SUMMARY OF GWAS COURSE

Matti Pirinen
University of Helsinki
2.5.2023

GENOME-PHENOME ASSOCIATION

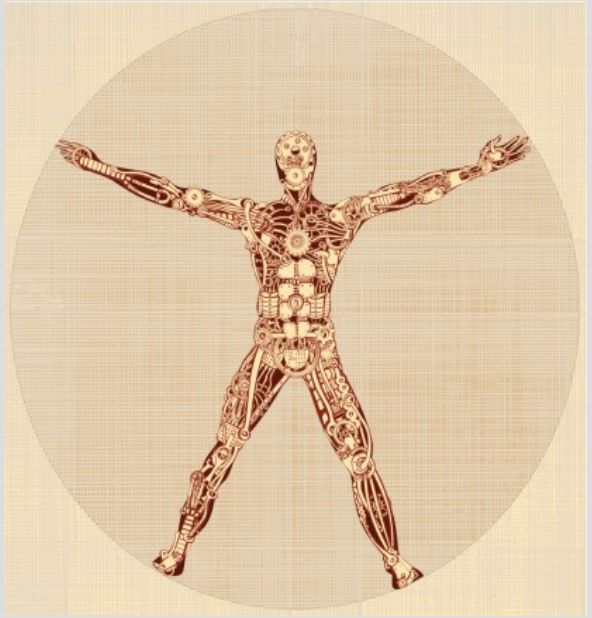
Genome

Phenome = all **phenotypes** combined

Variation in the Genome between individuals.

"genome-wide" studies consider variation in millions of positions

association ?



Measurable traits (blood pressure)

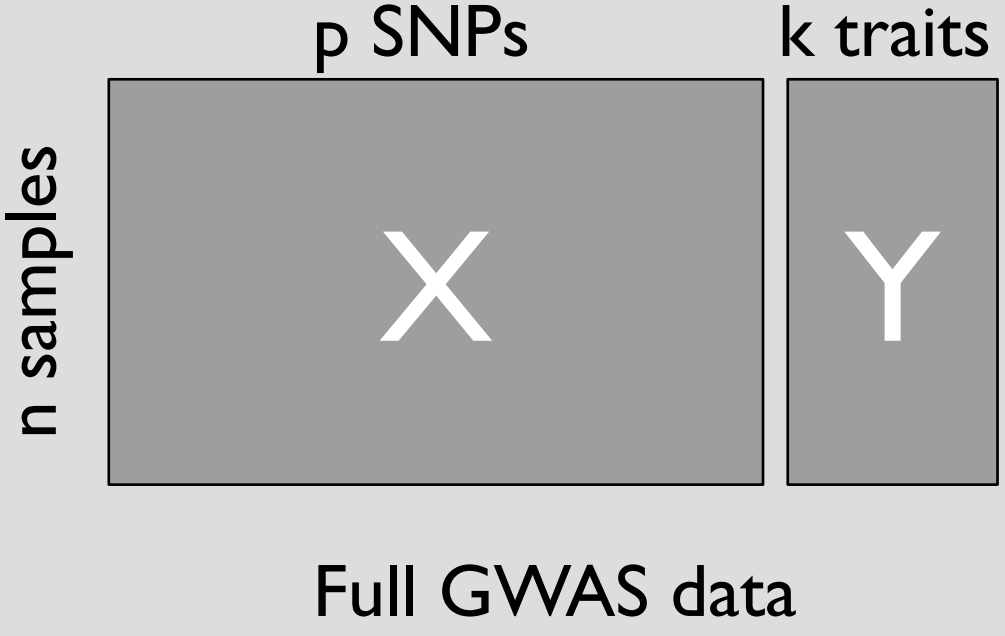
Disease status (MS-disease, diabetes)

Behavior (chronotype, smoking)

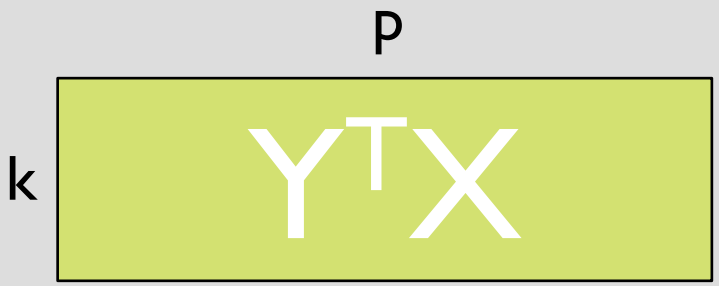
Statistical association can

- allow predicting one from the other
- suggest causal links between the two

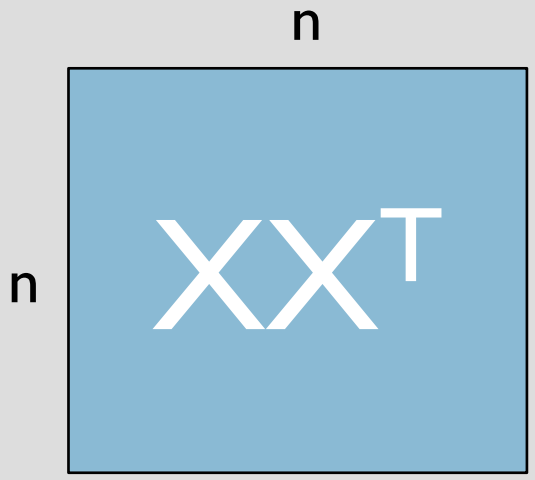
GWAS IN MATRICES



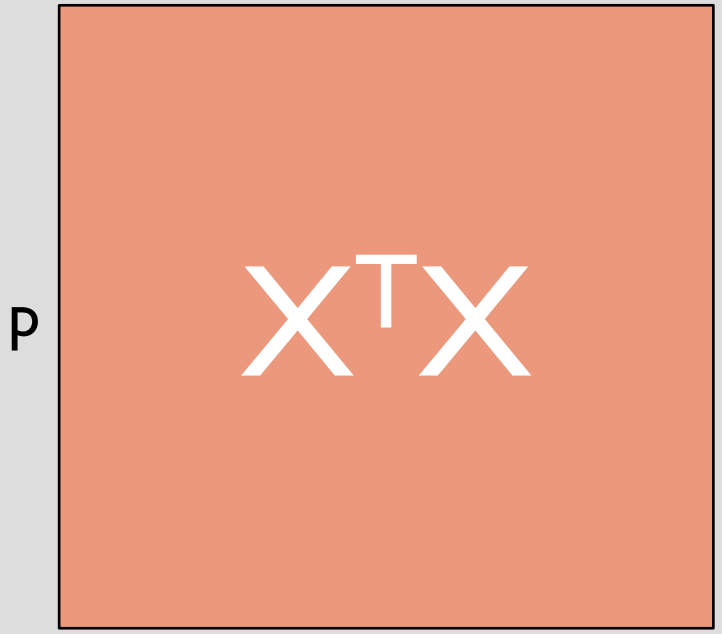
Full GWAS data



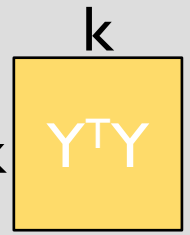
GWAS results (betas, SE, P-value)



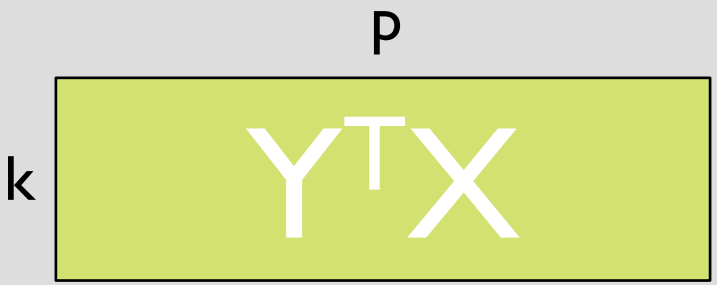
sample relatedness



LD matrix (correlation of variants)



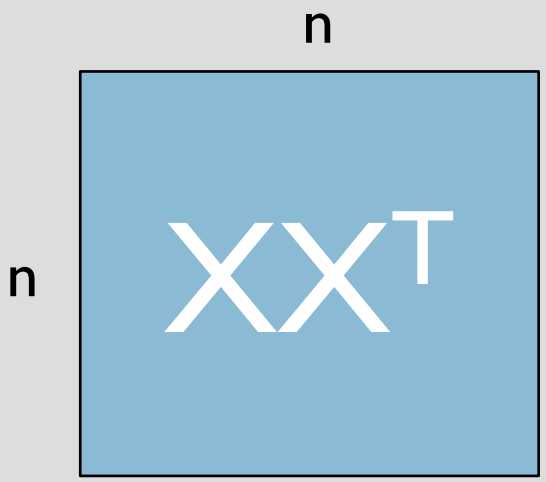
trait correlation



GWAS results (beta, SE, P-value)

Weeks 1-7:

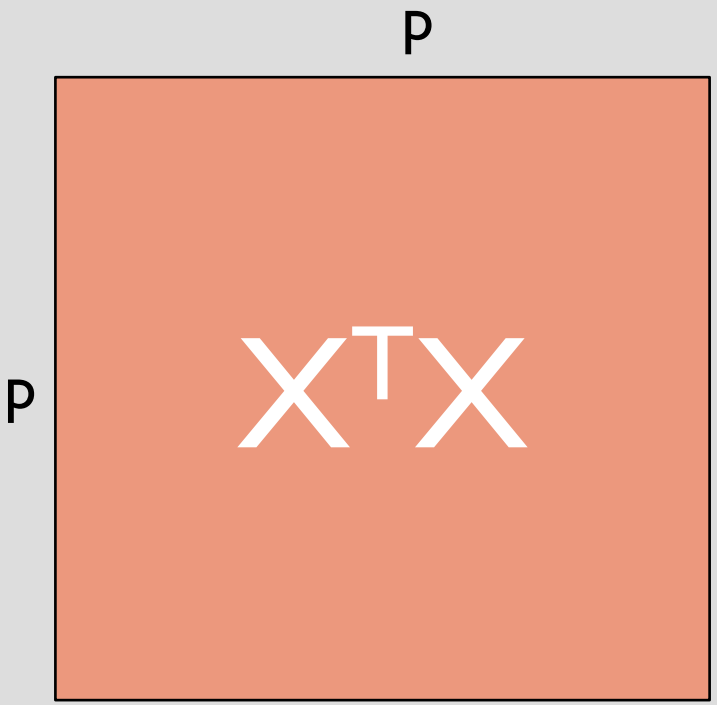
statistical inference,
statistical power,
confounders,
covariates,
summary statistics,
meta-analysis
polygenic scores



sample relatedness

Weeks 3, 5:

Relatedness & population structure
Heritability & mixed models



LD matrix (correlation of variants)

Weeks 4,5:

Haplotypes & linkage disequilibrium
Stepwise search & fine-mapping
LD-score regression

GWAS PARAMETERS

- β and $\hat{\beta}$, marginal effect size, scaled versions β^*
- λ and $\hat{\lambda}$, causal effect size, scaled versions λ^*
  - λ is also used for genomic control parameter in QQ-plots
- SE, standard error of effect sizes
- σ^2 error variance of linear regression model
- R^2 variance of phenotype explained by regression model
- τ^2 (prior) variance of a non-zero effect size in Bayesian models and in LD-score regression
- $\boldsymbol{R}$ LD-matrix of pairwise correlation between variants
- r LD between pair of variants and r^2 the squared LD
- h^2 heritability due to additive effects (for a variant, a region or whole genome)

STEPS OF A GWAS

Study design

1. Is the phenotype heritable?
2. Which set of samples is needed for a GWAS?

Running a GWAS

1. Regression model & covariates
2. Diagnostics

Downstream analyses

1. Conditional analyses & fine-mapping
2. (Other typical analyses we haven't studied on this course)

Replication & Meta-analysis

1. Does it replicate?
2. What is the combined evidence?
3. Relationship to other phenotypes?

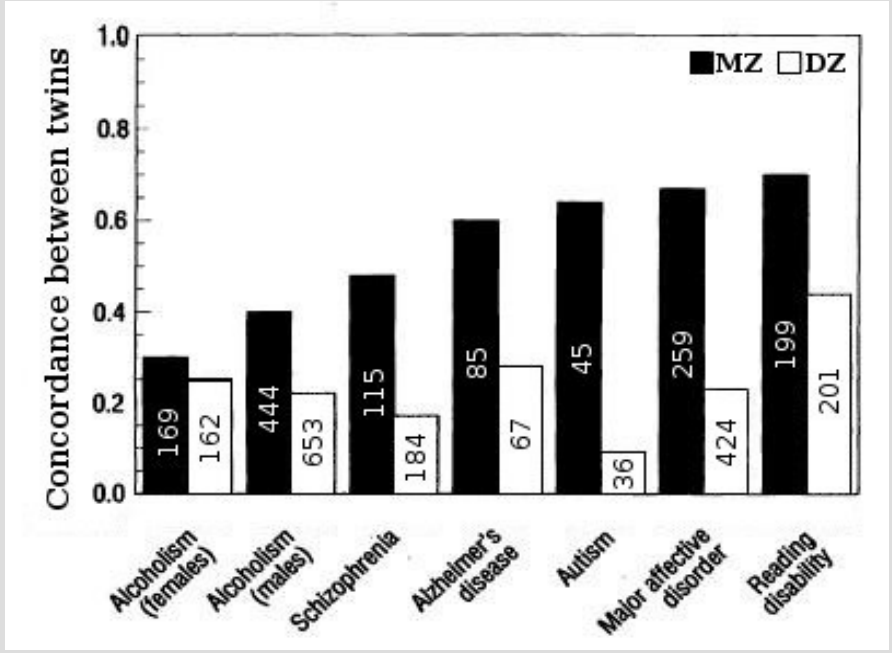
Further applications

1. Polygenic scores
2. (Mendelian randomization)

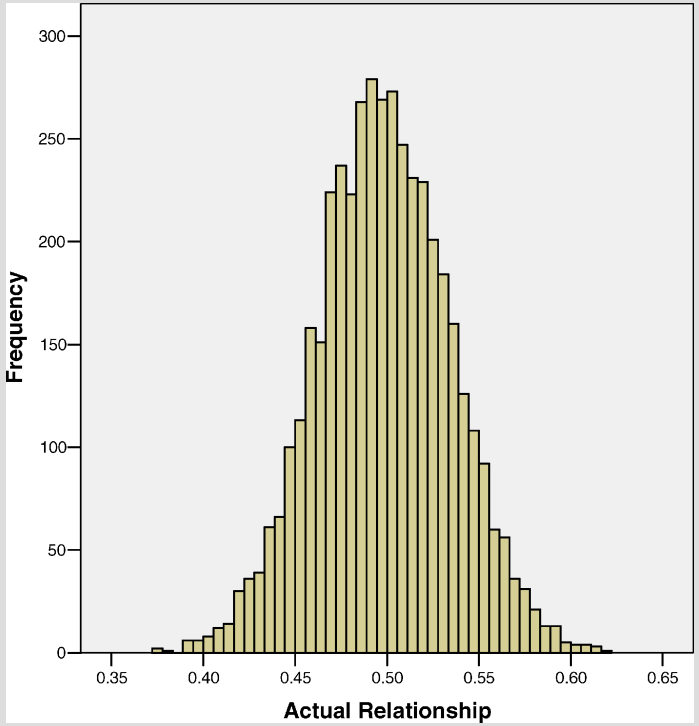
HERITABILITY

- Proportion of phenotypic variance explained by variation in genome
- Depends on the population and point of time because environmental variance can vary
  - Measurement accuracy can affect heritability
- Narrow sense heritability h^2: variance explained by the additive effects of the variants
  - Gives an upper bound for variance explained by polygenic scores
- Broad sense heritability H^2 : variance explained by all genetic variation

HOW HERITABILITY HAS BEEN ESTIMATED ?



