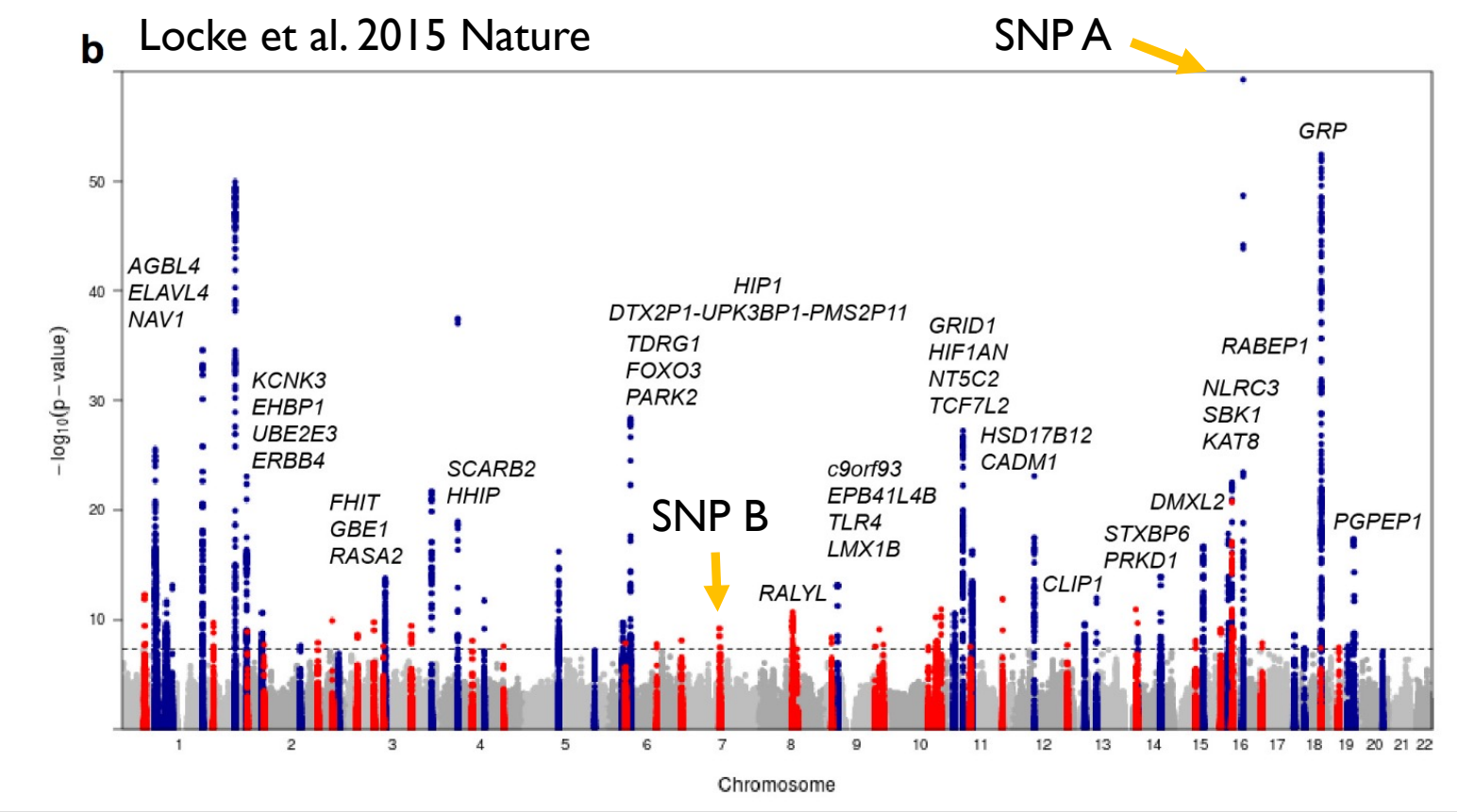
GWAS 3

Matti Pirinen
University of Helsinki
Week 2

WHICH VARIANTS BECOME "SIGNIFICANT"?

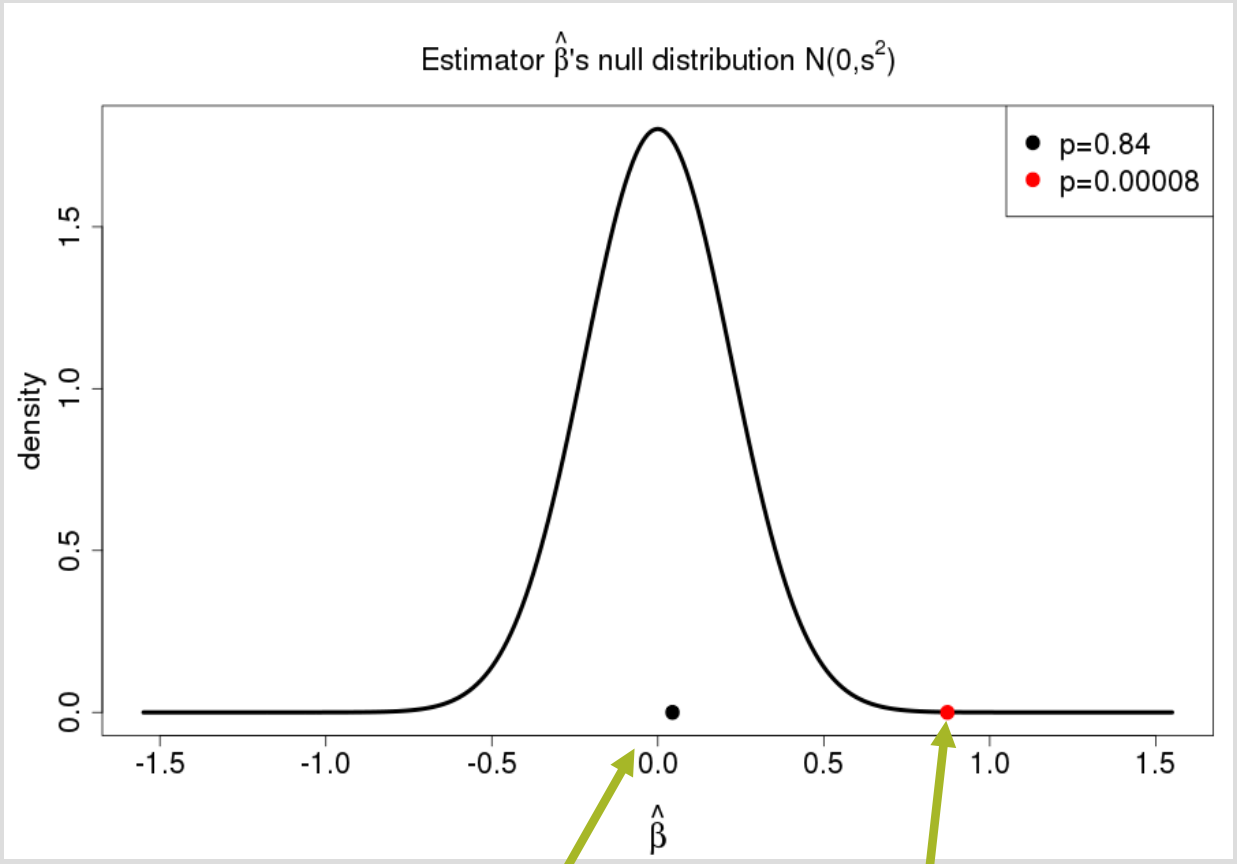


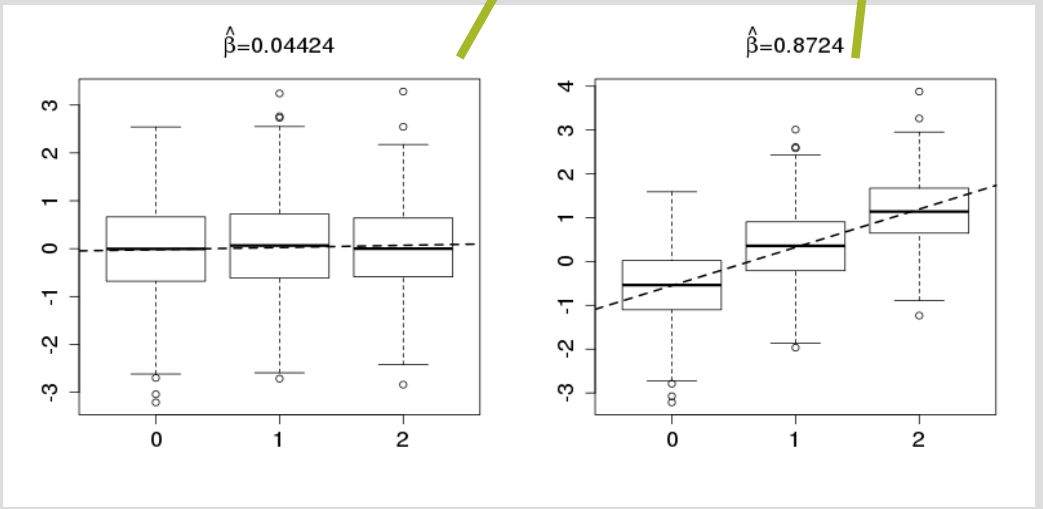
GWS = "genome-wide significant", typically $P < 5e\text{-}8$

