GWAS 2

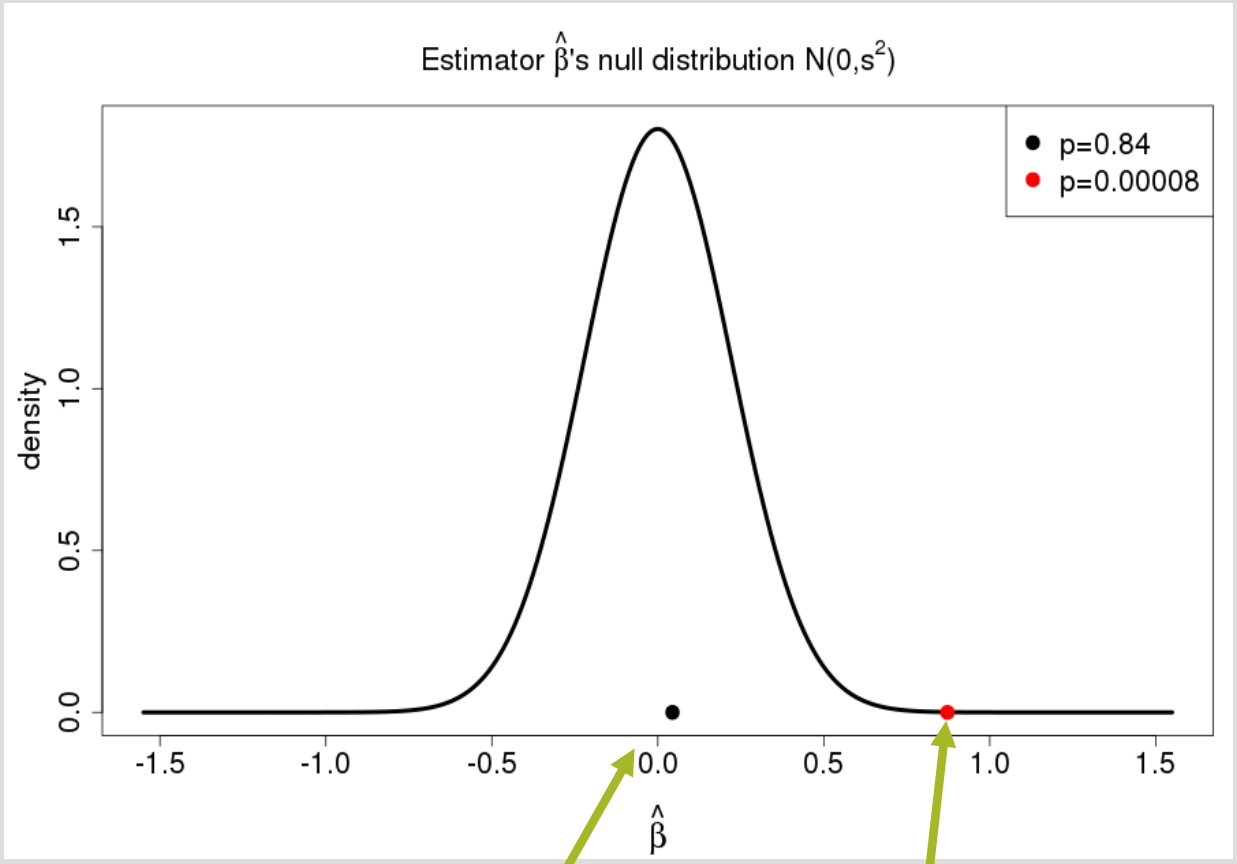
Matti Pirinen
University of Helsinki
March 2, 2023

