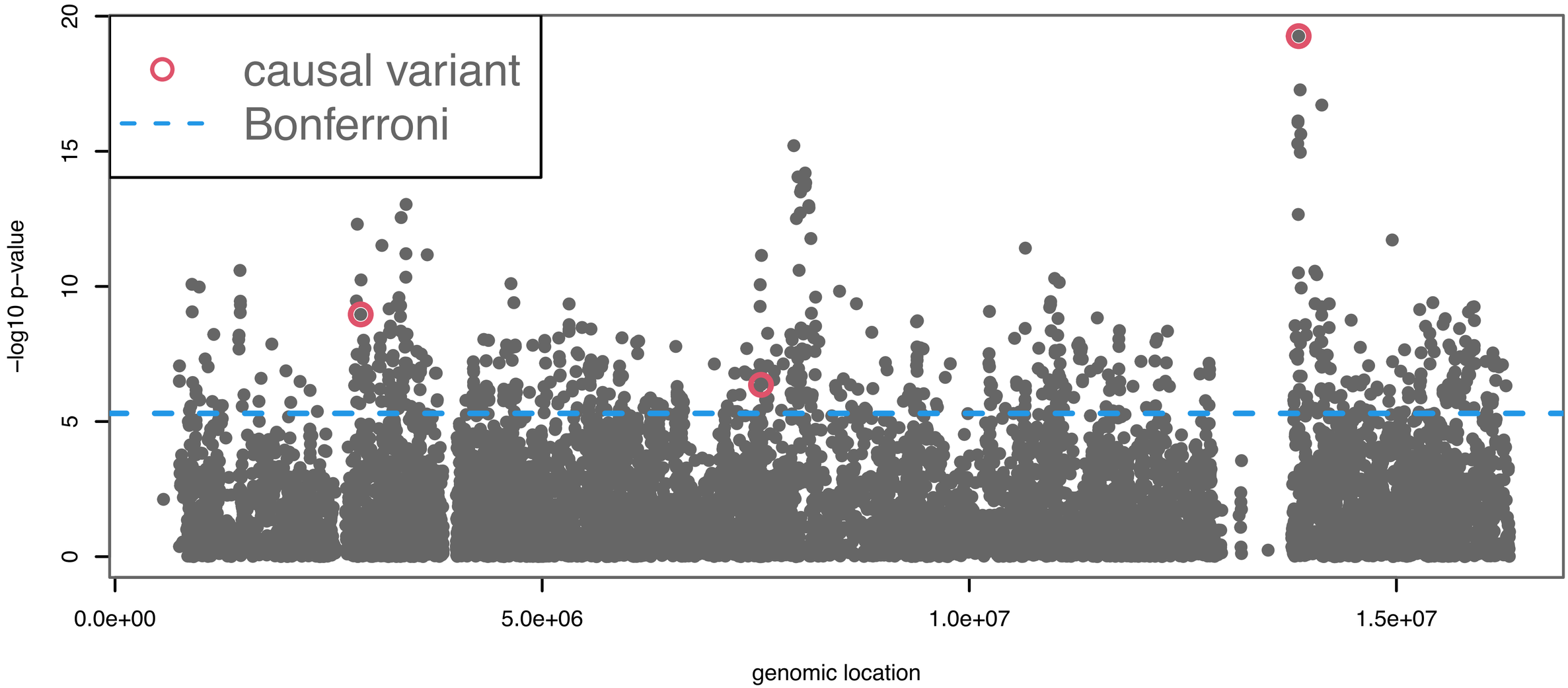
Simply there are so many causal variants

Bulik-Sullivan *et al. Nature Genetics* (2015)

Okay... do we have that many GWA-significant variants?



Okay... do we have that many GWA-significant variants?



How do we know there is an unknown group structure?

A common assumption of GWAS:

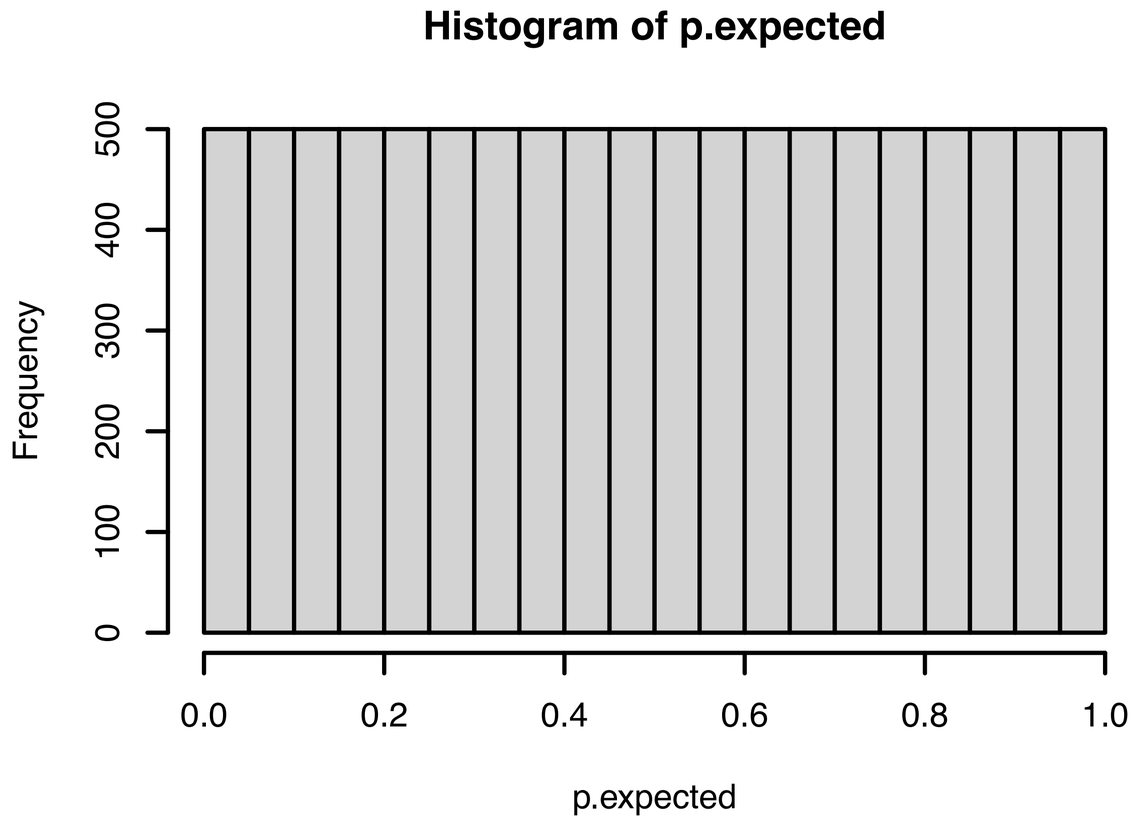
- Only a small portion of variants are significantly associated. Why?
- A majority of p-values under the null will follow Uniform distribution.

```
nn <- length(.gwas$pvalue)
p.expected <- seq(1, nn)/(nn+1)
```

How do we know there is an unknown group structure?

A common assumption of GWAS:

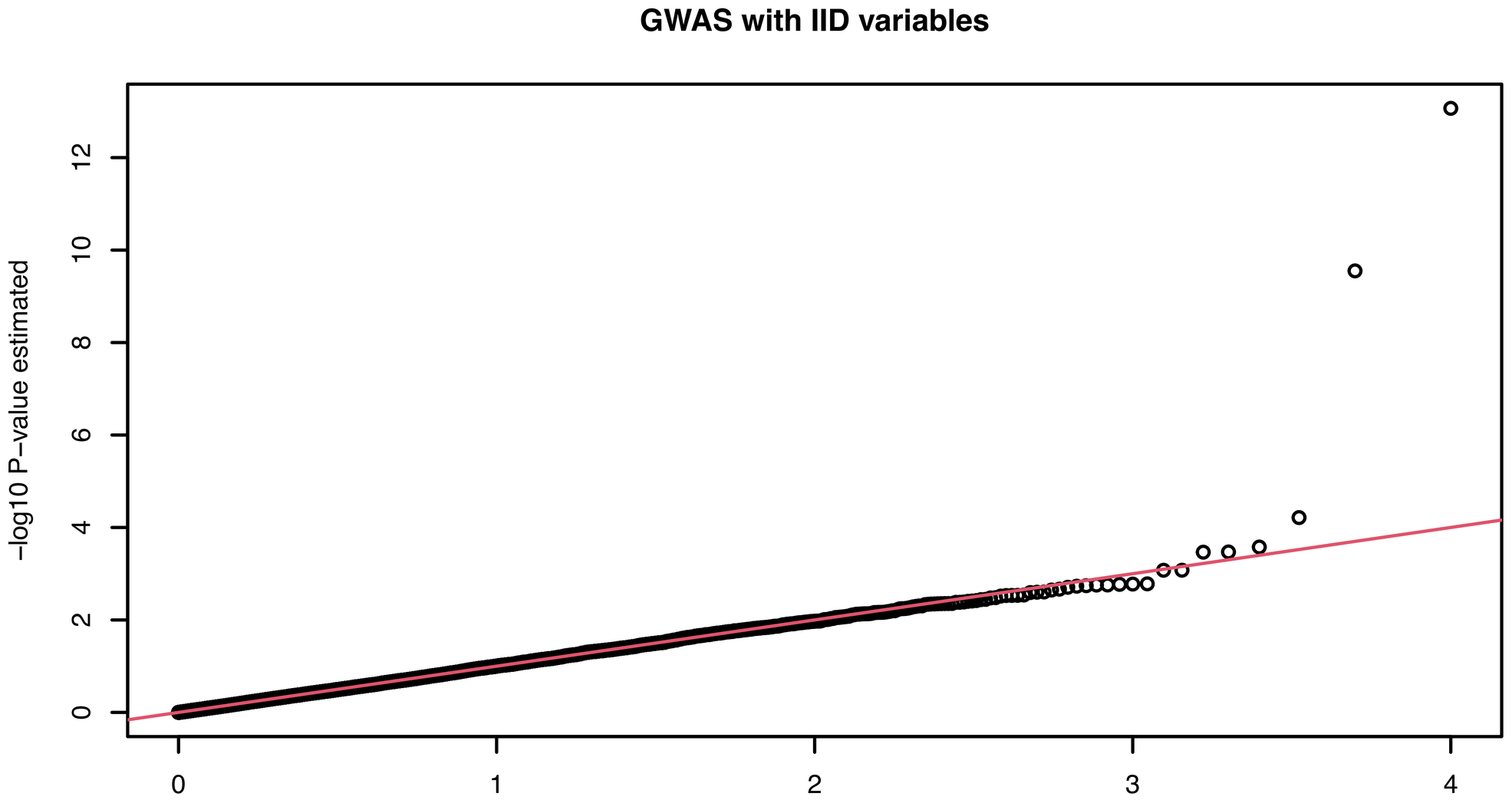
- Only a small portion of variants are significantly associated. Why?
- A majority of p-values under the null will follow Uniform distribution.




```
nn <- length(.gwas$pvalue)
p.expected <- seq(1, nn)/(nn+1)
```