- With stringent threshold of $P < 5e\text{-}8$ we have only few false positives
- What about true positives?
- Do we always find SNP A having $P < 5e\text{-}8$ when we do a GWAS on BMI?
- What about SNP B?
- Which properties affect whether a SNP with true effect becomes GWS?
- What is the probability that SNP A / B becomes GWS?
  - This is called "statistical power" to detect the SNP as associated

P-VALUE

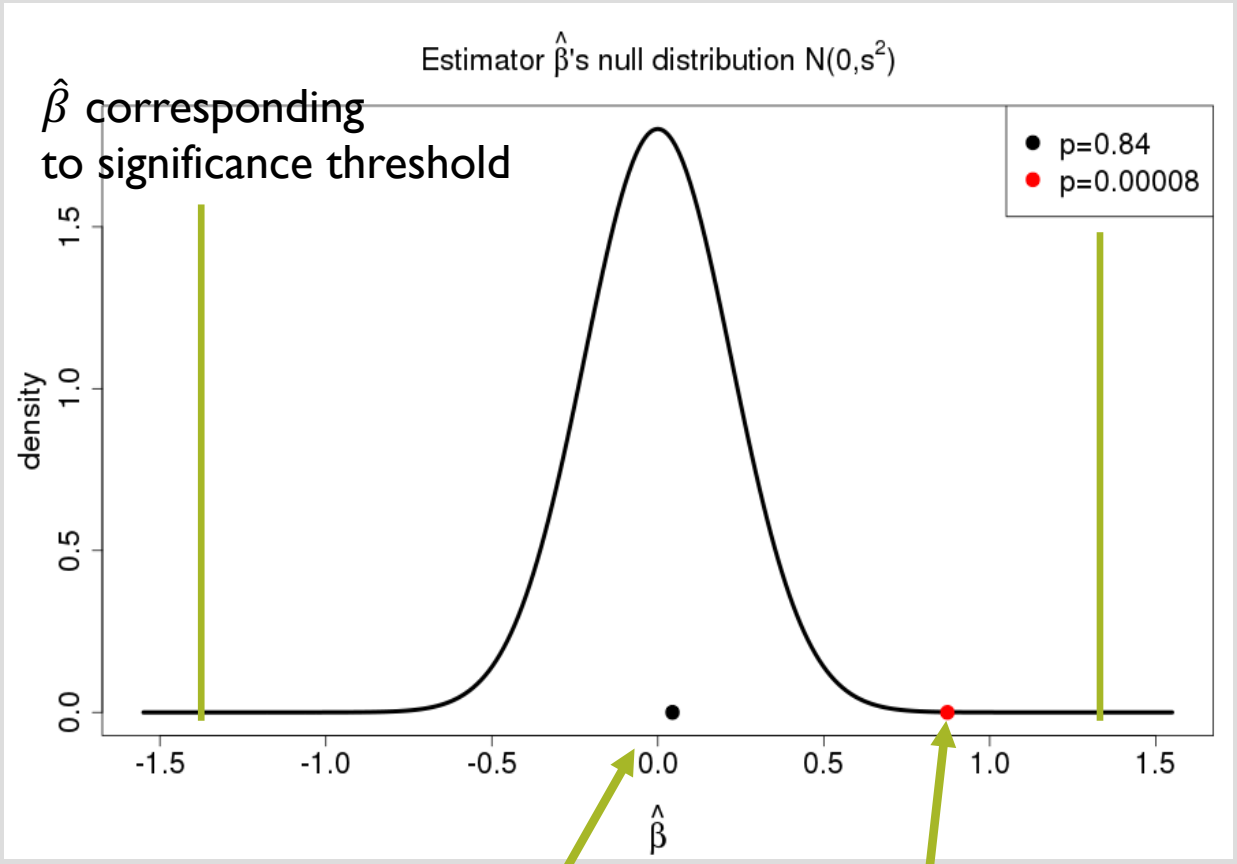
- Is the observed estimate $\hat{\beta}$ plausible if true slope $\beta = 0$?
- P-value: Probability that "by chance" we get at least as extreme value as we have observed, if true $\beta = 0$
- P = 0.84: No evidence for deviation from null
- P = 8e-5: Unlikely under the null → maybe not null

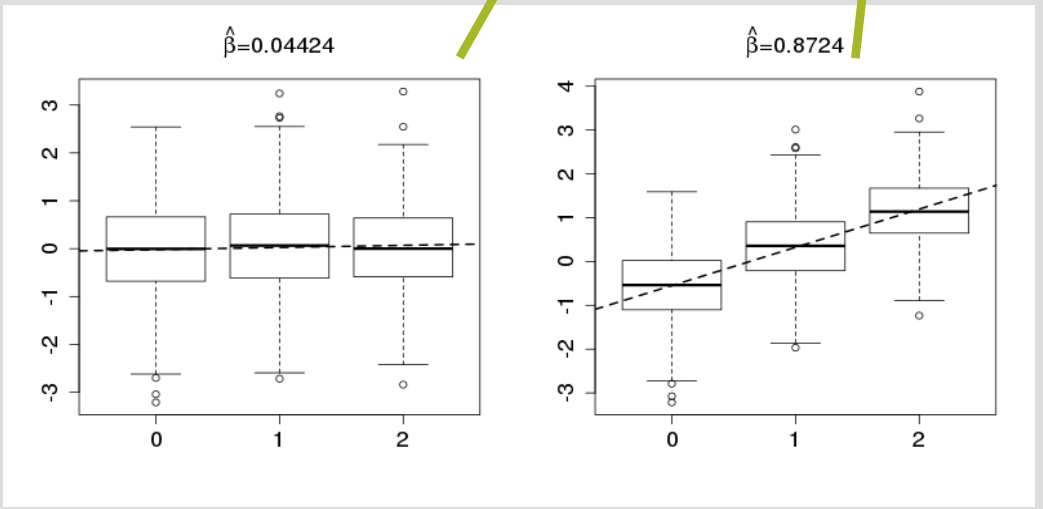




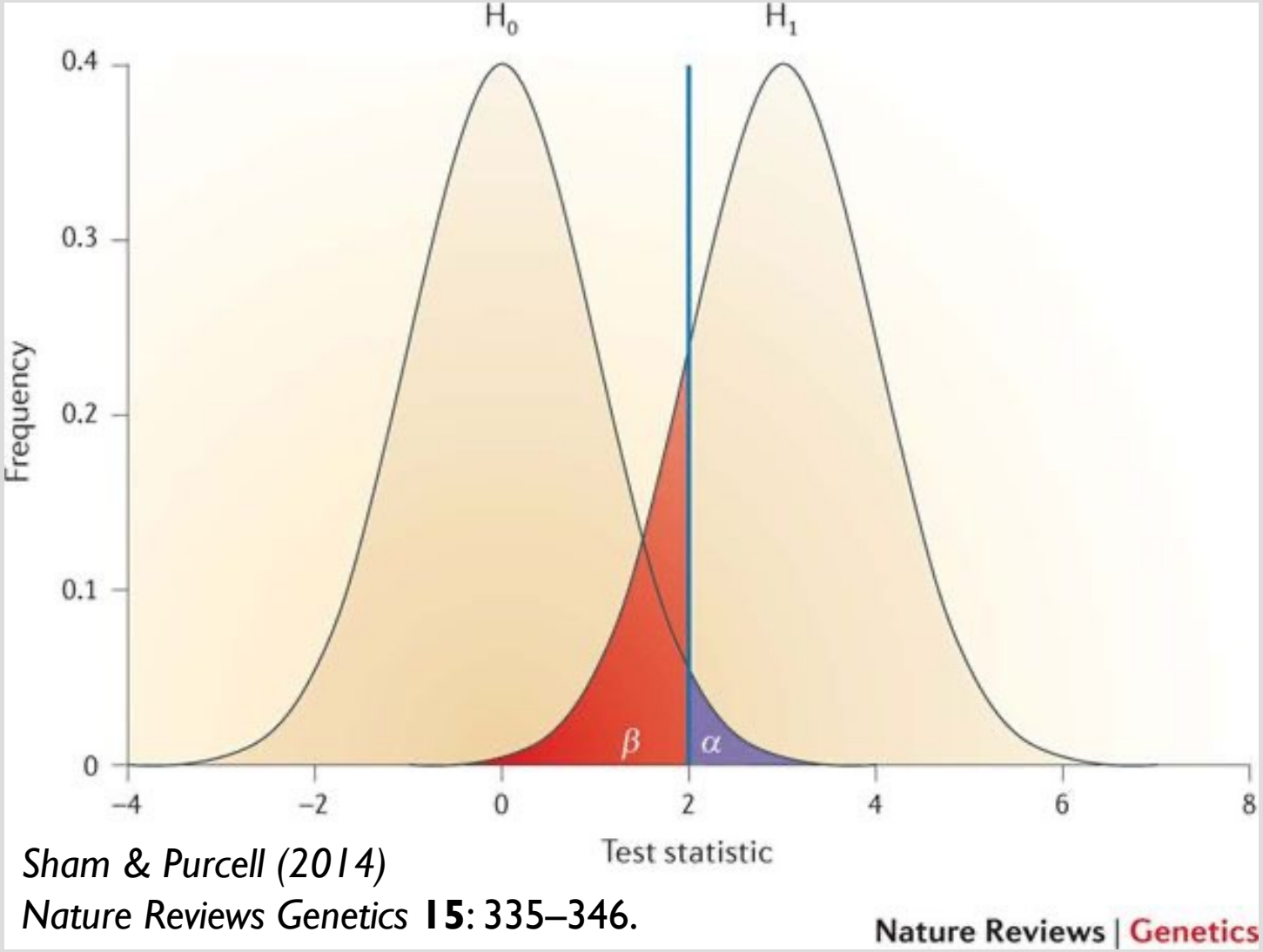
SIGNIFICANCE THRESHOLD & STATISTICAL POWER

- 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 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. This means dragging blue line to right.