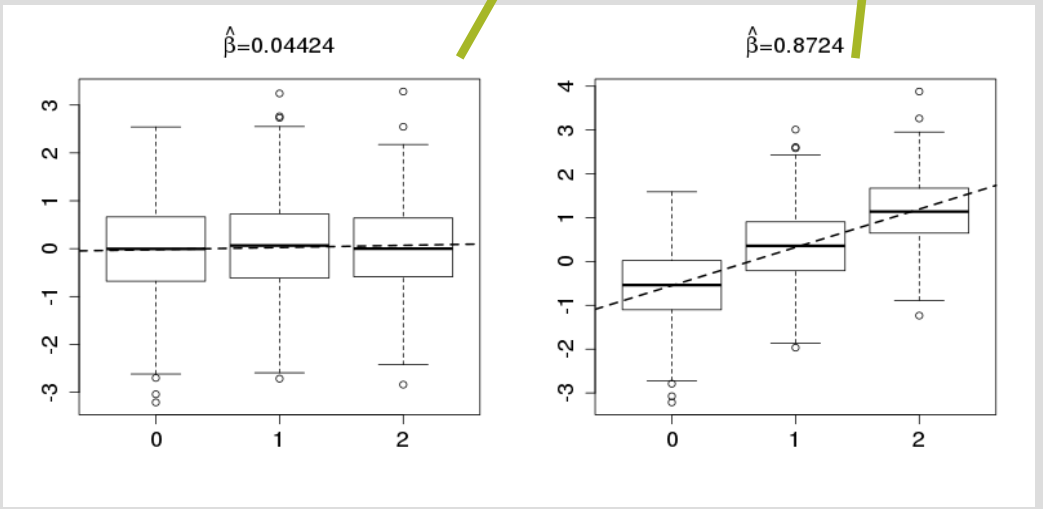
GWAS STATISTICS $\hat{\beta}$ AND SE

- Assuming additive model, β is the difference in mean phenotype between genotype classes 0 and 1, and it is also the difference between classes 1 and 2
  - For QTs the difference is measured on phenotypic scale, often in units of standard deviation of the phenotype
  - For disease traits, the difference is measured on the scale of logarithm of odds of disease
  - We never know the "true" β but can only get an estimate $\hat{\beta}$ from the data with some uncertainty
- Assuming reasonable sample sizes (say MAF > 0.1% and N > 100), standard error (SE) of $\hat{\beta}$ describes the uncertainty of the estimate
  - 95% confidence interval for $\hat{\beta}$ results by putting ~2 SEs around the estimate
  - Technically, SE is an estimate of the standard deviation of the sampling distribution of $\hat{\beta}$

P-VALUE

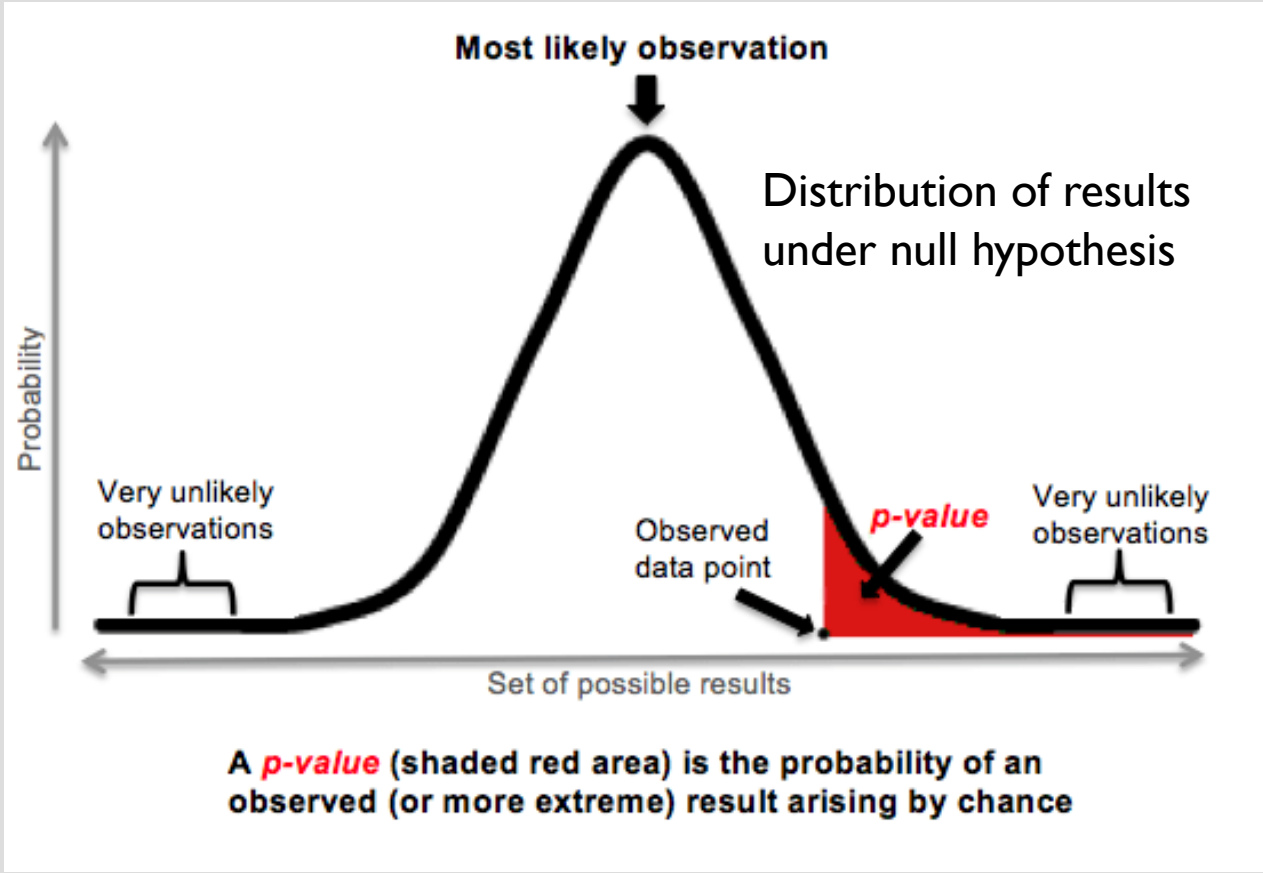
- Is the observed slope plausible if true slope = 0 ?
- P-value: Probability that "by chance" we get at least as extreme value as we have observed, if true slope = 0
- P = 0.84: No evidence for deviation from null
- P = 8e-5: Unlikely under the null → maybe not null