Compare concordance in monozygotic twins (share full genome) to that of dizygotic twins (share ~50%). Under (strong) assumptions, the difference estimates heritability.

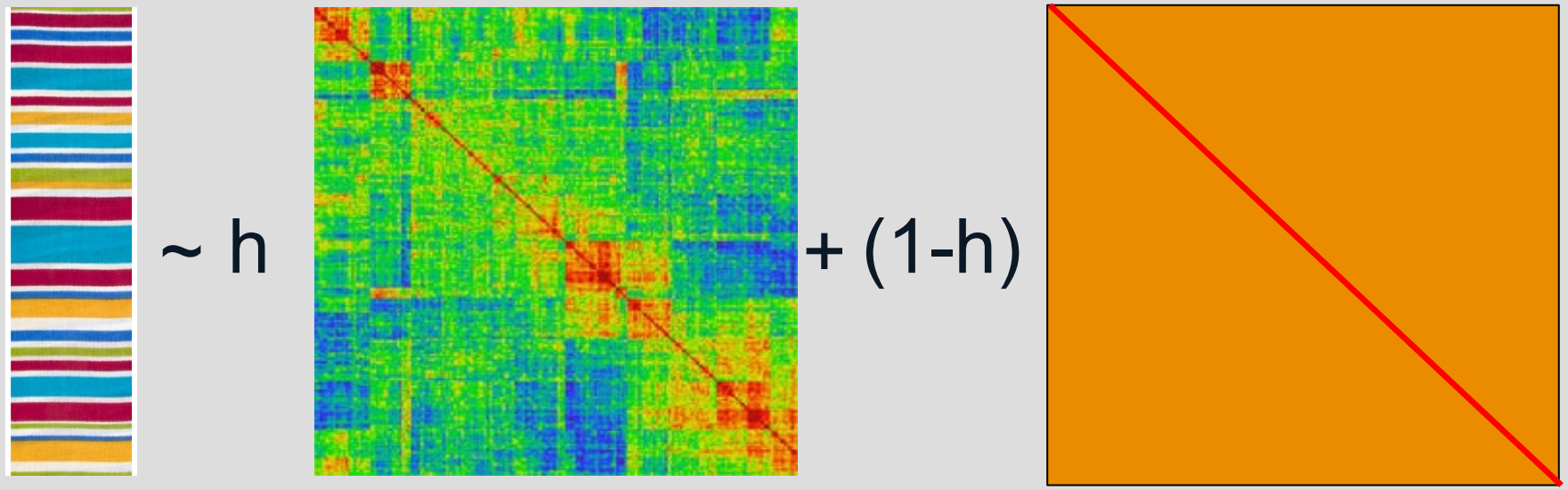


Do full-sib pairs that share more of their genomes also have more similar phenotypes?
Heritability estimate for height from 3375 pairs of sibs was 0.80 (0.46 – 085). (Visscher et al. 2006 PLoS Genetics)

LINEAR MIXED MODEL TO ESTIMATE HERITABILITY

$$\mathbf{Y} \sim \mathcal{N}(0, h\mathbf{R} + (1-h)\mathbf{I})$$

For height in Finns
we estimate h ~ 50%



Y

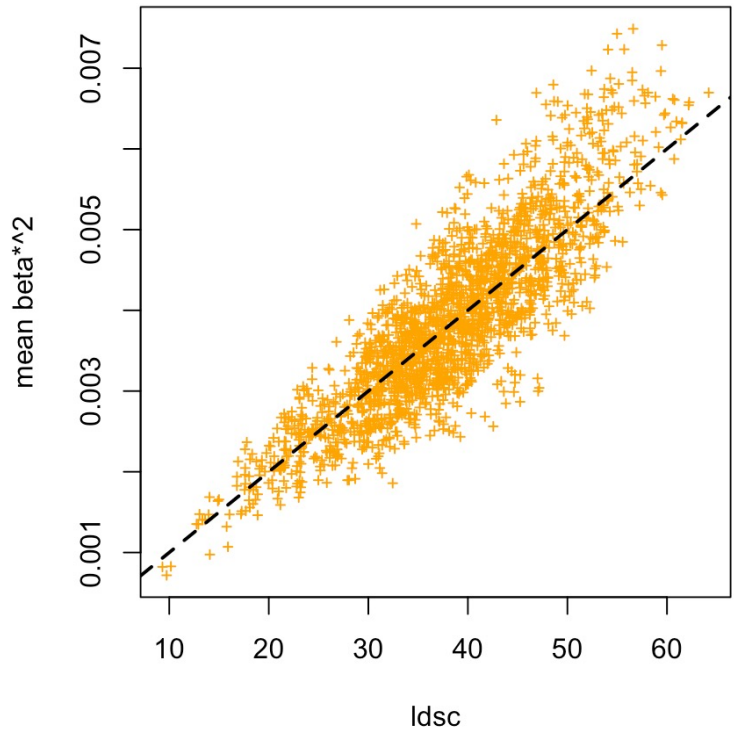
SNP relatedness

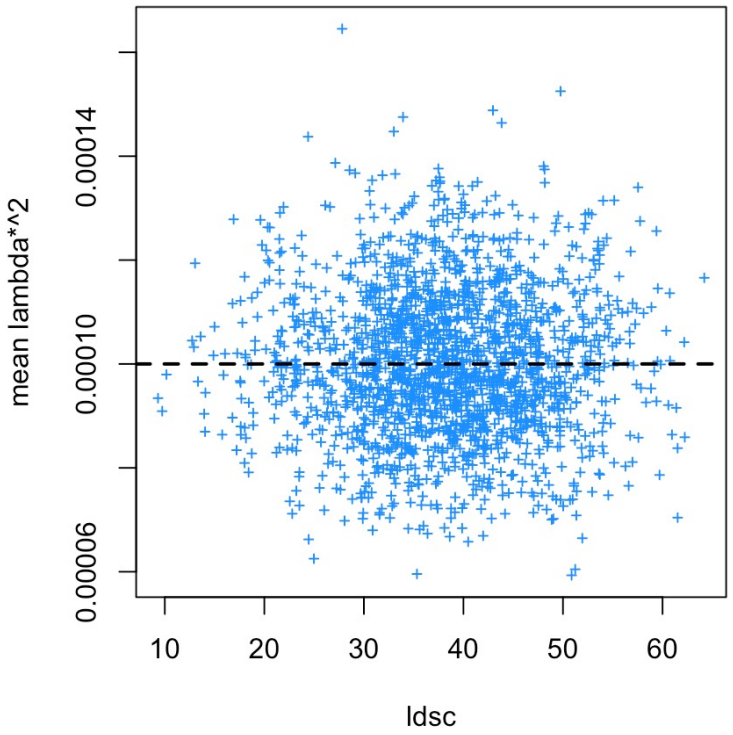
Independent environment

variation in phenotype ~ pattern of genetic similarity + random noise uncorrelated between individuals

Parameter h measures how well phenotypic variation is explainable by pattern of genetic similarity

LD-SCORE REGRESSION

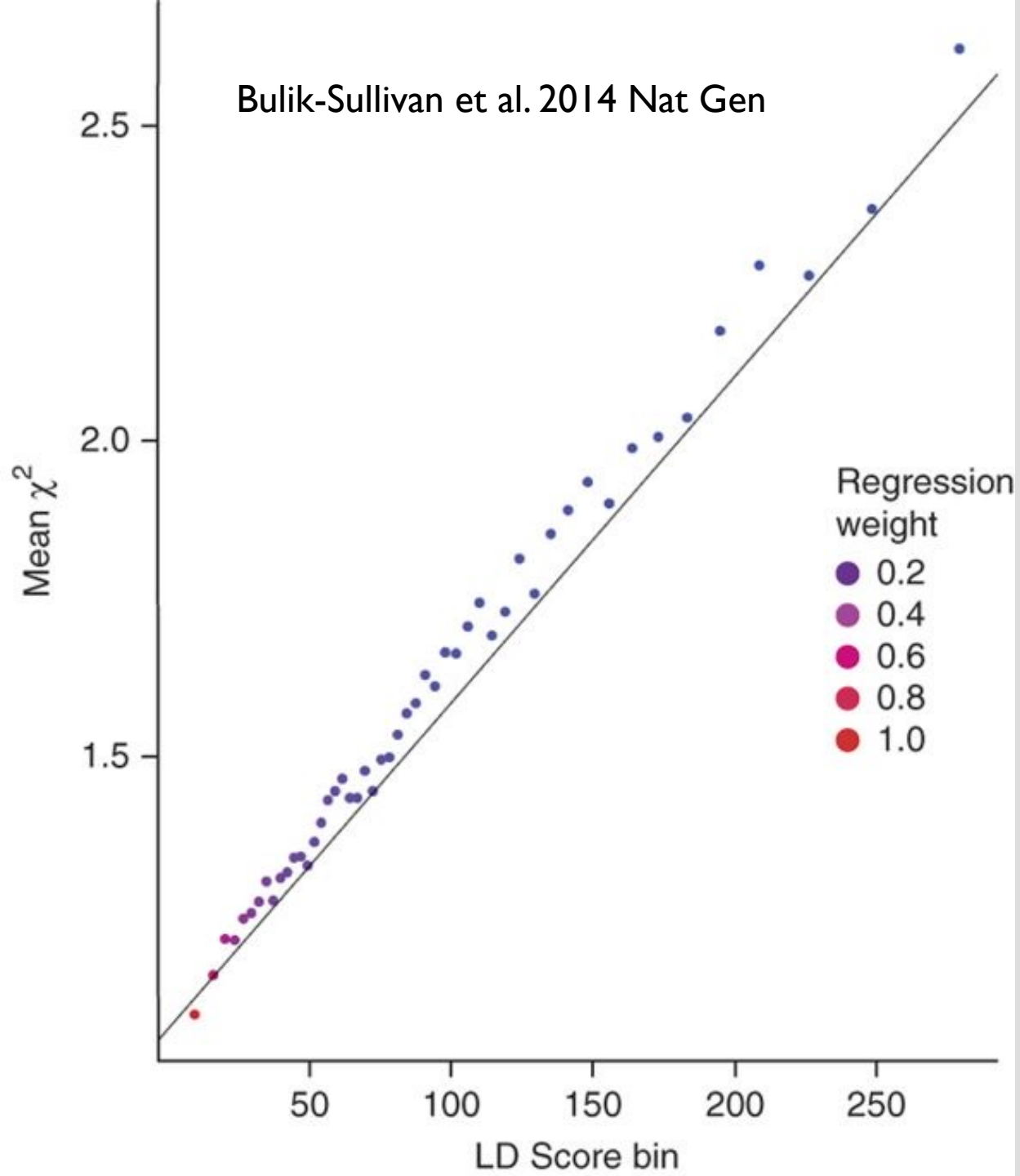




Now there is a clear difference! For the squared marginal effects we expect that the slope of the regression line estimates τ^2 and intercept estimates 0. The squared causal effects are independent of LDSC and their expectation equals to their variance τ^2, so the line has intercept τ^2 and slope 0.

Under extreme polygenicity, where each variant contributes an effect size from $\mathcal{N}(0, h^2/p)$, we estimate that $h^2 = p\,\tau^2$.

$r_{l+}^2 = \sum_{k=1}^{p} r_{lk}^2$ is the **LD-score** of SNP l.



