



**Association
tests**

**from the
Health Data Science
Sandbox**



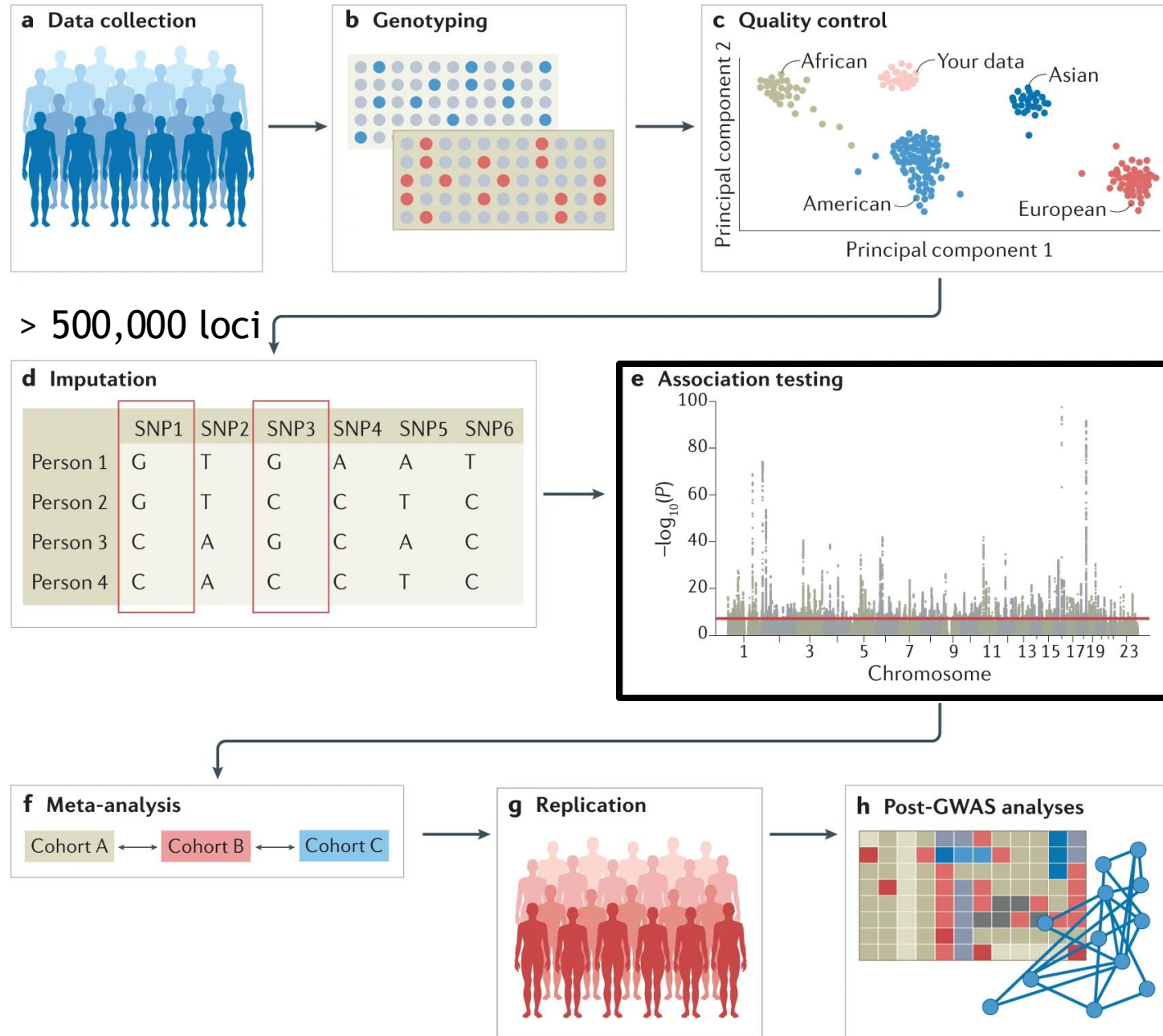
Alba Refoyo Martinez, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN





Types of association testing

Assessing the results:

- Make sure things went okay
- Identify associated SNPs.

GWAS with the Genomics Sandbox



Today's topics

Association tests

- Single SNP tests
 - Case-control (logistic model)
 - Quantitative traits (linear model)
- GWAS
 - Imputation
 - Interpreting GWAS results
 - Effect sizes & P-values thresholds
 - Visualization
 - Tools



Single SNP tests

- Test to identify a genetic variant that affect a trait
- Evidence from previous studies
- Monogenic traits
- Example: **PCSK9**

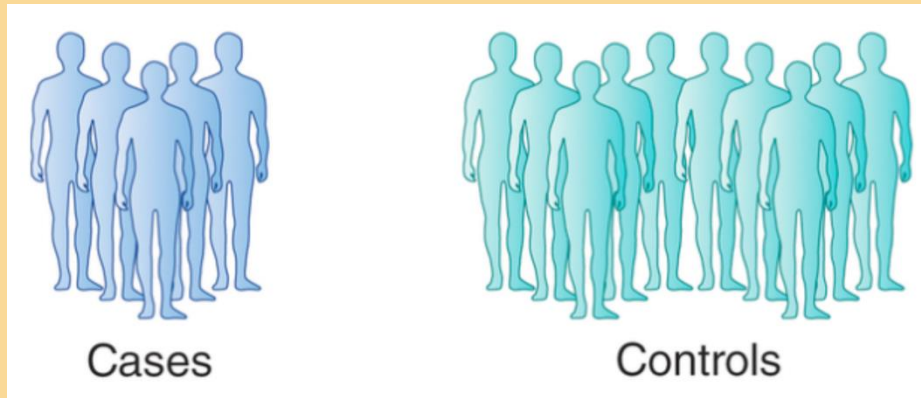
GWAS

- Scan the genome for variants that affect a given trait
- > 500,000 SNPS (no prior evidence required)
- Likely not to have the casual SNP
- Polygenic traits

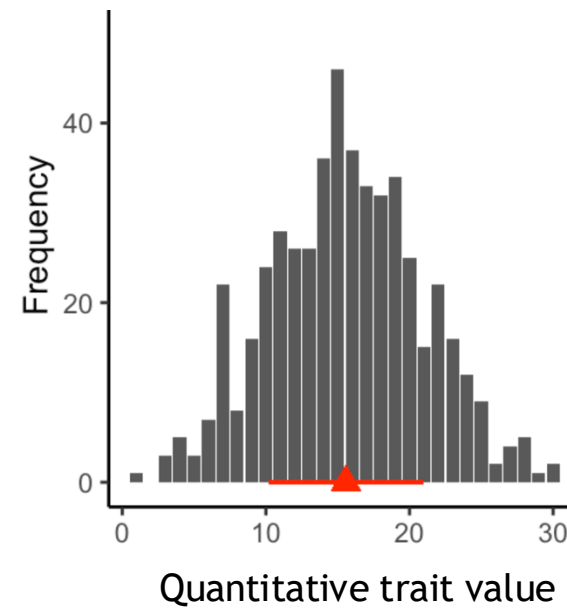


Single-SNP tests

Disease traits



Quantitative traits



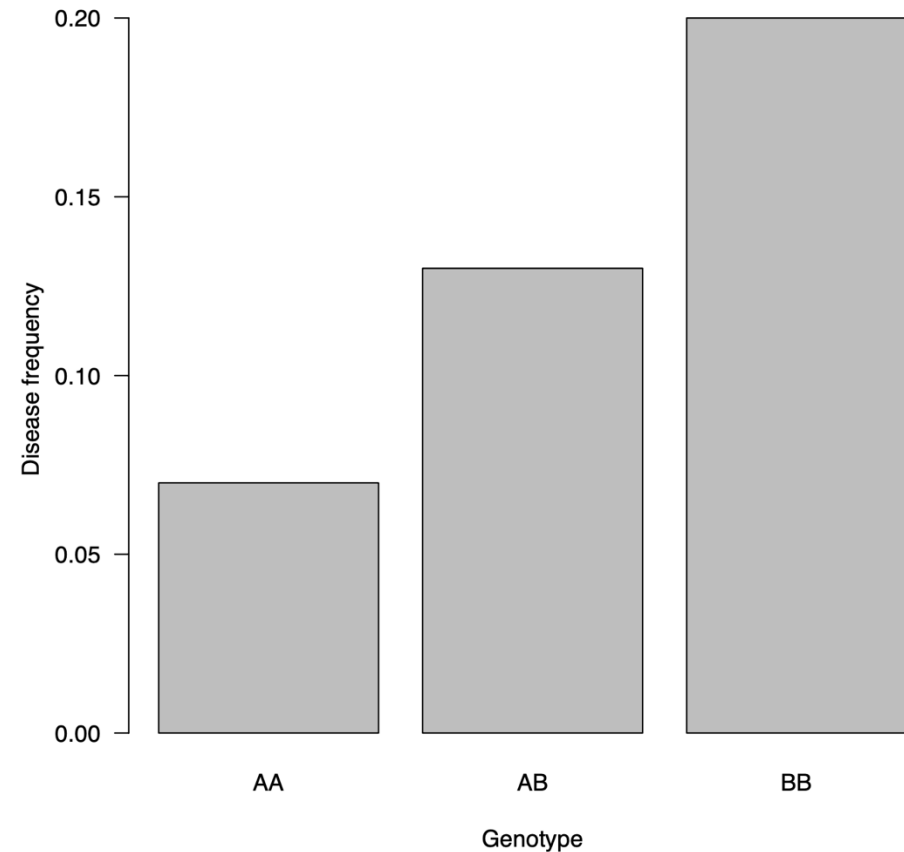
Association testing

How do we test if a genetic variant potentially has an effect on a disease?

Idea: Statistical test to evaluate the association between the variant and disease status (e.g. case/control)

If the variant affects the trait, we expect differences in disease frequencies across genotypes (variant B)

Genotype AA are less likely to have the disease



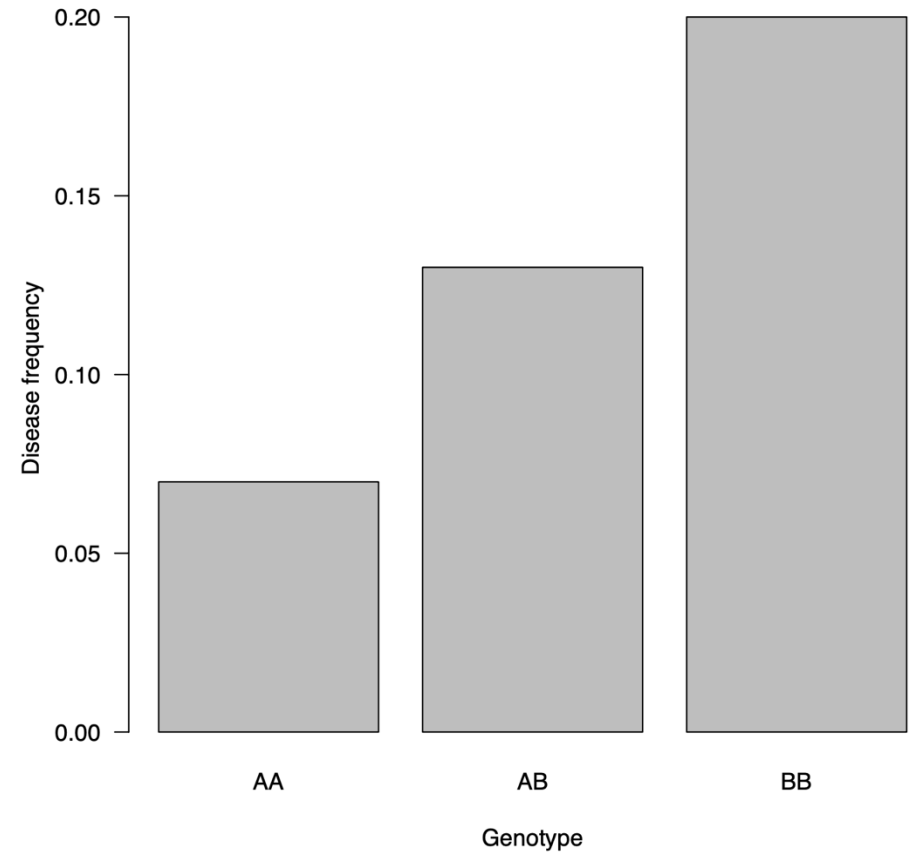
Association testing

How do we test if a genetic variant potentially has an effect on a disease?

Idea: Statistical test to evaluate the association between the variant and disease status (e.g. case/control)

If the variant affects the trait, we expect differences in disease frequencies across genotypes

Approach: test null hypothesis, H_0 , of no association between the variant and disease (i.e., they are independent)



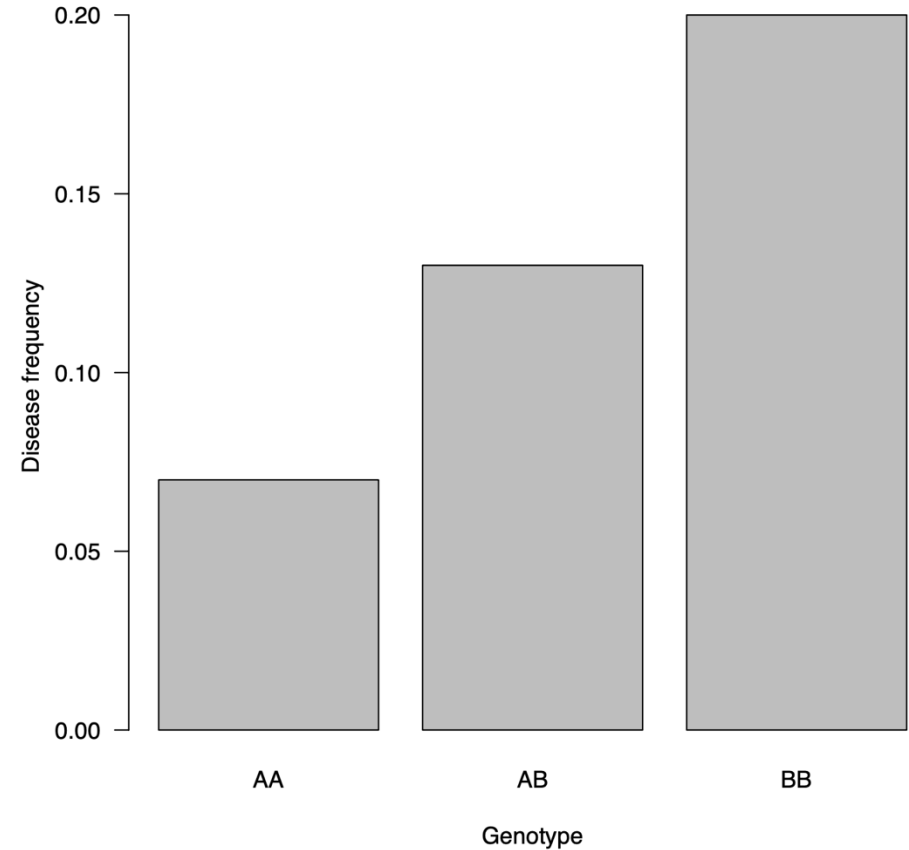
Association testing

Probability of disease given phenotype:

- $P(D|AA) = 0.07$,
- $P(D|AB) = 0.13$,
- $P(D|BB) = 0.20$

We can model the probability of disease of an individual

$$P(D|g)$$



Testing for association between disease and genotype

We can use a logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{AB}x_{AB} + \beta_{BB}x_{BB}$$

where β_s are regression coefficients (effect sizes),
and the x_s are determined by the genotype of an individual

Genotypes	x_{AB}	x_{BB}
AA	0	0
AB	1	0
BB	0	1

We can rephrase the logistic regression model into a probability

$$p = \frac{\exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}{1 + \exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}$$



Example

What is the probability of disease for the different genotypes?

$$p = \frac{\exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}{1 + \exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}$$

Genotypes	X_{AB}	X_{BB}
AA	0	0
AB	1	0
BB	0	1

For an individual with genotype **AA**?

For an individual with genotype **AB**?



Example

What is the probability of disease for the different genotypes?

$$p = \frac{\exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}{1 + \exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}$$

Genotypes	x_{AB}	x_{BB}
AA	0	0
AB	1	0
BB	0	1

For an individual with genotype **AA**?

$$p(D|AA) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

For an individual with genotype **AB**?

$$p(D|AB) = \frac{\exp(\beta_0 + \beta_{AB})}{1 + \exp(\beta_0 + \beta_{AB})}$$



Association tests (full genotype model)

$$p = \frac{\exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}{1 + \exp(\beta_0 + \beta_{AB} x_{AB} + \beta_{BB} x_{BB})}$$

If there is no association, then the probability of disease is the same regardless of the genotype $p(D|AB) = p(D|AA) = p(D|BB)$

The logistic regression null model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0, \text{ assuming } \beta_{AB} = \beta_{BB} = 0.$$

The **likelihood ratio test (LRT)** compares the likelihood of the data under these two models (2 df) (e.g. ANOVA test in R)

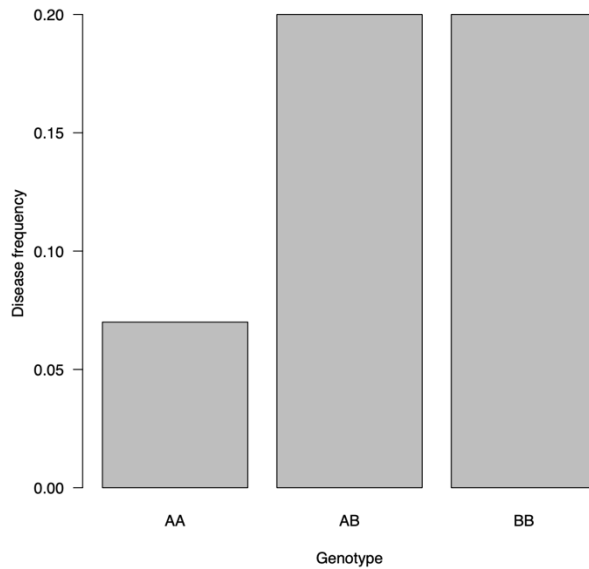


Simpler logistic regression models

Assuming recessive, dominant or additive genetic effects...

Dominant

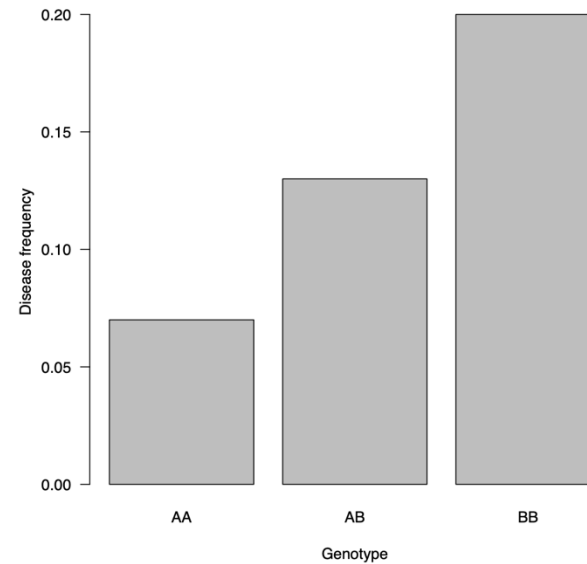
Genotypes	X_D
AA	0
AB	1
BB	1



$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_D x_D$$

Additive

Genotypes	X_A
AA	0
AB	1
BB	2



$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_A x_A$$



Model assumptions, degrees of freedom and power

Which single-SNP test is the most appropriate to use?

- If a sub-model is correctly specified, a test with fewer degrees of freedom has greater power than the full genotype test (avoids unnecessary parameters).
- However, if the model is severely miss-specified, the test may lose power.
- A slightly miss-specified model with fewer degrees of freedom can still outperform a fully correct test with more degrees of freedom in terms of power (avoids fitting noise or random fluctuations in the data).