P-VALUE

- P-value: Probability of getting at least as extreme data set in terms of effect size estimate as the one that has been observed assuming that the true effect size is 0, i.e., assuming that the deviation of the observed effect size from 0 is just due to statistical sampling variation.
- "At least as extreme" can have different definitions
  - One-tailed (Figure) or two-tailed (default)

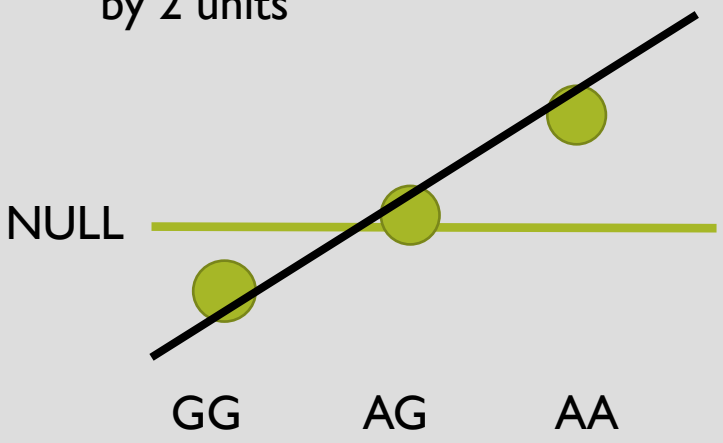


images.realclear.com

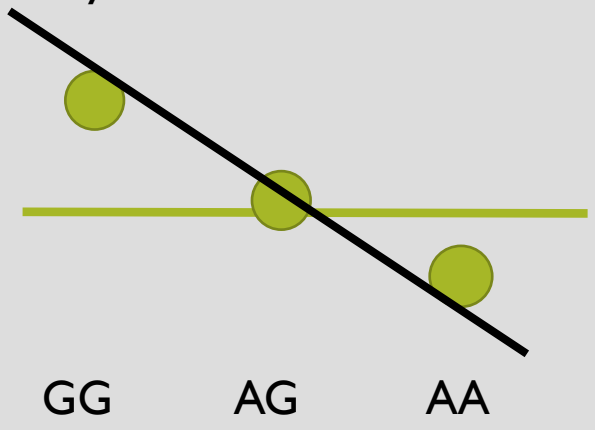
WHY USE TWO-SIDED P-VALUES?