LDSC ON SCHIZOPHRENIA GWAS RESULTS

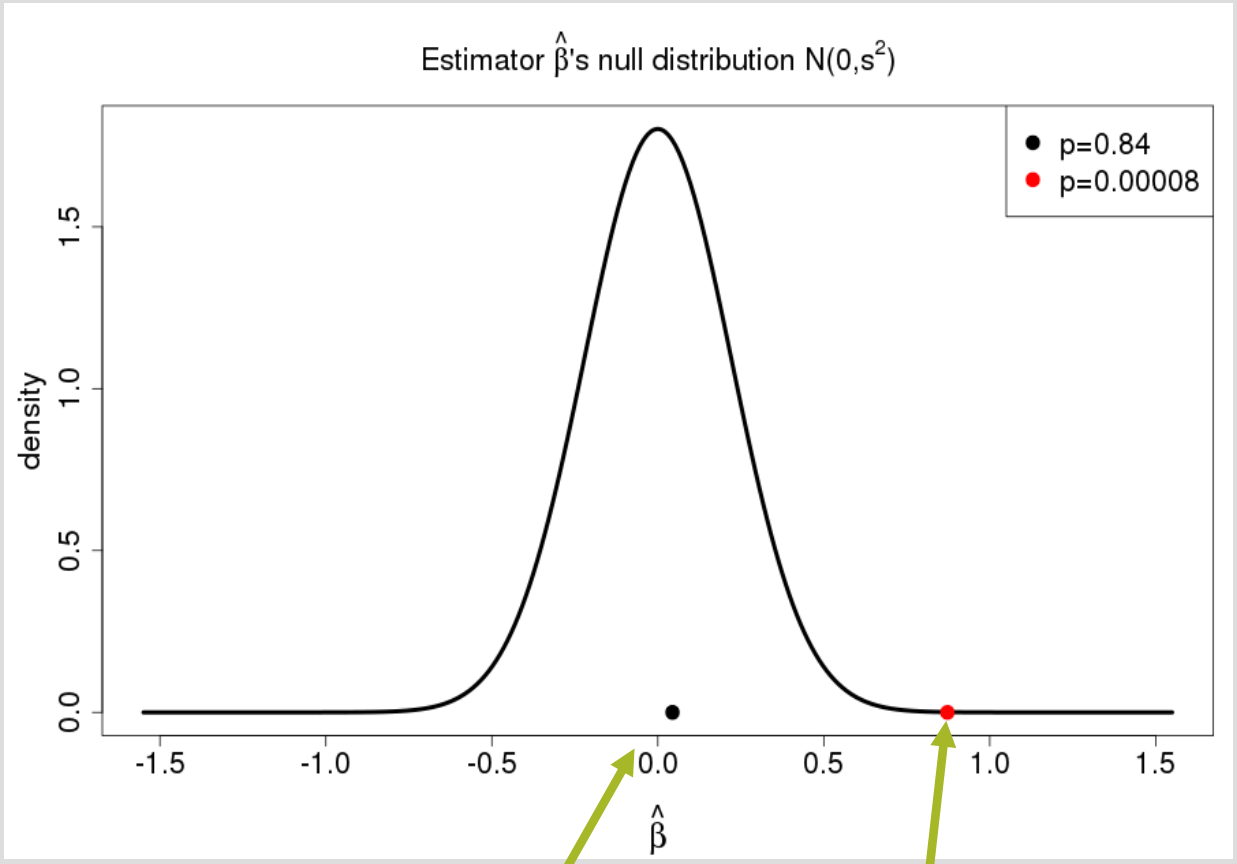
Each point represents an LD score quantile, where the *x* coordinate of the point is the mean LD score of variants in that quantile and the *y* coordinate is the mean χ^2 statistic of variants in that quantile in the most recent schizophrenia meta-analysis. Colors correspond to regression weights, with red indicating large weight and blue indicating small weight. The black line is the LD score regression line. The line appears to fall below the points on the right because this is a weighted regression in which the points on the left receive the largest weights.

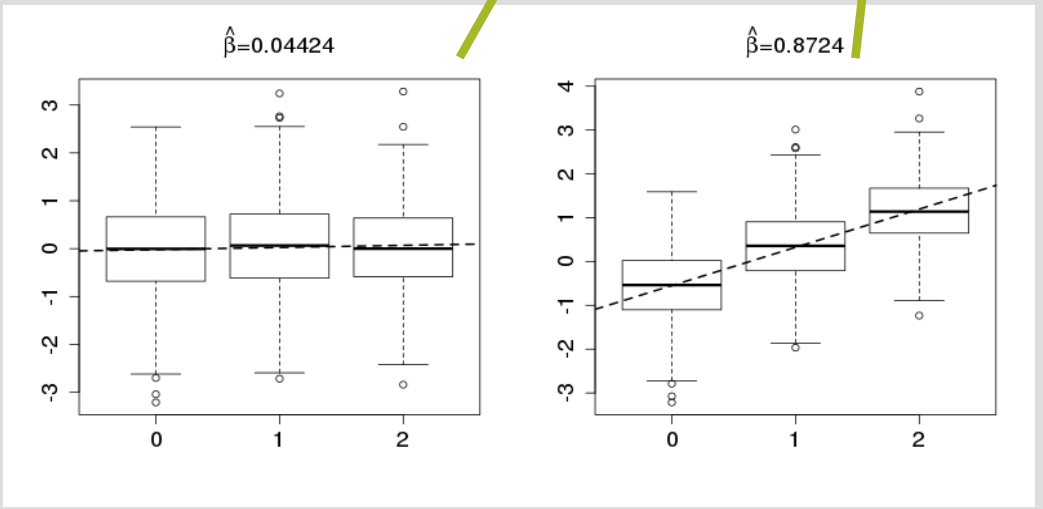
WHICH SET OF SAMPLES FOR A GWAS?

- Definition of phenotype
  - Measurement process for quantitative traits?
    - Measurement accuracy, measurement bias
  - Case and control definitions for binary traits?
    - Selection bias?
    - Bias from different processing of cases and controls
- Statistical power
  - Which kind of effects could / should we found?

REMINDER: P-VALUE

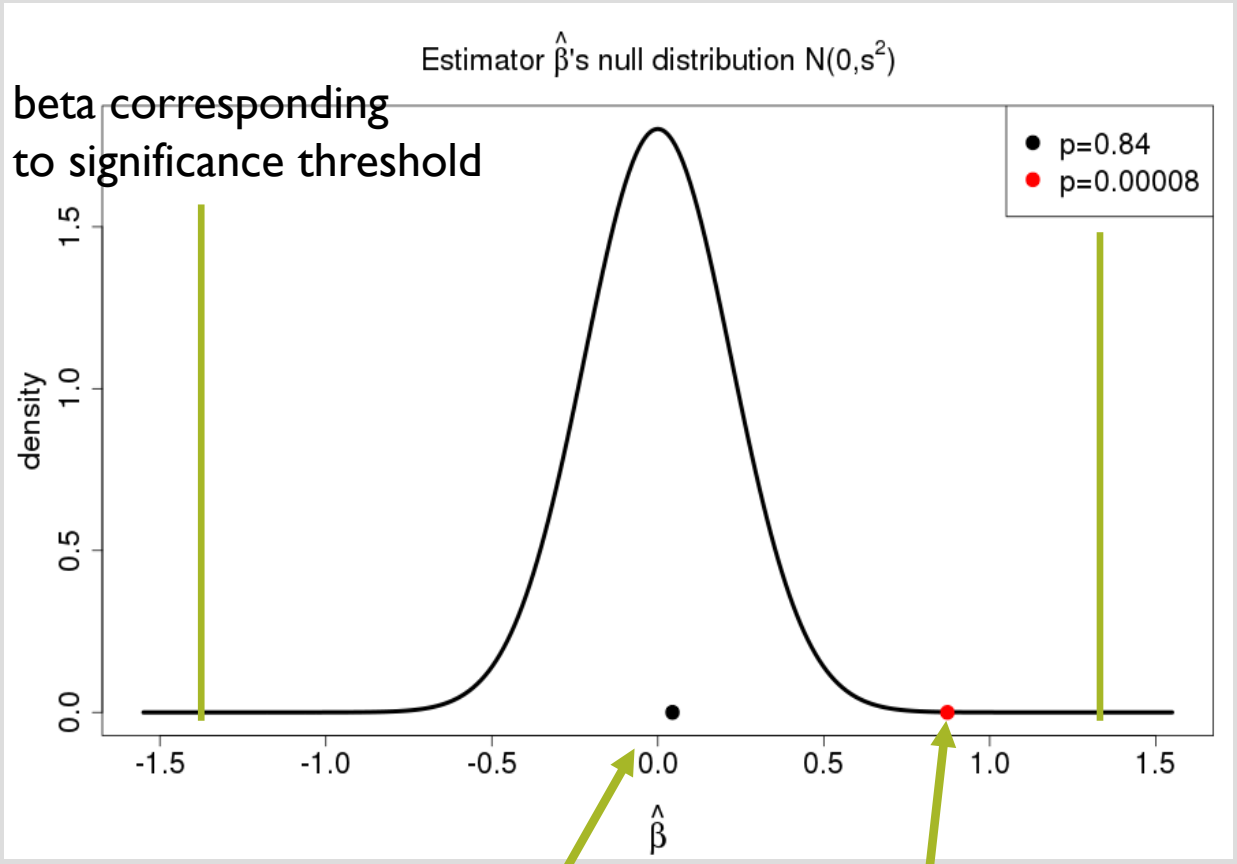
- Is the observed slope plausible if true slope = 0 ?
- *P*-value: Probability that we get at least as extreme estimate as we have observed, if true slope = 0
- *P* = 0.84: No evidence for deviation from null
- *P* = 8e-5: Unlikely under the null → maybe not null

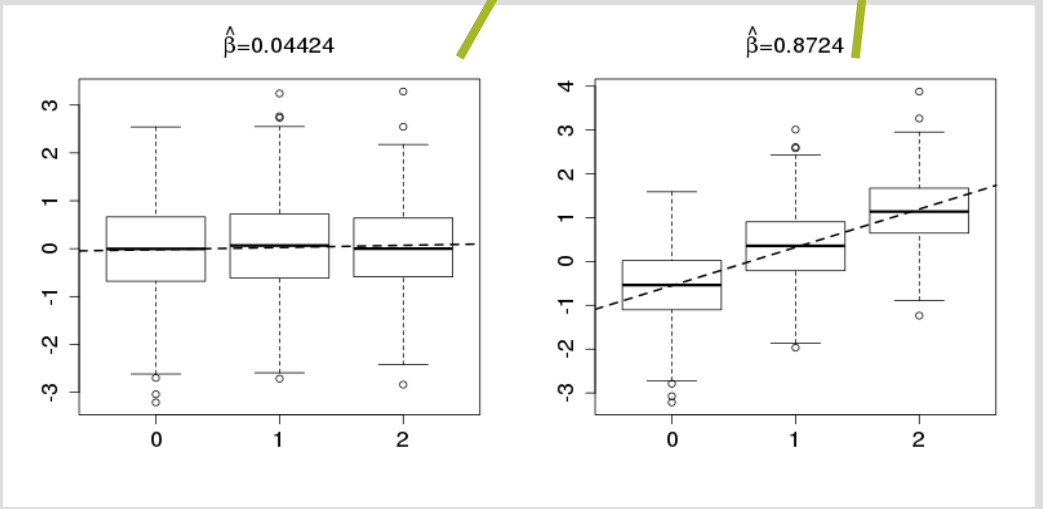




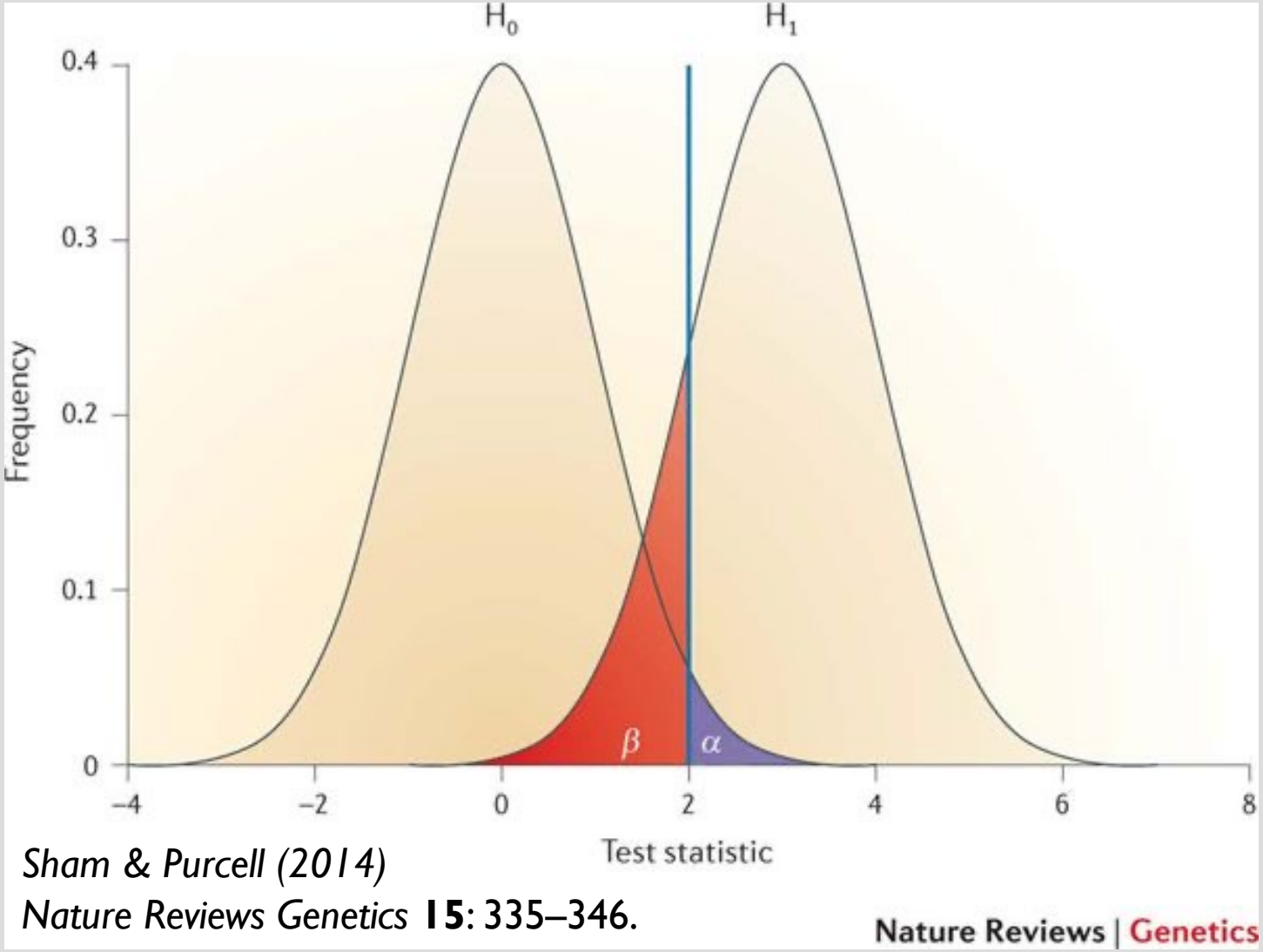
SIGNIFICANCE THRESHOLD & POWER

- Significance threshold α = Probability that a null variant has *P*-value $\leq \alpha$
- What is the probability that a non-null variant has *P*-value $\leq \alpha$?
  - Depends on the properties of the variant and study
  - Is called statistical power of the significance test





TYPE I AND TYPE II ERRORS AND POWER



Sham & Purcell (2014)
Nature Reviews Genetics **15**: 335–346.