Generalised logistic model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

- The regression coefficients, β_s represent the effect sizes while x_s are covariates
- Additional covariates, both discrete and continuous, can be included to adjust the model for **confounding factors** such as **sex, population structure, or batch effect**.

The design matrix consists of columns, x_s , which for single-SNPs:

Genotypes	Additive	Dominant	Recessive	Full	
	X_A	X_D	X_D	X_{AB}	X_{BB}
AA	0	0	1	0	0
AB	1	0	0	1	0
BB	0	1	0	0	1

Different nested models can be compared using ANOVA.



Effect sizes - odds ratio

How many times higher the **odds** of disease are for exposed individuals compared to unexposed individuals?

	Cases	Controls
Exposed (aa)	100	100
Not exposed (AA or Aa)	400	3600

$$OR = \frac{ODD_{Exposed}}{ODD_{Not\ exposed}} = \frac{\frac{P(Case|Exposed)}{P(Control|Exposed)}}{\frac{P(Case|Not\ exposed)}{P(Control|Not\ exposed)}} = \frac{\frac{100/200}{100/200}}{\frac{400/4000}{3600/4000}} = 9$$

The definition of "exposed" depends on the model. For example, in a recessive model, an individual is considered exposed if their genotype is aa.



Odds ratio from the logistic regression

For example, in the dominant model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_D x_D$$

where β_D is the effect size

The OR from the model is

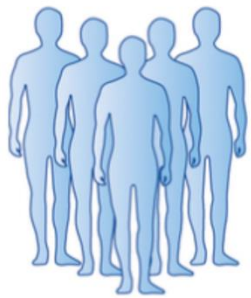
$$OR = \frac{ODD_{AB/BB}}{ODD_{AA}} = \frac{\frac{p_{AB/BB}}{1 - p_{AB/BB}}}{\frac{p_{AA}}{1 - p_{AA}}} = \frac{\exp(\beta_0 + \beta_{DD})}{\exp(\beta_0)} = \mathbf{\exp(\beta_D)}$$

Genotypes	X_D
AA	0
AB	1
BB	1

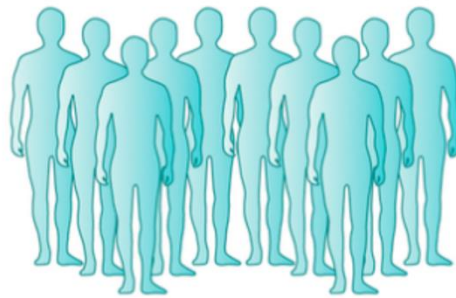


Single-SNP tests

Disease traits

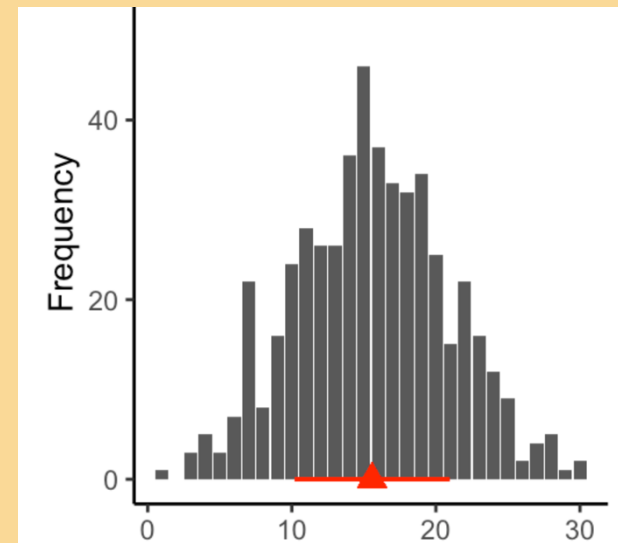


Cases



Controls

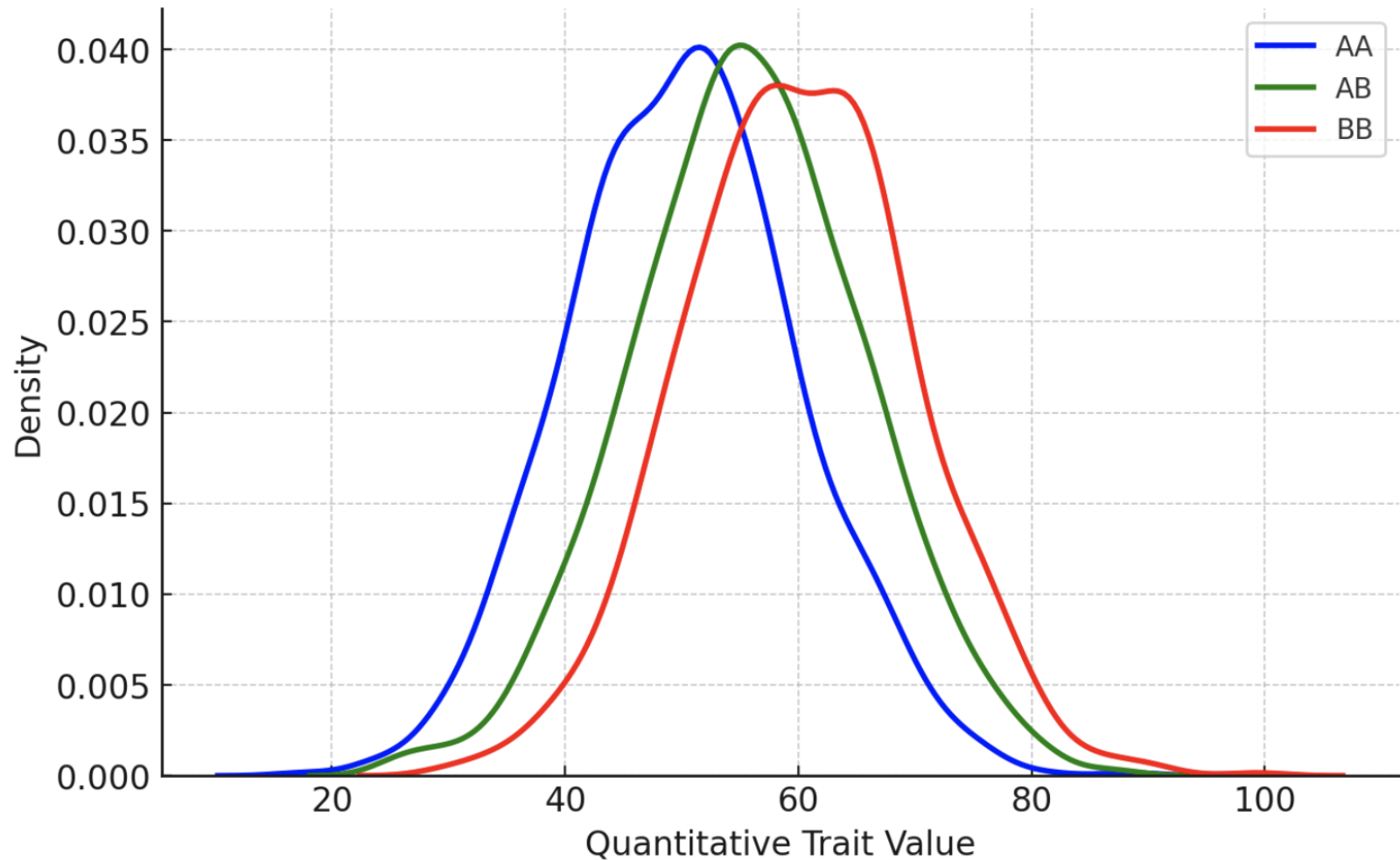
Quantitative traits



Quantitative trait value



If a variant is associated with a trait, we would expect different trait distributions:



Linear regression

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Genotype of individual i
↑
effect size
↓

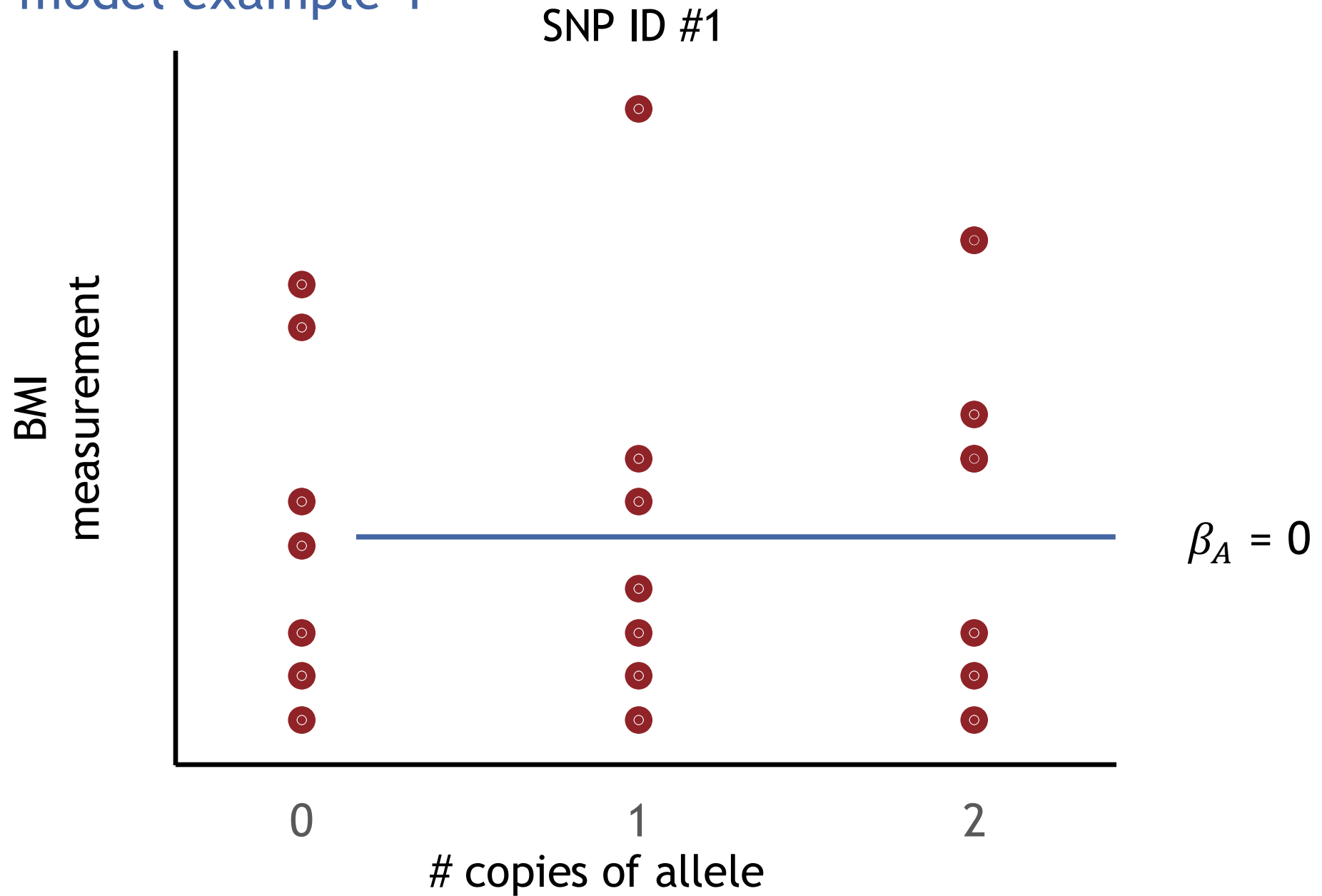
For a simple additive inheritance model we have

$$E(y) = \beta_0 + \beta_A x_A \longrightarrow x_A \text{ is the \# copies of the variant (0, 1 or 2)}$$

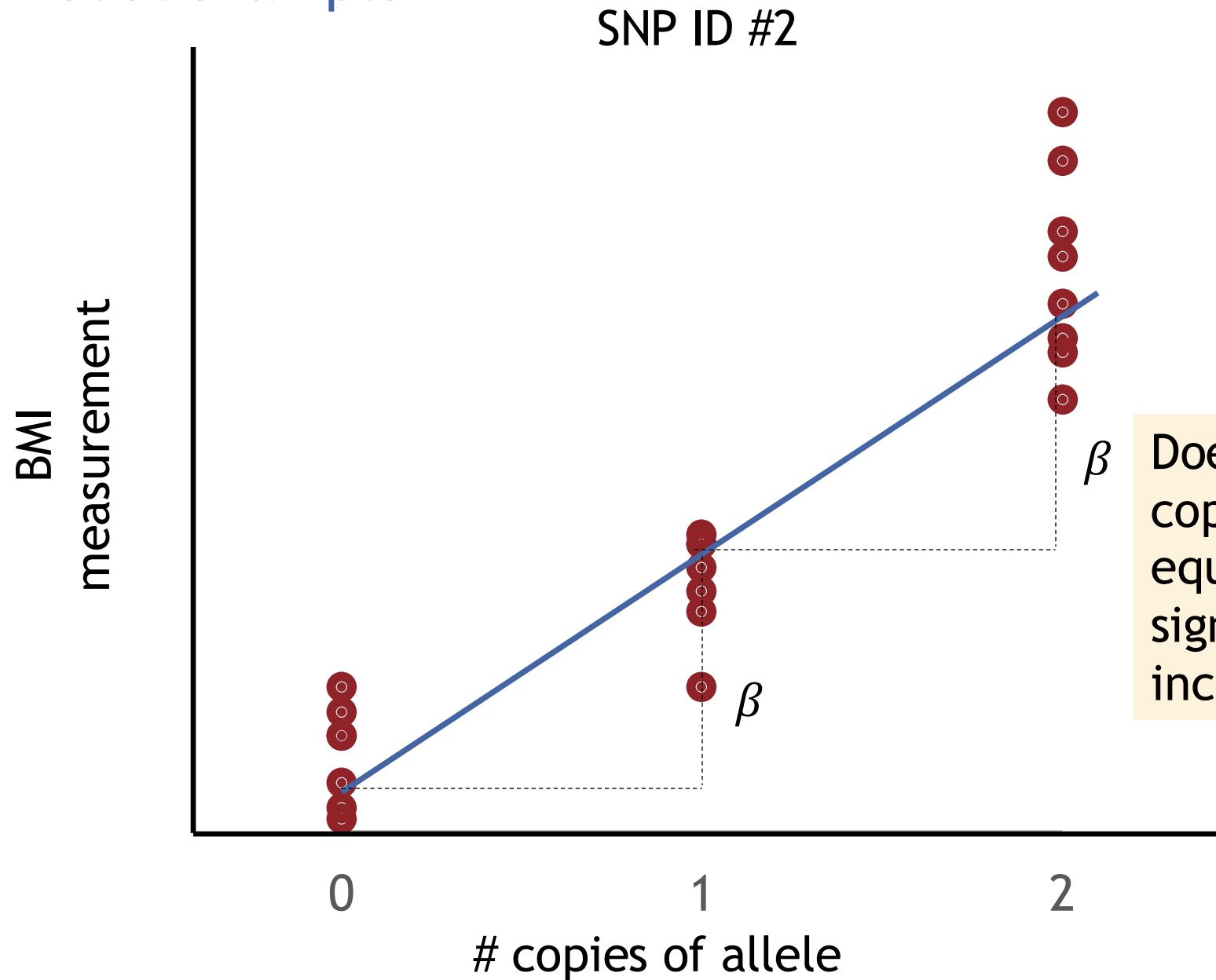
Test if $\beta_A = 0$ (no association between the variant and the trait)



Linear model example 1



Linear model example 2

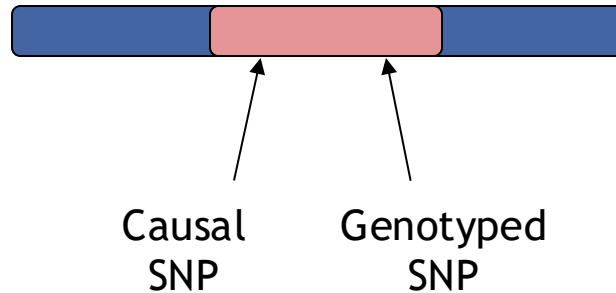


Does having more copies of allele #2 equate to a significant linear increase in BMI?



Limitations

Other loci are highly correlated (LD) with the causal variant -> also “associated” with the trait



All association tests assume that individuals are independent (unrelated) and from an homogenous (unstructured) population -> any violation can lead to false positive

QC and appropriate modelling is key!



Power of the associations tests

Will your study answer your research question? Key: power

- # of samples
- Test of choice (linear regression, mixed models...)
- Inheritance model (e.g. recessive)
- Effect sizes (strength of association, the larger the higher)
- Significance level or rejection criteria (e.g. $\alpha < 0.05$)
 - In GWAS, multiple testing
- Allele frequencies (MAF > 5%)
- Phenotypic variance explained (R^2) -> complex traits

→ The **lower** the alpha, the **larger** the sample size required to maintain power.



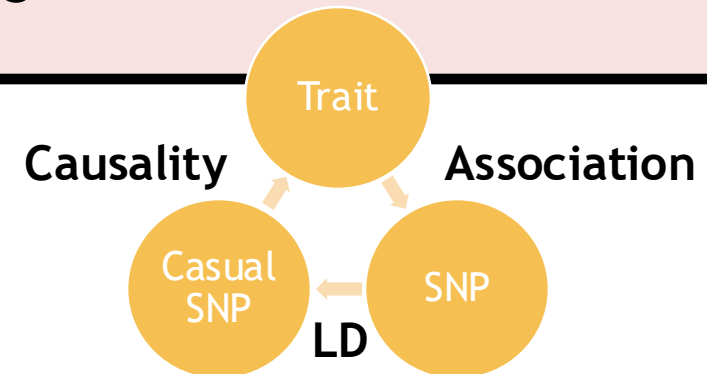
How do we test if a genetic variant potentially has an effect on a disease?