- What is ”at least as extreme data set as what we have observed”?
  - Depends on our null hypothesis
  - Typically, null is that slope $\beta = 0$, and then allele A increasing (and G decreasing) phenotype by 2 units is equally ”extreme” as A decreasing (and G increasing) by 2 units

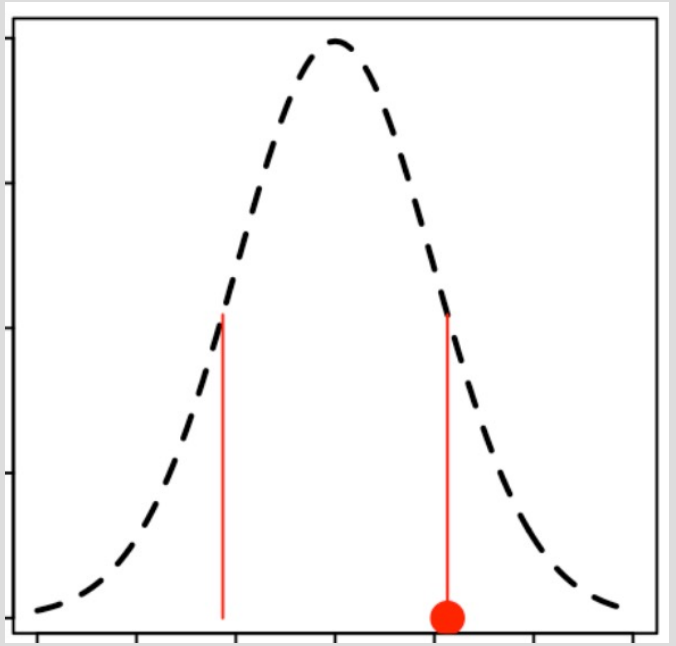
A increases (and G decreases) by 2 units



A decreases (and G increases) by 2 units



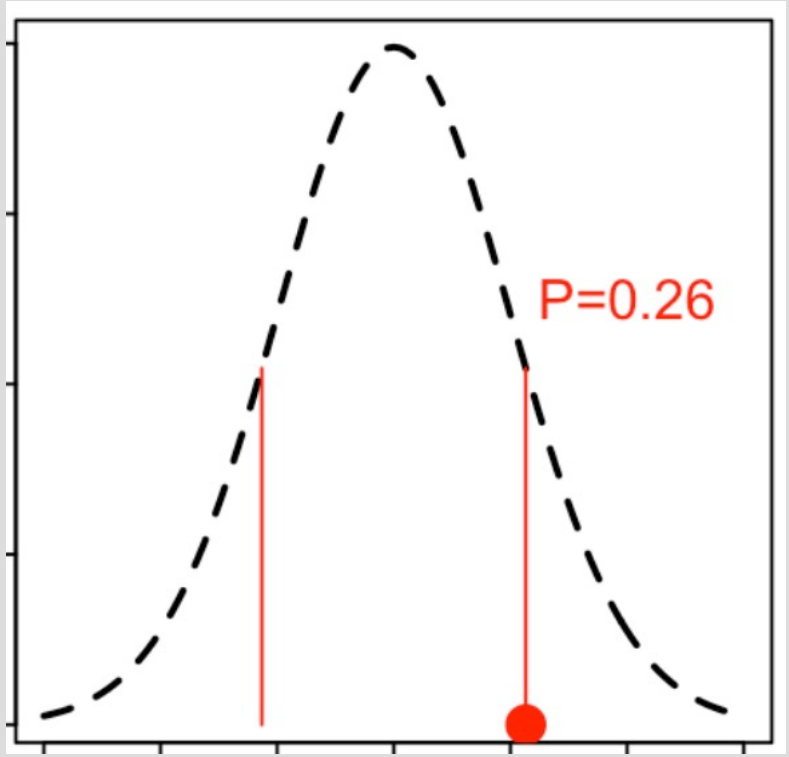
Distribution of results under null hypothesis