The probability distributions of test statistic under H_0 and H_1, the critical threshold for significance (blue line), the probability of type 1 error (α; purple) and the probability of type 2 error (β; red).

Type I error: "false positive", wrongly reject H_0 when H_0 holds. Making significance level very low **avoids** Type I errors.

We can lower α by dragging blue line to right.

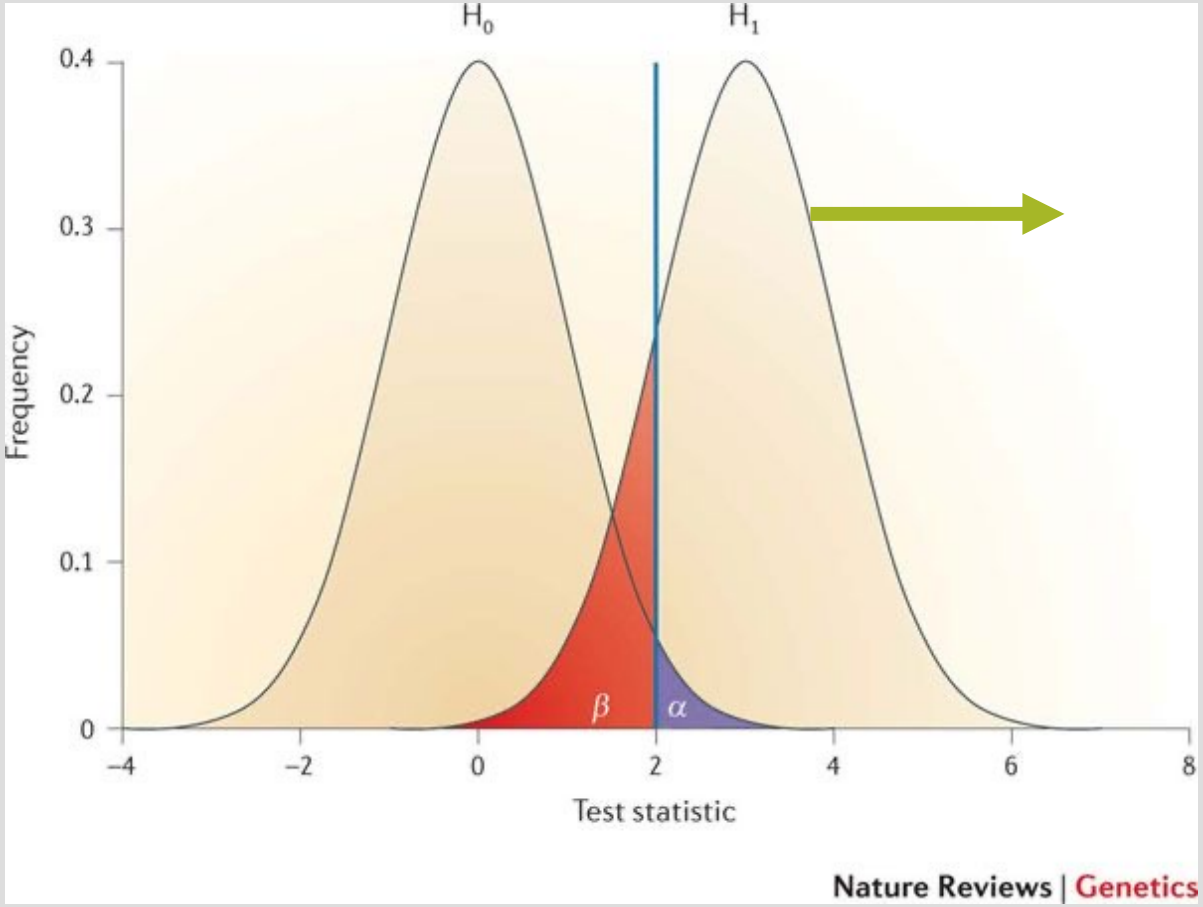
Type II error: "false negative", wrongly accept H_0 when H_0 is not true.
Making significance level very low **creates** Type II errors.

Power = 1- β = P(reject H_0 | H_1 true).

WALD TEST

- Assuming that the GWAS model is correct (i.e., there are no biases), the regression coefficient estimator $\hat{\beta} \sim N(\beta, SE^2)$
- Wald statistic $\mathrm{z} = {\hat{\beta}}/{SE} \sim N\left(\frac{\beta}{SE}, 1\right)$
  - z ~ *N*(0,1) under the null ($\beta = 0$), and this is how we compute P-values
  - Under the alternative hypothesis, the mean of the distribution of z depends on true β and *SE*
- Chi-square statistic $z^2 \sim \chi_1^2\left(\mathrm{NCP} = {\beta^2}/{SE^2}\right)$, where NCP is the "non-centrality parameter"
  - General definition: When $Y \sim N(\mu, \sigma^2)$ then $\frac{Y^2}{\sigma^2} \sim \chi_1^2\left(\mathrm{NCP} = {\mu^2}/{\sigma^2}\right)$
  - $z^2 \sim \chi_1^2$ under the null, i.e., the central (NCP = 0) chi-square distribution with 1 df

$$Z = {}^{\hat{\beta}}/_{SE} \sim N\left(\frac{\beta}{SE}, 1\right)$$



- The alternative's test statistic distribution will move farther from the null distribution when $|\beta|/SE$ grows
- For a fixed significance threshold, the power will thus increase as $|\beta|$ increases or as SE decreases
- Makes sense:
  - "Larger effects are easier to find"
  - "More precise estimates help separating real effects from noise"

FORMULAS FOR SE

- Liner model GWAS has $\mathrm{SE} \approx \frac{\sigma}{\sqrt{2\,n\,f\,(1-f)}}$
- Logistic model GWAS has $\mathrm{SE} \approx \frac{1}{\sqrt{2\,n\,\phi\,(1-\phi)\,f\,(1-f)}}$
- σ is the error variance
- n is the total sample size
- f is the minor allele frequency
- ϕ is the proportion of cases among all samples

FORMULAS FOR NCP = β^2/SE^2

- Liner model GWAS has NCP $\approx 2\,n\,f\,(1-f)\,\beta^2/\sigma^2$
- Logistic model GWAS has NCP $\approx 2\,n\,\phi\,(1-\phi)\,f\,(1-f)\,\beta^2$
- σ is the error variance
- n is the total sample size
- f is the minor allele frequency
- β is the effect size
- ϕ is the proportion of cases among all samples

STEPS OF A GWAS

Study design

1. Is the phenotype heritable?
2. Which set of samples is needed for a GWAS?

Running a GWAS

1. Regression model & covariates
2. Diagnostics

Downstream analyses

1. Conditional analyses & fine-mapping
2. (Other typical analyses we haven't studied on this course)

Replication & Meta-analysis

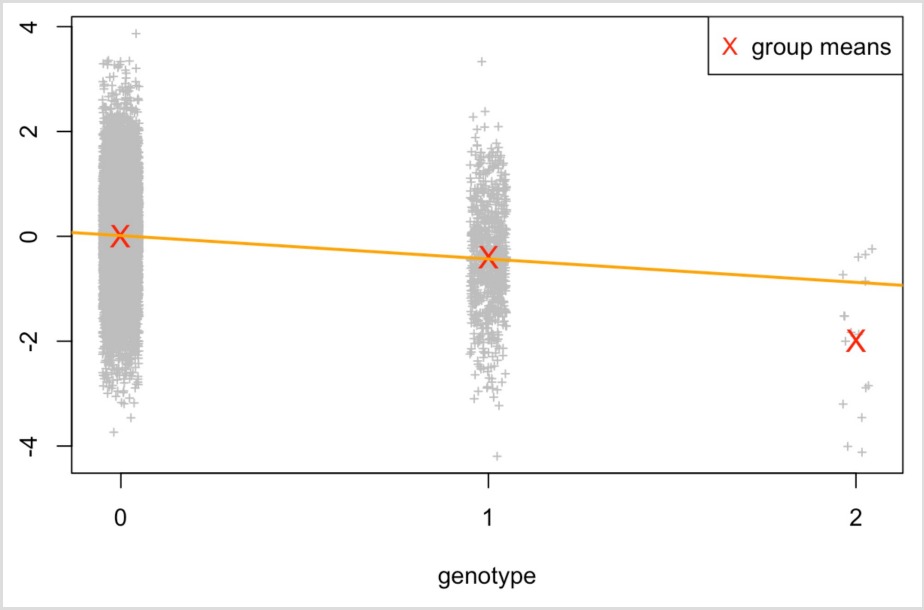
1. Does it replicate?
2. What is the combined evidence?
3. Relationship to other phenotypes?

Further applications

1. Polygenic scores
2. (Mendelian randomization)

REGRESSION MODEL

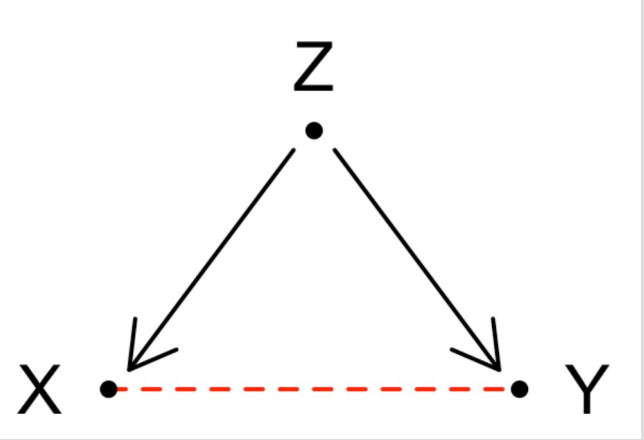
- Linear/logistic regression $y \sim \mu + z^T\gamma + x\,\beta$ where
  - y is the phenotype
  - μ is baseline trait value in quantitative traits / log-odds in diseases
  - z are the covariates and γ their effects
  - x is the genotype (0,1 or 2) and β additive effect per one copy of allele 1
- With logistic regression, all computations are done on the log-odds scale but results are often reported on the odds-ratio scale




```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01358    0.01032   1.316    0.188
## x           -0.44553    0.03570 -12.480   <2e-16 ***
```

CONFOUNDING

- We want to study X-Y relationship but if there are associations between some 3rd variable Z and both X and Y, then Z may cause an observable X-Y association even if there is no **direct/causal** relationship between X and Y
  - Z is **confounder** of X-Y association
  - We can remove (some part of) confounding by adjusting the model for Z
- Geography is a typical confounder in GWAS because it affects both genetics and phenotypes
  - We can estimate population structure by PCA and include it in a regression model
  - Relationship matrix can be included in a linear mixed model as a random effect to account for genetic relatedness (both population structure and close relatedness)



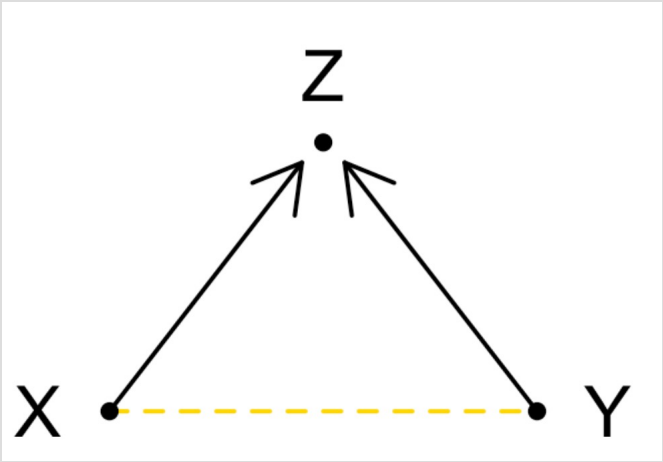
Frequencies
Case | Control

0.35 | 0.35

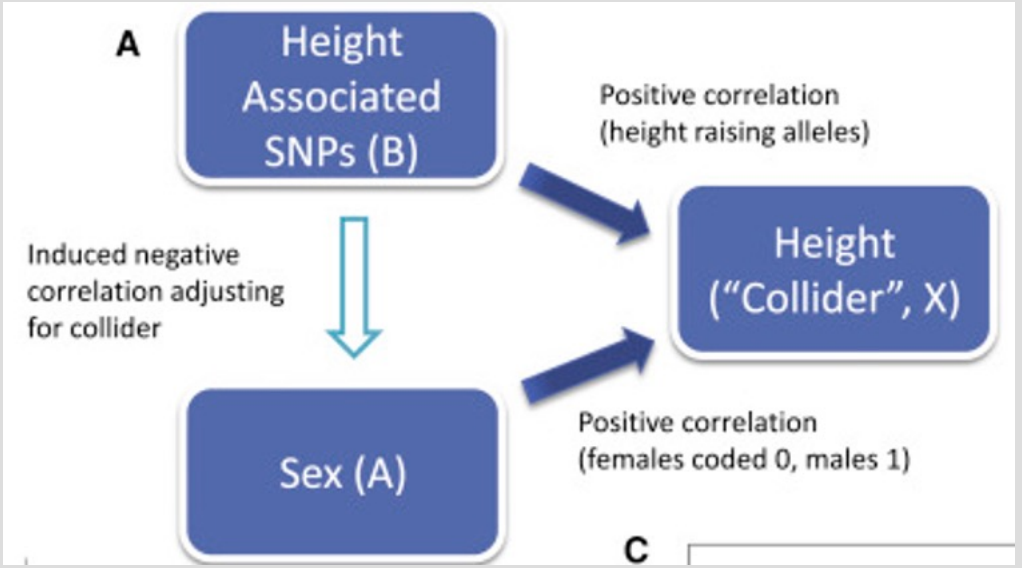
Sample frequencies:
0.32 | 0.26

0.23 | 0.23

COLLIDER BIAS



- If a available covariate is caused by both the outcome Y and the predictor X, then adjusting for the covariate will cause an association between X and Y even if X and Y are independent in the general population
  - Such collider bias associations are not of interest to us so we want to avoid them