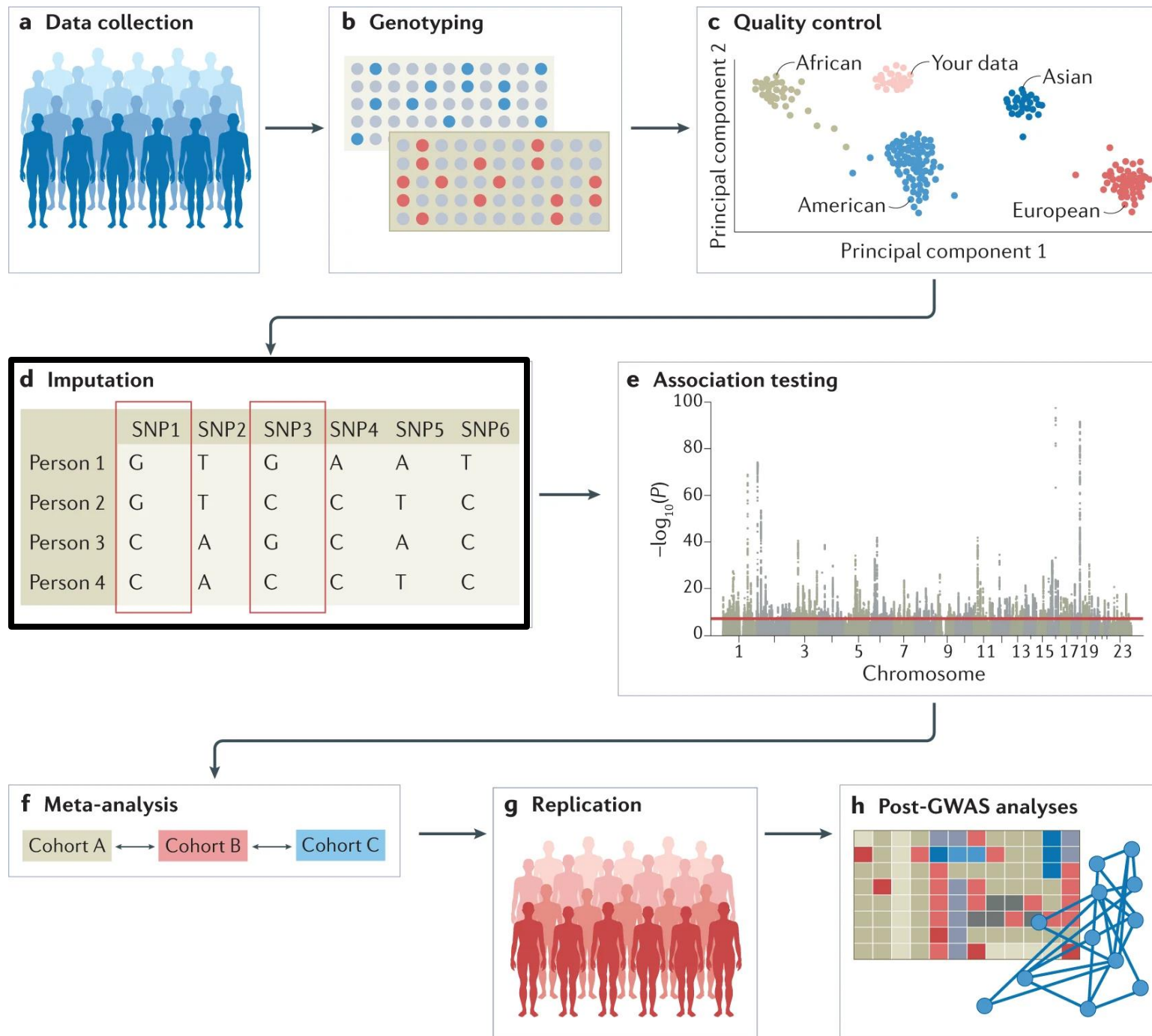
Single SNP tests

- Test to identify a genetic variant that affect a trait
- Evidence from previous studies
- Monogenic traits

GWAS

- Scan the genome for variants that affect a given trait
- > 500,000 SNPS (no prior evidence required)
- Likely not to have the casual SNP
- Polygenic traits





Software

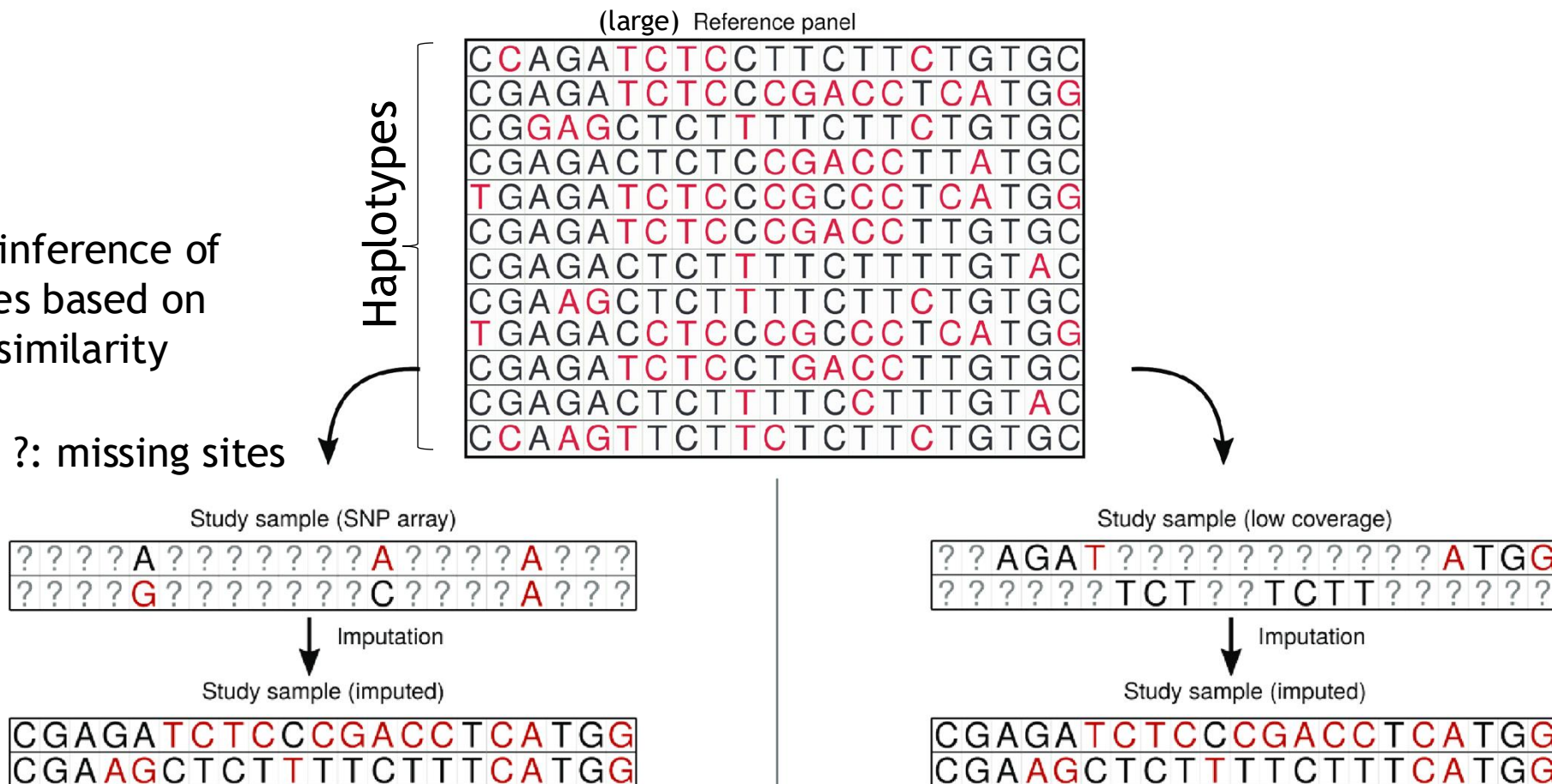
- Beagle
- SHAPEIT
- IMPUTE5



Imputation NGS

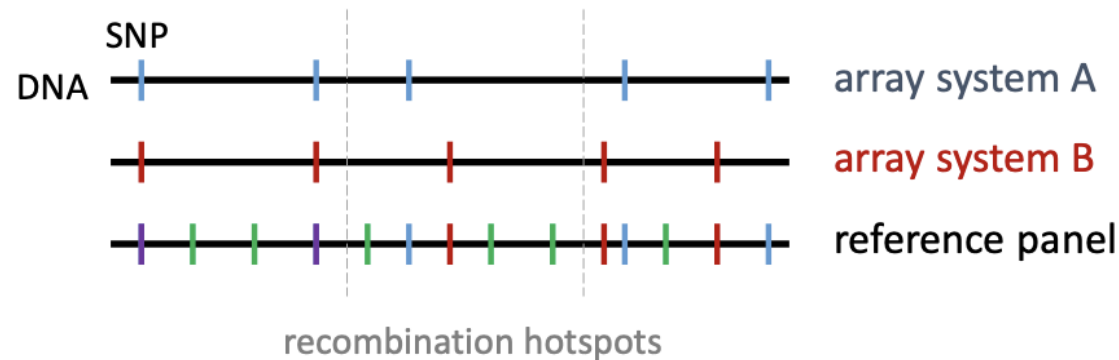
Extensive reference panels
(e.g. 1000 Genomes Project with > 2,500 genomes)

Statistical inference of
missing sites based on
haplotype similarity



Why imputation in GWAS?

- To allow comparison across GWAS studies
- To perform meta-analysis with other samples on other chips
- To fine map - i.e. run association at variants we have not genotyped
- To improve call rate - i.e. increase the number of variants available for poorly genotyped samples (not ideal)



Association testing

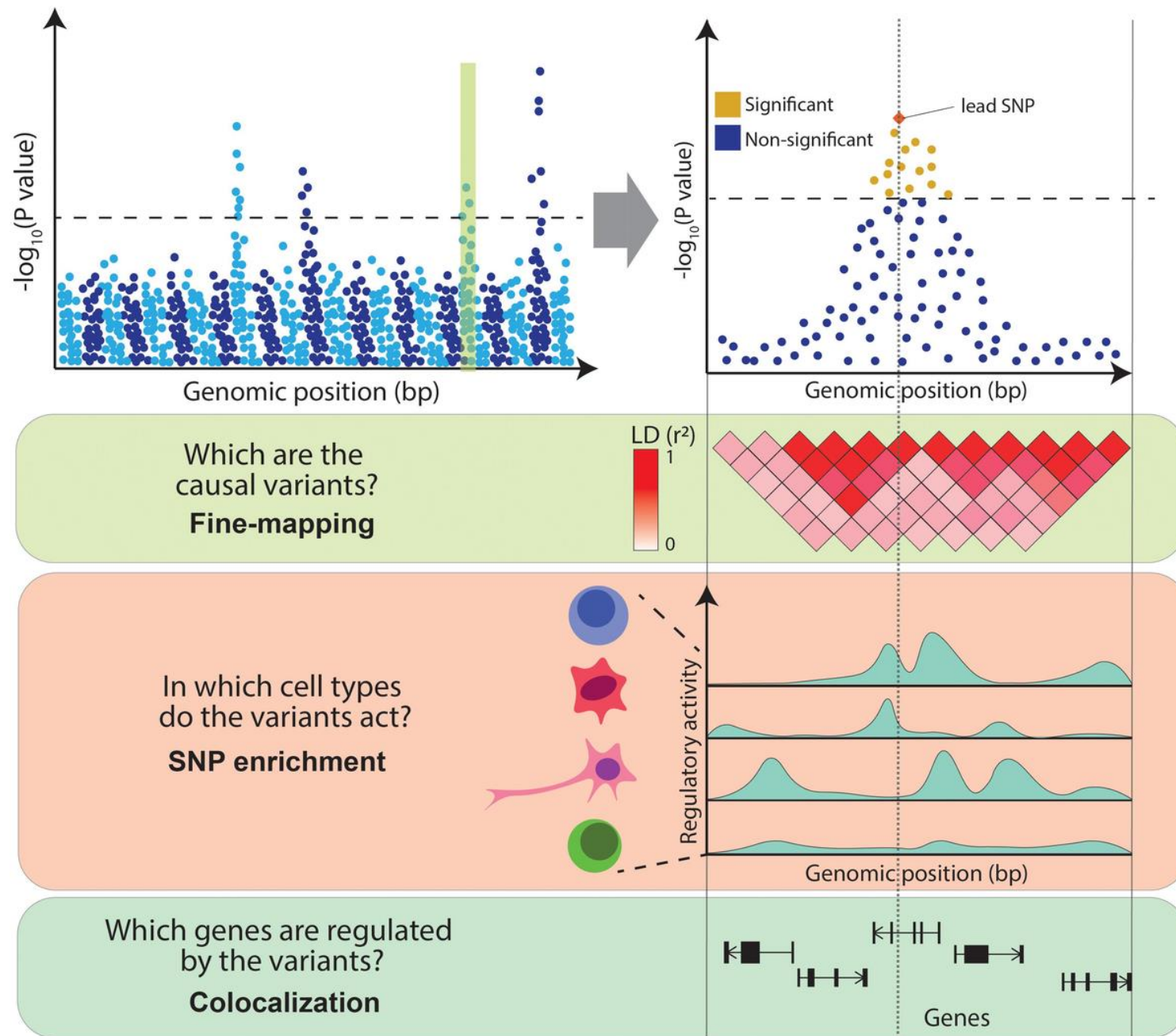
- Generalized linear models ✓
- Penalized multiple regressions
- Bayesian methods

More complex models try to take care of the limitations GWAS studies might have

How can we interpret GWAS results?

What is the strength and reliability of the associations in our GWAS results?





Which variant has the highest association?

Are they all in high-LD? (beta and p-value very similar)

Variable levels of regulatory activity across cell types

Genes within the associated locus



Effect sizes

Strength of association

Odds Ratio

Binary traits

- $OR = 1 \rightarrow$ No effect on the trait.
- $OR > 1 \rightarrow$ The allele increases the odds of having the trait.
- $OR < 1 \rightarrow$ The allele decreases the odds of having the trait.

$$OR = \exp(\beta_X)$$

Beta

Quantitative traits

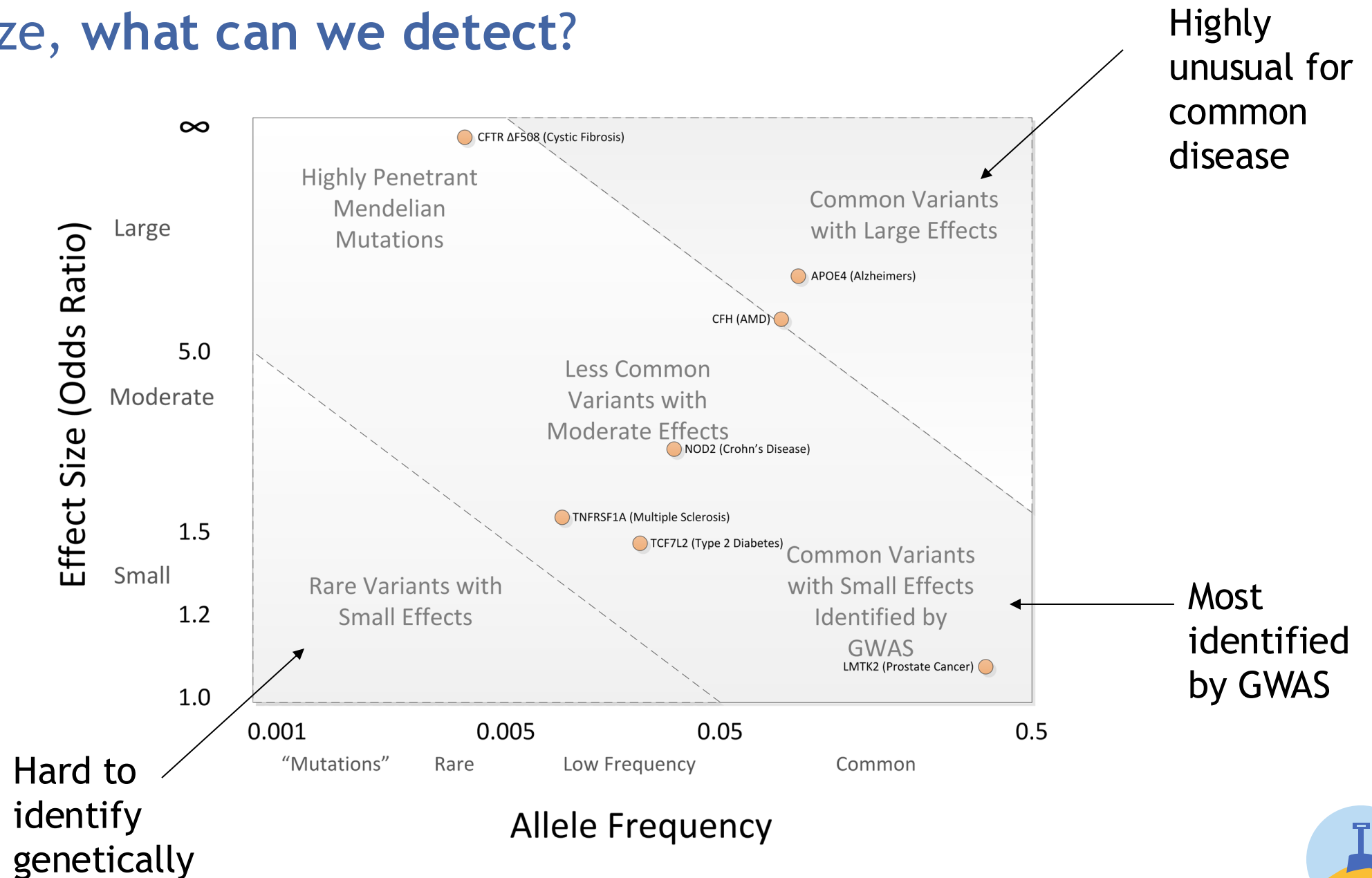
Change in the trait value per additional copy of the allele.

- $\beta > 0 \rightarrow$ The allele increases the trait value.
- $\beta < 0 \rightarrow$ The allele decreases the trait value.

E.g.: An OR of 1.12 means we will expect to see a 12% increase in the odds of having Parkinson's disease for a one unit increase in allele copy.



Effect size, what can we detect?