Type II error: "false negative", fail to "reject 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).

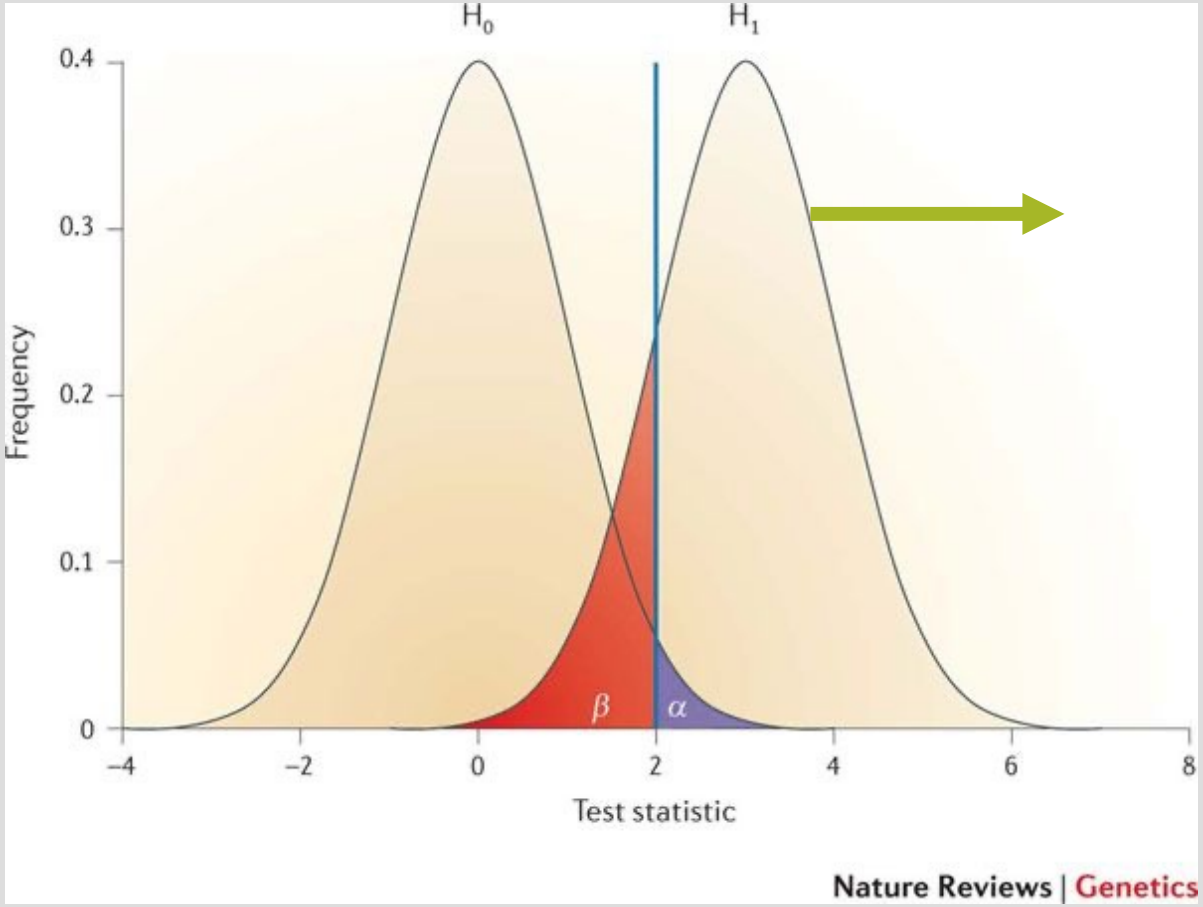
HYPOTHESIS TESTING TERMINOLOGY

- H_0 (NULL HYPOTHESIS): Variant has no association to phenotype
- Significance testing at significance level α: "Reject H_0" if *P*-value < α, where *P*-value is calculated under H_0
- If α is defined before the experiment, then the proportion of false rejections of H_0 out of all true H_0 would be α in repeated experiments
  - By making α small (say 5e-8) we can protect from false positive findings (Type I errors) but increase false negative findings (Type II errors)
  - By keeping α larger (say 0.05) we have more statistical power to reject H_0 (avoid Type II error) but we are more likely to make a false positive finding (Type I error)
- H_1 (ALTERNATIVE HYPOTHESIS): Variant has a non-zero association to the phenotype
  - In power calculations we must be specific about H_1 (What are MAF, N, effect size?)

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 $z = {\hat{\beta}}/{SE} \sim N\left(\frac{\beta}{SE}, 1\right)$
  - 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 $z^2 \sim \chi_1^2\left(\text{NCP} = {\beta^2}/{SE^2}\right)$, where NCP is the "non-centrality parameter"
  - General definition: When $Y \sim N(\mu, \sigma^2)$ then $\frac{Y^2}{\sigma^2} \sim \chi_1^2\left(\text{NCP} = {\mu^2}/{\sigma^2}\right)$
  - $z^2 \sim \chi_1^2$ when $\beta = 0$, 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"

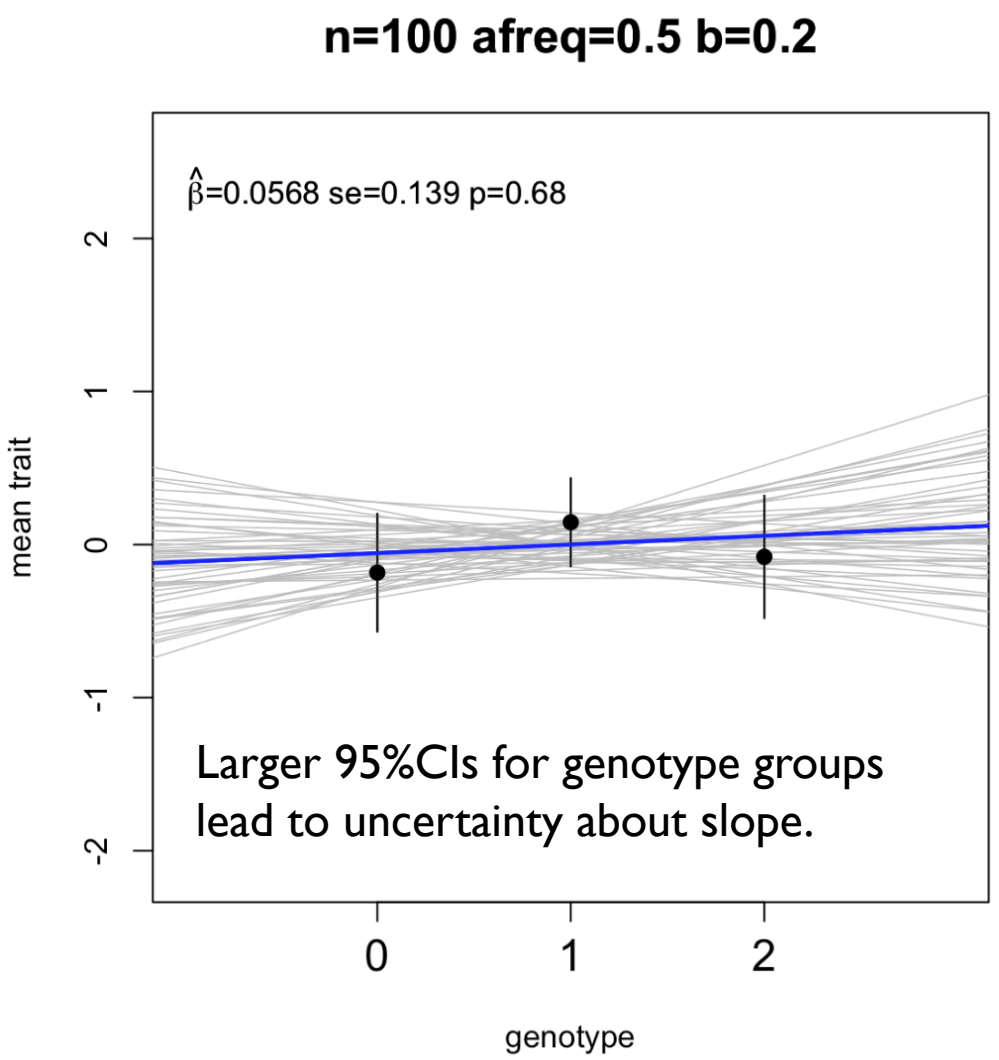
FORMULAS FOR SE

- Linear 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 standard deviation
- n is the total sample size
- f is the minor allele frequency
- ϕ is the proportion of cases among all samples
  - $n\,\phi\,(1-\phi)$ is an "effective sample size" $n\,\phi\,(1-\phi)$ in case-control GWAS