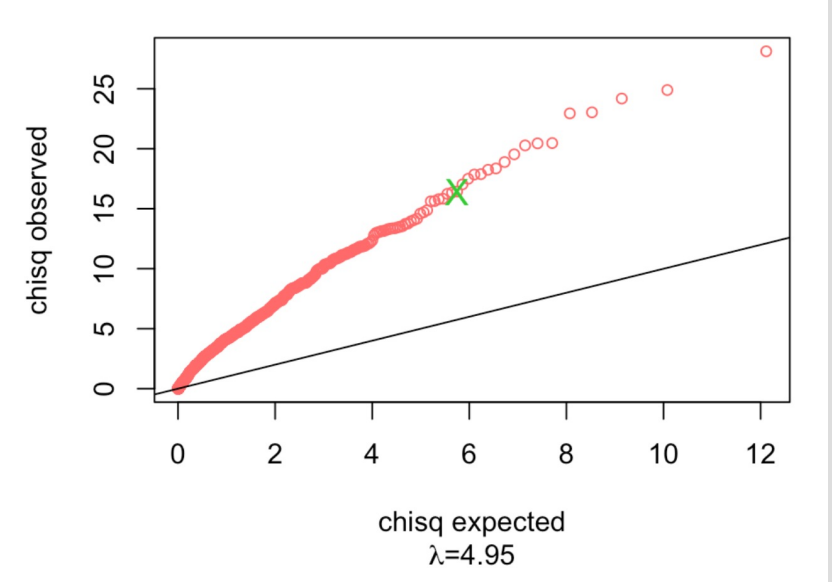
INDEPENDENT COVARIATES

- We consider two models when X and W are independent
  - Model M: $Y \sim \mu + X\beta + W\gamma$
  - Model M': $Y \sim \mu' + X\beta'$
- In linear model $\beta = \beta'$ and model M gives more precise estimate
- In logistic model $|\beta'| \leq |\beta|$ but model M' gives more precise estimate
- In population data, model M is more powerful than model M'
- In case-control data, power depends on prevalence
  - If prevalence < 2%, model M' is typically more powerful
  - If prevalence > 10%, model M is typically more powerful

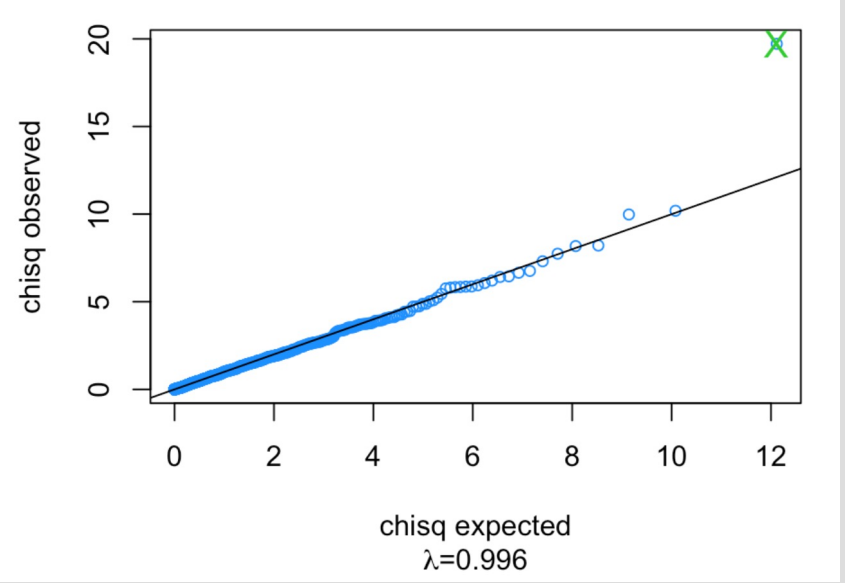
QQ-PLOT

- Shows the observed distribution of test statistics (chi-square or –log10(P-value)) against the null distribution as an ordered scatter plot
- Above diagonal means inflation, i.e., larger than expected association signal
- If inflation is present widely across the genome, some bias may be present
  - But polygenicity also causes inflation in large data sets
- Genomic control parameter (λ) computed as ratio of median statistics

Inflated QQ-plot

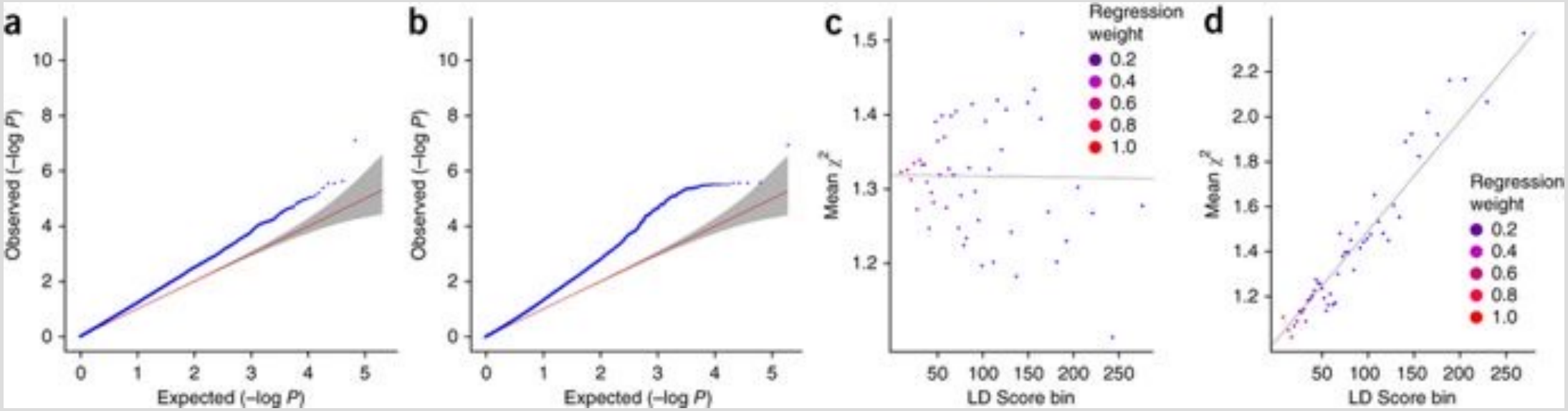


Non-inflated QQ-plot



Bulik-Sullivan et al.
2014 Nat Gen

LD SCORE REGRESSION



(**a**) Quantile-quantile plot with population stratification (λ_{GC} = 1.32, LD score regression intercept = 1.30). (**b**) Quantile-quantile plot with a polygenic genetic architecture where 0.1% of SNPs are causal (λ_{GC} = 1.32, LD score regression intercept = 1.006). (**c**) LD score plot with population stratification. Each point represents an LD score quantile, where the *x* coordinate of the point is the mean LD Score of variants in that quantile and the *y* coordinate is the mean χ^2 statistic of variants in that quantile. Colors correspond to regression weights, with red indicating large weight. The black line is the LD score regression line. (**d**) LD score plot as in **c** but with polygenic genetic architecture.

STEPS OF A GWAS

Study design

1. Is the phenotype heritable?
2. Which set of samples is needed for a GWAS?

Running a GWAS

1. Regression model & covariates
2. Diagnostics

Downstream analyses

1. Conditional analyses & fine-mapping
2. (Other typical analyses we haven't studied on this course)

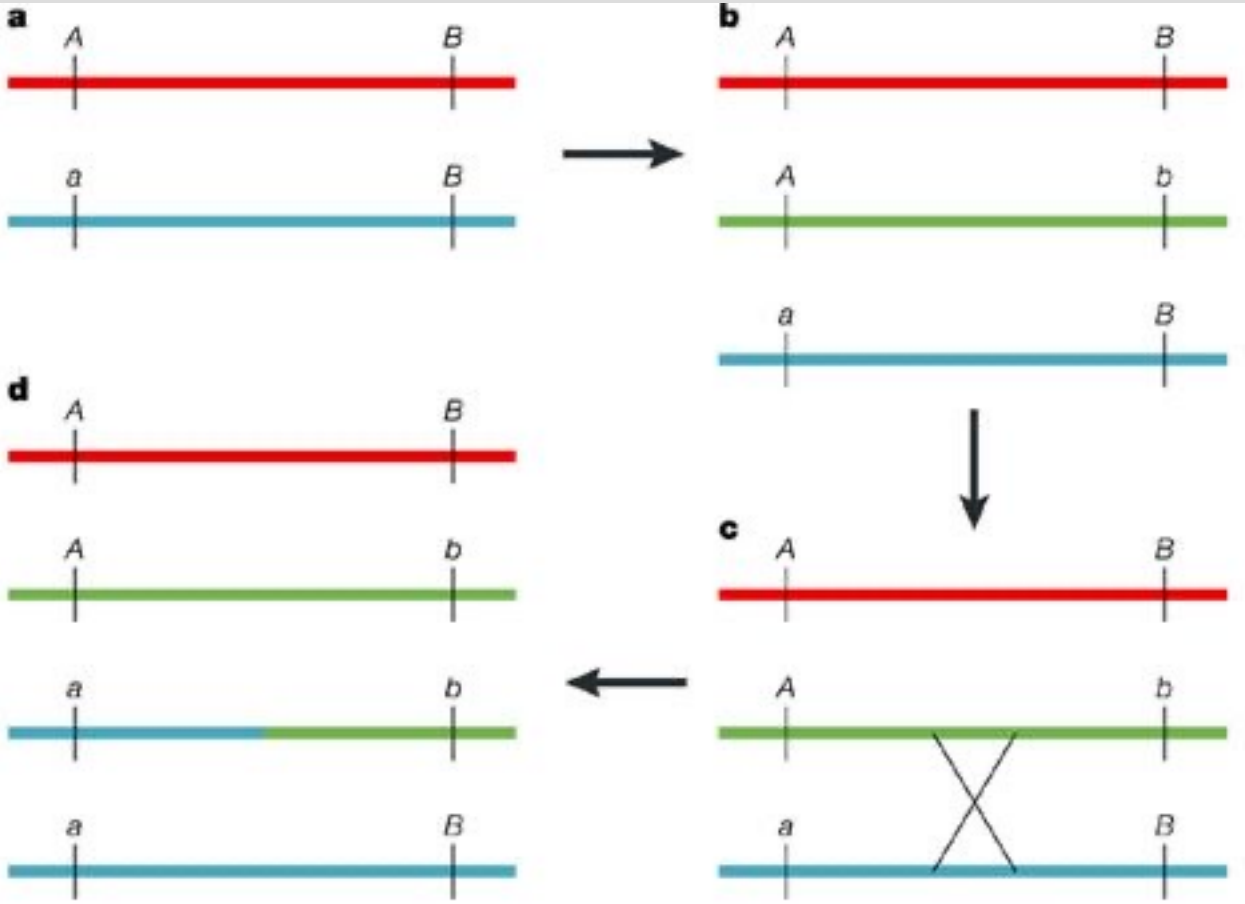
Replication & Meta-analysis

1. Does it replicate?
2. What is the combined evidence?
3. Relationship to other phenotypes?

Further applications

1. Polygenic scores
2. (Mendelian randomization)

LINKAGE DISEQUILIBRIUM



Ardlie et al. Nat Rev Gen 2002

a | At the outset, there is a polymorphic locus with alleles *A* and *a*. **b** | When a mutation occurs at a nearby locus, changing an allele *B* to *b*, this occurs on a single chromosome bearing either allele *A* or *a* at the first locus (*A* in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The *b* allele will always be found on a chromosome with the *A* allele at the adjacent locus. **c** | The association between alleles at the two loci will gradually be disrupted by recombination. **d** | This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (*a*, *b*) increases in frequency.

LDPAIR

CEU
(Central Europe)

rs7837688 chr8:128539360	rs4242382 chr8:128517573 A	rs4242382 chr8:128517573 G		
G	0	180	180	(0.909)
T	16	2	18	(0.091)
	16 (0.081)	182 (0.919)	198	

Haplotypes

G_G: 180 (0.909)
T_A: 16 (0.081)
T_G: 2 (0.01)
G_A: 0 (0.0)

Statistics

D': 1.0
R^2: 0.8791
Chi-sq: 174.0659
p-value: <0.0001

rs7837688(G) allele is correlated with rs4242382(G) allele
rs7837688(T) allele is correlated with rs4242382(A) allele

LWK
(Kenya)

rs7837688 chr8:128539360	rs4242382 chr8:128517573 A	rs4242382 chr8:128517573 G		
G	40	139	179	(0.904)
T	5	14	19	(0.096)
	45 (0.227)	153 (0.773)	198	

Haplotypes

G_G: 139 (0.702)
G_A: 40 (0.202)
T_G: 14 (0.071)
T_A: 5 (0.025)

Statistics

D': 0.0464
R^2: 0.0008
Chi-sq: 0.1541
p-value: 0.6946

rs7837688 and rs4242382 are in linkage equilibrium

D' is a normalized version of D that has maximum of 1.

From LDpair
https://ldlink.nci.nih.gov/

MARGINAL EFFECT AT A NON-CAUSAL SNP

Marginal effect at SNP A is a linear combination of the causal effects of all variants in LD with A, where the weights are the correlations with A (after scaling the genotypes).

$$\beta_A^* = \lambda_A^* + r_{A1}\,\lambda_1^* + r_{A2}\lambda_2^* + r_{A3}\lambda_3^* + r_{A4}\lambda_4^* + \ldots$$

* denotes **scaled effect**: the allelic efect multiplied by $\sqrt{2f(1-f)}$, where f is MAF of the SNP

$$\boldsymbol{\beta}^* = \boldsymbol{R}\boldsymbol{\lambda}^* \quad \text{or equivalently} \quad \boldsymbol{\lambda}^* = \boldsymbol{R}^{-1}\boldsymbol{\beta}^*$$

where **R** is the LD-matrix of pairwise correlations of the variants.