2-sided P-value = sum of the two tail probabilitites

P-VALUE

- Small P-value tells that the observation would have been unlikely if there was no real non-zero effect
  - Small P-value can arise because of a real difference ☺
  - **OR** because an unlikely event has happened without a real difference ☹

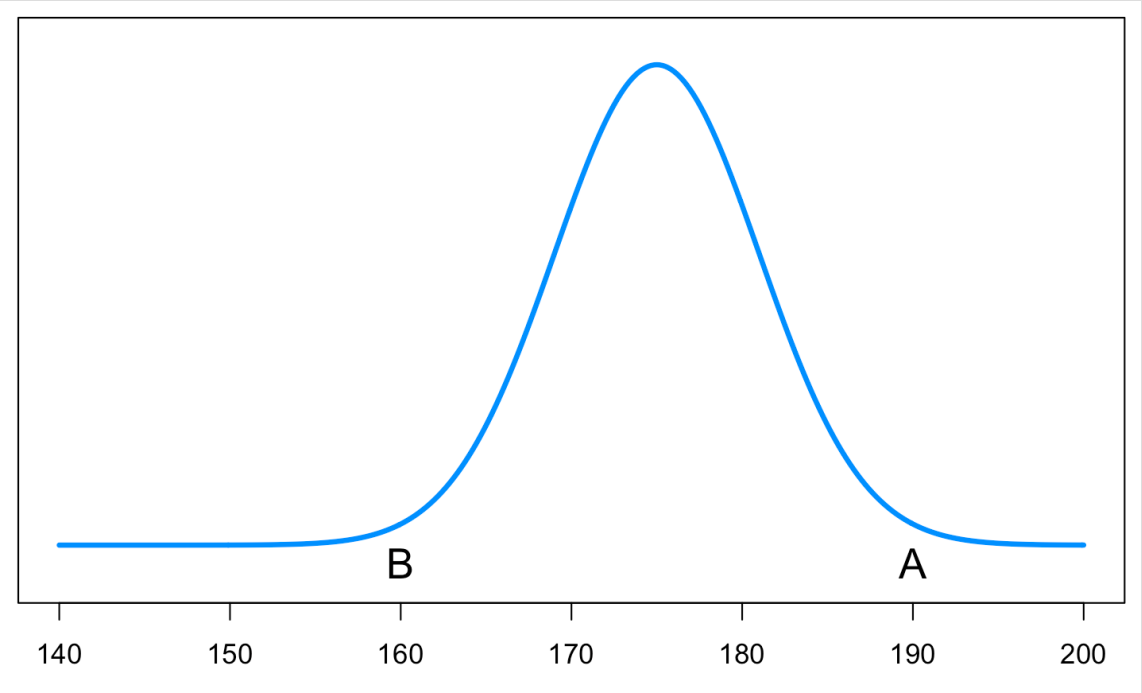


BUT P-VALUE IS NOT PROBABILITY OF THE NULL HYPOTHESIS

- To talk about probabilities of models or hypotheses we must first specify all possible competing models and then compare them against each other