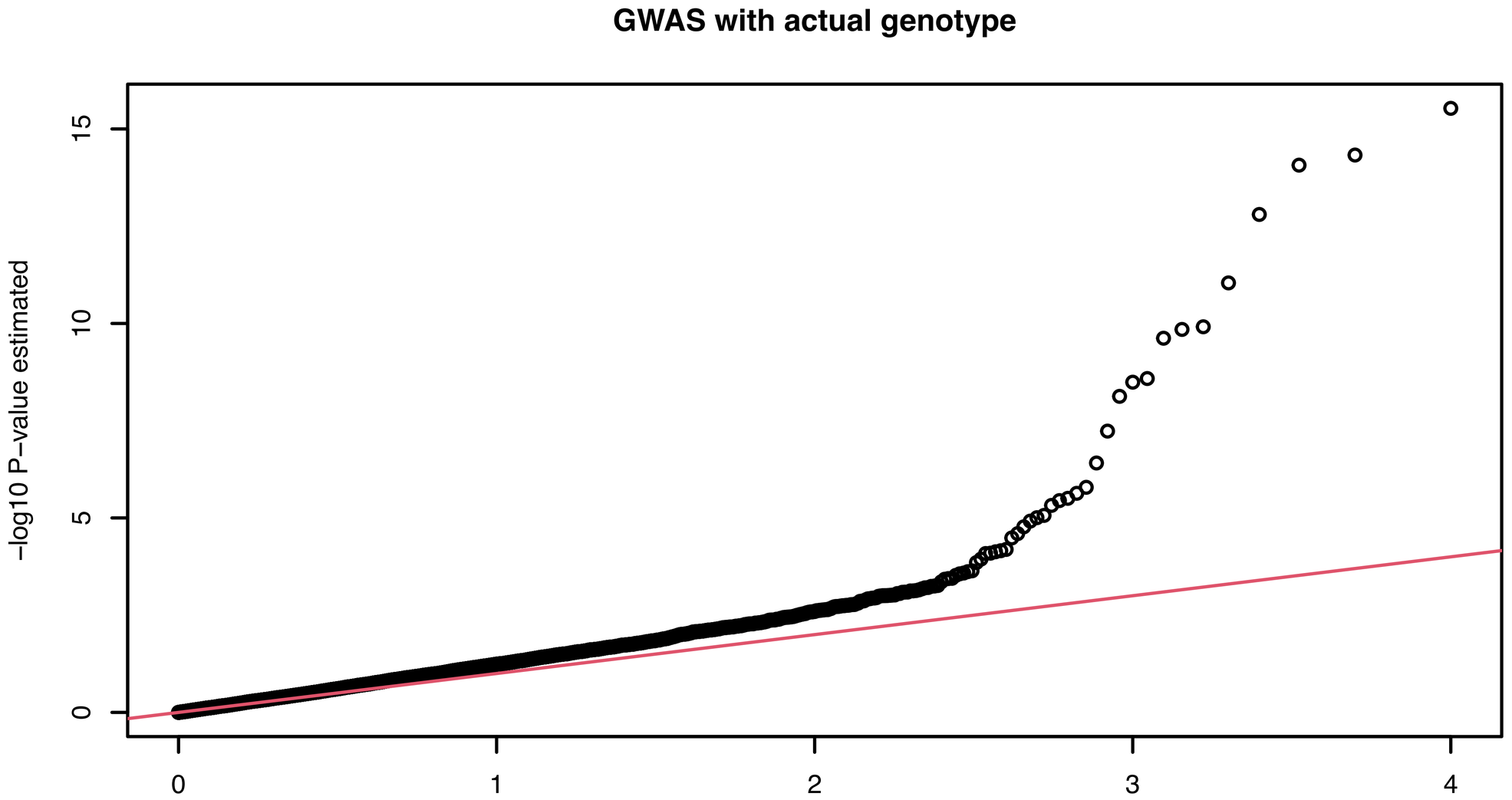
Quantile-quantile plot as a diagnostic tool (log-scale)

Use `qqplot(pvalue1, pvalue2)`



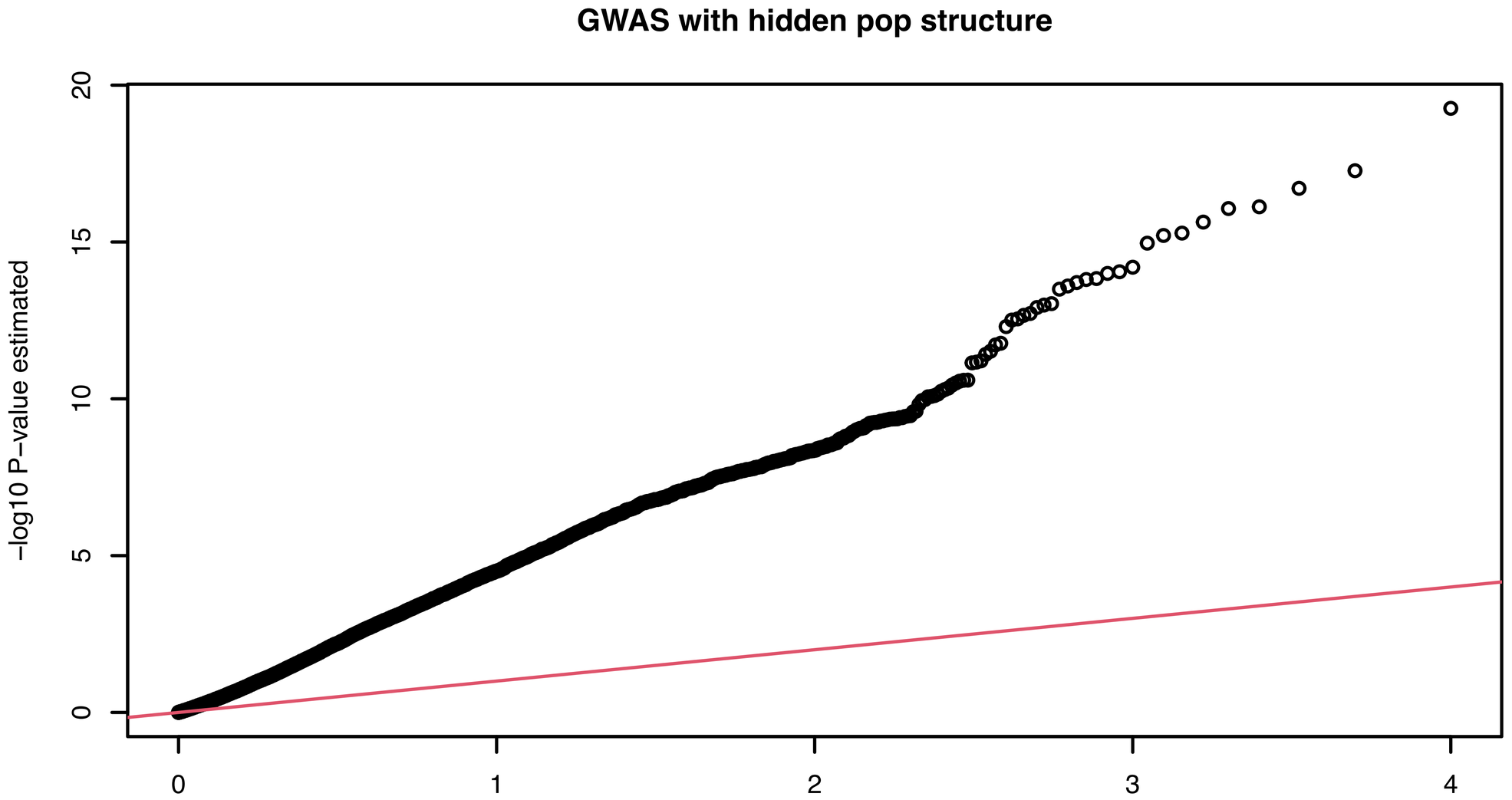
Quantile-quantile plot as a diagnostic tool (log-scale)

Use `qqplot(pvalue1, pvalue2)`



Quantile-quantile plot as a diagnostic tool (log-scale)

Use `qqplot(pvalue1, pvalue2)`



How can we ascertain population/ancestry group structures?