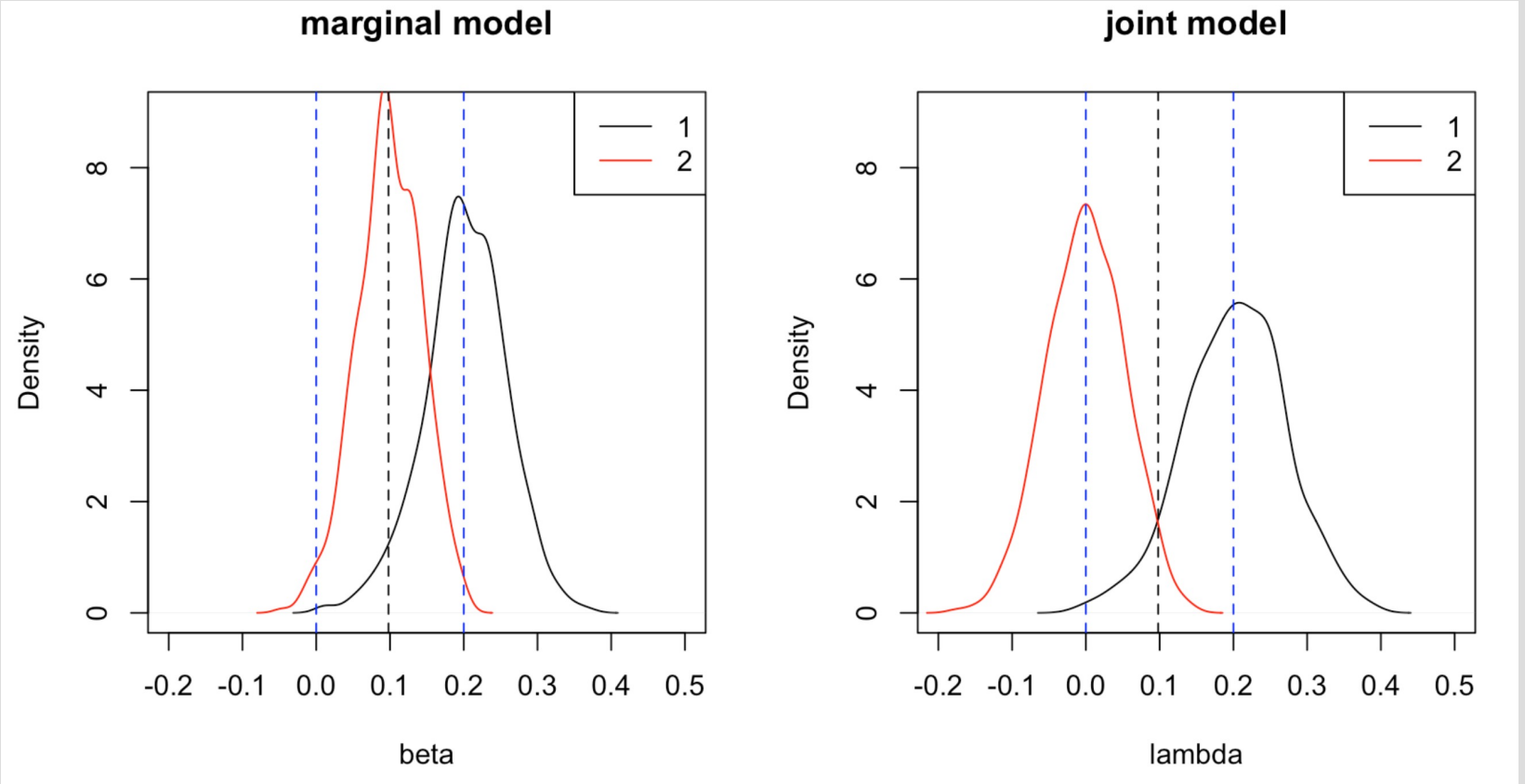
Simulation scenario where causal effects were

$$\lambda_1 = 0.2$$
$$\lambda_2 = 0.2$$

and LD was $r = 0.6$.
MAFs were 0.2 and 0.4.

Marginal effects are then

$$\beta_1 = 0.2$$
$$\beta_2 = 0.6 \cdot 0.2 = 0.18.$$

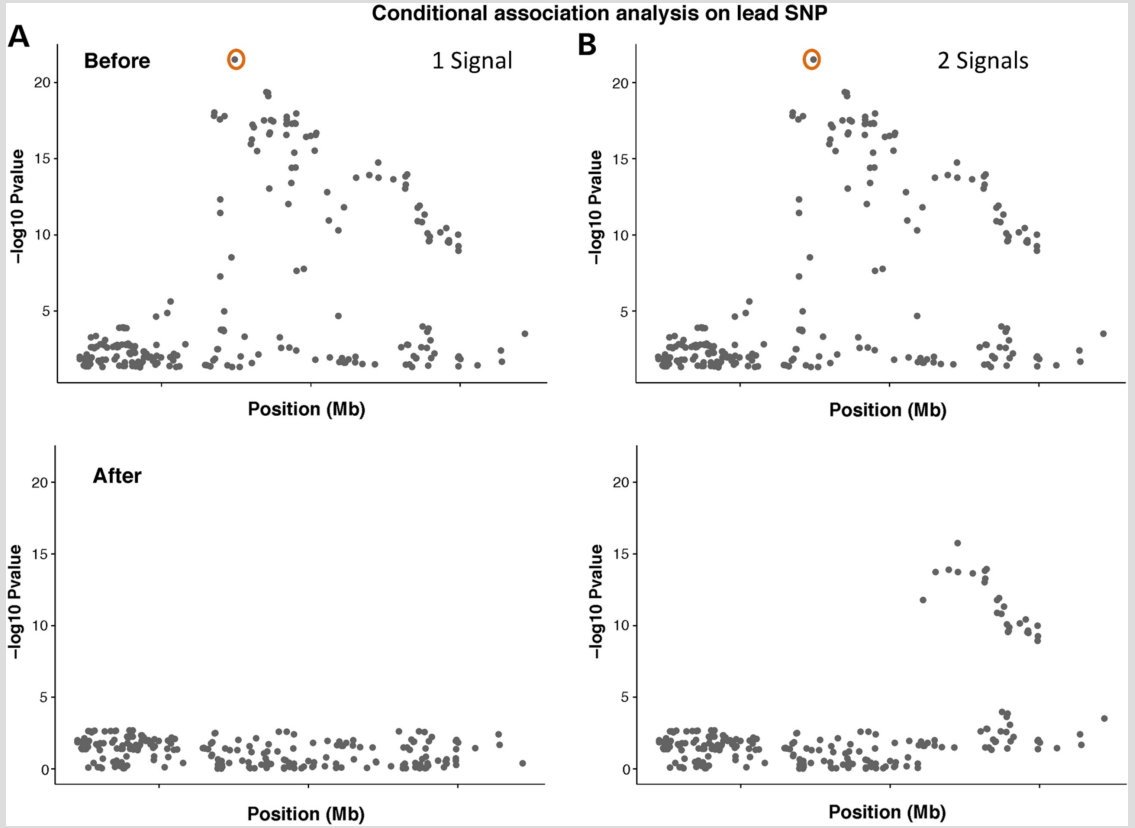


Consequences

- Lowest P-value need not be for a causal variant, especially when there are many causal variant in LD with each other
- Non-causal variants can tag the causal variants and show the signal even if the causal variant was not included in the analysis.

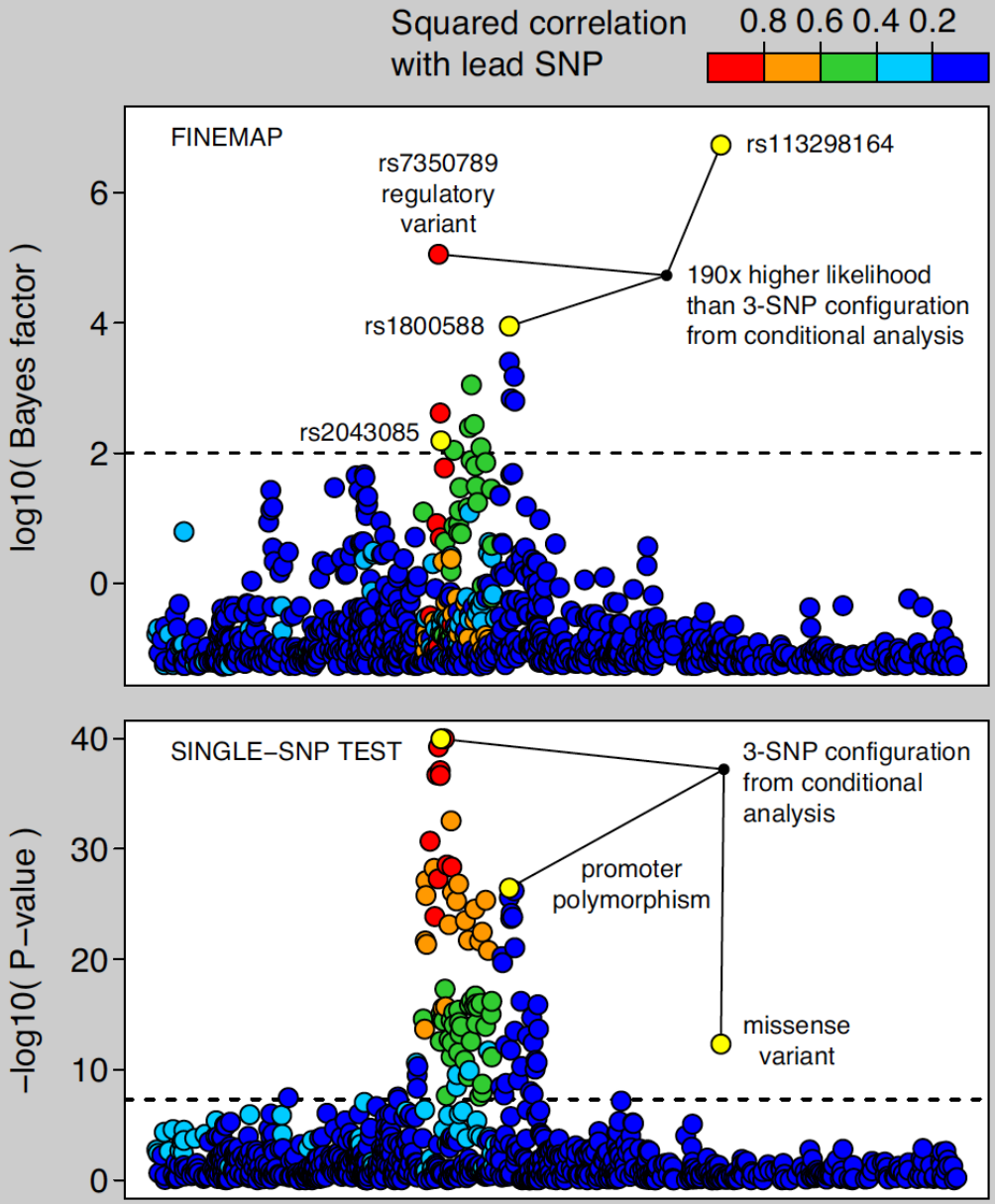
STEPWISE FORWARD SEARCH

- Starts by conditioning on the lowest P-value
- Continues until no additional variant reaches pre-defined P-value threshold
- + Informs about multiple causal variants accounting for LD
- − Does not necessarily find the optimal configuration
- − Completely ignores the uncertainty of the possible causal configurations



Spain & Barrett 2016

15q21/*LIPC* association with HDL cholesterol



Christian Benner

Surakka et al.
Nat. Genet. 2015

NATIONAL INSTITUTE FOR HEALTH AND WELFARE
FINLAND

FINRISK STUDY
20000 indivduals

FINE-MAPPING ASSUMING 1 CAUSAL VARIANT

- If there is exactly one causal variant in the region and it is among the genotyped variants, then the posterior probability of being causal is proportional to the single-SNP marginal Bayes factor of association (ABF from GWAS4)
- This idea can be extended to fine-mapping each independent signal of the region after we have conditioned on the other signals in the region when we have computed the GWAS statistics (betas and SEs) that are used in calculating ABFs
- For multiple causal variants, we use methods such as FINEMAP or SuSiE