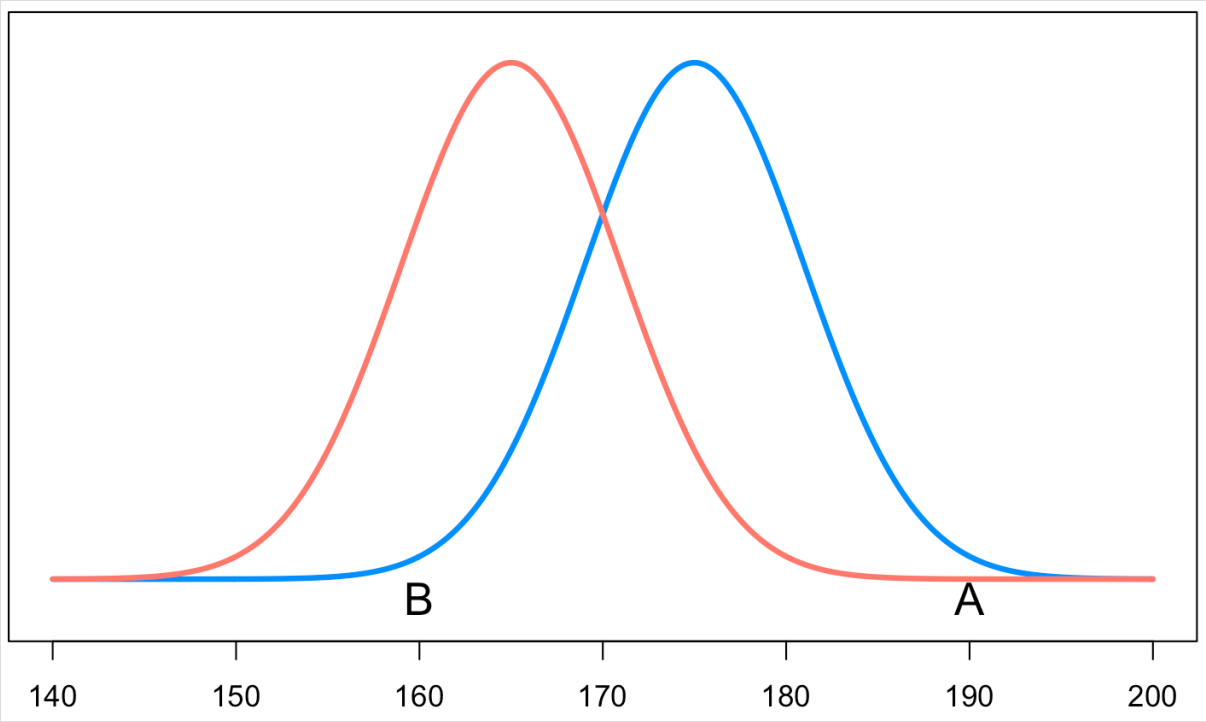
P-VALUE IS NOT PROBABILITY OF THE NULL HYPOTHESIS

- Suppose we want to predict probabilistically whether an individual is a male
- We have observed heights for individuals A (190cm) and B (160 cm)
- Male population has mean 175 cm and SD of 6 cm
- P-values of both A and B are 0.00620 under the null that individual is male



P-VALUE IS NOT PROBABILITY OF THE NULL HYPOTHESIS

- We have observed heights for individuals A (190cm) and B (160 cm)
- Male population has mean 175 cm and SD of 6 cm
- Female population has mean 165 cm and SD of 6 cm
- What is the probability that A / B are male?



Answer:

A: 99.6% probability of being male

B: 6% probability of being male

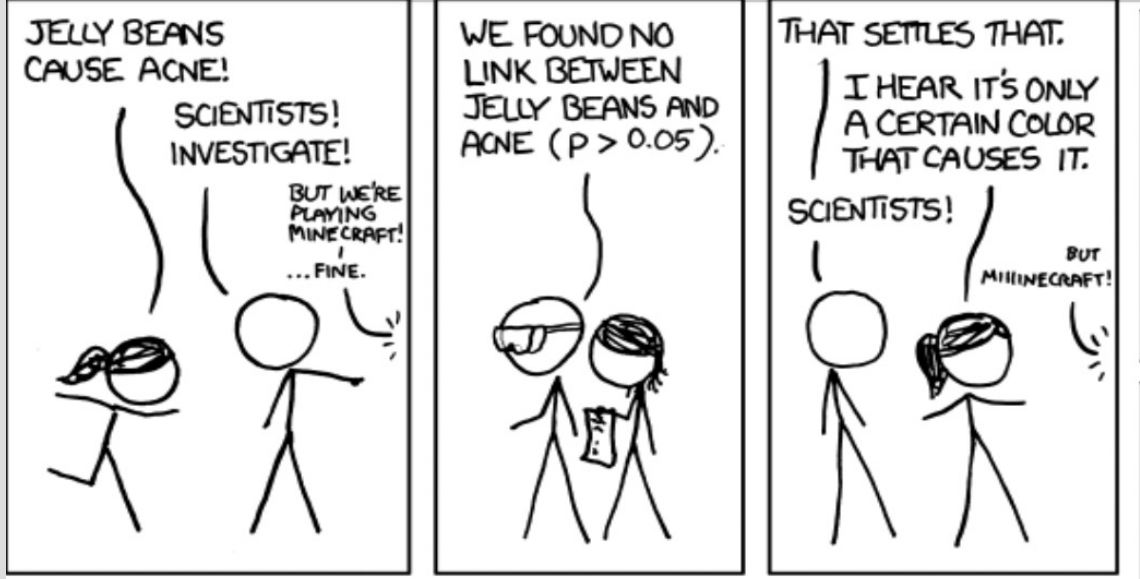
(assuming that *a priori* males and females were equally likely options)