We will discuss more in the unsupervised learning lectures.

```
pca <- prcomp(t(X), rank=3) # rows
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
dim(pca$rotation)
```

```
## [1] 2490    3
```

How can we ascertain population/ancestry group structures?

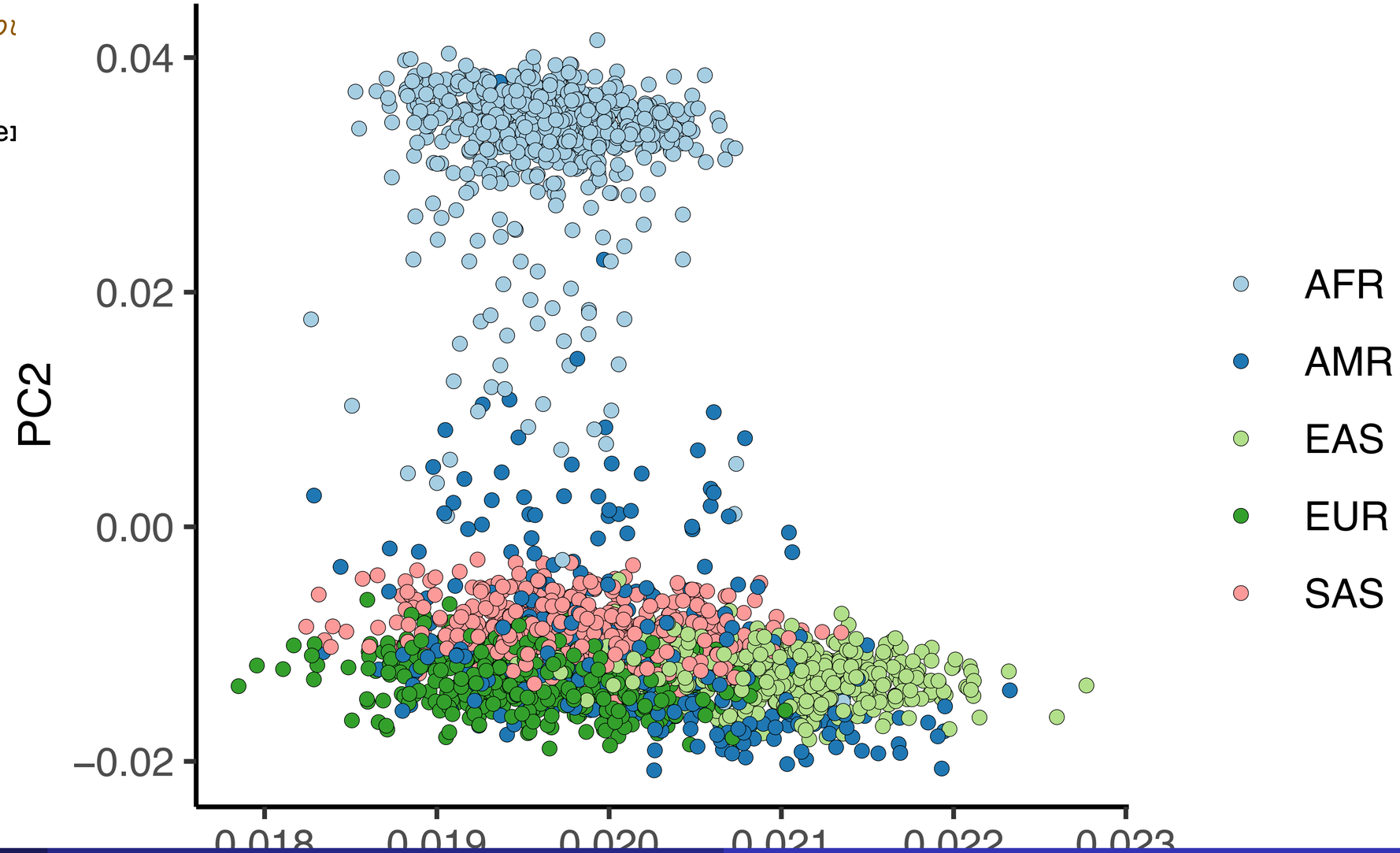
We will discuss more in the unsupervised learning lectures.

```
pca <- prcomp(t(X), rank=3) # rou
names(pca)
```

```
## [1] "sdev"     "rotation" "cer
```

```
dim(pca$rotation)
```

```
## [1] 2490    3
```



How can we ascertain population/ancestry group structures?

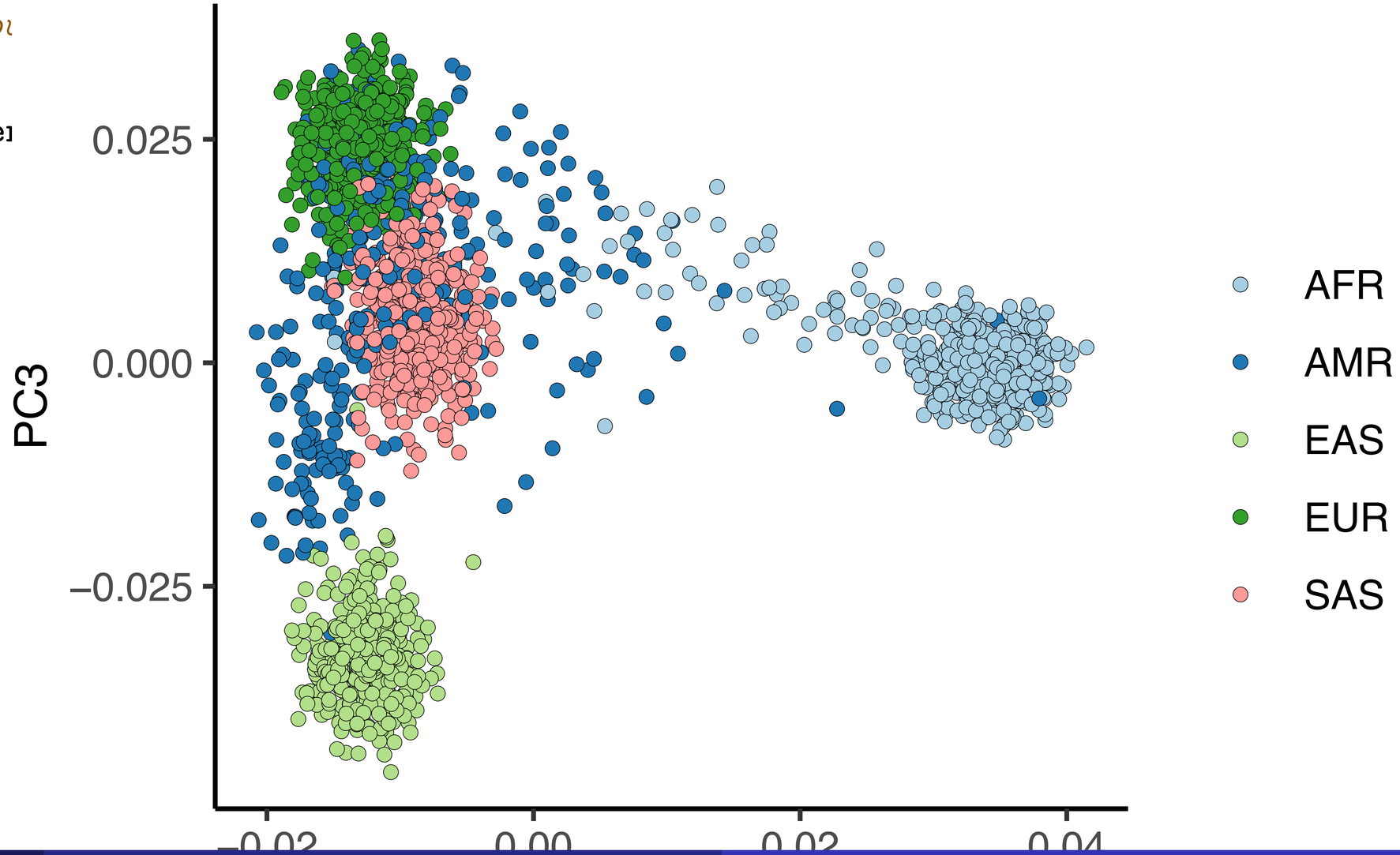
We will discuss more in the unsupervised learning lectures.

```
pca <- prcomp(t(X), rank=3) # rou
names(pca)
```

```
## [1] "sdev"     "rotation" "cer
```

```
dim(pca$rotation)
```

```
## [1] 2490    3
```



Can we quickly fix this? A little bit...

```
.svd <- rsvd::rsvd(X, 10)
Y.adj <- lm(Y ~ .svd$u)$residual
.gwas.adj <- col_cor_pearson(X, Y.adj)
```

Recall:

$$g\left(\underset{\text{phenotype}}{Y}\right) \sim \underbrace{X_k\beta_k}_{\text{true genetic effect}} + \underbrace{\sum_g \mathbf{U}_g\gamma_g}_{\text{estimate this by SVD}} + \epsilon$$

Side note: SVD captures principal components

$$X = UDV^\top$$

Side note: SVD captures principal components

$$X = UDV^\top$$

What is this?