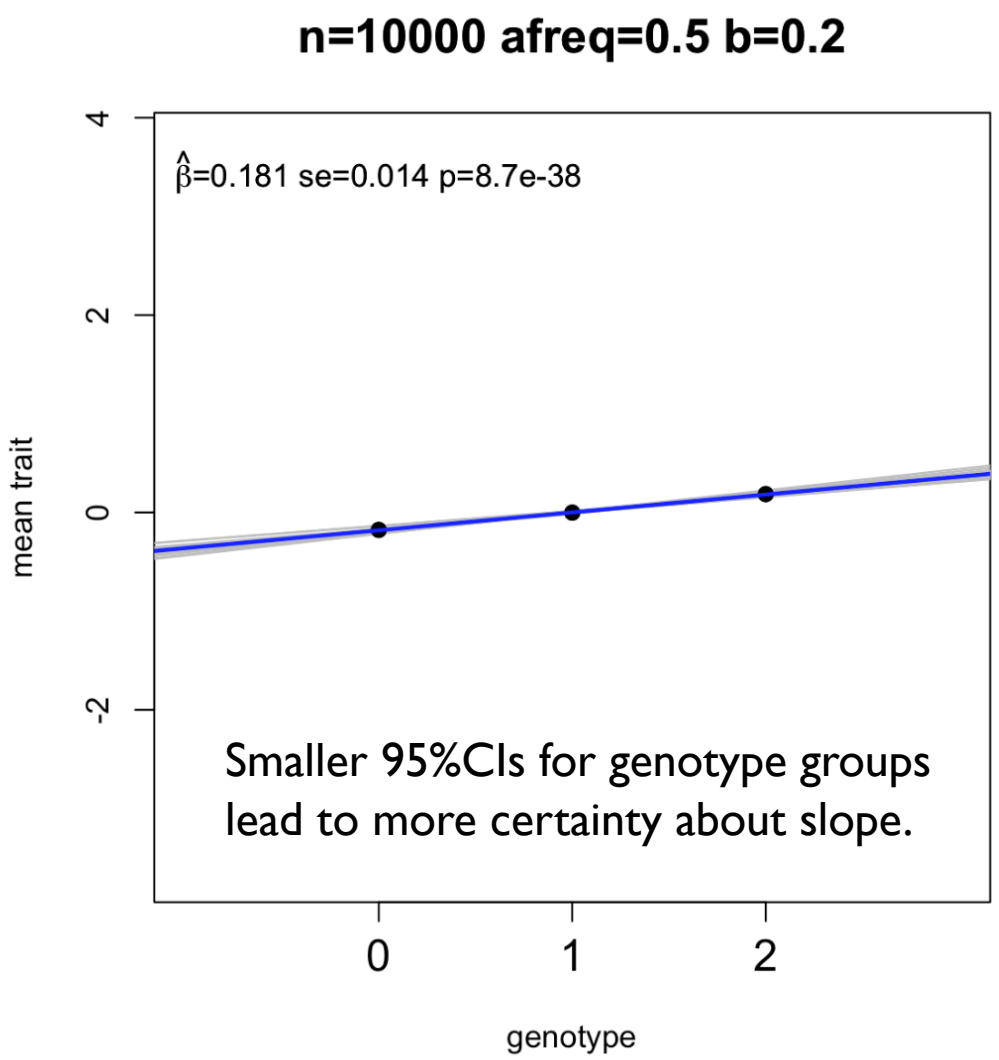
FORMULAS FOR NCP = β^2/SE^2

- Linear 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

WHY DOES N INCREASE POWER?