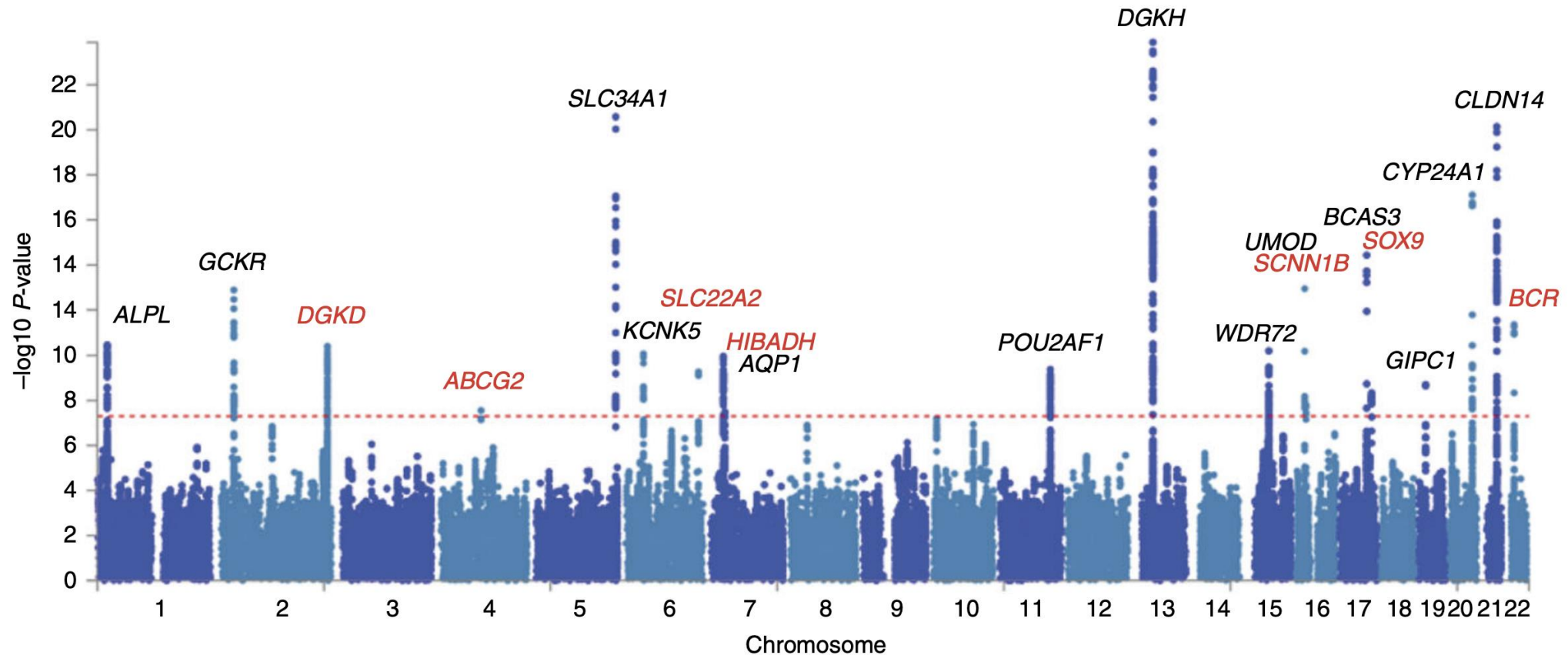


Manhattan plots: p-values

Significance of the association

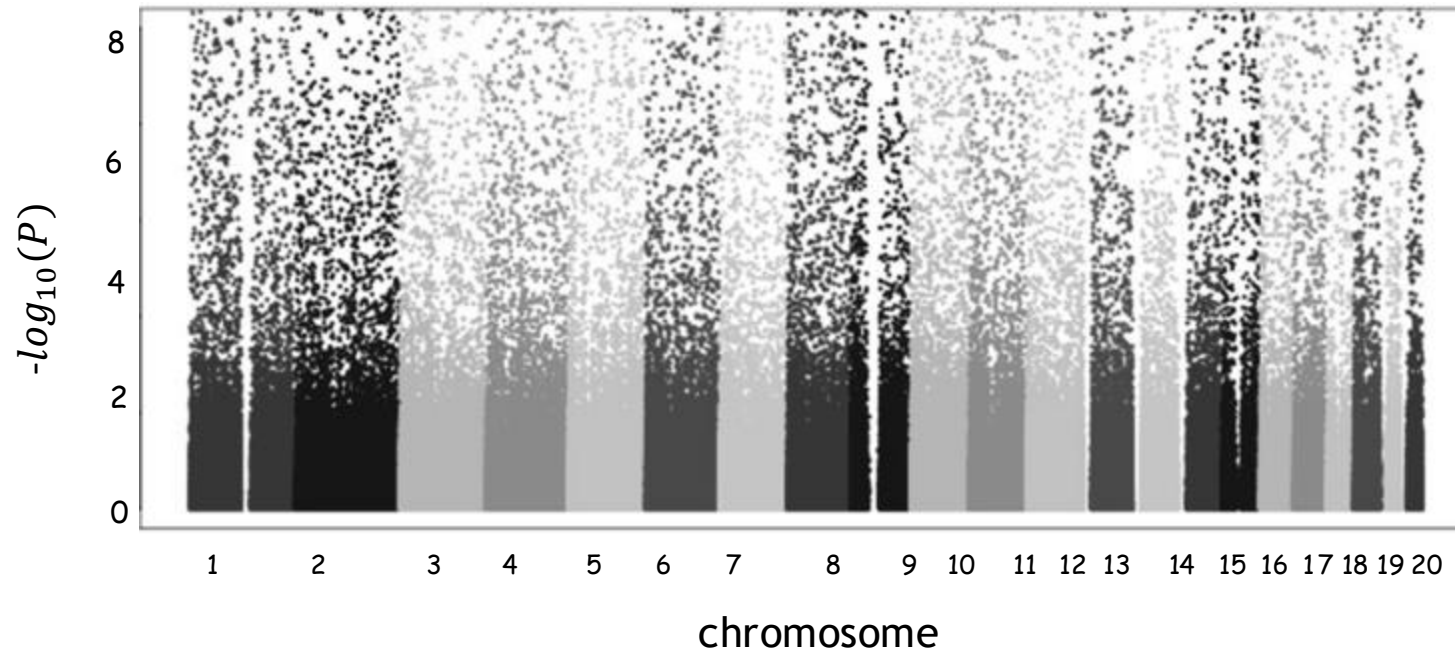
Useful to visualize our GWAS results and identify potentially associated regions

The height of the spike represents the strength of the association with the phenotype



What are the consequences of poor quality control?

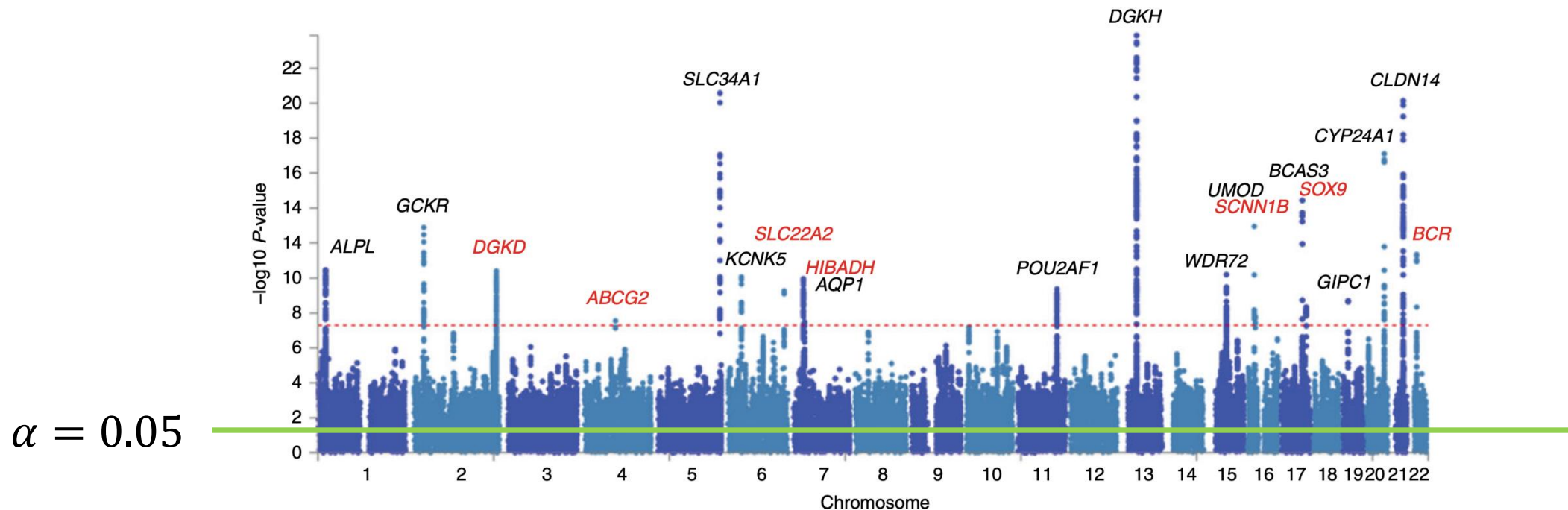
We would expect association study results to indicate very few associations between SNPS a specific trait!



Multiple testing and p-values

Significance of the association

- Common threshold for a single test ($\alpha = 0.05$)
- In the plot: $-\log(0.05) = 1.3$
- $>100,000$ tests \rightarrow Inflation of false positives



Multiple testing and p-values

Significance of the association

- Common threshold for a single test ($\alpha = 0.05$)
- >100000 tests → Inflation of false positives
→ We need **deflation of p-value**

How do we avoid false discovery?

Multiple test adjustments

- **Bonferroni correction:** $\alpha_{new} = \frac{\alpha}{\#tests}$
- **Permutations:** multiple tests with subsets of data



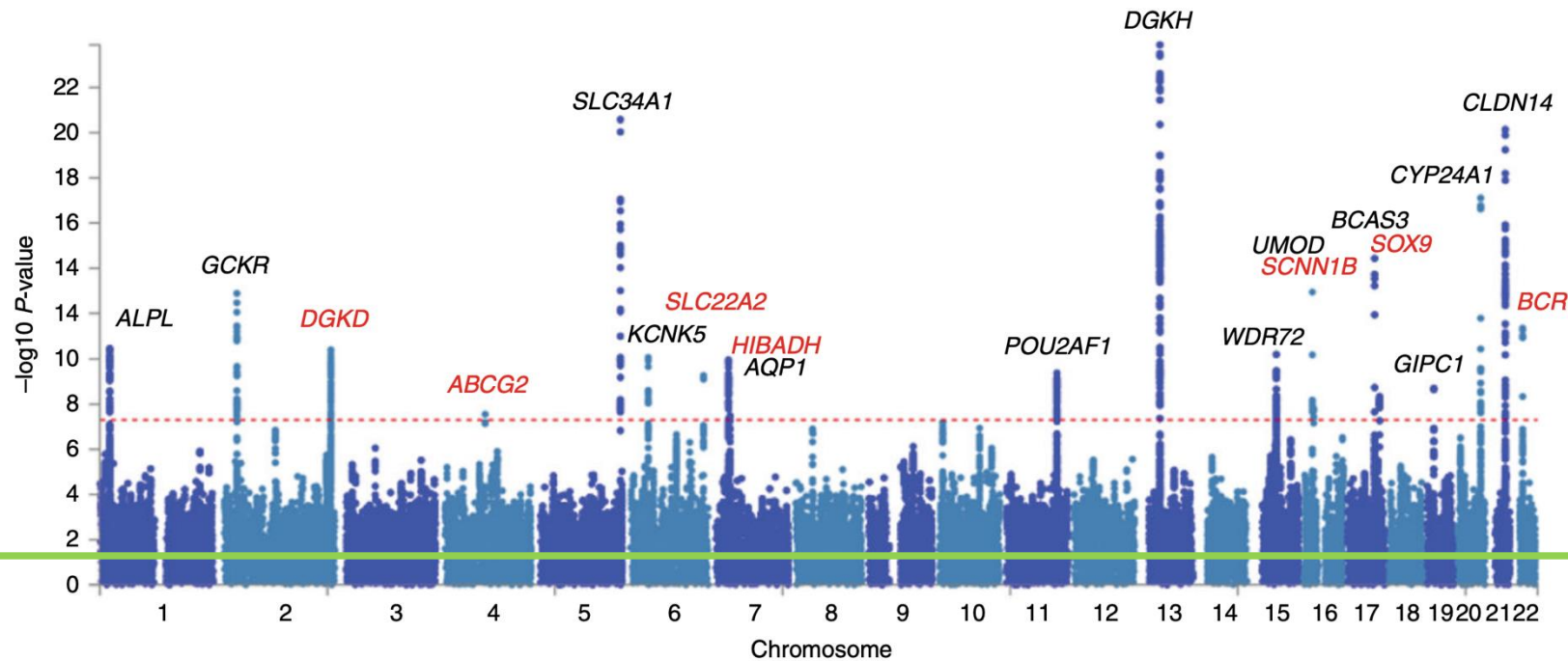
Multiple testing and p-values

Significance of the association

Bonferroni correction

Approx. 1 million independent tests: $P < 5 \times 10^{-8}$ “genome-wide” significance

$$\alpha = 0.05$$



QQ-plot

Significance of the association

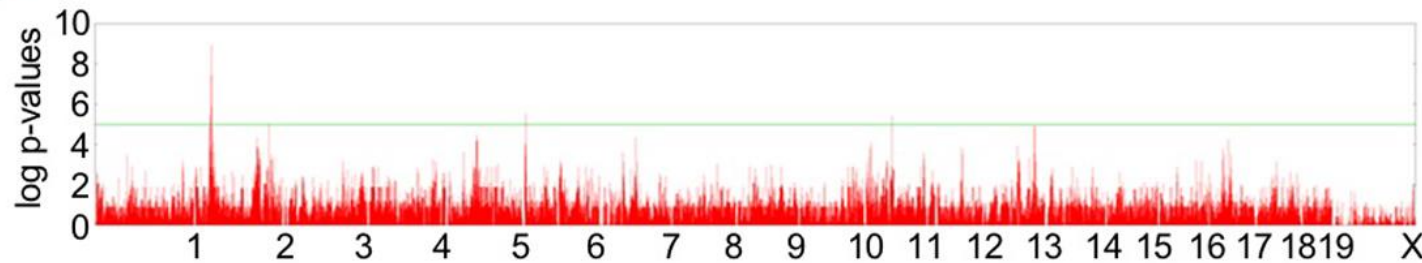
How likely the observed association is to be real?

Let's look at some examples...



What do we expect to see?

A GENOME-WIDE ASSOCIATION MAP

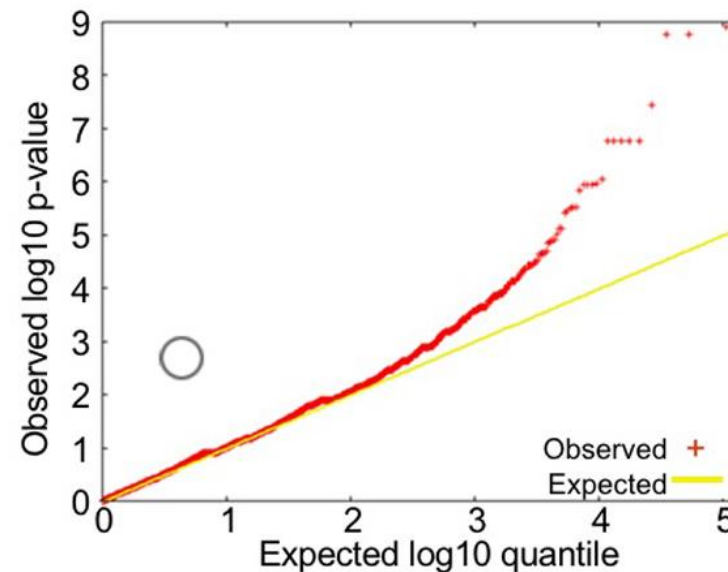


Significant threshold
A few SNPs over the significant threshold.

Most are not associated with the phenotype

A few show stronger signals than expected at the tail

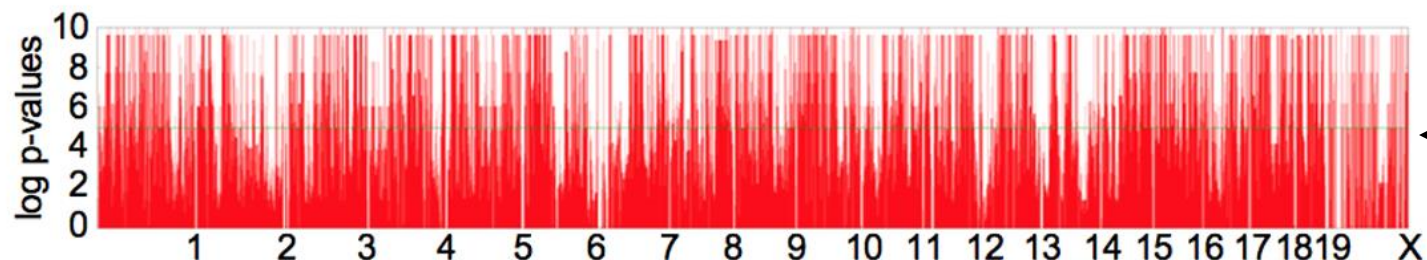
C Q-Q PLOT



Expected distribution of p -values in a typical (A) Manhattan plot, (B) cumulative p -value distribution, and (C) Q-Q plot. Circles in (B) and (C) denote where the median p -value (red line) falls on the graph in comparison to the expected median p -value (yellow line). Here, the median falls close to 0.5, suggesting that population structure is not affecting association results or has been corrected for in the model. Q-Q, quantile-quantile. <https://doi.org/10.1371/journal.pgen.1007309.g004>

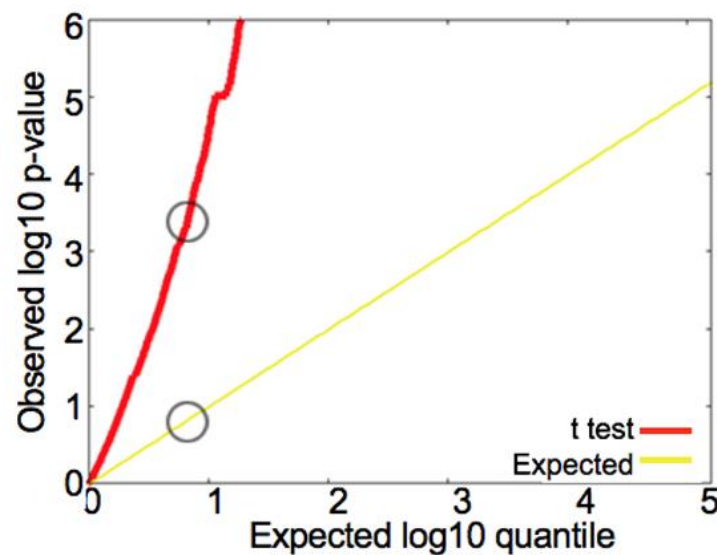


A GENOME-WIDE ASSOCIATION MAP



← Significant threshold

C Q-Q PLOT



Log-transformed

Extreme amount of SNPs crossing this line (50% showing “significant association”)

Observed distribution in a (A) Manhattan plot, (B) cumulative p -value distribution, and (C) Q-Q plot. Circles in (B) and (C) indicate where the median p -value falls on the plot compared to where it is expected. Here, there is a substantial deviation between the red and yellow lines due to **inflation of false positive associations** for the body weight phenotype. Q-Q, quantile-quantile. <https://doi.org/10.1371/journal.pgen.1007309.g005>





~ 30 min



GWAS5-AssociationTesting.ipynb



- Linear regression models using PLINK
- Correcting for multiple testing
- Visualization:
 - Manhattan
 - Q-Q plots



Choose the Bash kernel



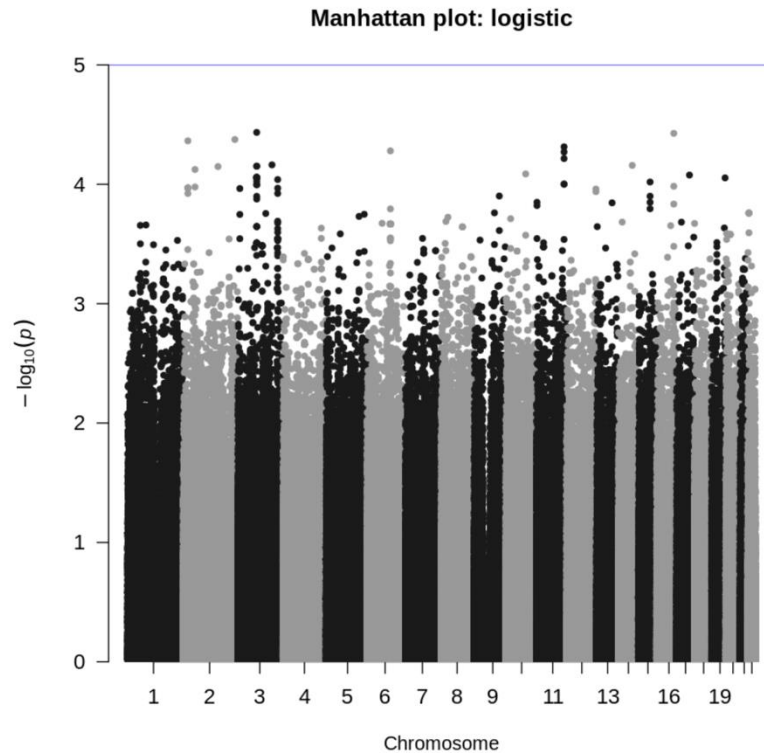
Choose the R-GWAS kernel

Solutions

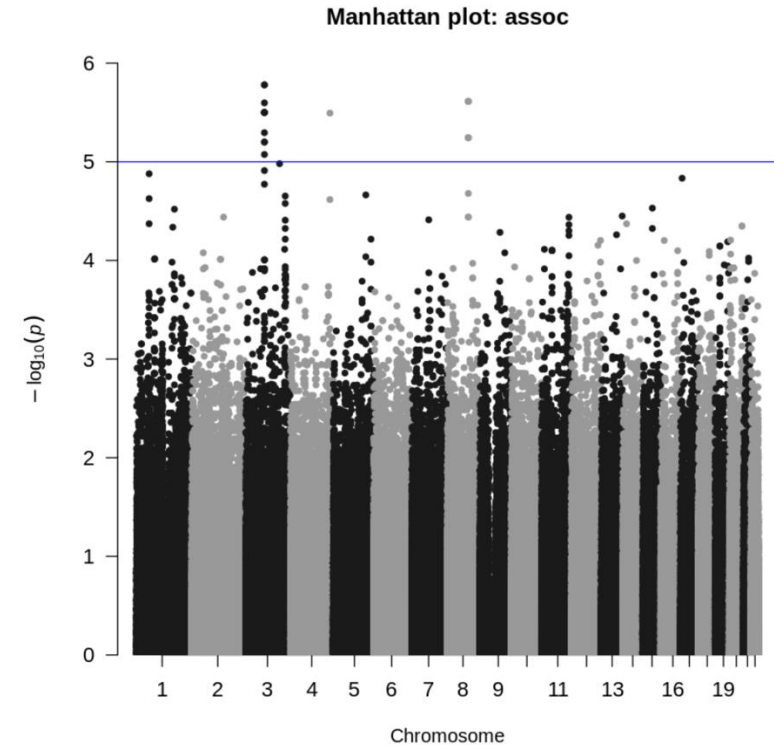
- Problems/Issues/Comments?