$$\frac{1}{n}X^\top X = \frac{1}{n}VDU^\top UDV^\top = \frac{1}{n}VD^2V^\top$$

variant × variant

Side note: SVD captures principal components

$$X = UDV^\top$$

What is this?

$$\frac{1}{n}X^\top X = \frac{1}{n}VDU^\top UDV^\top = \frac{1}{n}VD^2V^\top$$

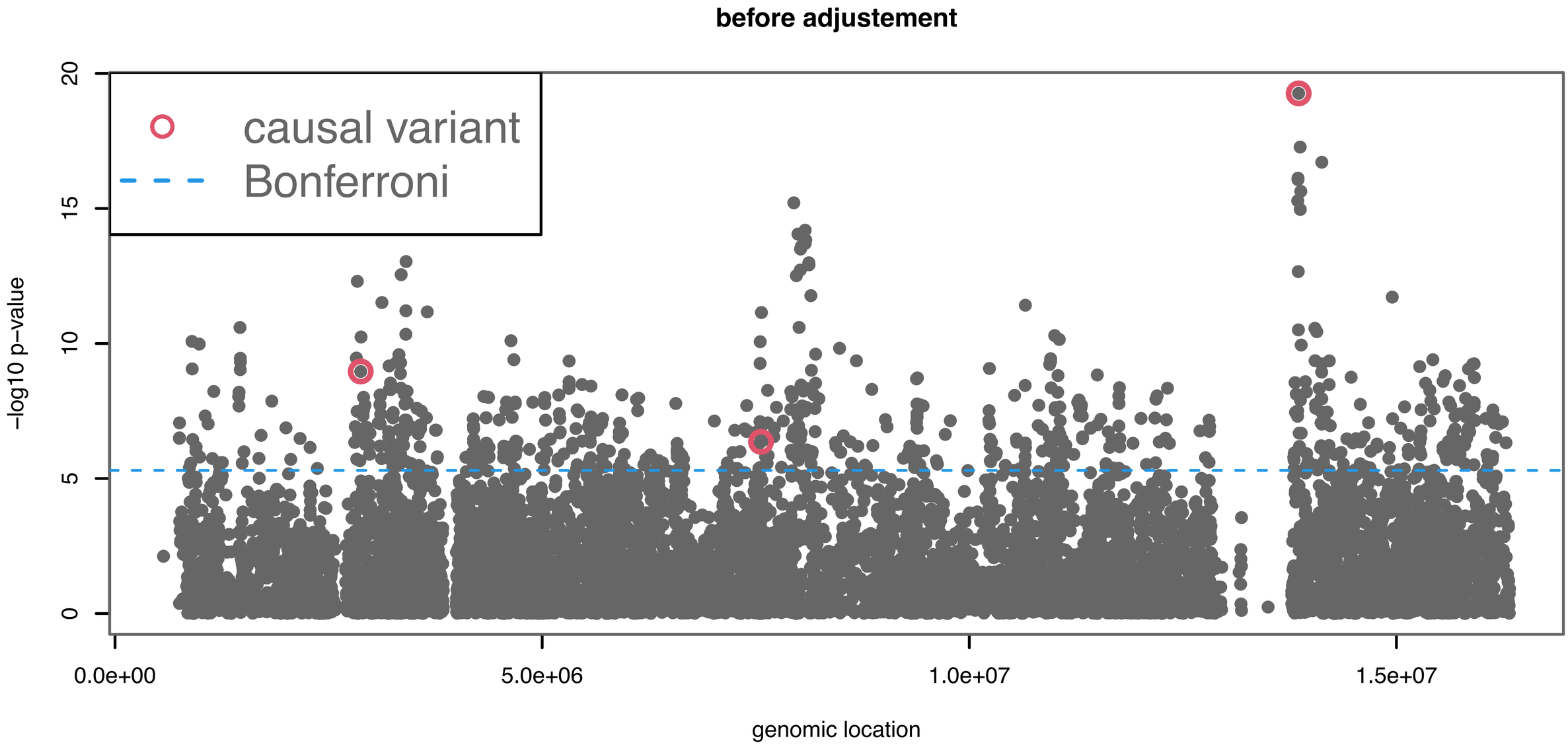
variant × variant

What is this?

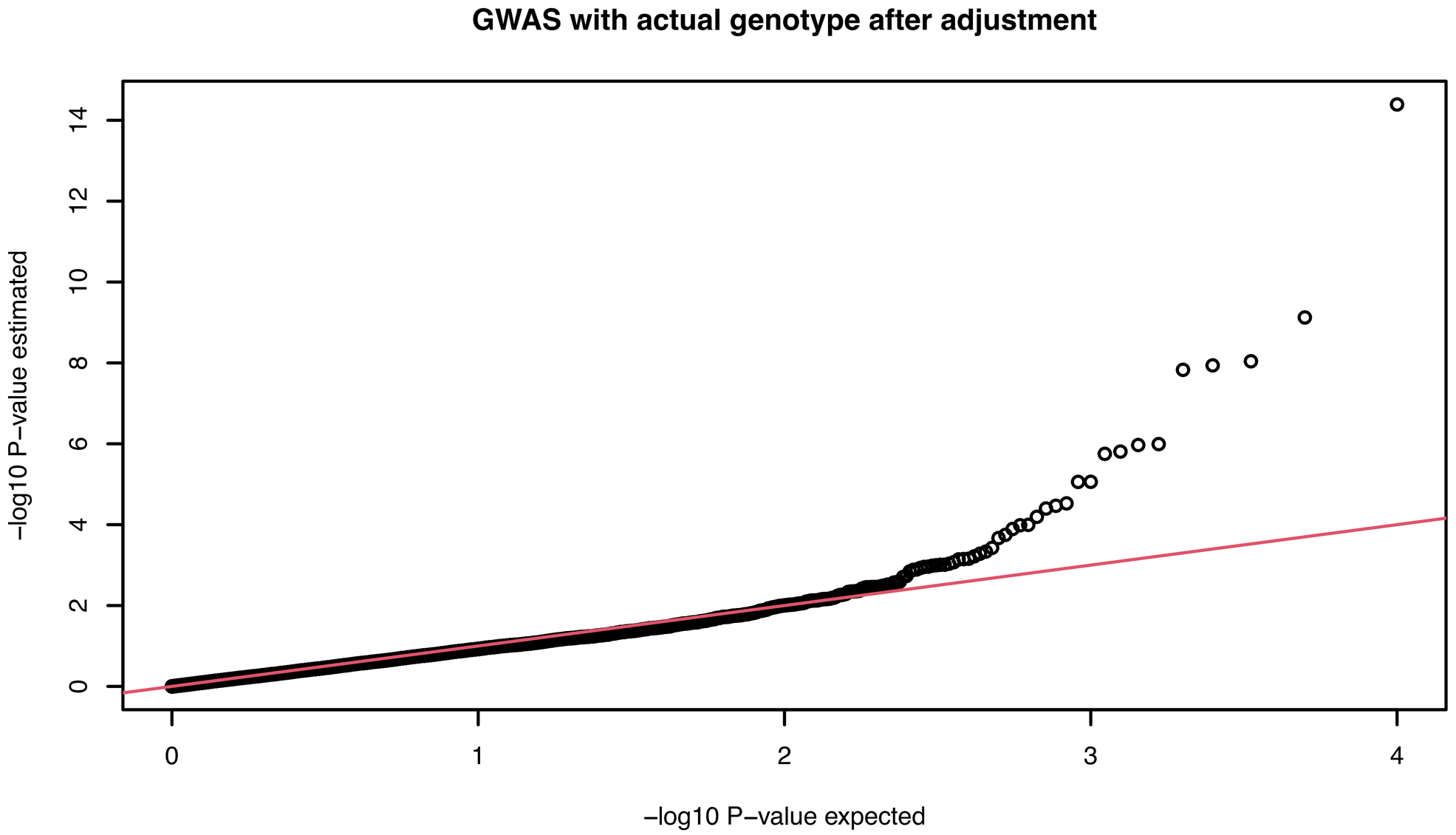
$$\frac{1}{n}XX^\top = \frac{1}{n}UDV^\top VDU^\top = \frac{1}{n}UD^2U^\top$$