STEPS OF A GWAS

Study design

1. Is the phenotype heritable?
2. Which set of samples is needed for a GWAS?

Running a GWAS

1. Regression model & covariates
2. Diagnostics

Downstream analyses

1. Conditional analyses & fine-mapping
2. (Other typical analyses we haven't studied on this course)

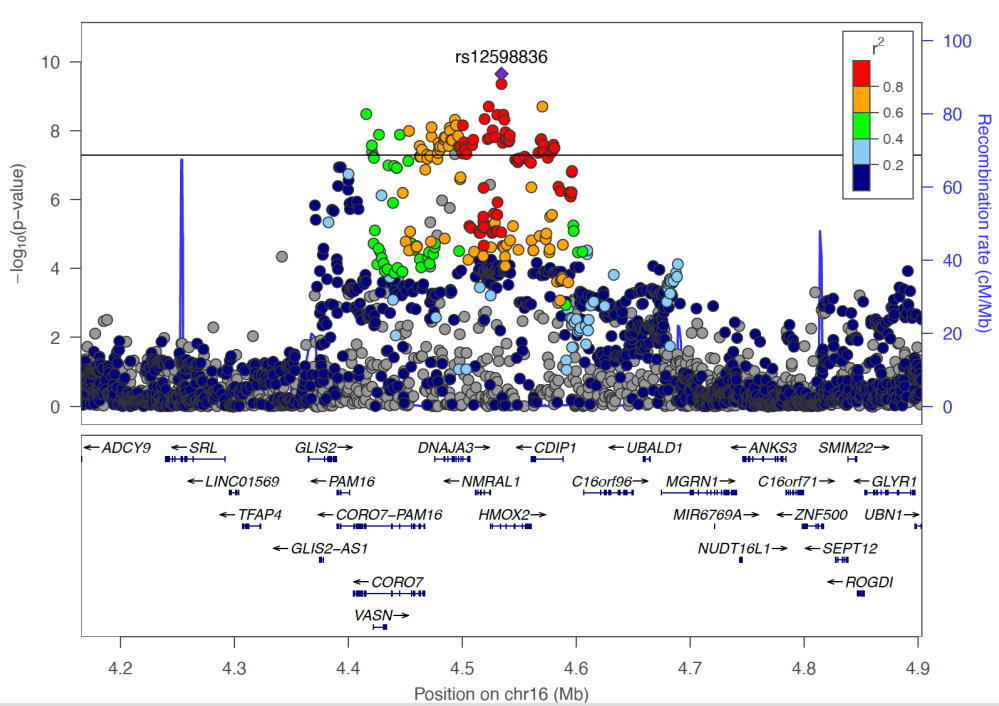
Replication & Meta-analysis

1. Does it replicate?
2. What is the combined evidence?
3. Relationship to other phenotypes?

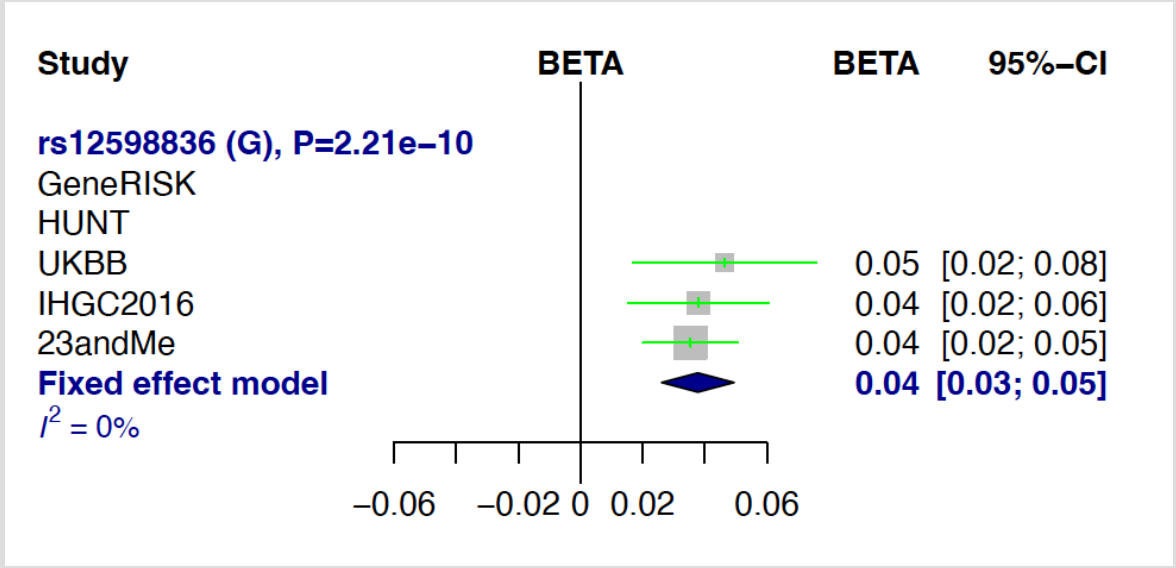
Further applications

1. Polygenic scores
2. (Mendelian randomization)

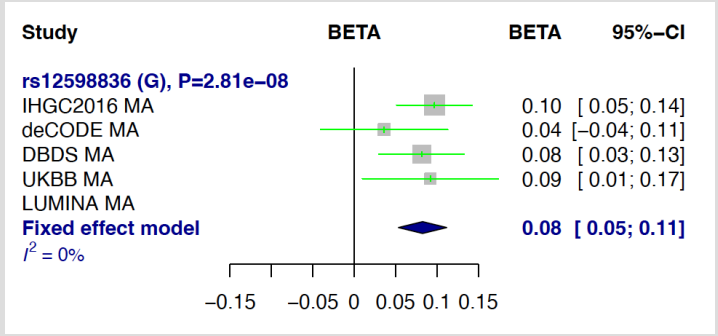
REPLICATION & META-ANALYSIS



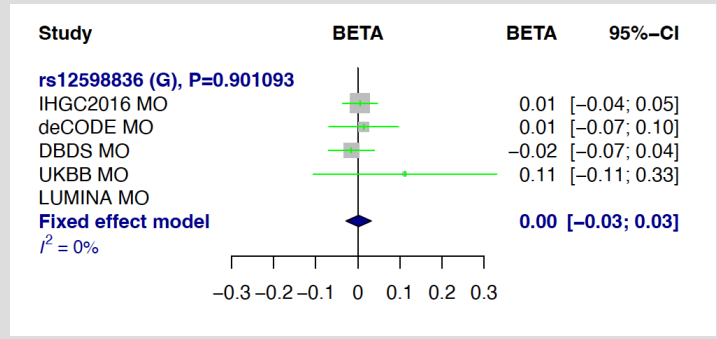
Top-SNP of migraine association in HMOX2 gene.
Is the signal consistent across studies?



Yes it is. What about in subtypes of migraine?



Migraine with aura: effect stronger.



Migraine without aura: zero effect.

INVERSE VARIANCE WEIGHTED (IVW) FIXED-EFFECT (F) ESTIMATOR

$$\hat{\beta}_{l,F} = \frac{w_{1l}\hat{\beta}_{1l} + \ldots + w_{Kl}\hat{\beta}_{Kl}}{w_{1l} + \ldots + w_{Kl}}$$

studies 1,…, K

$$\mathrm{SE}_{l,F} = (w_{1l} + \ldots + w_{Kl})^{-\frac{1}{2}},$$

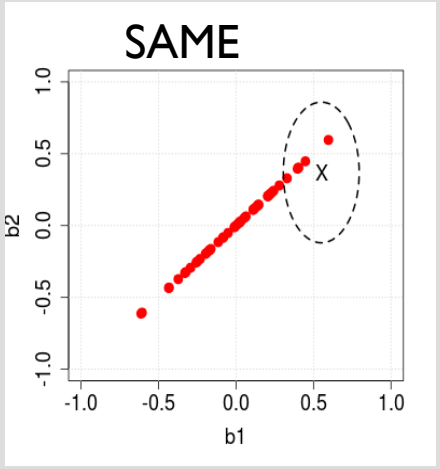
where the weight

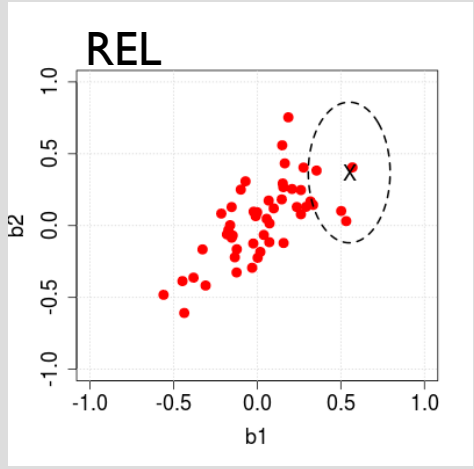
$$w_{kl} = \frac{1}{\mathrm{SE}_{kl}^2}$$ is the inverse-variance of study k.

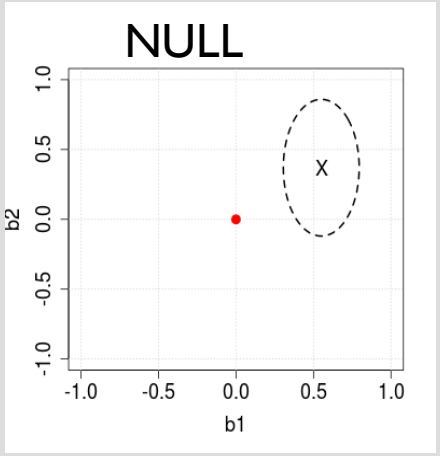
- Each study is weighted by its precision (= inverse of the variance)
- Precision of the combined estimate is the sum of the precisions of the contributing estimates
- For binary outcomes, $\hat{\beta}$ is on the log-odds scale as in logistic regression output, **not** on the odds-ratio scale

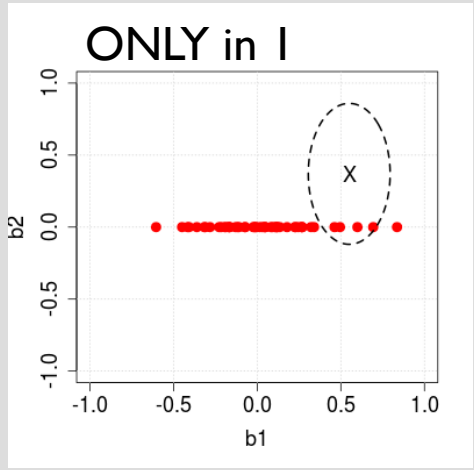
(BAYESIAN) MODEL COMPARISONS

- Specify how different models would produce observed summary statistic data
- Combine likelihood functions with the prior probability of the models to get posterior probability of models

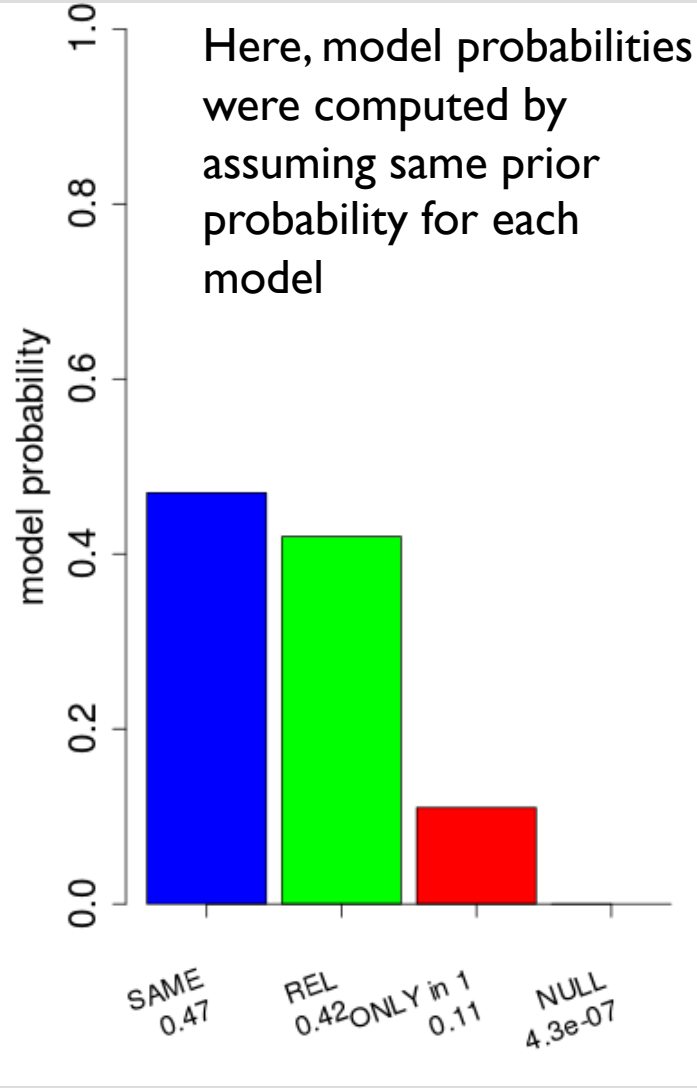








Here, model probabilities were computed by assuming same prior probability for each model