Comparing two GWAS approaches

Which model is more appropriate in GWAS?



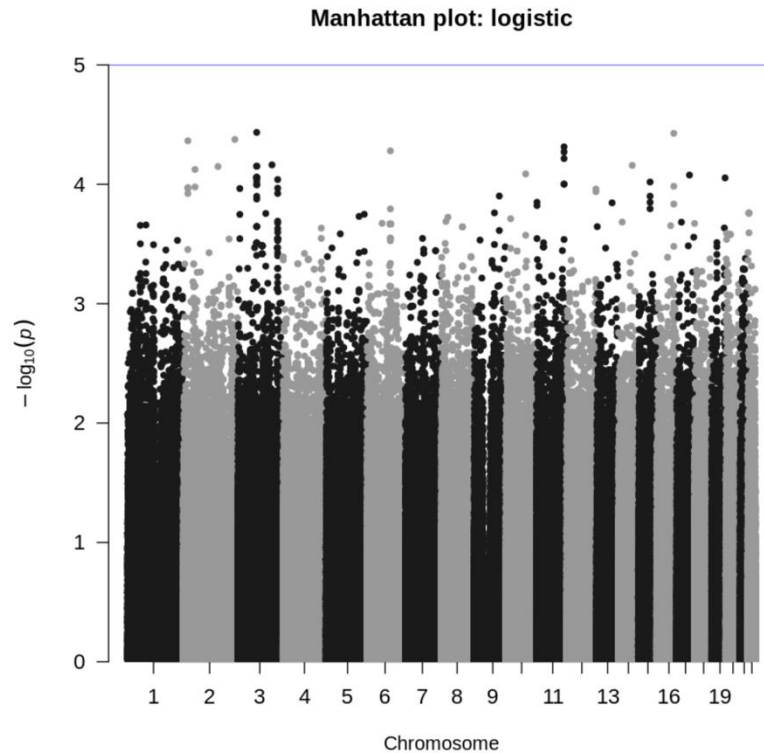
Are we being too conservative?



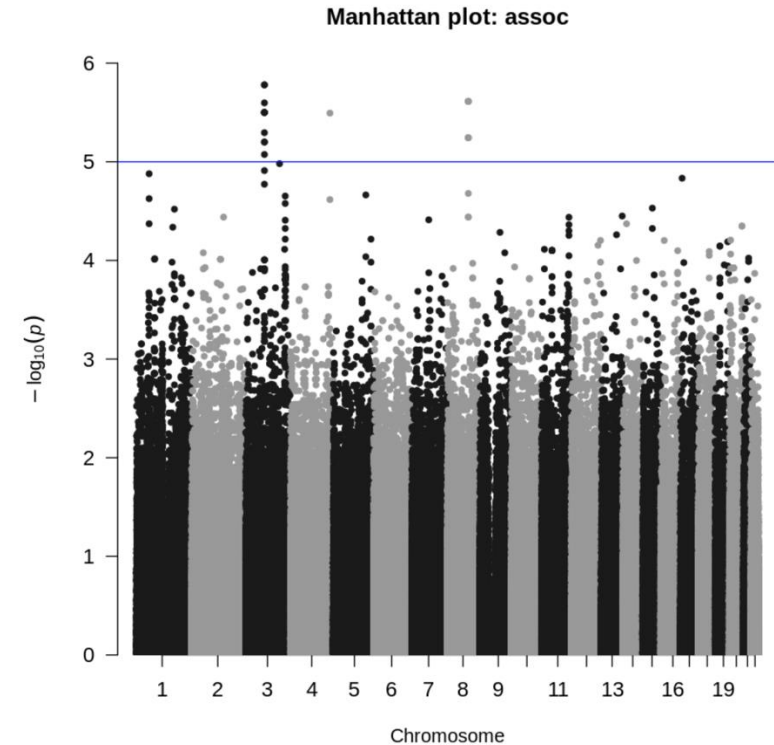
What might more extreme p-values in GWAS indicate?

Comparing two GWAS approaches

Which model is more appropriate in GWAS?



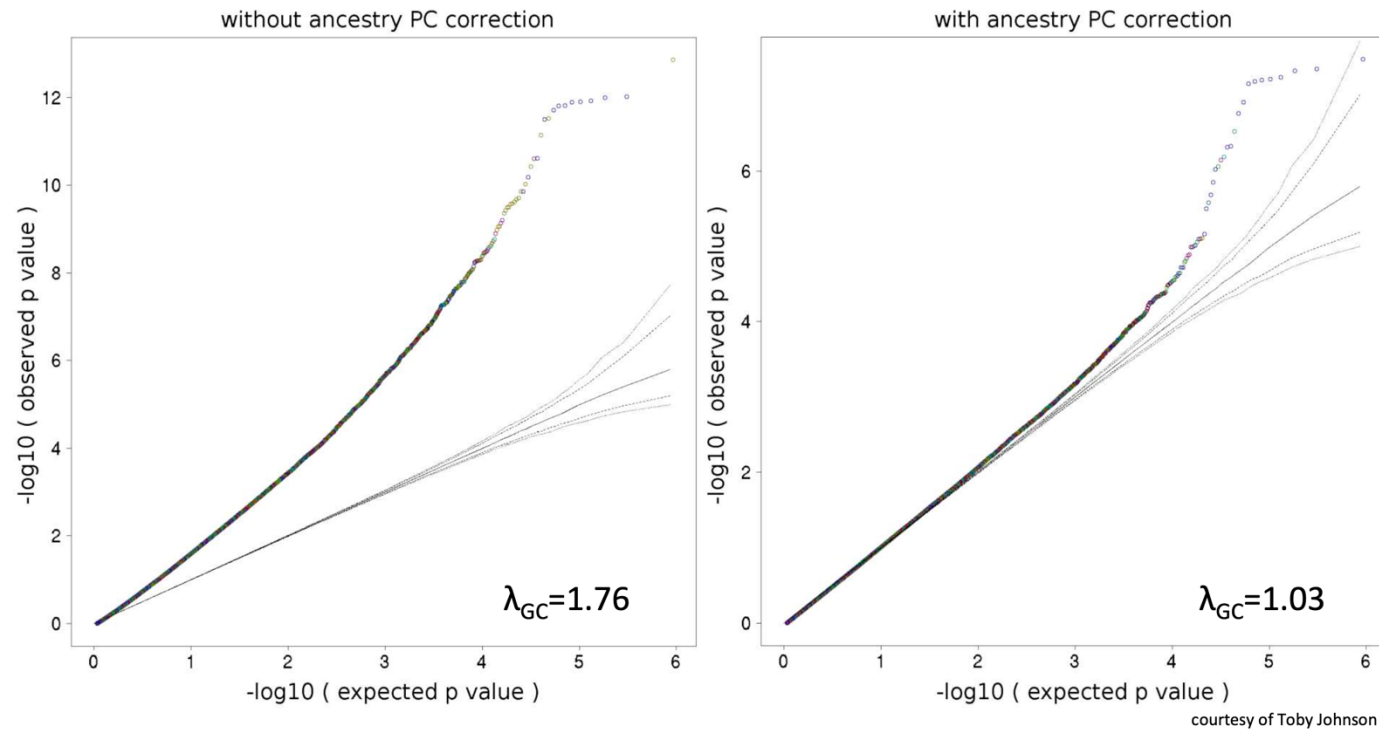
Stringent adjustments could lead to higher false negatives



More extreme p-values might indicate inflation due to confounding!

Q-Qplots

This is what we would expect before and after adjusting for population structure...



We might not have enough power

GWAS software

- **PLINK**
 - (-) limited flexibility on imputed data using allele dosage information (no hard calls)
- **BOLT-LMM**
 - (+) accounts for relatedness, very fast for large datasets
 - (-) linear regression only, genotyped and imputed data required (two-step approach)
- **Regenie**
 - (+) accounts for population structure and relatedness (on most cases), rare variant analysis possible, very fast for large datasets, analysis of multiple traits at once
 - (-) two-step analysis
- **Quicktest**
 - (+) very fast, can calculate GxE effects
 - (-) not accounting for relatedness
- **SAIGE**
 - (+) accounts for relatedness, very fast for large datasets, rare variant analysis possible (gene-based tests)
- **Raremetalworker**
 - (+) accounts for relatedness, rare variant analysis possible
 - (-) linear regression only, requires specific software for meta-analysis (Raremetal)



GWAS software Links

- PLINK
 - <https://www.cog-genomics.org/plink/1.9/> (v1.9)
 - <https://www.cog-genomics.org/plink/2.0/> (v2.0)
- Quicktest
 - <https://wp.unil.ch/sgg/program/quicktest/>
- BOLT-LMM
 - <https://www.hsph.harvard.edu/alkes-price/software/>
- SAIGE
 - <https://github.com/weizhouUMICH/SAIGE>
- Raremetalworker
 - <https://genome.sph.umich.edu/wiki/RAREMETALWORKER>
- Regenie
 - <https://rgcgithub.github.io/regenie/>



Additional important validations?

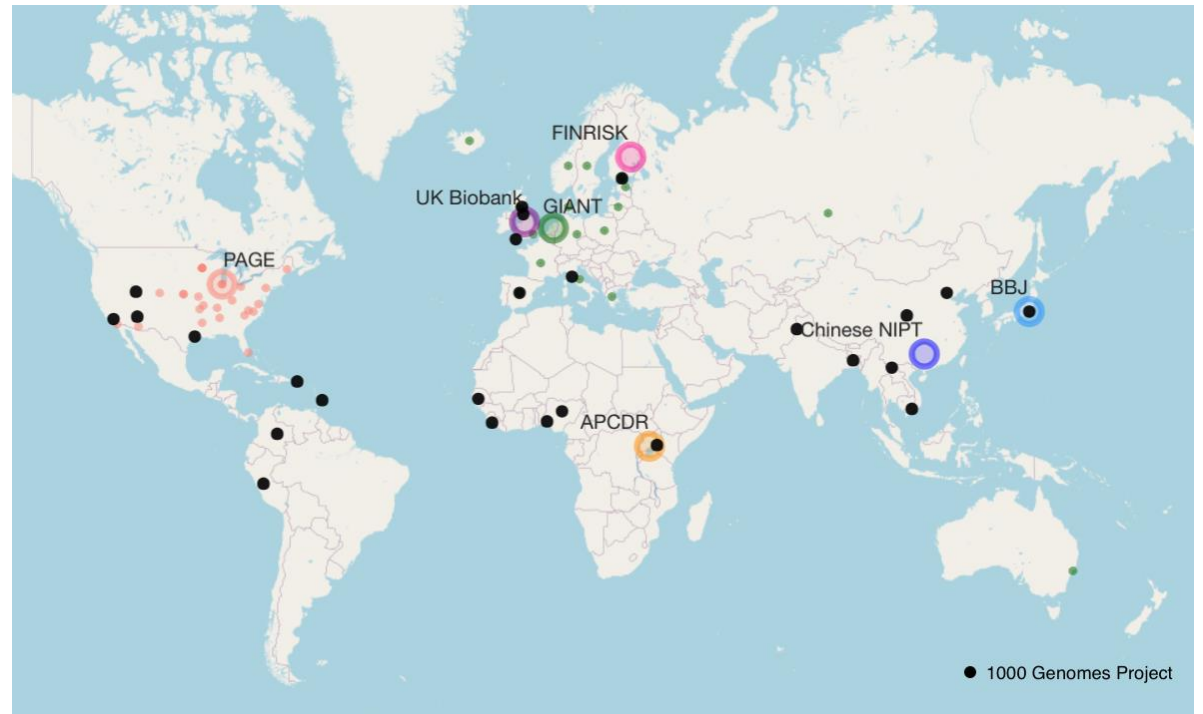
- Are there signs of something being wrong?
- Consequences of bad QC in downstream analysis
- Case-study

Evolutionary Biology

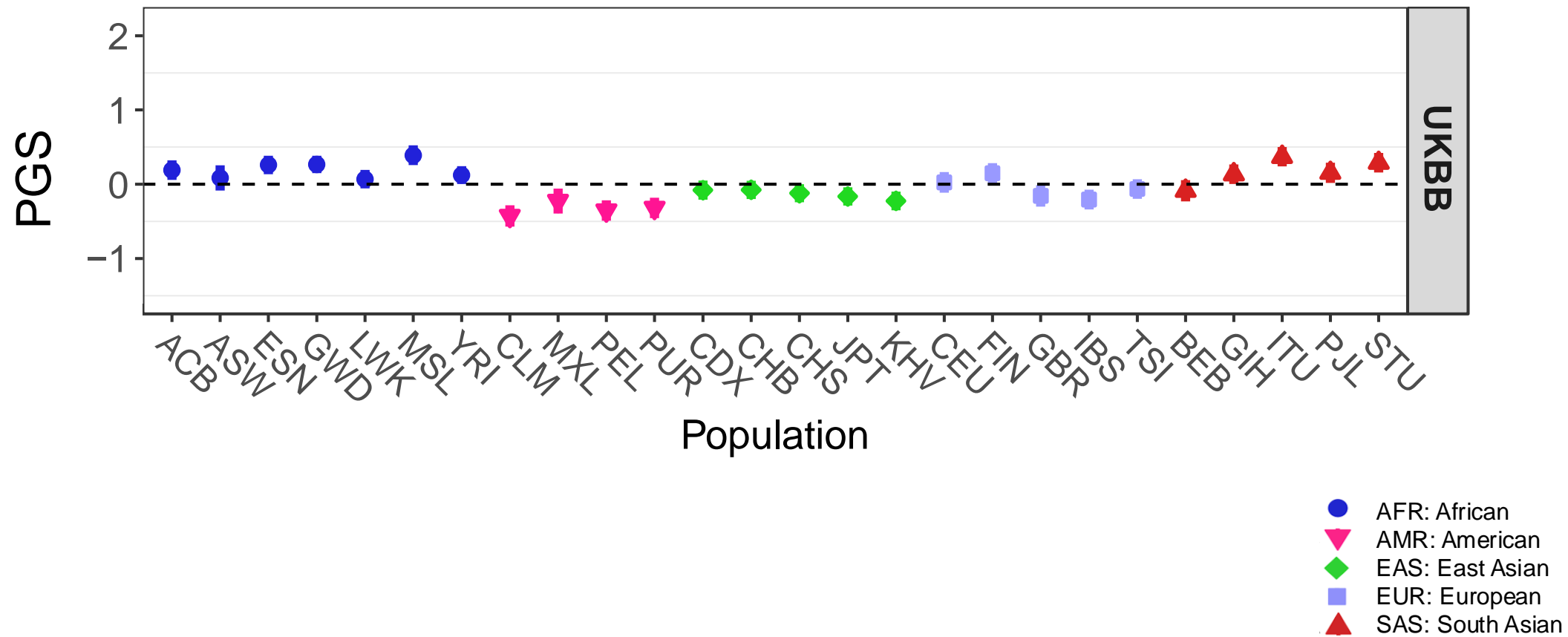
How robust are cross-population signatures of polygenic adaptation in humans?

Refoyo-Martínez, Alba¹ ; Liu, Siyang^{2, 3}; Jørgensen, Anja Moltke⁴ ; Jin, Xin²; Albrechtsen, Anders³ ; Martin, Alicia R.^{5, 6, 7} ; Racimo, Fernando¹  

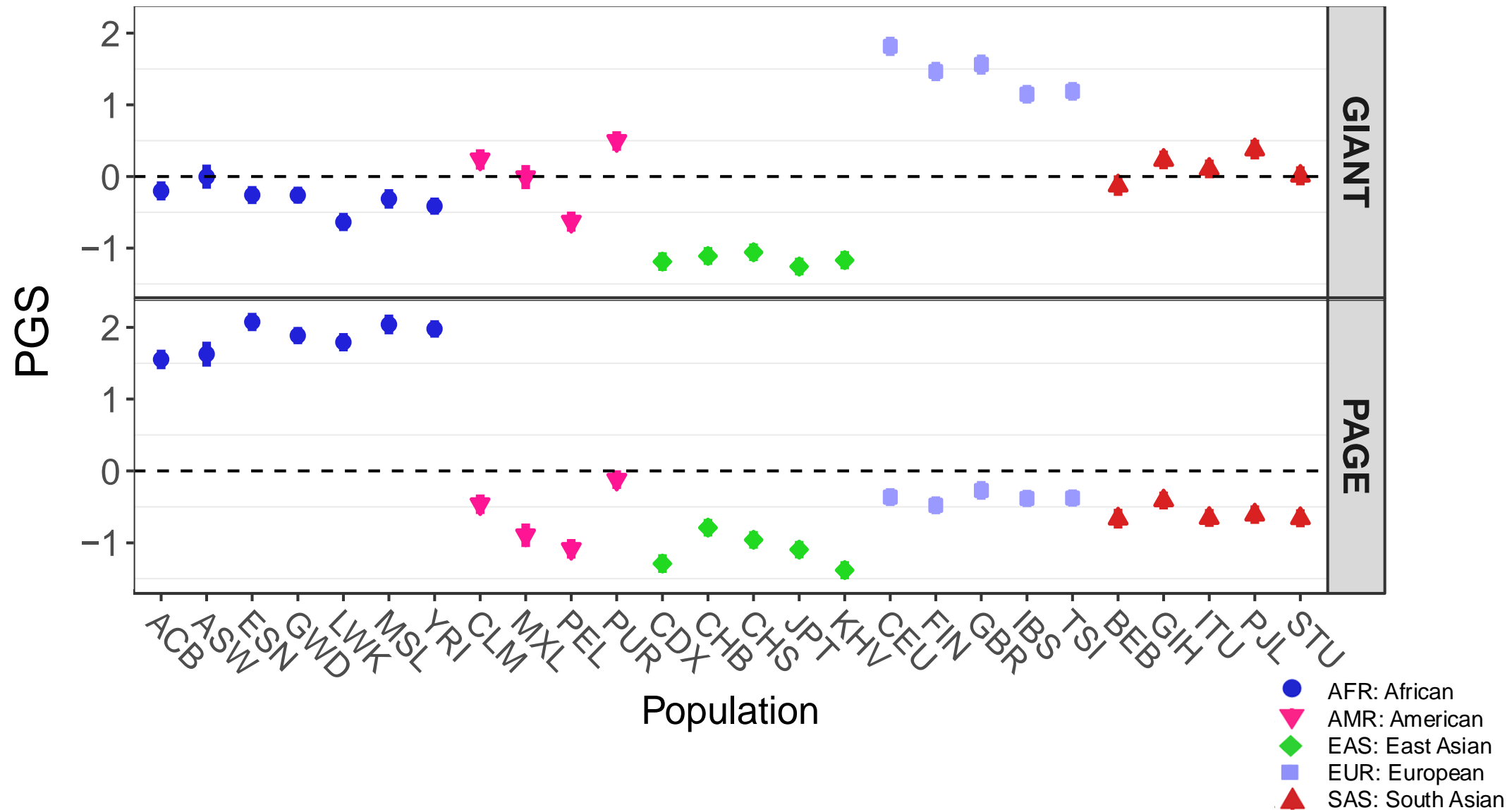
10.24072/pcjournal.35 - Peer Community Journal, Volume 1 (2021), article no. e22.