sample × sample

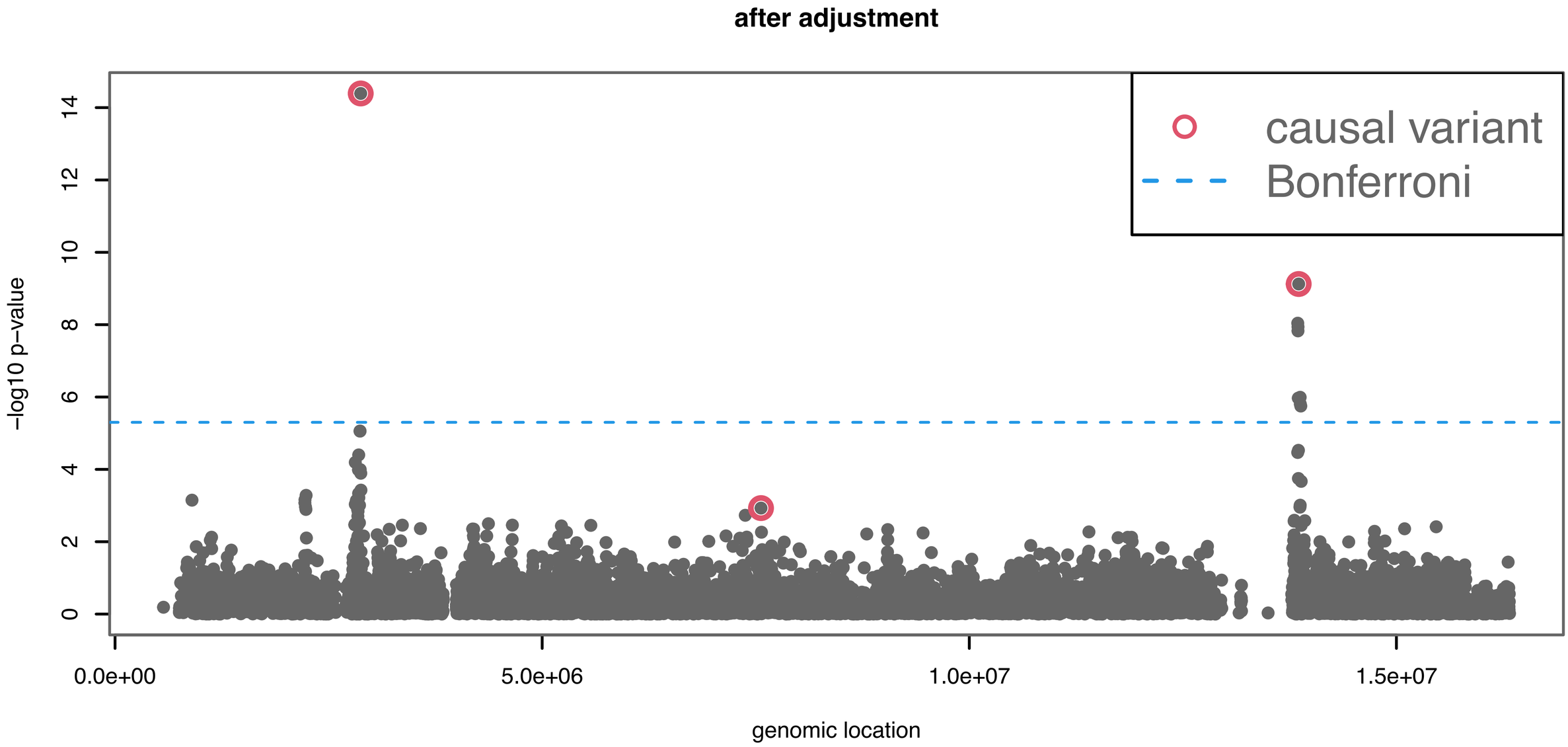
Simply regressing out population PCs can reduce biases



Simply regressing out population PCs can reduce biases



Simply regressing out population PCs can reduce biases



A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \qquad \beta_j \qquad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects

A linear regression model:

$$\mathbf{y} = \mathbf{x}_j \quad \beta_j \quad + \epsilon$$

a fixed genetic effect

What are we missing? Can we assume homo-scedasticity, i.e.,

$$\epsilon \overset{?}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with a random effect:

$$\mathbf{y} = \mathbf{x}_j \, \beta_j + \mathbf{u} + \epsilon$$

fixed; random effect

*N*ote: There is no specific parameterization for this $n \times 1$ random vector $\mathbf{u}$. Now, we assume:

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)

A linear model with population-driven random effects - 2

We want to capture unwanted population, cohort-specific random effects by $n \times 1$ vector $\mathbf{u}$ and **remove** since our **goal** is to estimate the fixed genetic effect of a particular variant j.

$$\mathbf{y} = \underset{\text{goal}}{\mathbf{x}_j \beta_j} + \underset{\text{remove}}{\mathbf{u}} + \epsilon$$

1. Note that $\mathbf{u}$ shouldn't be tied to a particular variant (by definition)
2. Also, the covariation of $\mathbf{u}$ is primarily driven by relatedness among individuals, not the variants.

$$\mathbf{u} \sim \mathcal{N}\left(\mathbf{0}, \tau^2 K\right), \quad K \approx \frac{1}{n} X X^\top$$

A linear mixed effect model (LMM) to test associations while adjusting population structure

We can define a hierarchical model:

$$\mathbf{y}|X, \beta, \mathbf{u}, \sigma \sim \mathcal{N}(X\beta + \mathbf{u}, \sigma^2 I) \quad (1)$$

$$\mathbf{u}|\tau, K \sim \mathcal{N}(\mathbf{0}, \tau^2 K) \quad (2)$$

If we integrate out $\mathbf{u}$,

$$\mathbf{y}|X, \beta \sim \mathcal{N}\left(\mathbf{y} \middle| X\beta, \underbrace{\tau^2 K}_{\text{genetic-relatedness matrix}} + \underbrace{\sigma^2 I}_{\text{irreducible}}\right)$$

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps

Why using LMM instead of regressing out confounding factors?

- It is hard to distinguish between causative vs. confounding effects
- Cumbersome computation required for matrix factorization or other latent variable modelling on a large genotype matrix
- We many not have a large matrix to learn about non-genetic confounders...
- One LMM estimation can substitute multiple matrix factorization steps
- We may have a good idea about relationships induced by random effects!

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + \ldots}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + \ldots}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\ \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + \ldots}_{\text{random effects}})\right)$$

A key research question in LMM: What covariance matrix?

If there were many types of random effects,

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\mathbf{u} + \mathbf{w} + ...}_{\text{random effects}} + \underbrace{\epsilon}_{\text{unknown}}$$

We would need to many covariance matrices:

$$\mathbf{y}|\cdot \sim \mathcal{N}\left(X\beta,\ \sigma^2(I + \underbrace{\delta_u K_u + \delta_w K_w + ...}_{\text{random effects}})\right)$$

If we only care about variance decomposition $\beta_j \sim \mathcal{N}(0, \tau)$:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0},\ \sigma^2\left(\underbrace{\frac{\sigma^2_{\text{genetic}}}{n} XX^\top}_{\text{observed genetic}} + I + \underbrace{\delta_u K_u + \delta_w K_w + ...}_{\text{random effects}}\right)\right)$$

Should we worry about "over-fitting" in LMM?

An equivalent question for PCA-based confounder adjustment:

How many PCs to adjust in GWAS?
Can we include a candidate SNP in the GRM K matrix?

Should we worry about "over-fitting" in LMM?

An equivalent question for PCA-based confounder adjustment:

How many PCs to adjust in GWAS?
Can we include a candidate SNP in the GRM K matrix?

For each chromosome $c \in \{1, \ldots, 22, \mathrm{X}, \mathrm{Y}\}$, build a leave-one-chromosome-out (LOCO) kinship matrix, say K_{-c}:

$$\mathcal{N}\left(\mathbf{y} | \mathbf{x}_j \beta_j, \sigma^2(\delta K_{-c} + I)\right)$$

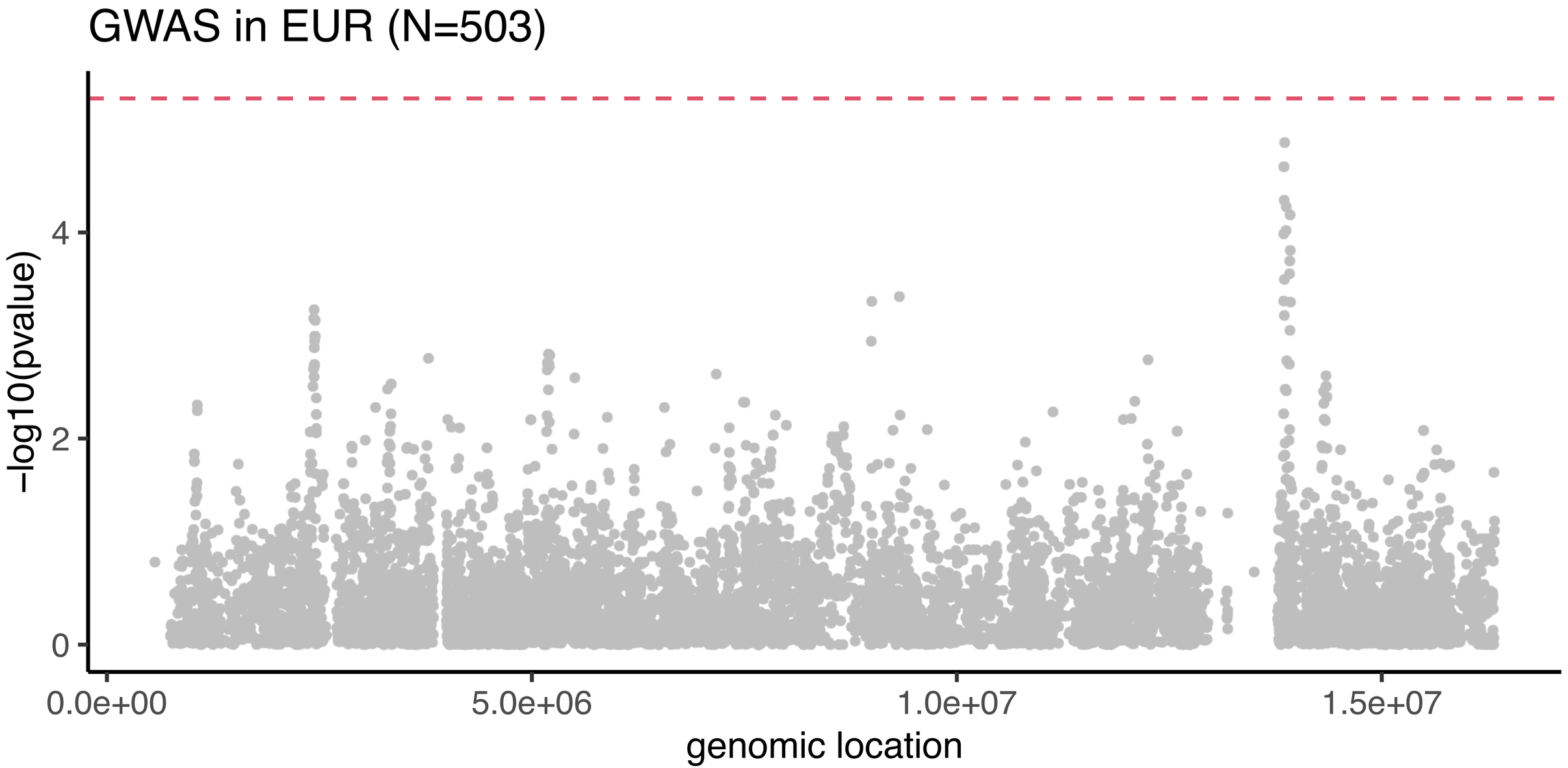
Yang, *et al.*, *Nature Genetics* (2014)