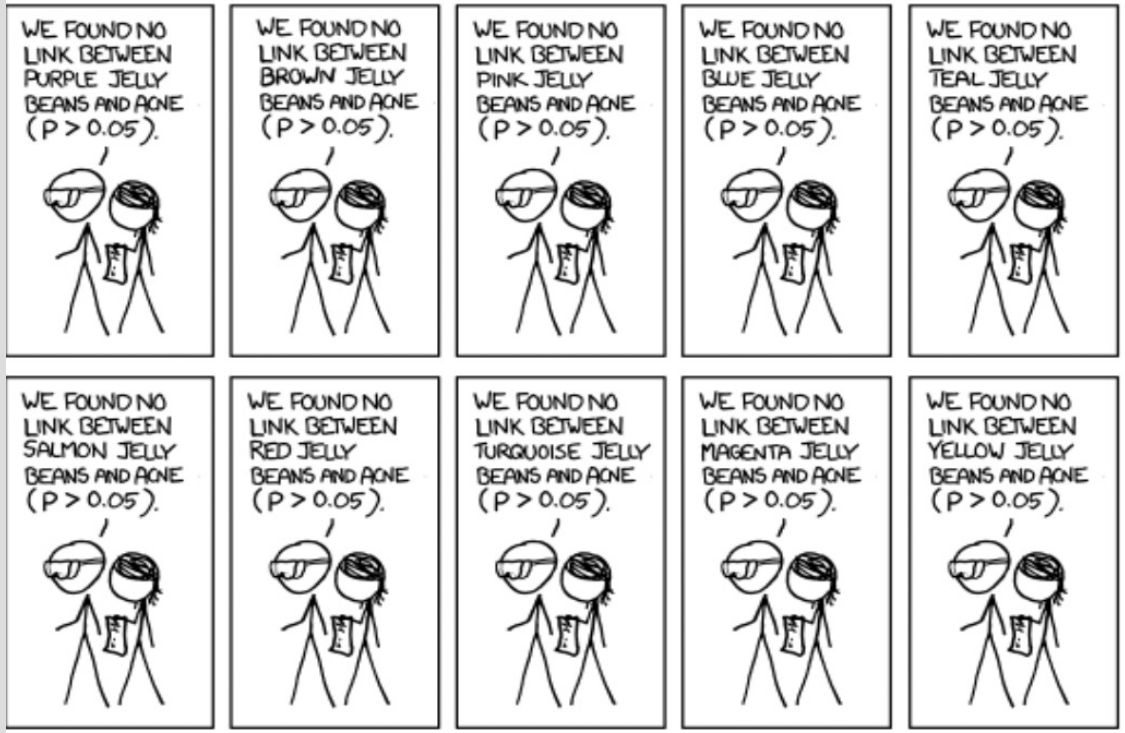
While P-value was small (0.0062) and equal for both A and B, the probability of the "NULL" is completely different