Polygenic scores discordance when using effect sizes from different cohorts



Polygenic scores discordance when using effect sizes from different cohorts

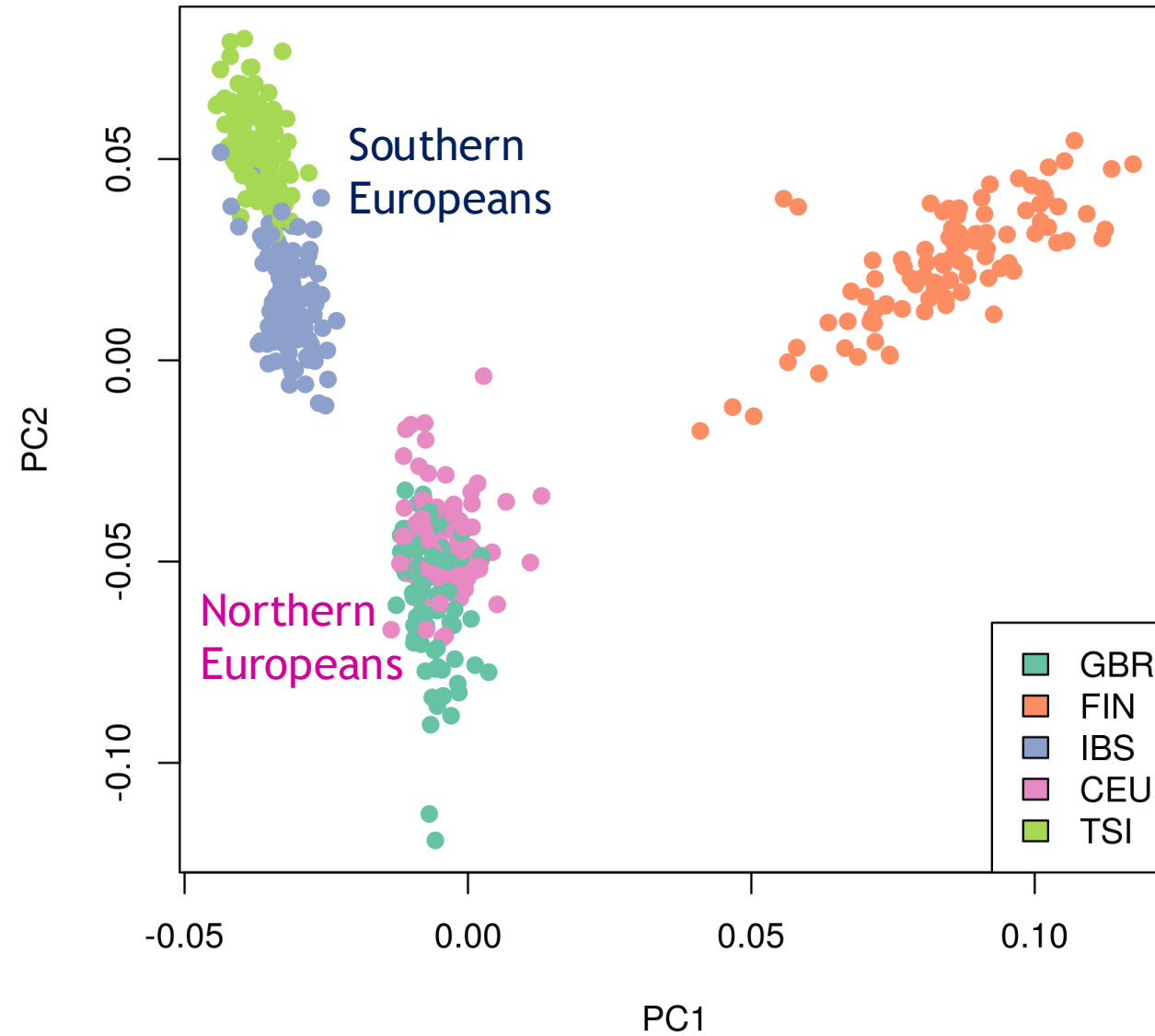


- All association tests assume independent samples—has this assumption been violated?
- Is the accuracy of the scores influenced by the ancestry of the GWAS panel? More on the lack of transferability in the next lecture
- Was rigorous quality control (QC) performed?
- Is there any overlap between the GWAS panel and the target sample?
- Since accuracy depends on the P-value threshold, did we test multiple thresholds and select the optimal one?

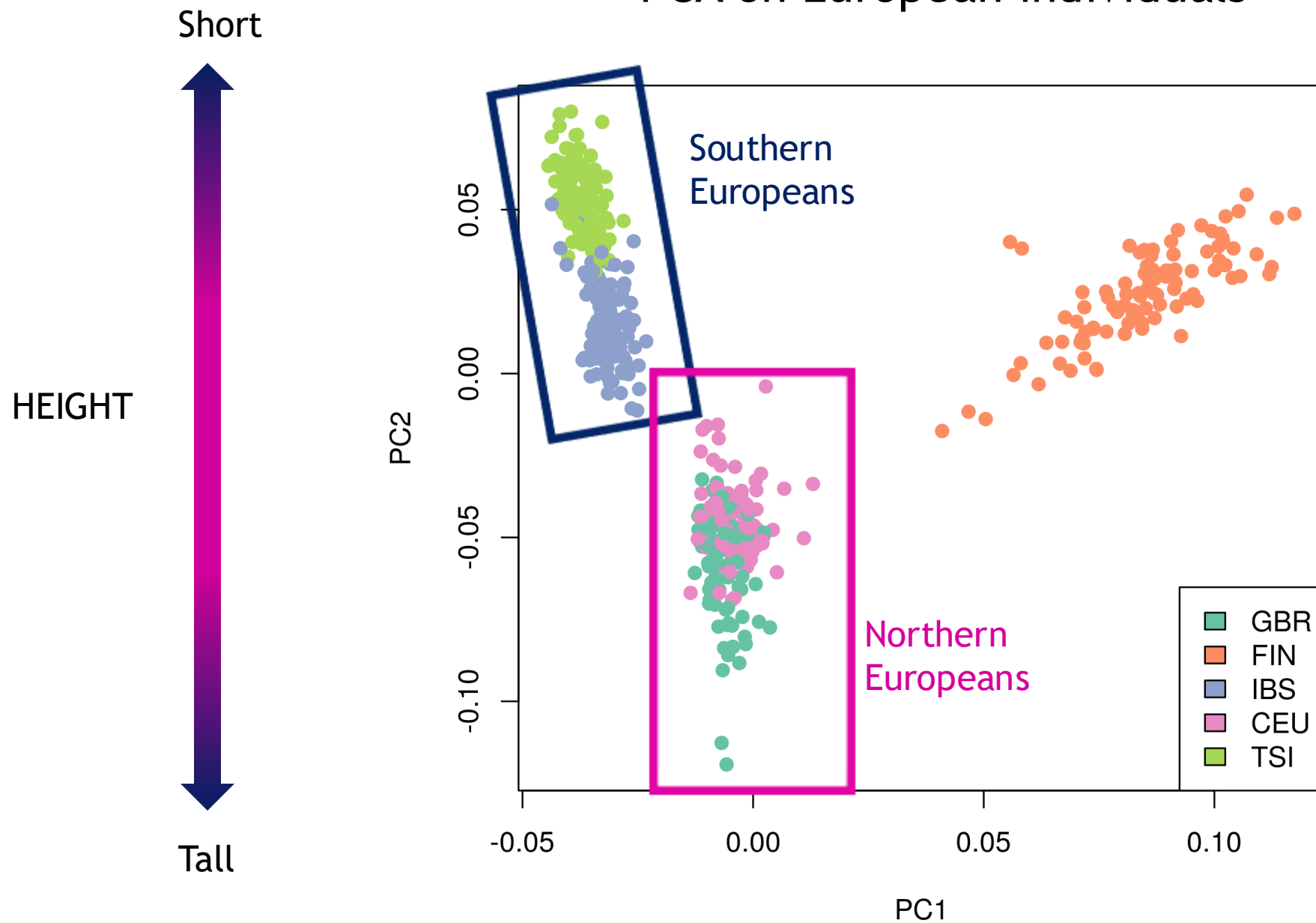
Are the differences driven by population stratification?



PCA on European individuals

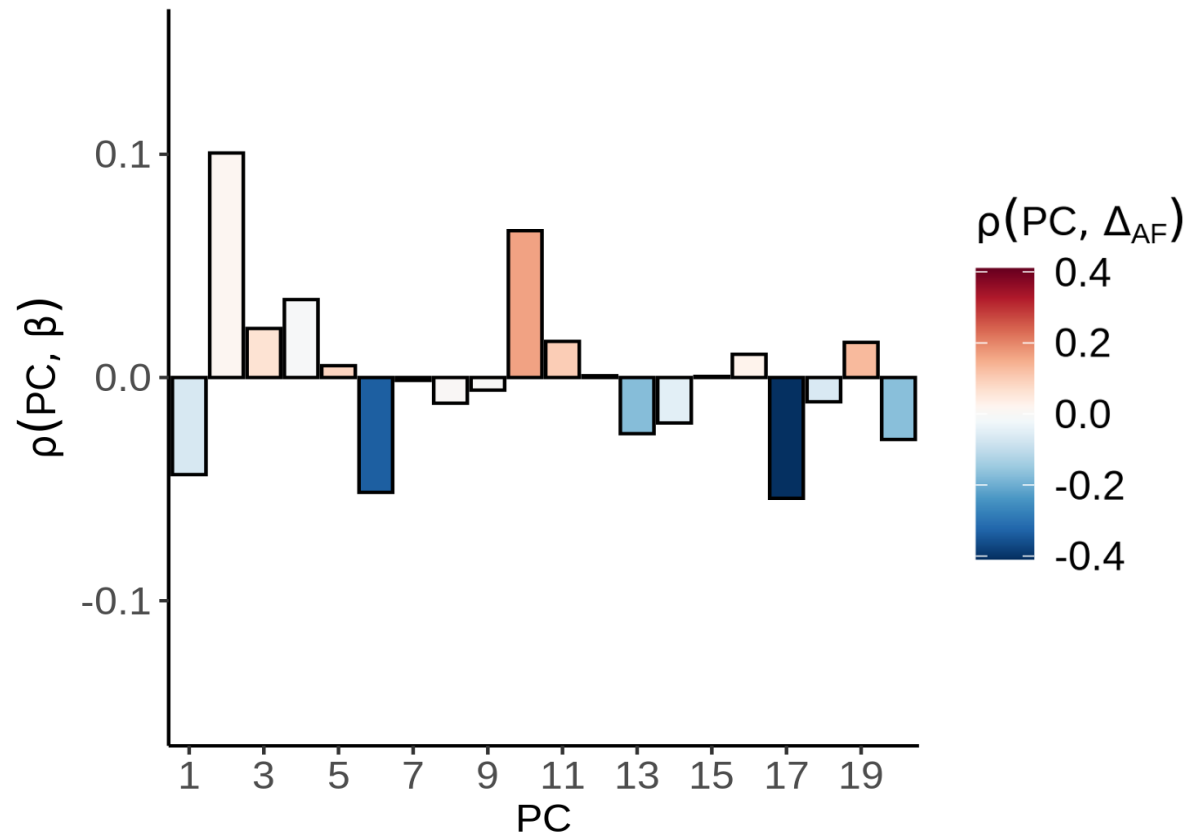


PCA on European individuals



Population structure along PCA axes

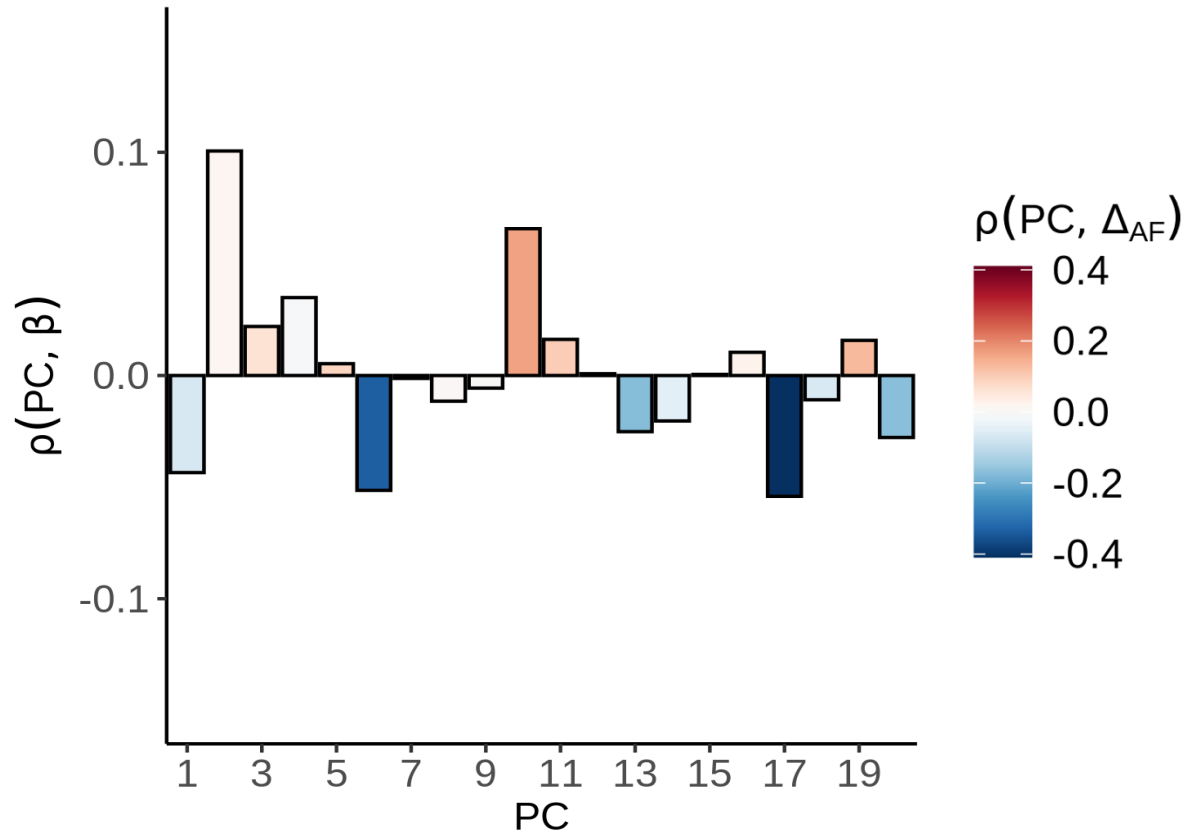
Effect sizes (β) GIANT



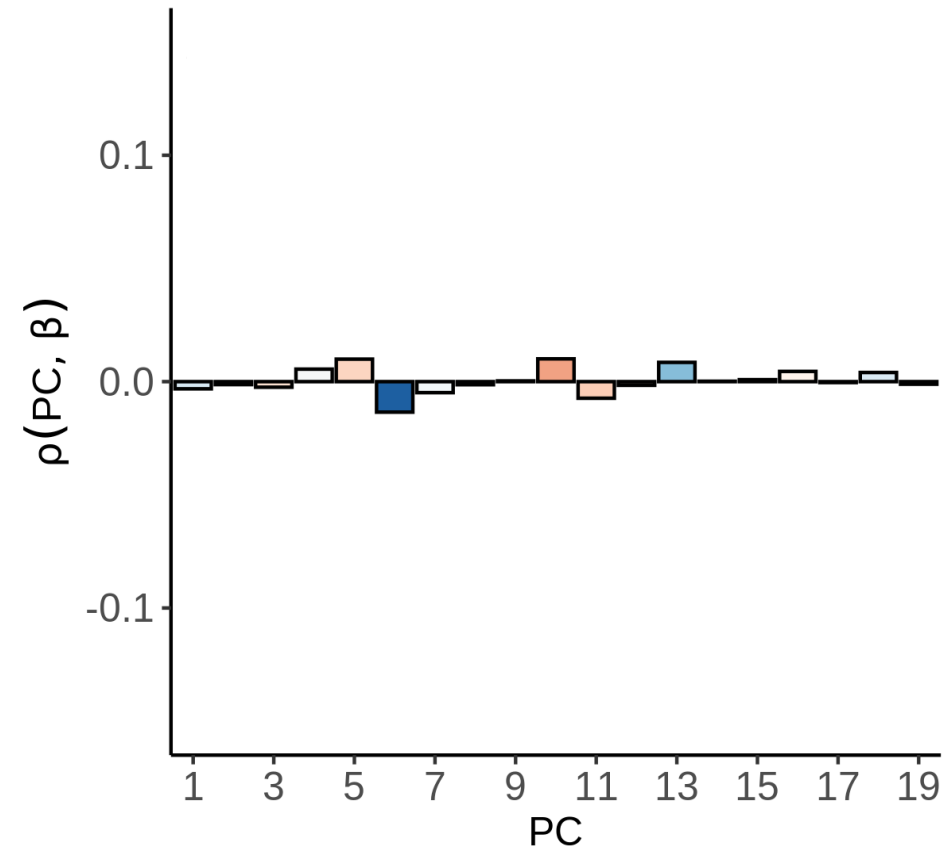
Population structure along PCA axes

Population stratification reduced using large-scale datasets with relatively homogeneous ancestries

Effect sizes (β) GIANT



Effect sizes (β) UK Biobank



Adjusting effect sizes can increase PRS accuracy using LD information from a external reference panel

LDpred

(Vilhjalmsson et al. AJHG 97:576-592)



 ~ 15 min



GWAS5b-PopulationStratification.ipynb



- Visualize potential residual population stratification in the discovery GWAS

Why is important? This will have an effect on your downstream analysis.



Choose the Bash kernel



Choose the R-GWAS kernel

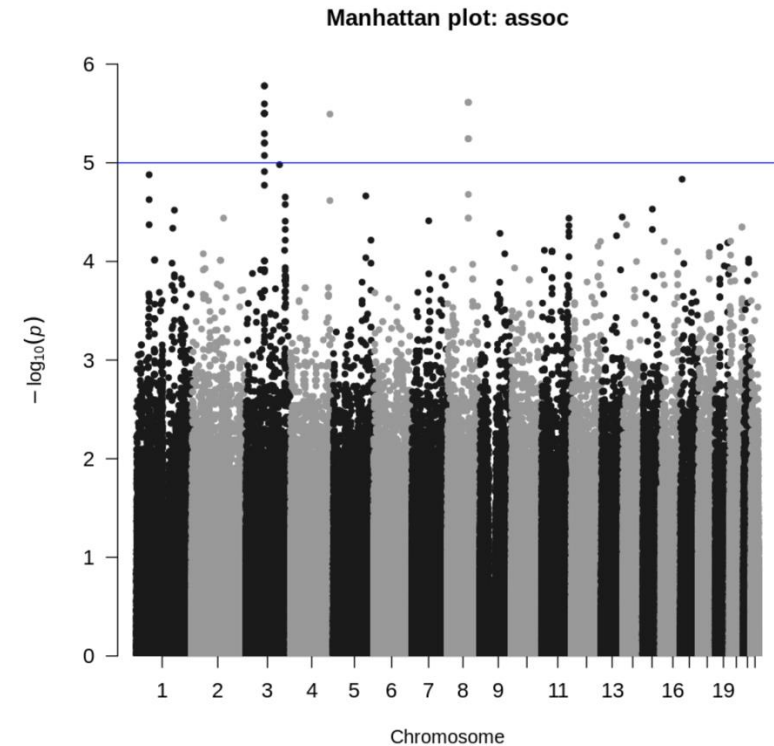
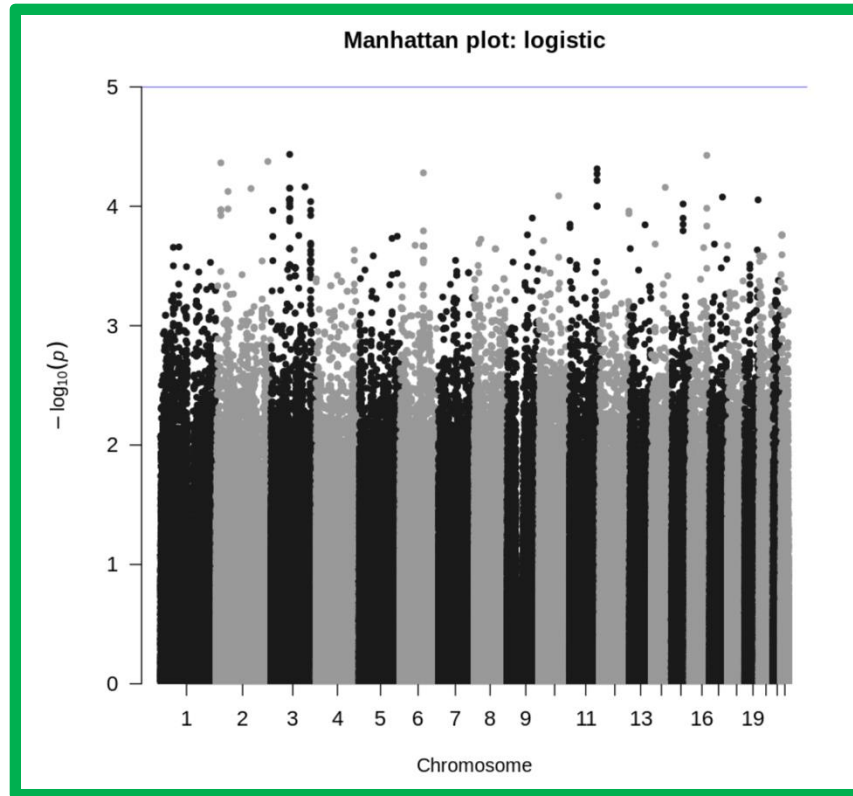
Solutions

- Problems/Issues/Comments?