Tucker, Price, Berger, *Genetics* (2014)

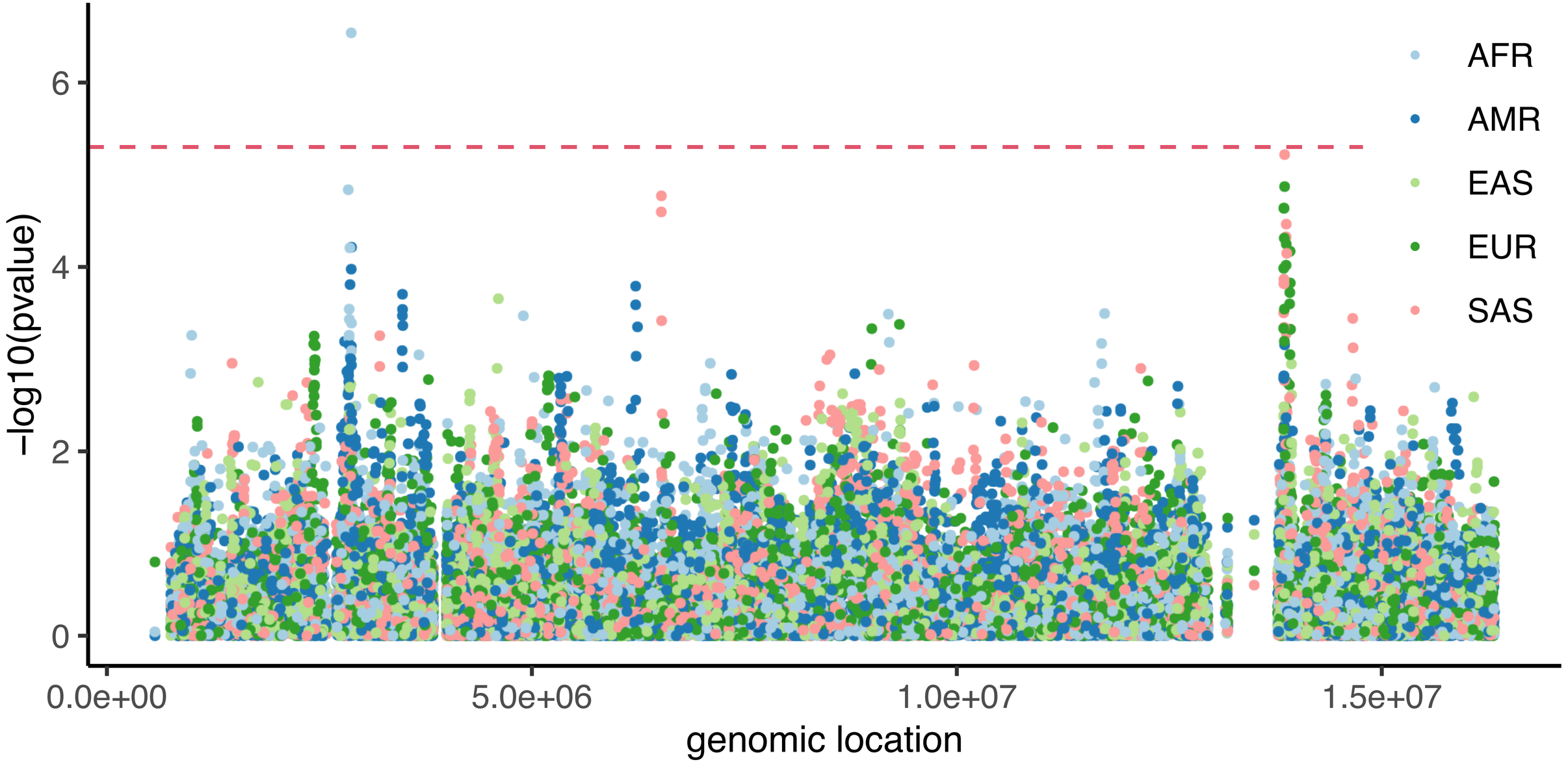
Today's lecture

1. Mapping disease-specific locations in Genome
2. How GWAS can be interpreted wrongfully
3. Combining evidence from multiple studies
4. Appendix

Suppose we have GWAS summary statistics within each AG



Suppose we have GWAS summary statistics within each AG



Meta analysis: How can we combine multiple studies?

For each population/study k:

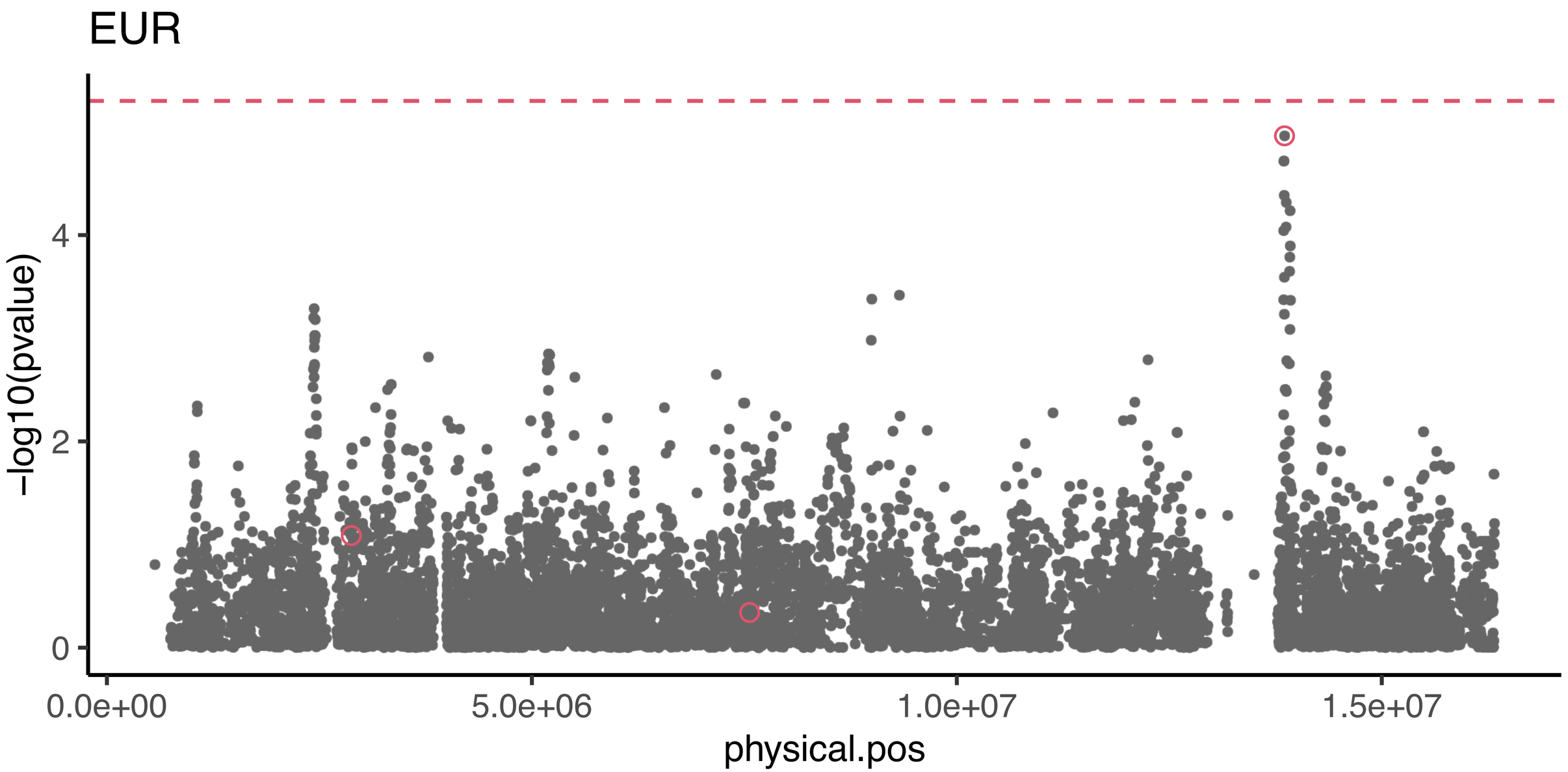
- $\hat{\beta}_{jk}$: effect size for a variant j on a study k
- $s_{jk} = \mathsf{SE}(\beta_{jk})$: standard error for a variant j on a study k
- $w_{jk} = 1/s_{jk}^2$: inverse variance