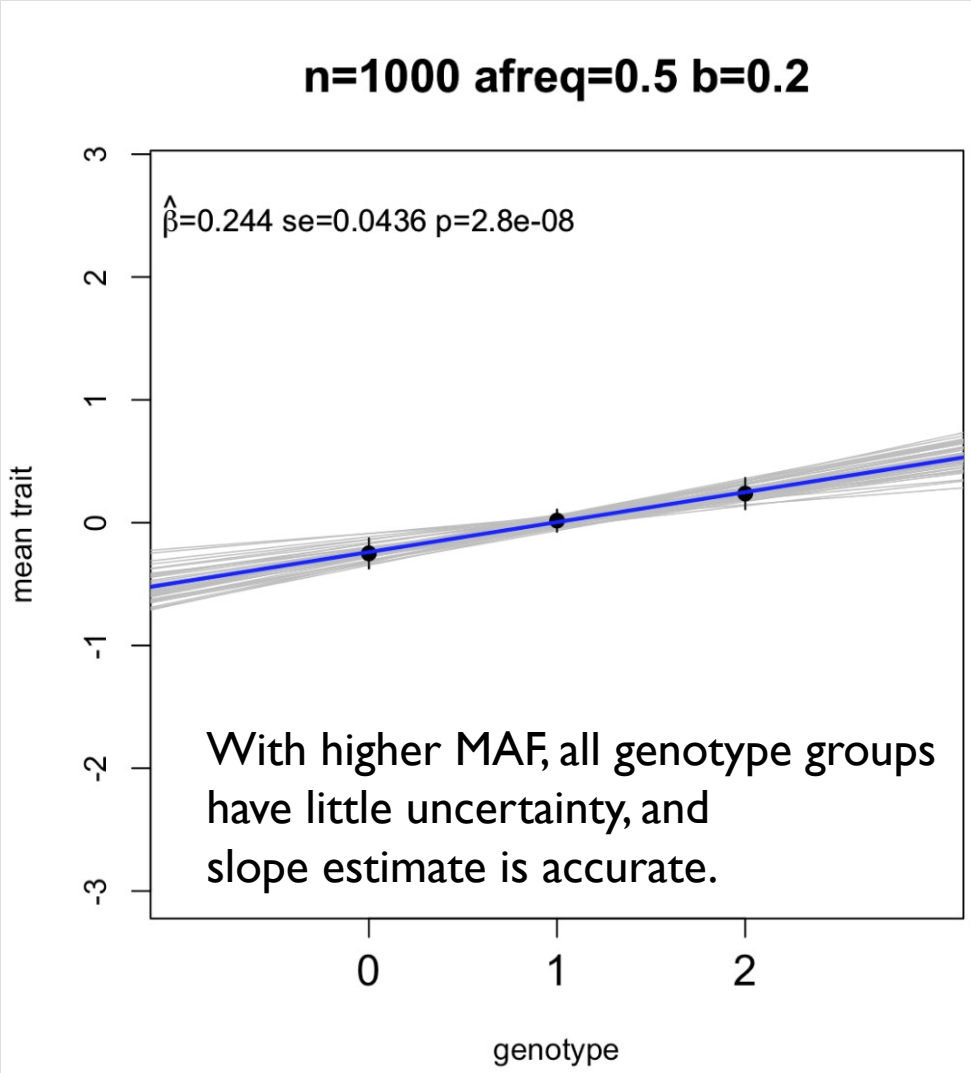


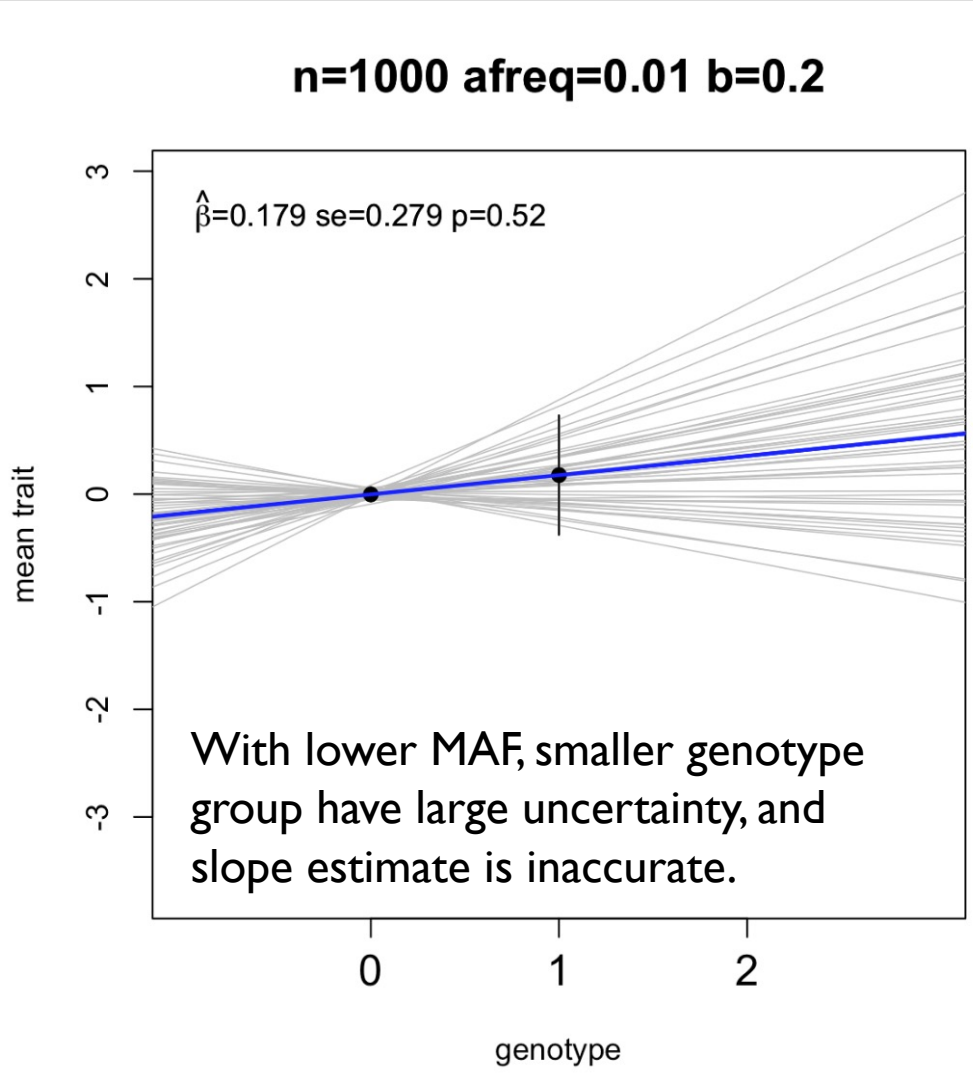
We are unsure whether slope is positive



We are confident that the slope is positive

WHY DOES MAF INCREASE POWER?

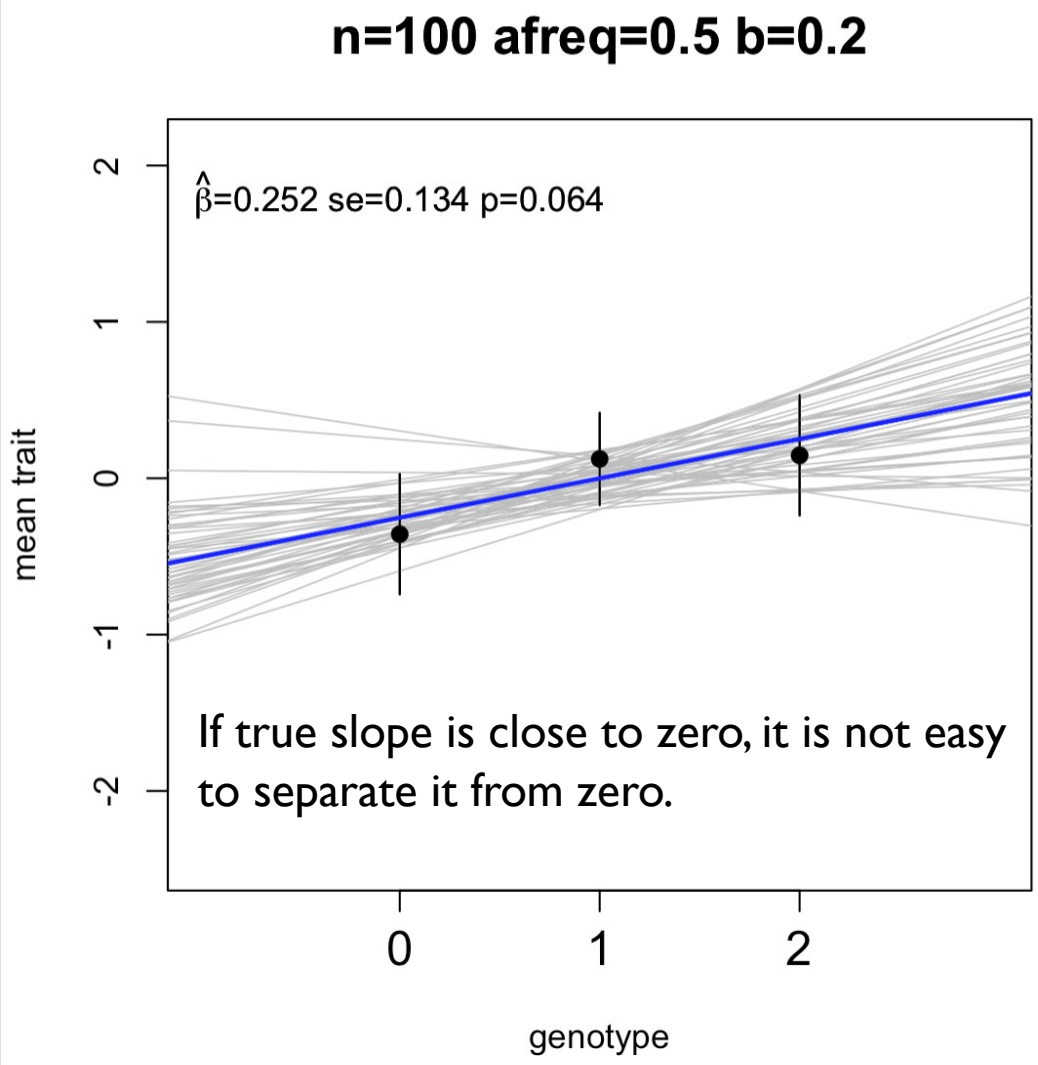




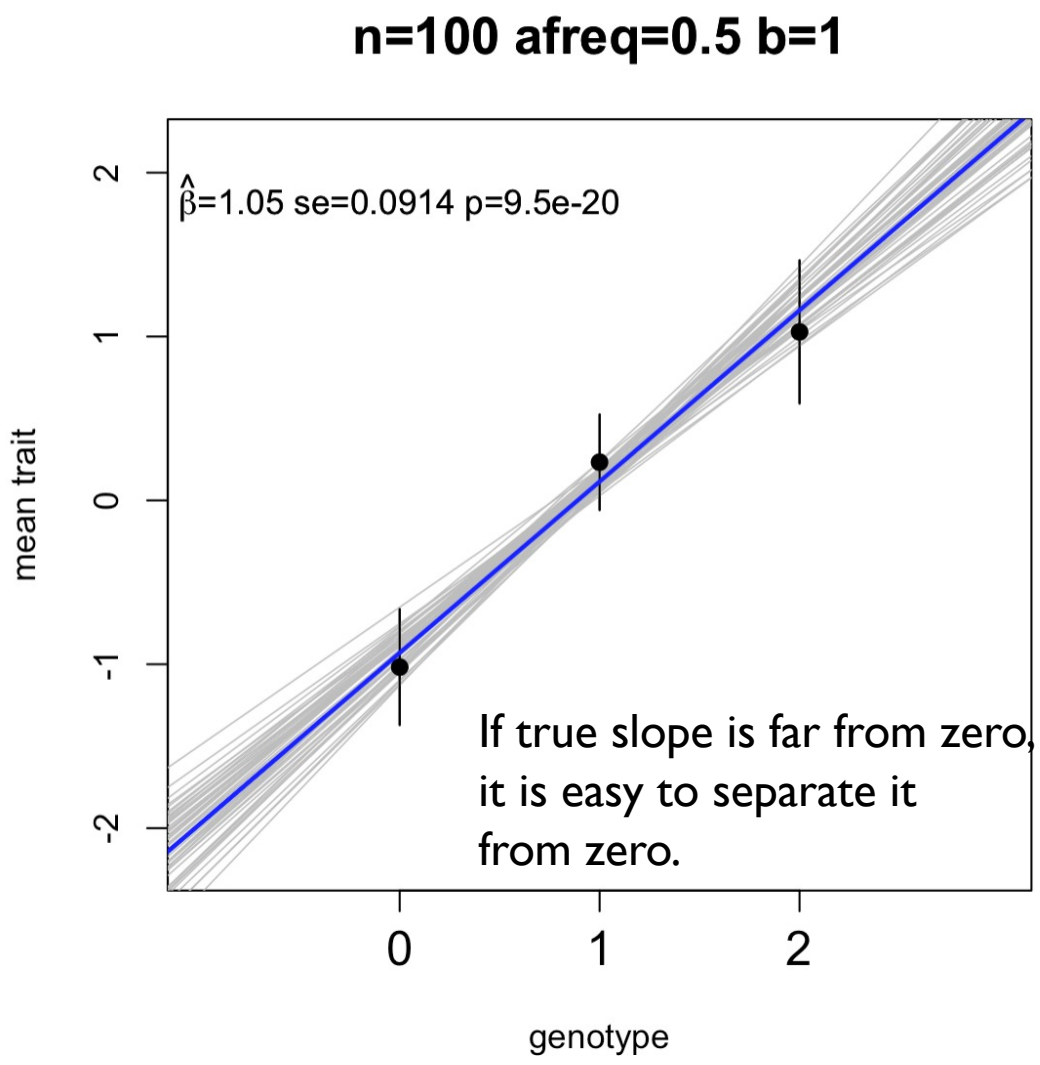
We are confident that the slope is positive

We are unsure whether the slope is positive

WHY DOES $|\beta|$ INCREASE POWER?