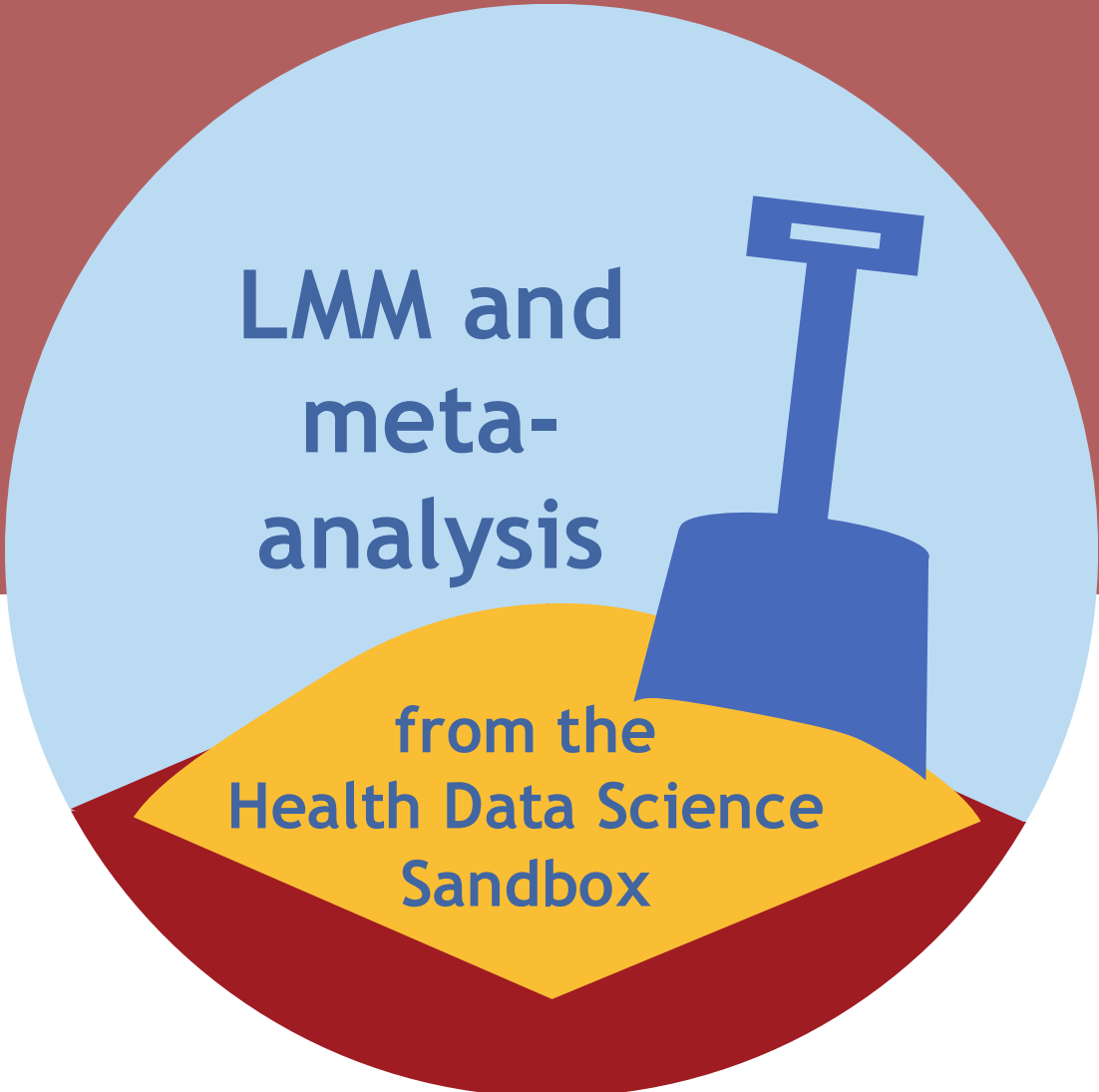
Solutions

Comparing two GWAS approaches

Which model is more appropriate in GWAS?



What else can we do? Run a linear mixed model!



**LMM and
meta-
analysis**

**from the
Health Data Science
Sandbox**



Samuele Soraggi, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN



GWAS with the Genomics Sandbox



Today's topics

Association tests

- GWAS
 - Linear mixed models
 - Tools
- Meta-analysis



Linear Mixed Models (LLMs)

Challenges in traditional GWAS

- Standard GWAS uses **linear regression**:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$y = X\beta + \epsilon, \quad \text{with } \epsilon \sim N(0, \sigma^2 I)$$

- **Population structure and relatedness** introduce false positives
 - The model is missing terms to describe their effect
- E.g. Height differences between populations can confound results

Linear Mixed Models (LLMs)

The random effect term

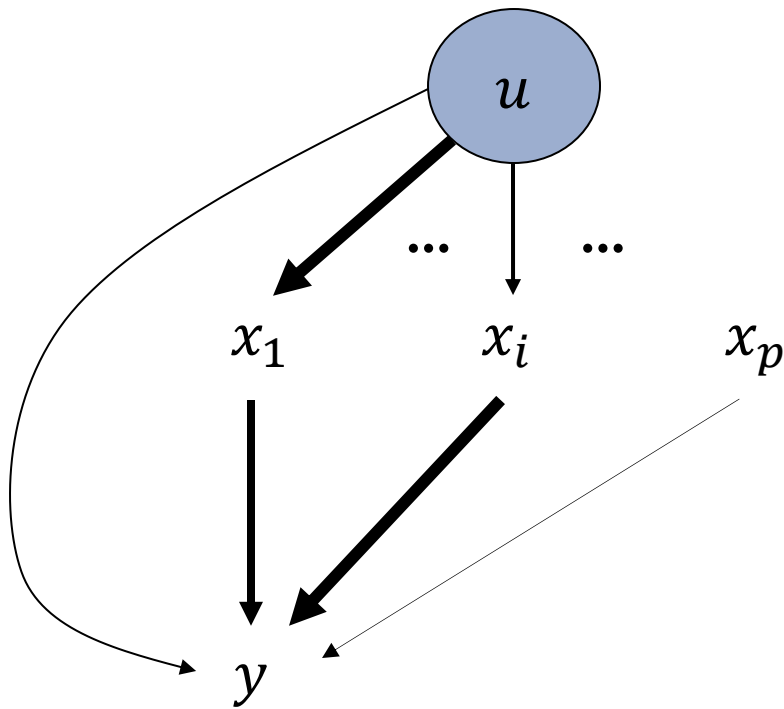
$$y = \underbrace{X\beta}_{\text{Fixed effect}} + \underbrace{u}_{\text{Polygene or Random effect}} + \underbrace{\epsilon}_{\text{Residual}}$$

*with $\epsilon \sim N(0, \sigma^2 I)$,
 $u \sim N(0, Z)$
 Z cov. matrix of r.e.*

Linear Mixed Models (LLMs)

The random effect term

How does the r.e. term acts in our model?



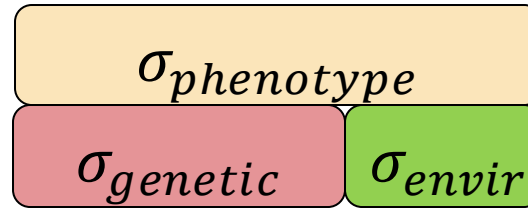
Pop.structure, ...

SNPs affected by u

Trait affected by $u + \text{SNPs}$

Linear Mixed Models (LLMs)

variances to consider



$$y = X\beta + \epsilon$$

To include genetic effect

$$y = X\beta + u + \epsilon,$$

$$Z = K = \text{Kinship matrix}$$
$$y = X\beta + u + \epsilon$$

$$\epsilon \sim N(0, \sigma_e^2 I) \quad (\text{residual env. effect})$$
$$u \sim N(0, \sigma_g^2 K) \quad (\text{genetic effect})$$

To include genetic effect
+
other covariates

$$Z_1 = K = \text{Kinship matrix}$$
$$Z_2 = Q = \text{covariates correlation}$$
$$y = X\beta + u_1 + u_2 + \epsilon$$

$$\epsilon \sim N(0, \sigma_e^2 I) \quad (\text{residual env. effect})$$
$$u_1 \sim N(0, \sigma_g^2 K) \quad (\text{genetic effect})$$
$$u_2 \sim N(0, \sigma_Q^2 Q) \quad (\text{covariates effect})$$

Linear Mixed Models (LLMs)

In practice - fill in the variables

$$y = X\beta + u_1 + u_2 + \epsilon \rightarrow \mathbf{y} = \mathbf{X}'\boldsymbol{\beta}' + \mathbf{u}$$

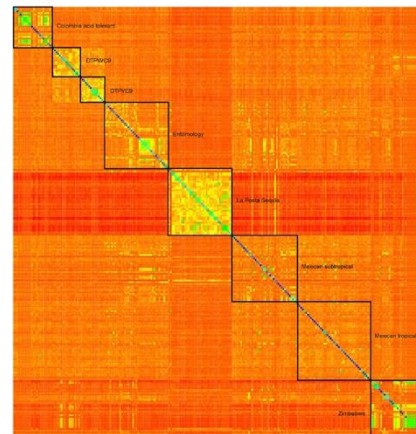
$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} & PC_{11} & PC_{12} & \dots \\ x_{21} & x_{22} & \dots & \dots & PC_{21} & PC_{22} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kn} & PC_{k1} & PC_{k2} & \dots \end{pmatrix}$$

$\underbrace{\hspace{10em}}_{\text{n SNPs (normalized) k samples}} \quad \underbrace{\hspace{10em}}_{\text{Other covariates (PCA, ...)}}$

$$\mathbf{u} \sim N(0, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) = N(0, \mathbf{V})$$

Total covariance matrix

Residual environmental effect and noise



Kinship matrix
(finer relatedness structure)
Easily calculated with plink, GCTA, ...

Linear Mixed Models (LLMs)

In practice - parameters not directly calculated, heritability

Very innocent-looking formula

$$y = X'\beta' + u$$
$$u \sim N(0, \sigma_g^2 K + \sigma_e^2 I)$$



Heritability comes into play

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{\sigma_g^2}{\sigma_{phenotype}^2}$$

Heritability = variance proportion explained only by genetic variance.

The fundamental parameter for phenotype prediction

What about the variances σ_g^2, σ_e^2

Linear Mixed Models (LLMs)

Some approaches

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta}' + \mathbf{u}$$

$$\mathbf{u} \sim N(0, V)$$

$$h^2 = \frac{\sigma_g^2}{\sigma_y^2} \rightarrow h^2 \sigma_y^2 = \sigma_g^2$$

y usually normalized so
 $\sigma_y^2 = 1$

BOLT-LMM
(Loh *et al.* 2015)

Optimizes

$$V = \sigma_g^2 K + \sigma_e^2 I$$

Through prior on sigma's.

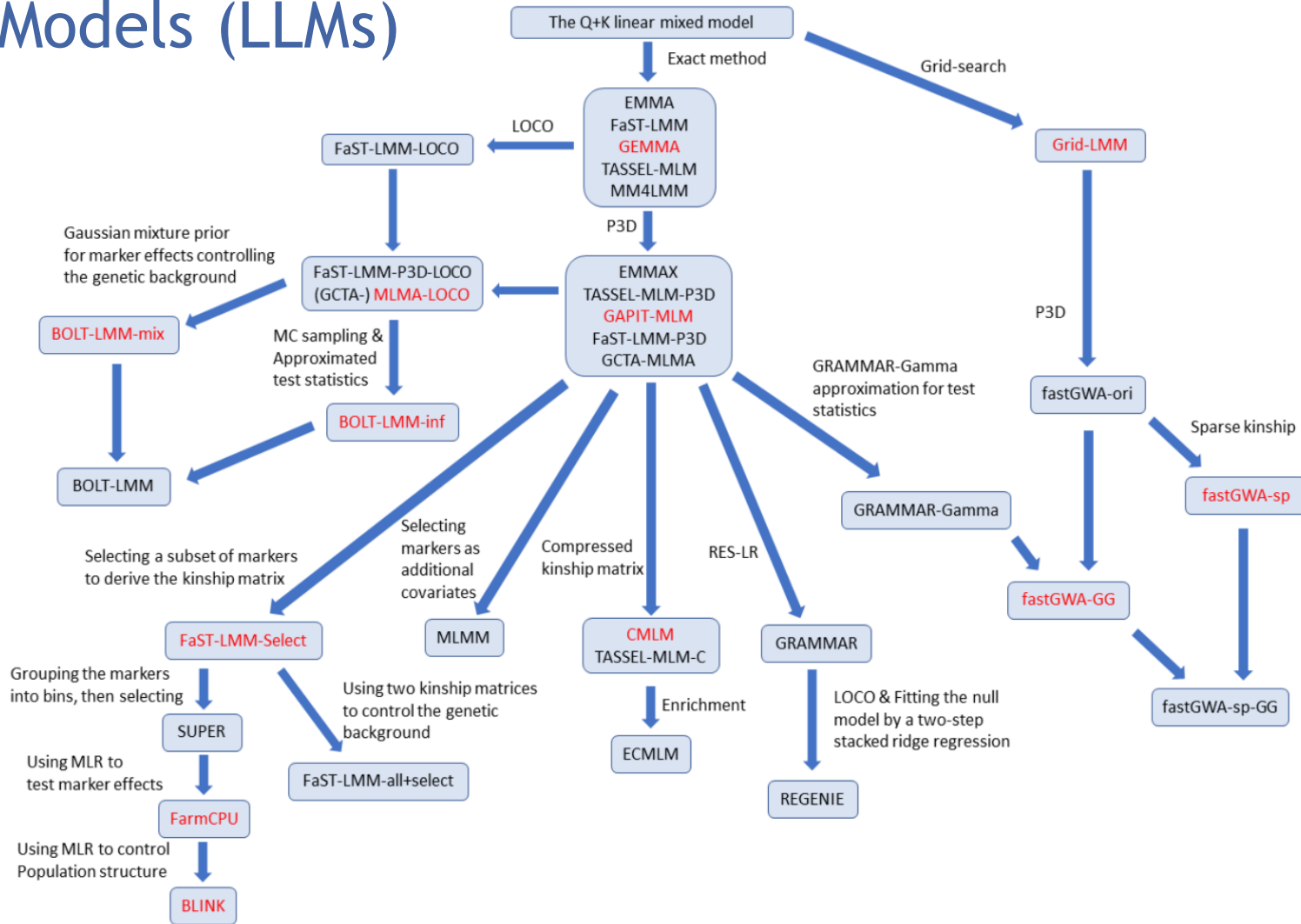
Then uses $h^2 = \frac{\sigma_g^2}{\sigma_y^2}$ to
define heritability.

Regenie
(Yang *et al.* 2011)

- Does not use K, but principal components
- Shrinks effect of SNPs to 0 to avoid overfitting
- Multiple other steps to avoid overfitting such as penalties and cross-validation
- Very fast and good for large studies with > Millions of SNPs

Linear Mixed Models (LLMs)

Some approaches



A phylogeny of 33 GWAS algorithms. If two algorithms are connected by an arrow, the target is based on the source with additional techniques indicated by the text. If two algorithms target the same algorithm, the target combines the techniques implemented by the two sources. P3D, population parameters previously determined; MC, Monte-Carlo; LOCO, leave-one-chromosome-out; MLR, multi-variate linear regression; RES-LR, using the residuals from the null model as the response to test marker effects in a simple linear model. From (Liu et al, 2023, bioArxiv. DOI 10.1101/2023.12.05.570105).

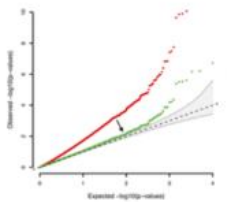
Beyond LLMs

New methods

New methods are

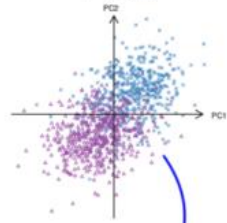
- Fast on large datasets
- Reliable in detecting association
- Use mixed models
- Have faster implementations

Genomic control



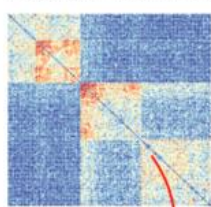
$$\hat{\beta} \rightarrow \hat{\beta}'$$

PCA



$$Y_i = \beta X_i + \gamma \bar{PC}_i + \epsilon_i$$

Mixed models



$$Y_i = \beta X_i + \eta_i + \epsilon_i$$

From basic Genomics Control (rescaling test statistics) to correcting through PCA only and to Mixed Models, of which LMMs are a special case. Credit Iain Mathieson.

Some examples

[LDAK-KVIK \(Hof and Speed, 2024\)](#)

Uses mixed models: often the preferred tool, are more flexible and can be more complex than LMMs. Faster and outperforming REGENIE, BOLT-LMM

[Quickdraw \(Loya et al, 2025\)](#)

Shrinks variant effects to increase association power, computationally efficient with variational inference and GPU calculations. It also uses mixed models.