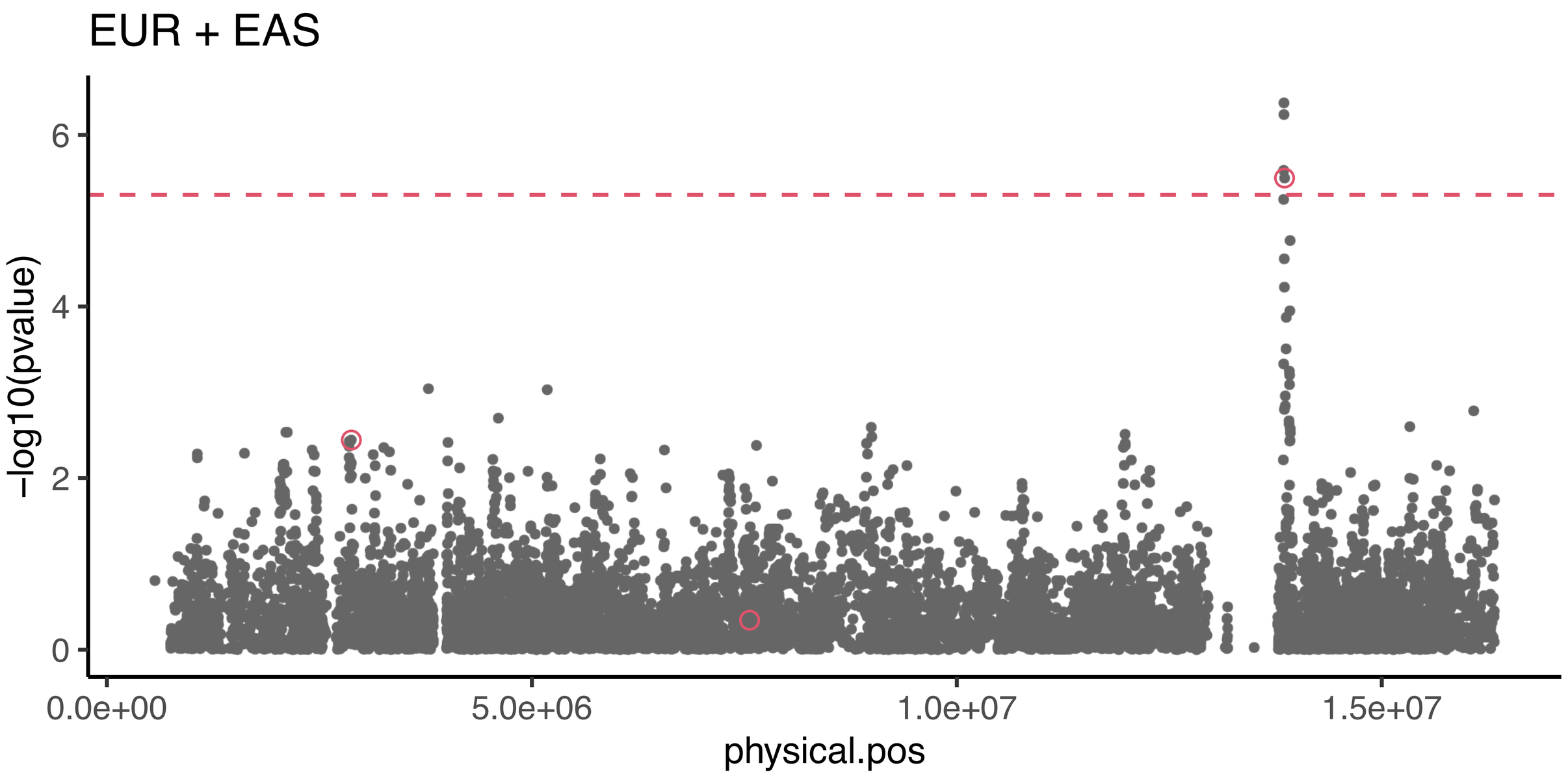
Meta analysis by inverse variance weighting

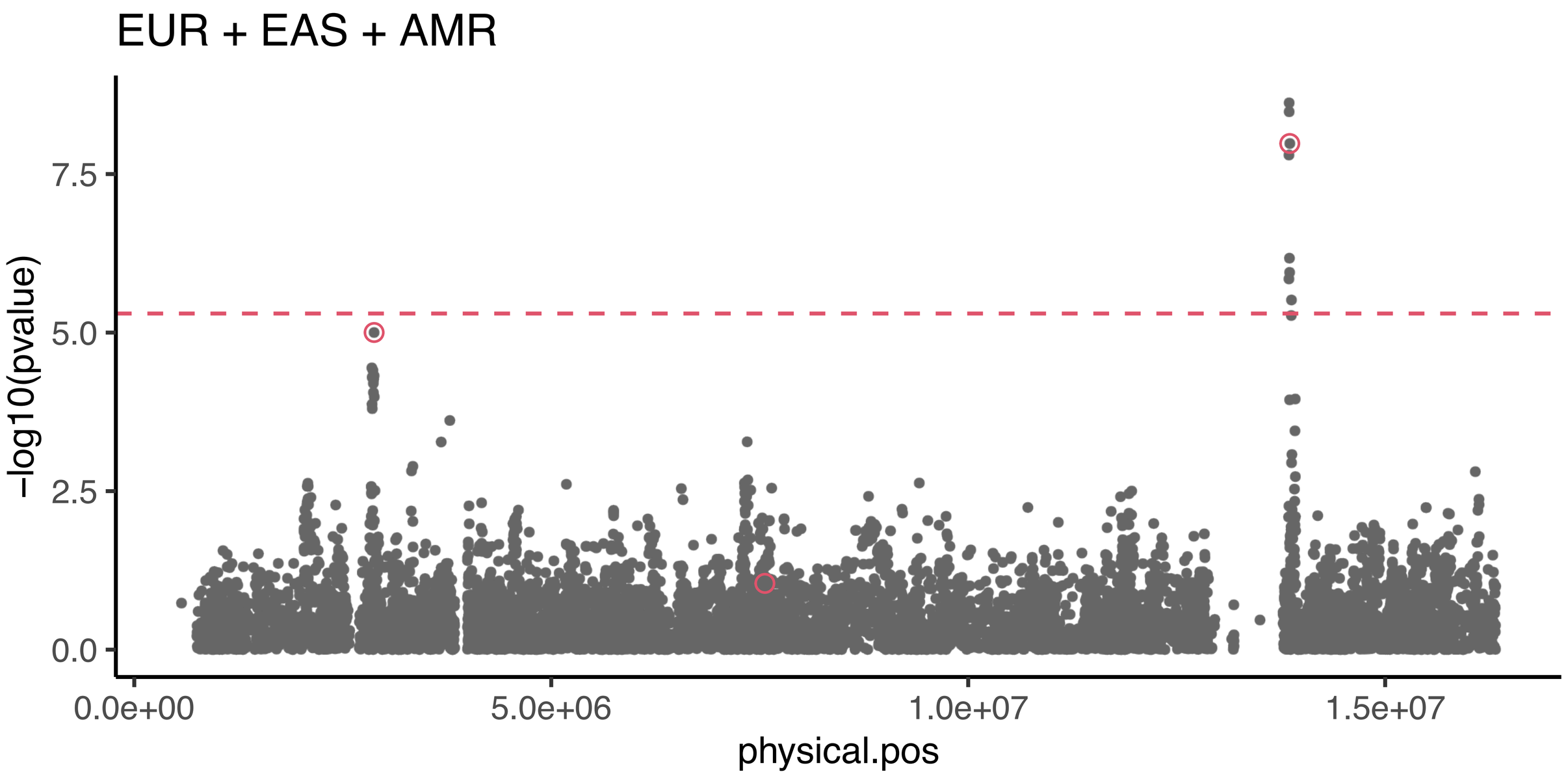
$$\hat{\beta}_j = \frac{\sum_k w_k \hat{\beta}_{jk}}{\sum_k w_k}, \qquad \mathsf{SE}(\hat{\beta}_j) = \frac{1}{\sqrt{\sum_k w_k}}$$