STEPS OF A GWAS

Study design

1. Is the phenotype heritable?
2. Which set of samples is needed for a GWAS?

Running a GWAS

1. Regression model & covariates
2. Diagnostics

Downstream analyses

1. Conditional analyses & fine-mapping
2. (Other typical analyses we haven't studied on this course)

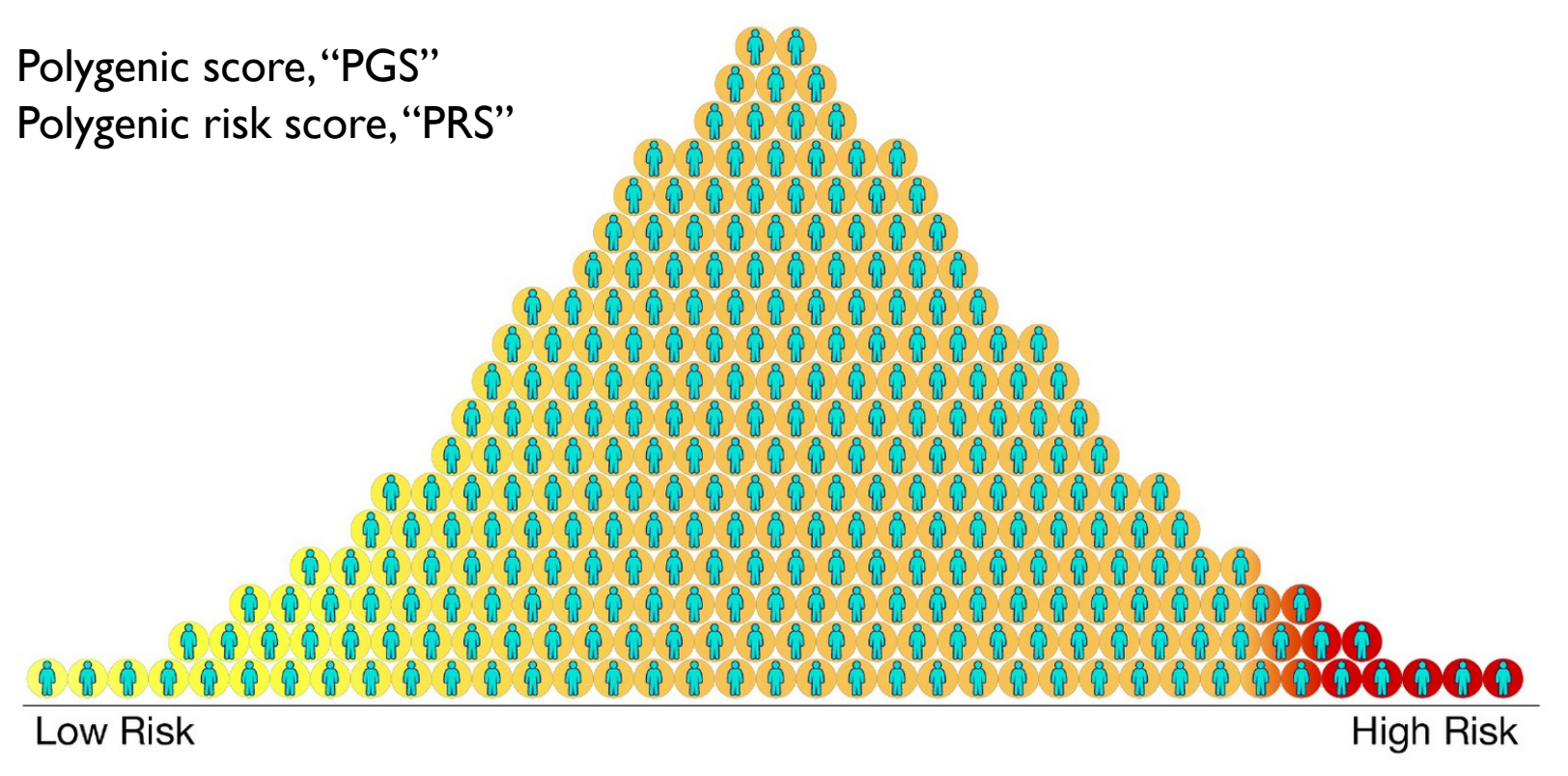
Replication & Meta-analysis

1. Does it replicate?
2. What is the combined evidence?
3. Relationship to other phenotypes?

Further applications

1. Polygenic scores
2. (Mendelian randomization)

POLYGENIC SCORES



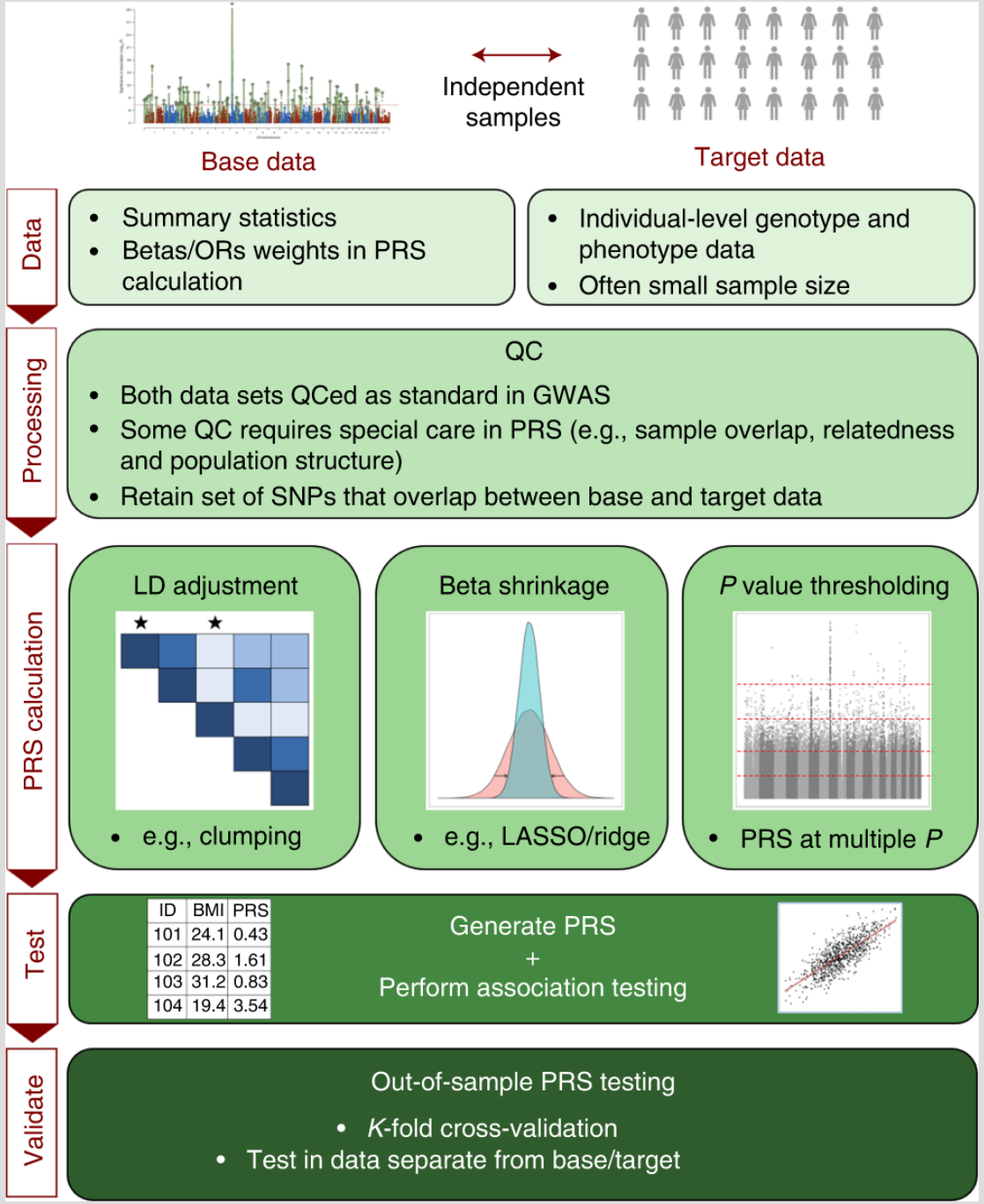
Use GWAS results to predict external individuals' risk for a disease from his/her genotypes.

Figure: NIH

GENERATING POLYGENIC SCORES

- Take allelic effect estimates ($\hat{\beta}_k$) from GWAS
  - Ideally causal effects estimated by multiple regression but often marginal effects used
- Take target individual's genotypes (g_{ik}) at variants $k = 1, \ldots K$
- Compute PRS for individual *i* as sum

$$PRS_i = \sum_{k=1}^{K} g_{ik}\beta_k$$



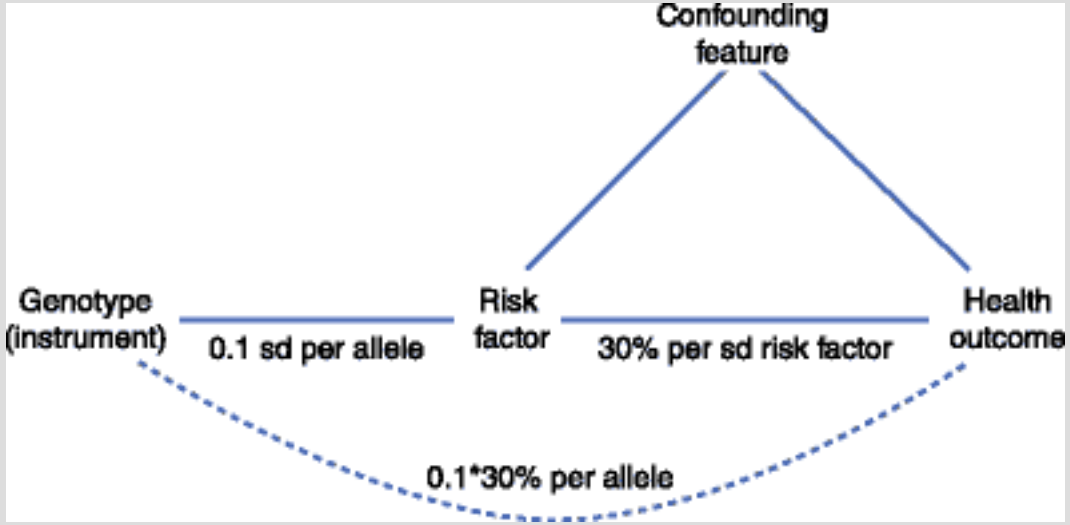
Choi et al. 2020 Nat Protocols

STANDARD PRS METHOD: CLUMPING & THRESHOLDING

- Consider only SNPs with GWAS P-value < P_{thr} , where P_{thr} is a threshold
- From two SNPs that are in LD > r^2 , choose the one with a smaller GWAS P-value
  - This forms "clumps" of "significant" SNPs in LD with each other and only picks the most "significant" ANP as the only representative of the clump
  - A light version of conditional analysis where no joint regression is used but r^2 value alone determines whether two SNPs have "independent signals"
- Use marginal allelic effect estimates in PRS calculation
- Tune parameters P_{thr} and r^2 in a validation set to optimize performance

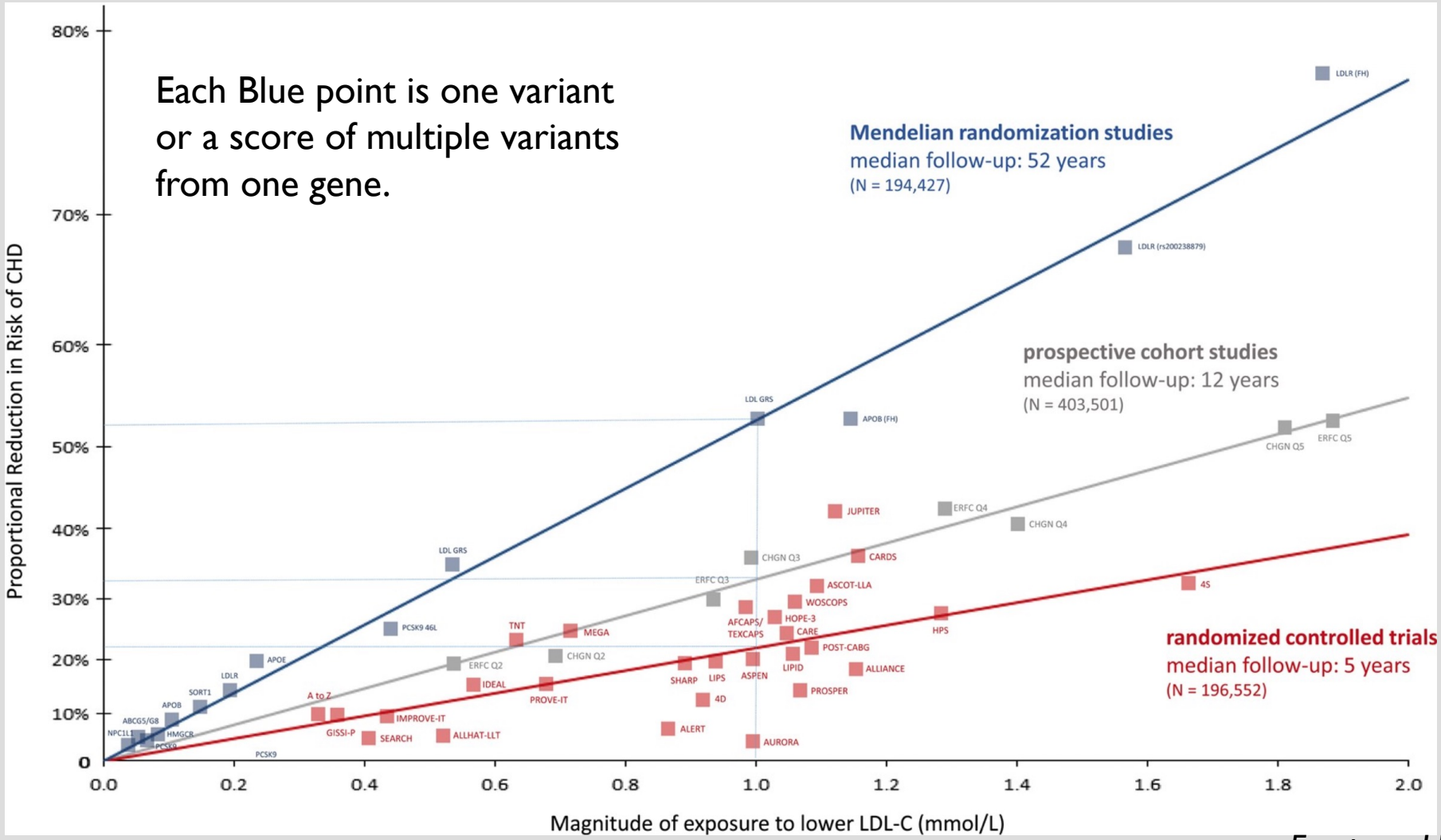
MENDELIAN RANDOMIZATION

- Is risk factor (LDL-C) causal for disease (CHD)?
- If yes, then any genetic variant that raises level of risk factor should also increase risk of disease
- If we see such pattern, then causal association is possible
  - but difficult to rule out that the same genetic variant couldn't affect also other things than the particular risk factor of interest



Current Hypertension Reports **14**, p. 29–37 (2012)

LDL-C AND CORONARY HEART DISEASE



Log-linear association per unit change in low-density lipoprotein cholesterol (LDL-C) and the risk of cardiovascular disease as reported in meta-analyses of Mendelian randomization studies, prospective epidemiologic cohort studies, and randomized trials.

The increasingly steeper slope of the log-linear association with increasing length of follow-up time implies that LDL-C has both a causal and a cumulative effect on the risk of cardiovascular disease.

European Heart Journal, 38, (32), 2017, P. 2459–2472

GWAS PARAMETERS

- β and $\hat{\beta}$, marginal effect size, scaled versions β^*
- λ and $\hat{\lambda}$, causal effect size, scaled versions λ^*
  - λ also used for genomic control parameter in QQ-plots)
- SE, standard error of effect sizes
- σ^2 error variance of linear regression model
- R^2 variance of phenotype explained by regression model
- τ^2 (prior) variance of a non-zero effect size in Bayesian models and in LD-score reg.
- $\boldsymbol{R}$ LD-matrix of pairwise correlation between variants
- r LD between pair of variants and r^2 the squared LD
- h^2 heritability due to additive effects (for a variant, a region or whole genome)