We are unsure whether the slope is positive

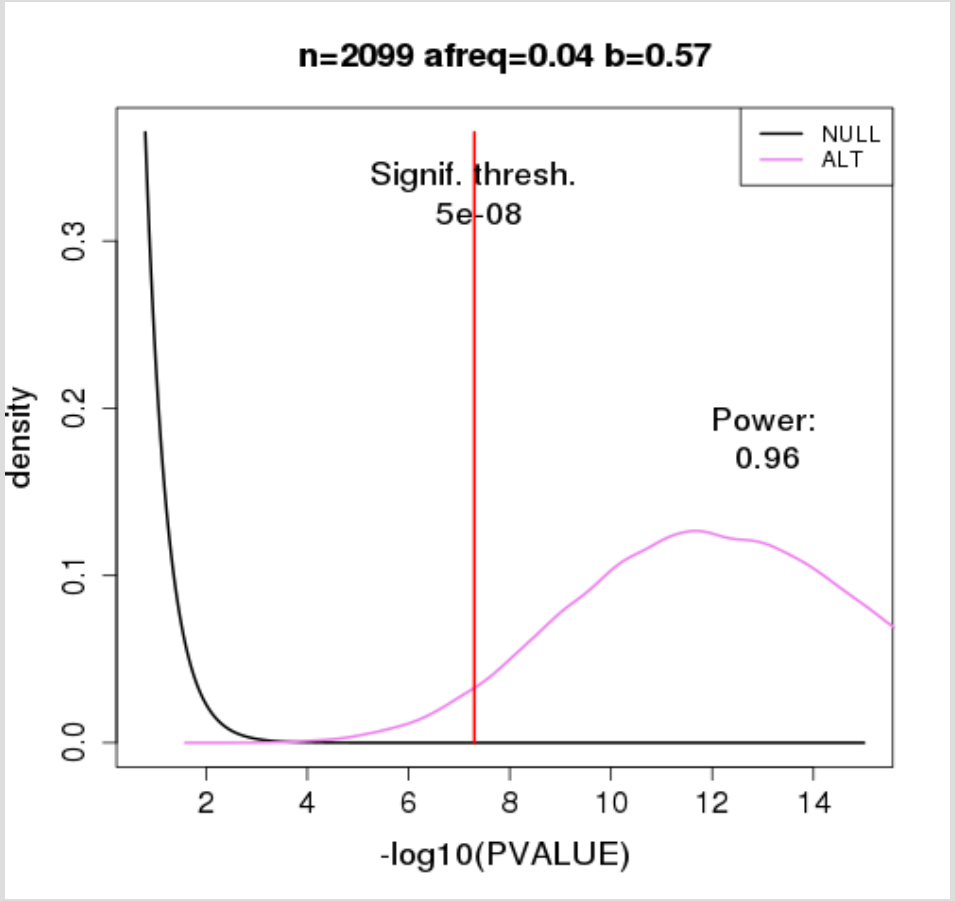


We are confident that the slope is positive

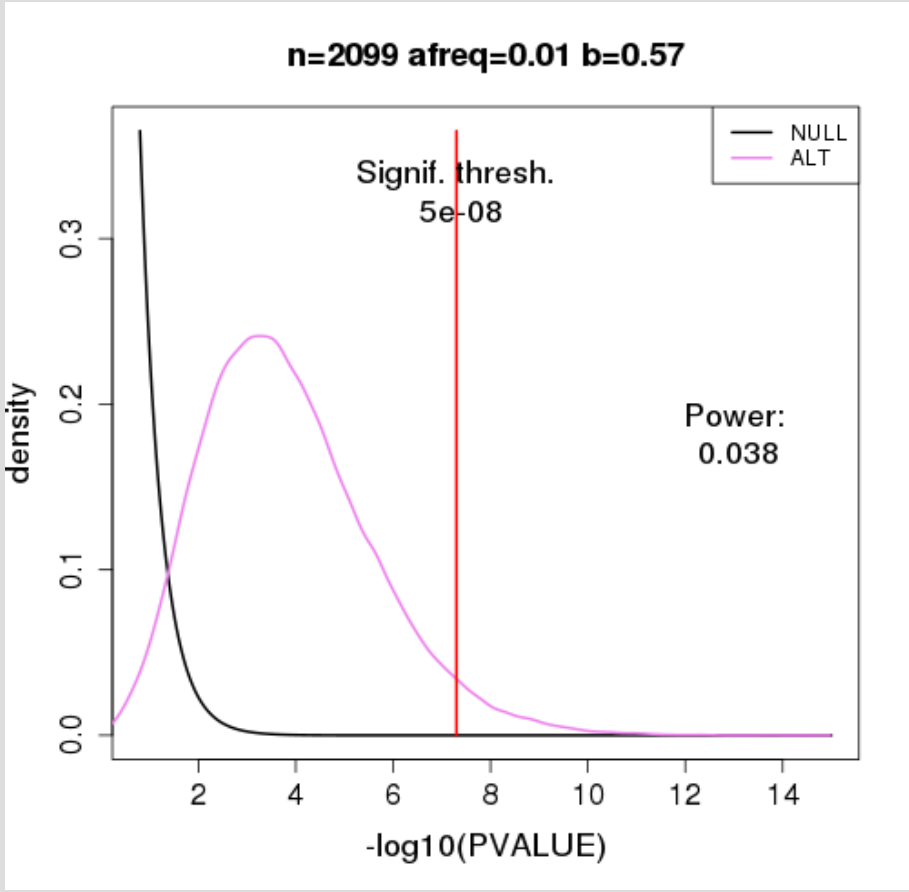
WHY DOES $\phi(1-\phi)$ INCREASE POWER?

- If we have a lot of controls, we know the control frequencies very accurately
- But if have only few cases, then we don't know the case frequencies accurately
- We cannot tell whether cases are different from controls unless we know accurately **BOTH** the case and the control frequencies
- Extreme setting: all samples are controls -- we learn nothing
- $N\ \phi\ (1-\phi)$ is the **effective sample size** of a case-control study
  - To make it large, we should have large *N* and ϕ close to 0.5
  - Often effective sample size is defined as $4\ N\ \phi\ (1-\phi)$, which is the total sample size of a study where count of cases equals count of controls and power is the same as the power observed in our study

PCSK9 VARIANT FROM MOTIVATION VIDEO



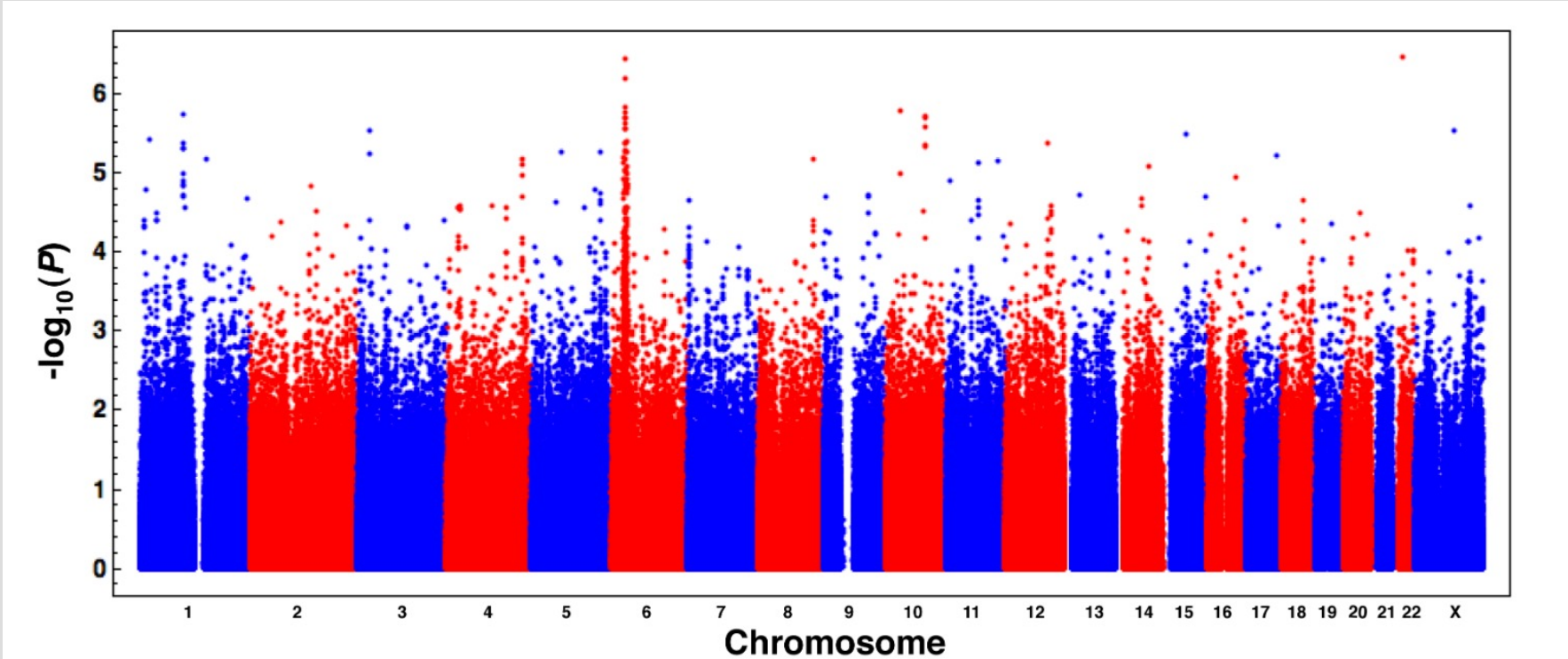
In Finland MAF = 4%:
We are almost certain to detect it with 2099 samples