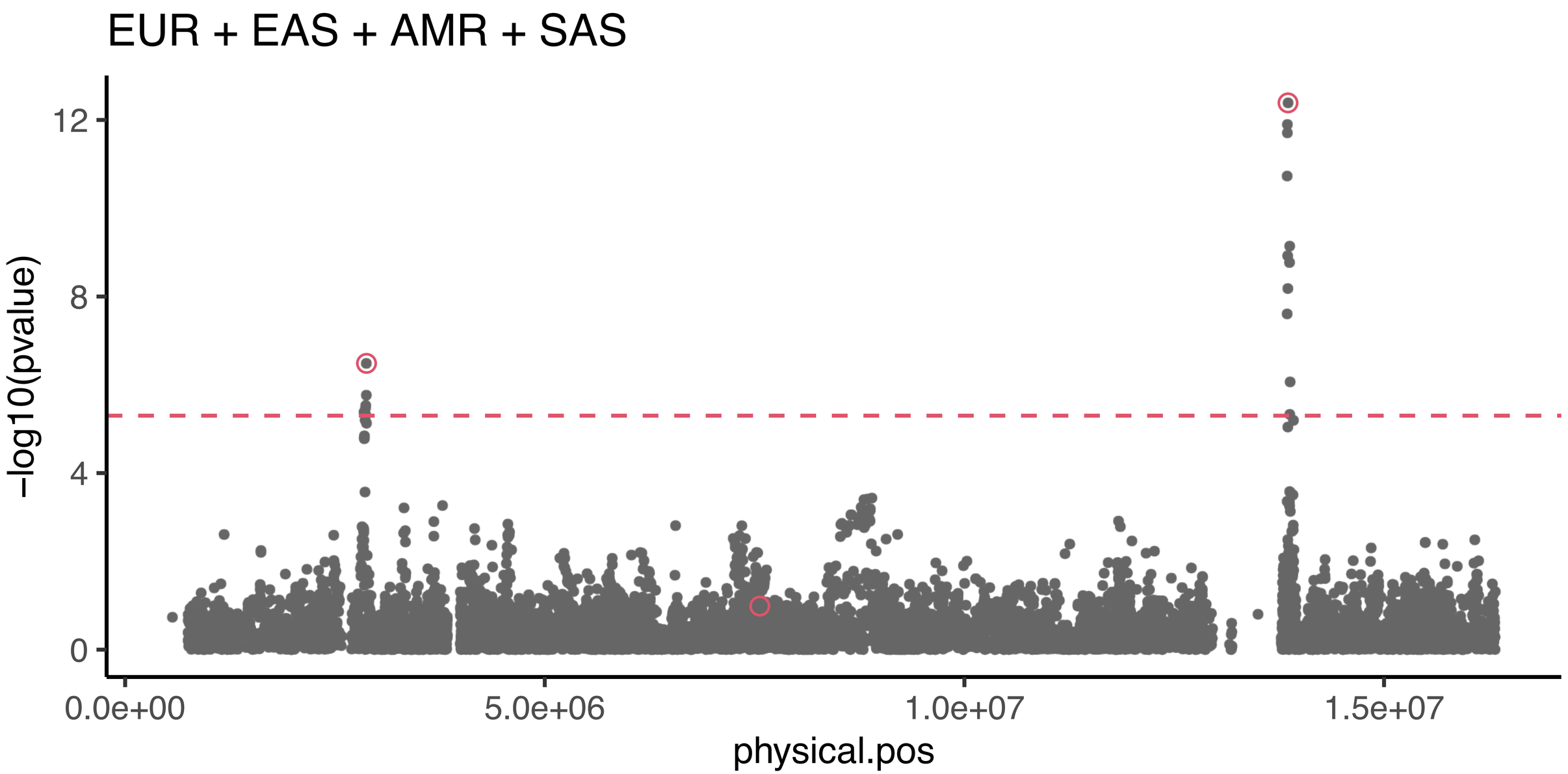
Meta analysis can boos statistical power

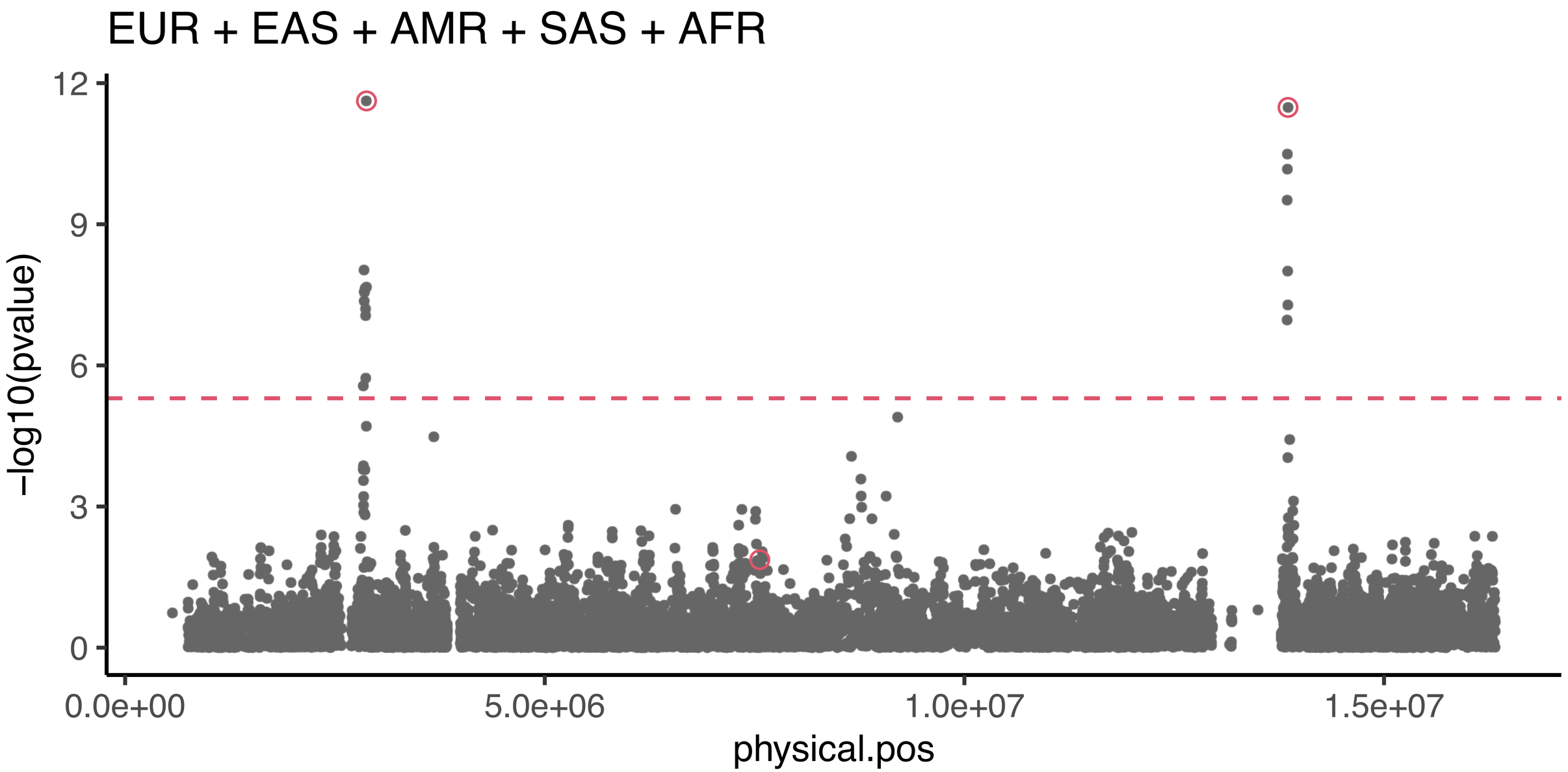
```
gwas.dt[, w := 1/(`stderr`^2 + 1e-16)]

meta.dt <-
    gwas.dt[!is.na(w),
            .(`mean` = sum(`mean.diff` * w)/sum(w),
              `se` = 1/sqrt(sum(w))),
            by = .(physical.pos)]

meta.dt[, `z` := `mean`/`se`]
meta.dt[, `pvalue` := 2*pnorm(abs(`z`), lower.tail=F)]
```


EUR
−log10(pvalue)
0
2
4
0.0e+00
5.0e+06
1.0e+07
1.5e+07
physical.pos


EUR + EAS
−log10(pvalue)
0
2
4
6
0.0e+00
5.0e+06
1.0e+07
1.5e+07
physical.pos


EUR + EAS + AMR
−log10(pvalue)
7.5
5.0
2.5
0.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
physical.pos


EUR + EAS + AMR + SAS
−log10(pvalue)
12
8
4
0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
physical.pos


EUR + EAS + AMR + SAS + AFR
−log10(pvalue)
12
9
6
3
0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
physical.pos

Common pitfalls of GWAS meta analysis

- Multiple GWAS results on similar yet slightly different phenotypes
- Different reference human genome
- Different array platforms (could cover different variants)
- Shared individuals
  - Some individuals can be shared among different studies
  - In UK Biobank data, many non-cancer individuals can be also non-neuro-degenerative disorder individuals.

Today's lecture

1. Mapping disease-specific locations in Genome
2. How GWAS can be interpreted wrongfully
3. Combining evidence from multiple studies
4. Appendix

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

We can resolve maximum likelihood estimate of the parameters, β, τ, σ,

$$\max \log \mathcal{N}(\mathbf{y} \,|\, X\beta, \sigma_2 \left(\delta K + I\right))$$

where $\tau^2 = \delta\sigma^2$.

We need to deal with this unfriendly form of likelihood:

$$-\frac{1}{2}\left(n \log(2\pi\sigma^2) + \log|I + \delta K| + \frac{1}{\sigma^2}\left[\mathbf{y} - X\beta\right]^\top \left(I + \delta K\right)^{-1}\left[\mathbf{y} - X\beta\right]\right)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underbrace{U^\top \mathbf{y}}_{\text{projected output}} \sim \mathcal{N}\Big(\underbrace{U^\top X}_{\text{projected genotype}} \beta, \sigma^2 U^\top (I + \delta K) U \Big)$$

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)

FaST Linear Mixed Model (Lippert *et al.* 2011)

Instead, we can transform the underlying distribution using spectral decomposition of the genetic-relatedness matrix (GRM),

$K = USU^\top$ where $U^\top U = I$, and S is a diagonal matrix.

$$\underset{\text{projected output}}{U^\top \mathbf{y}} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\beta, \sigma^2 U^\top (I + \delta K)U\right)$$

$$\text{(by the affine transformation)} \sim \mathcal{N}\left(\underset{\text{projected genotype}}{U^\top X}\beta, \sigma^2(I + \delta \underset{\text{diagonal matrix}}{S})\right)$$

- We can find β by weighted least square
- We can find σ^2 and δ by fixing β

Lippert, Listgarten, .. , Heckerman, *Nature Methods* (2011)