Meta studies

- Individual genetic variants often have **small effects**.
- **Large sample sizes** are required to detect novel associations.
- **Low minor allele frequency (MAF)** reduces statistical power.
- **Combining individual level data** is technically and administratively challenging (large dataset sizes, variations in study designs, and data protection constraints)

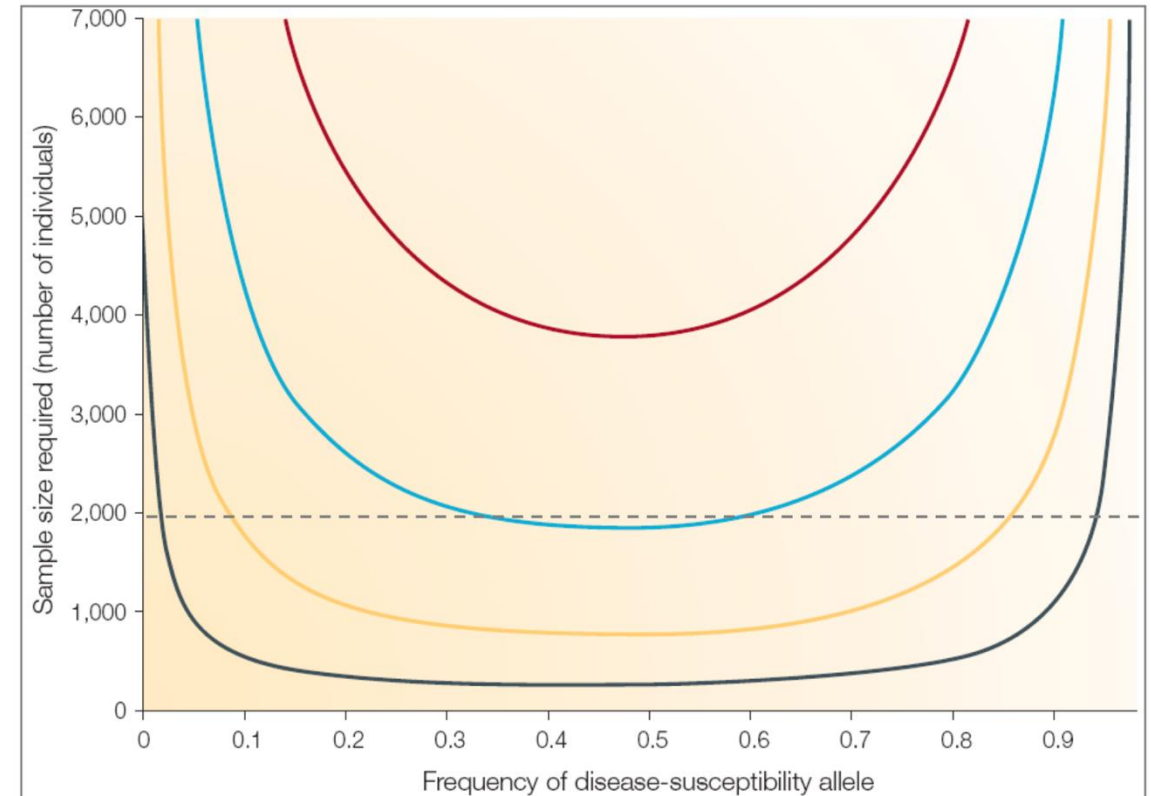


Figure 1 | **Effects of allele frequency on sample-size requirements.** The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown. Numbers shown are for a statistical power of 80% at a significance level of $P < 10^{-6}$, assuming a multiplicative model for the effects of alleles and perfect correlative linkage disequilibrium between alleles of test markers and disease variants.

Wang et al. 2005



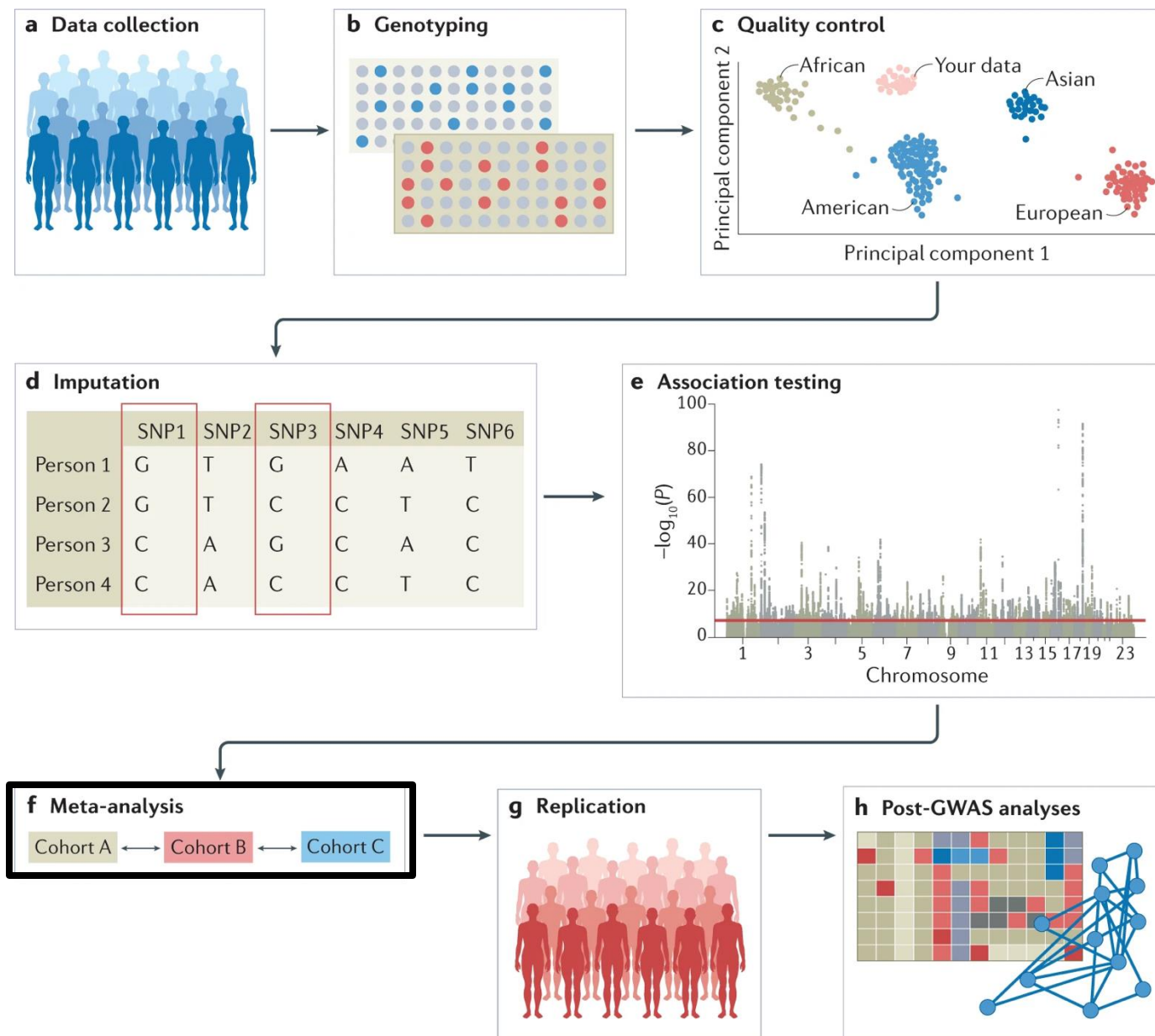
Meta studies

GWAS summary statistics are publicly available:

- Meta-studies integrate those summary statistics
- Increased statistical power as sample size increases

Softwares:

- METAL
- GWAMA
- MANTRA



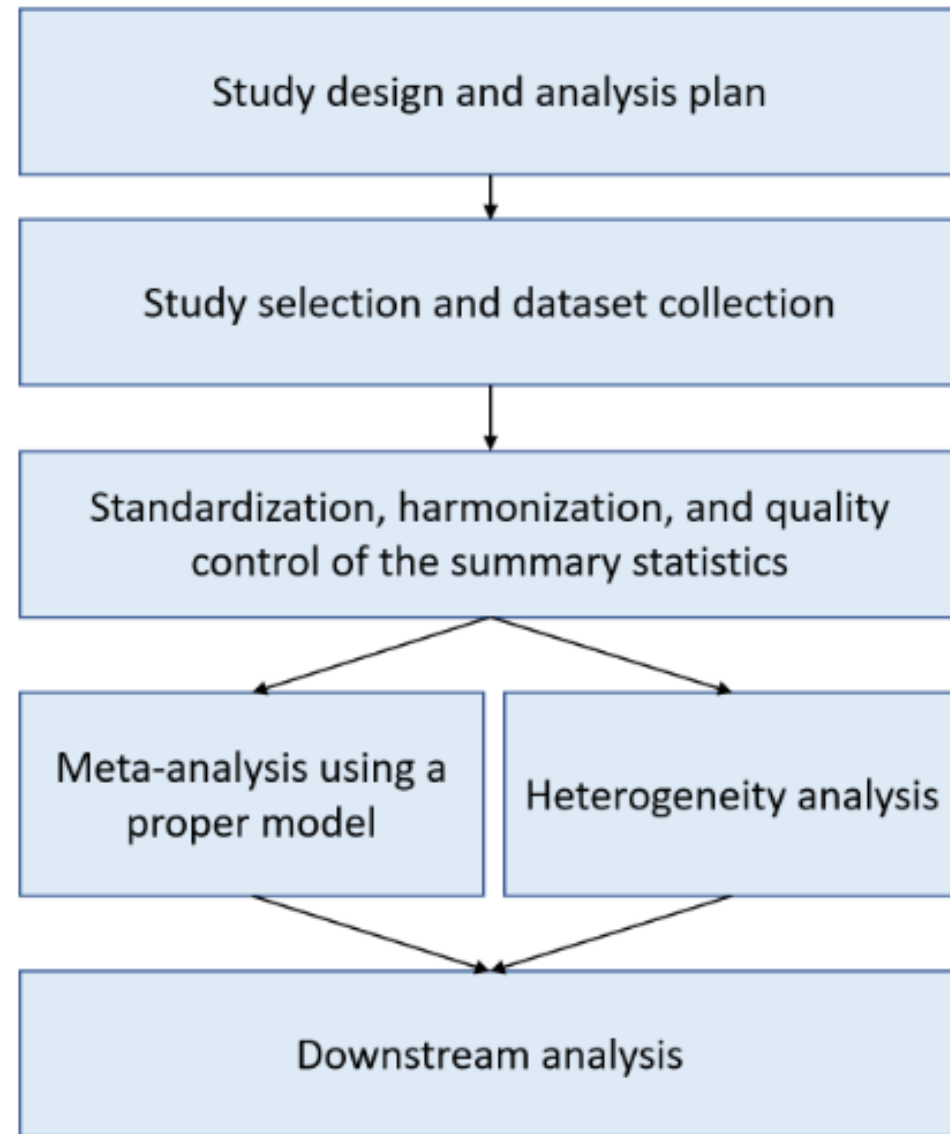
Meta studies

GWAS summary statistics are publicly available:

- Meta-studies integrate those summary statistics
- Increased statistical power as sample size increases

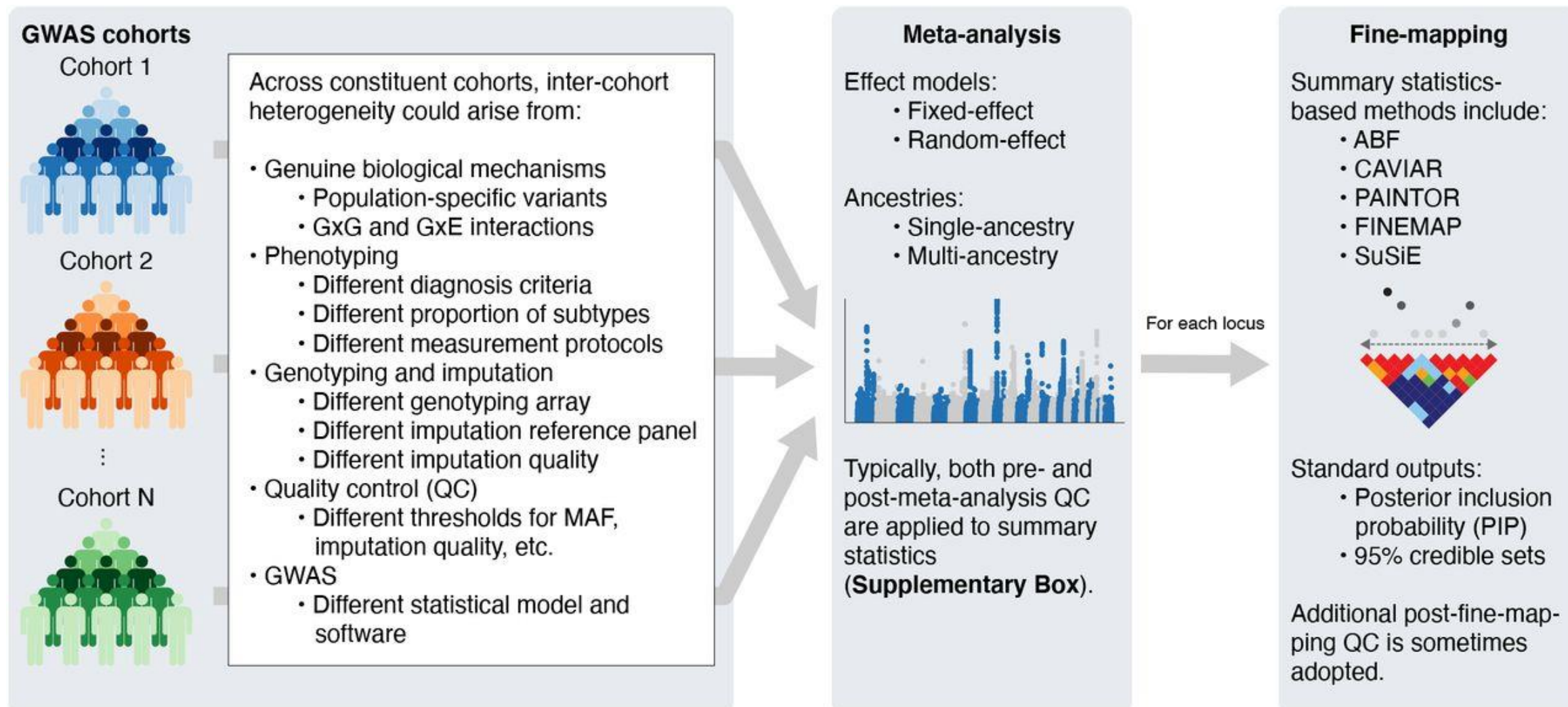
Softwares:

- METAL
- GWAMA
- MANTRA



Credit: Yunye He

Meta studies



Meta-analysis

Approaches

Fixed Effects

- Most commonly used and most powerful for discovery when assuming a consistent effect of each risk allele across datasets.
 - Inverse variance weighting is the most common method.
 - Sample size weighting (z-score based) is also widely used.

Random Effects

- Less common but useful for assessing the generalizability of associations.
- Estimates the average effect size and its uncertainty across different populations.

Bayesian Approaches (rarely used)



Meta studies

Quality control is crucial!

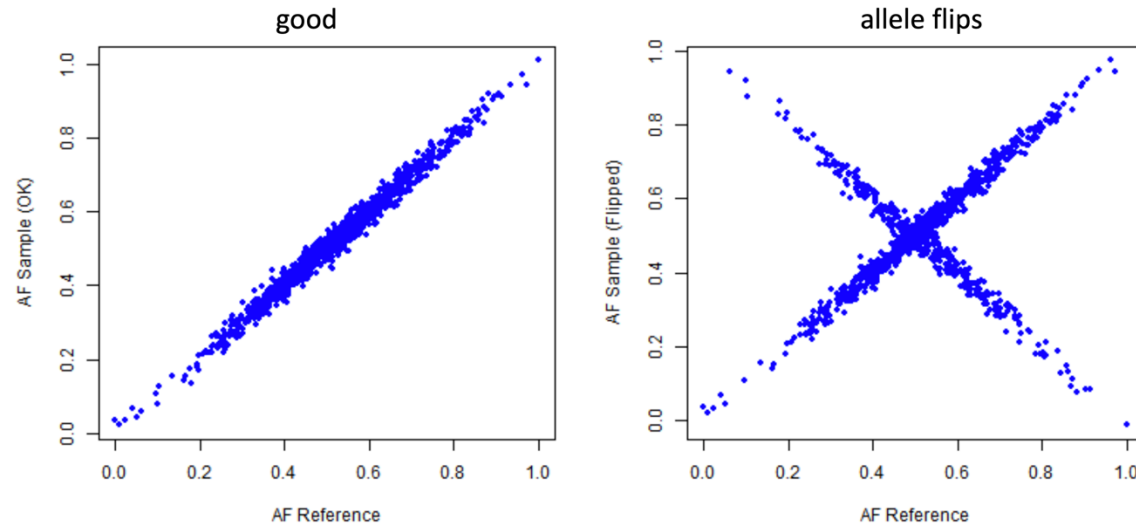
- Rigorous QC on the individual GWAS results
- Exclude rare variants and poorly imputed variants
- Control for population stratification and ancestry differences
- Verify input data and identify differences (tools: GWAtoolbox, EasyQC, GWASinspector)
- Harmonization of the data (effect allele polarization)
- Perform both fixed effects approaches and compare the results
- As in GWAS, QQ and Manhattan plots are important.



Quality control

Allele flipping

Effect allele must be the same across GWAS studies.
How does it look if the effect direction is not the same?



Meta-analysis software

Most commonly used software for common variant analysis: METAL

- Automatic strand flipping of non-ambiguous SNPs
- Calculation of max/min/mean allele frequency
- Inverse variance & sample size weightings
- Automatic genomic control correction
- Heterogeneity tests

Link: www.sph.umich.edu/csg/abecasis/metal/

Documentation: genome.sph.umich.edu/wiki/Metal_Documentation



Meta-analysis example!

Setup

Modify files to include:

- all information
- consistent **marker name**

Tools: WAtoolbox,
EasyQC, GWASinspector

Input: Script file

```
# Execute analysis on 2 studies
# GENOMICCONTROL ON
# SCHEME STDERR

#-- DESCRIBE AND PROCESS 1st FILE --
MARKER SNP
ALLELE REF_ALLELE OTHER_ALLELE
EFFECT BETA
PVALUE PVALUE
WEIGHT N
STDERR SE
PROCESS gwas1.txt.gz

#-- DESCRIBE AND PROCESS 2nd FILE --
MARKER SNP
ALLELE A1 A2
EFFECT EFFECT1
PVALUE pvalue
WEIGHT N
STDERR SE
PROCESS gwas2.txt.gz

OUTFILE META_GWAS1-2
MINWEIGHT 10000
ANALYZE HETEROGENEITY
```

Running METAL

META_GWAS1-2.TBL.INFO

This file contains a short description of the columns
meta-analysis summary file, named 'META_GWAS1-2.TBL'

Marker - this is the marker name
Allele1 - the first allele for this marker in the first file where it occurs
.
.
Input for this meta-analysis was stored in the files: # --> Input File 1 :
gwas1.txt.gz
--> Input File 2 : gwas2.txt.gz

META_GWAS1-2.TBL

MarkerName	Allele1	Allele2	Weight	Zscore	P-value	Direction
rs560887	t	c	6806	-7.075	1.491*10 ⁻¹²	---
rs853787	t	g	6806	6.691	2.221*10 ⁻¹¹	+++
rs853789	a	g	5339	-6.597	4.189*10 ⁻¹¹	?--
rs853773	a	g	6806	-6.132	8.662*10 ⁻¹⁰	---
rs537183	t	c	6806	6.007	1.887*10 ⁻⁹	+++
rs557462	t	c	6806	6.005	1.917*10 ⁻⁹	+++
rs502570	a	g	6806	-6.001	1.955*10 ⁻⁹	---
rs563694	a	c	6806	5.975	2.300*10 ⁻⁹	+++
rs475612	t	c	6806	-5.867	4.423*10 ⁻⁹	---
rs853781	a	g	6806	-5.844	5.092*10 ⁻⁹	---



JOURNAL ARTICLE

GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies FREE

Stephanie M. Gogarten ✉, Tushar Bhangale, Matthew P. Conomos, Cecelia A. Laurie, Caitlin P. McHugh, Ian Painter, Xiuwen Zheng, David R. Crosslin, David Levine, Thomas Lumley ... Show more

[Author Notes](#)

Bioinformatics, Volume 28, Issue 24, December 2012, Pages 3329–3331, <https://doi.org/10.1093/bioinformatics/bts610>

- Ensures consistency of input file columns
- Compares effect size distributions across cohorts
- Harmonized header and separator across input files
- Calculated effective N and corrects for genomic control