In Central Europe MAF = 1%:
We are almost certain to not detect it with 2099 sample

SCHIZOPHRENIA GWAS 1/3 2009

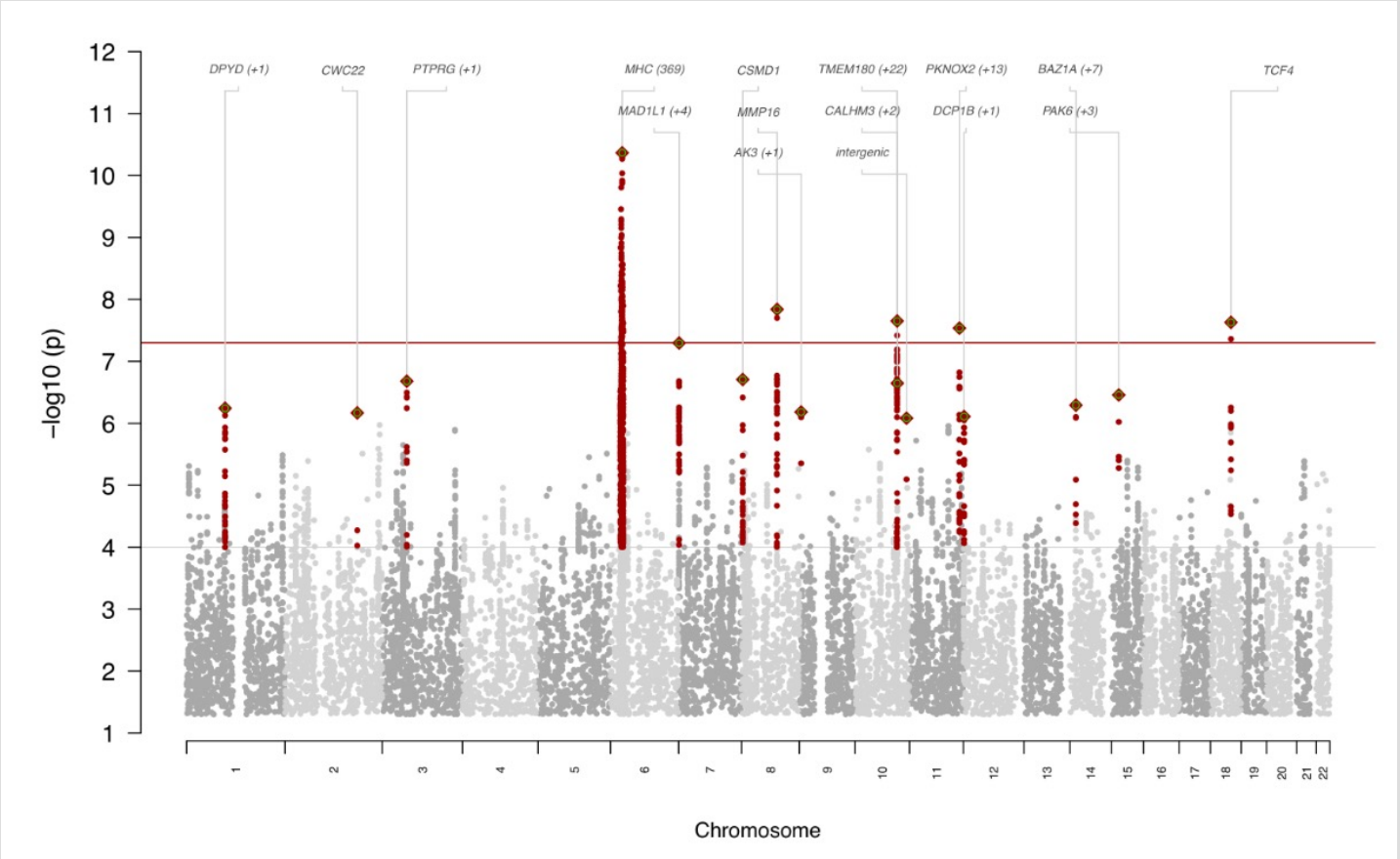
- 3,332 SZ cases and 3,587 controls at 1M SNPs
- No genome-wide significant findings
- Suggestive evidence for HLA-region on chr 6



Int'l SCZ consortium
Nature 2009

SCHIZOPHRENIA GWAS 2/3 2011

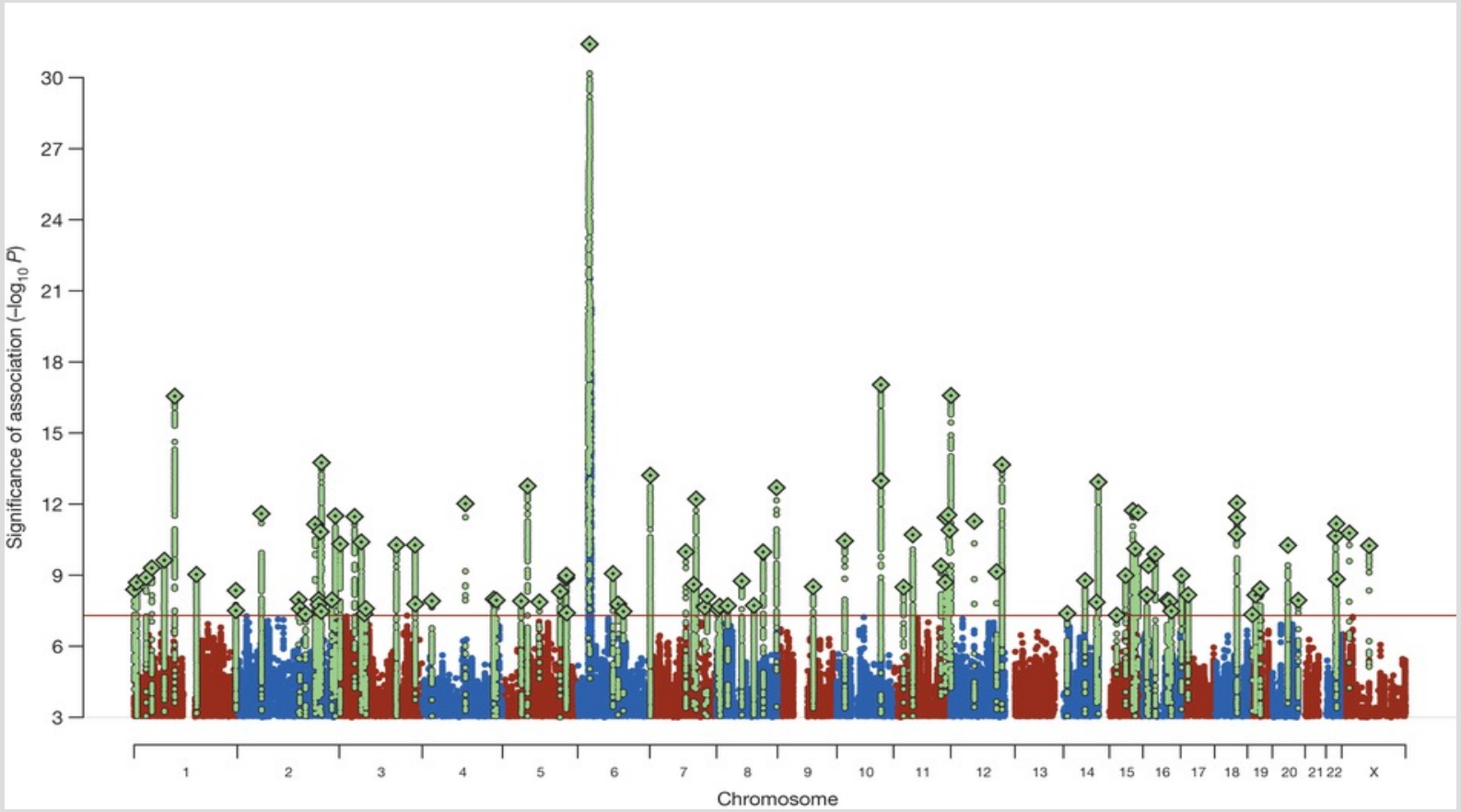
- 9,394 SZ cases and 12,462 controls at 1M SNPs
- 5 GWS loci



SCZ consortium
Nat Gen 2011

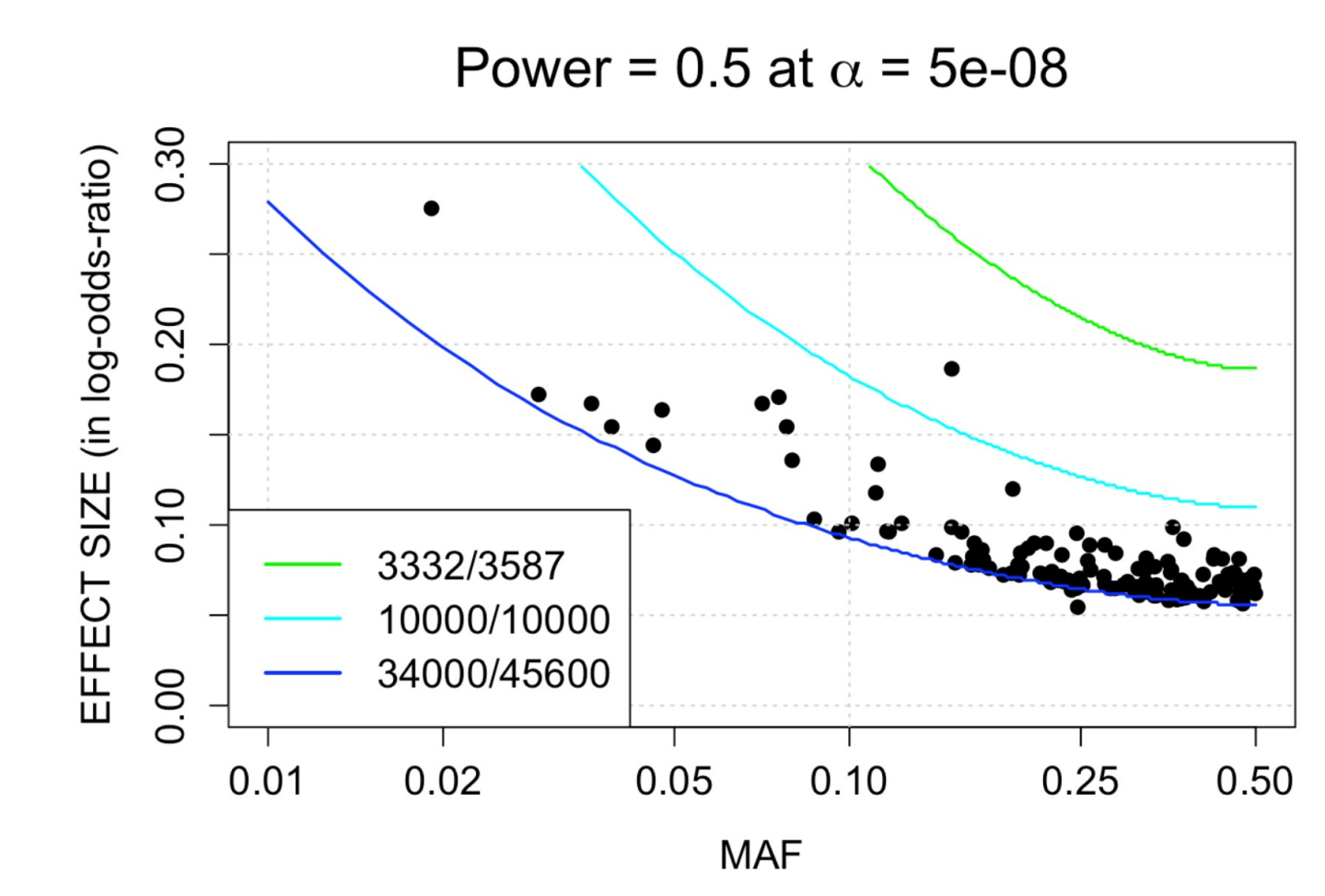
SCHIZOPHRENIA GWAS 3/3 2014

- 34,000 SZ cases and 45,600 controls at 9.5M SNPs
- 108 loci



Psychiatric genomics consortium
Nature 2014

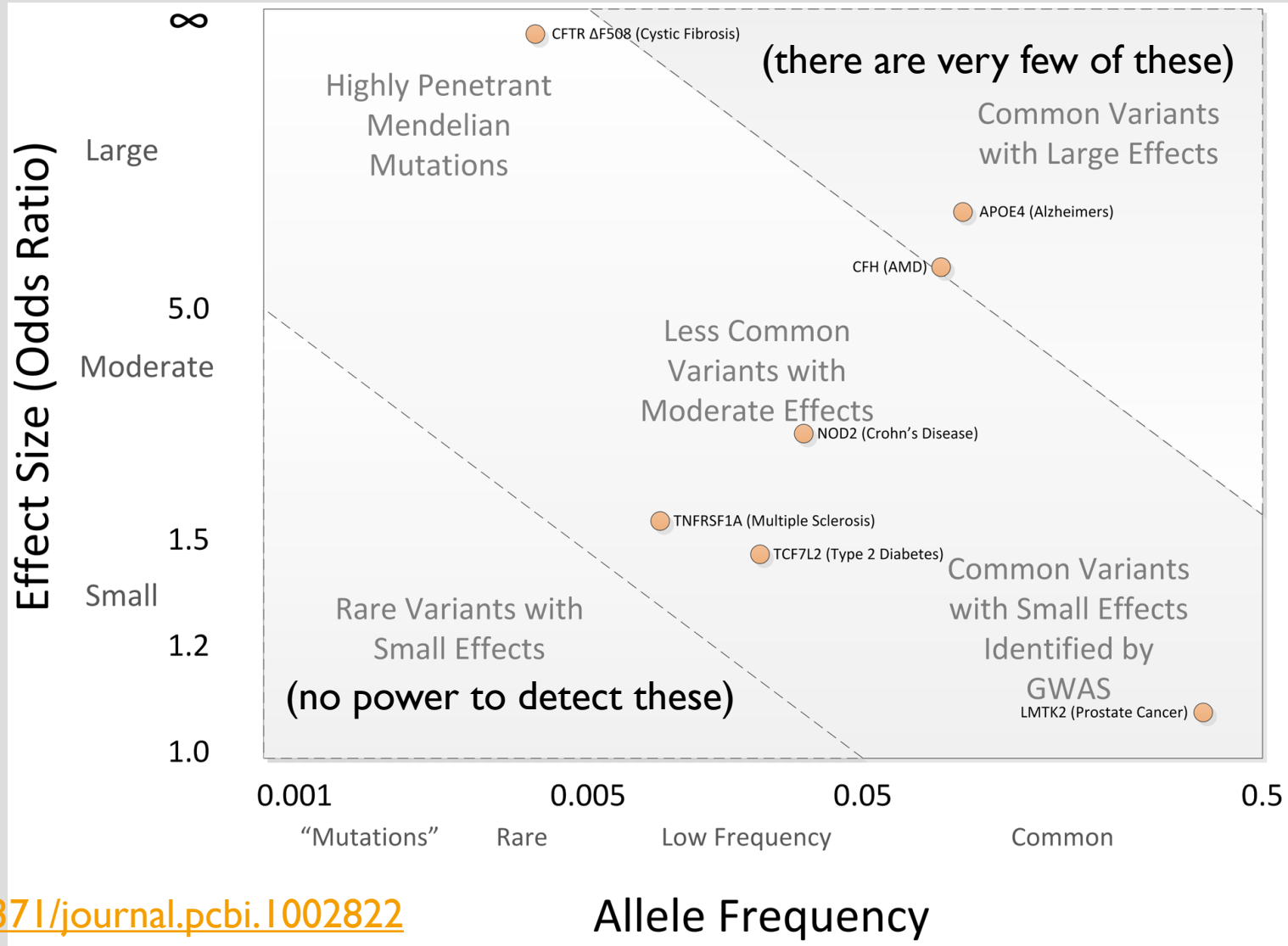
ABSENCE OF EVIDENCE IS NOT EVIDENCE OF ABSENCE



- Non-significant P-value does NOT exclude the existence of non-zero effect, it only excludes the existence of that large effects for which the power to detect them would had been close to one.

Three lines show where power is 0.5 for the 3 schizophrenia GWAS. Dots are now known SZ vaiants.
The first 2 GWAS were underpowered to find the kind of effects that exist for SZ.

EFFECT SIZE, MAF, AND REGION OF POWER



Disease associations are often conceptualized in two dimensions: allele frequency and effect size.
Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of the discovered genetic associations lie on the diagonal denoted by the dashed lines.

Bush & Moore
PLoS Genetics
https://doi.org/10.1371/journal.pcbi.1002822