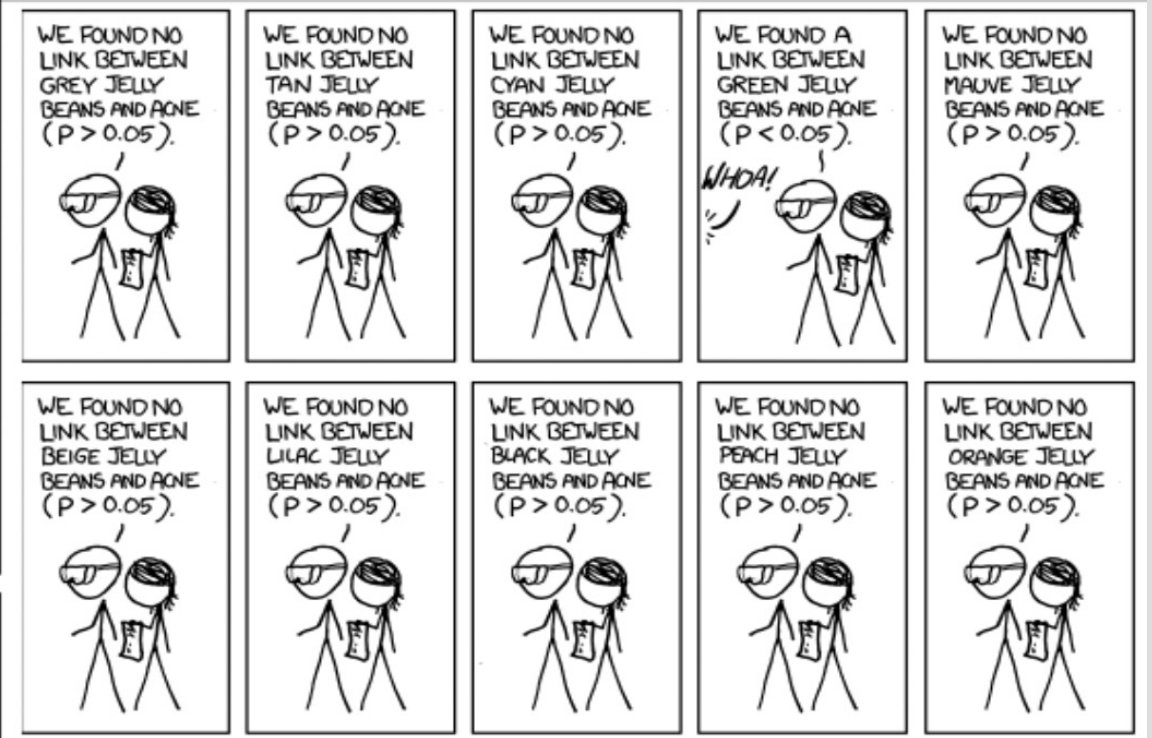
”STATISTICAL SIGNIFICANCE” IS NOT A PROOF BUT A HINT TO LOOK MORE

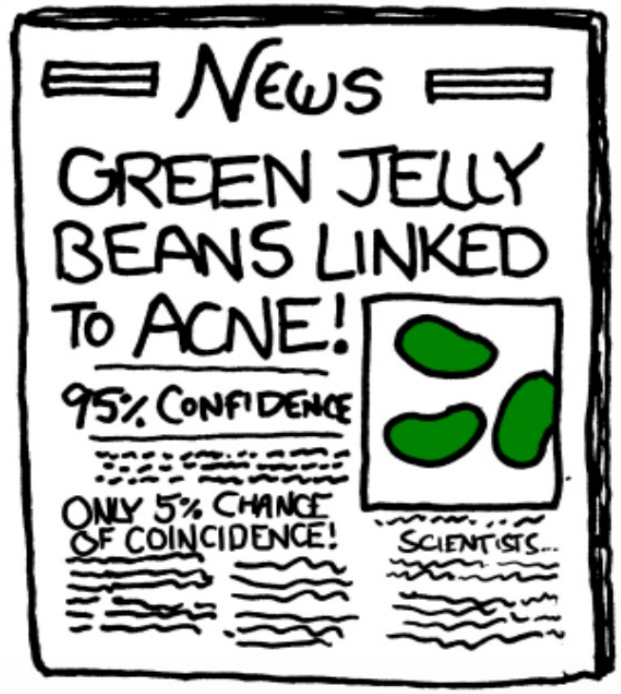
- All thresholds are artificial
  - Whether P = 0.04 or P = 0.06 should have little difference for interpretation of results
    - It is a problem if scientists considered 0.04 being ”significant” while 0.06 being ”not significant” to mean that there was indeed any kind of ”significant” difference between the two results
- But thresholds are a practical tool to handle large data sets and that’s why we use them

MULTIPLE TESTING









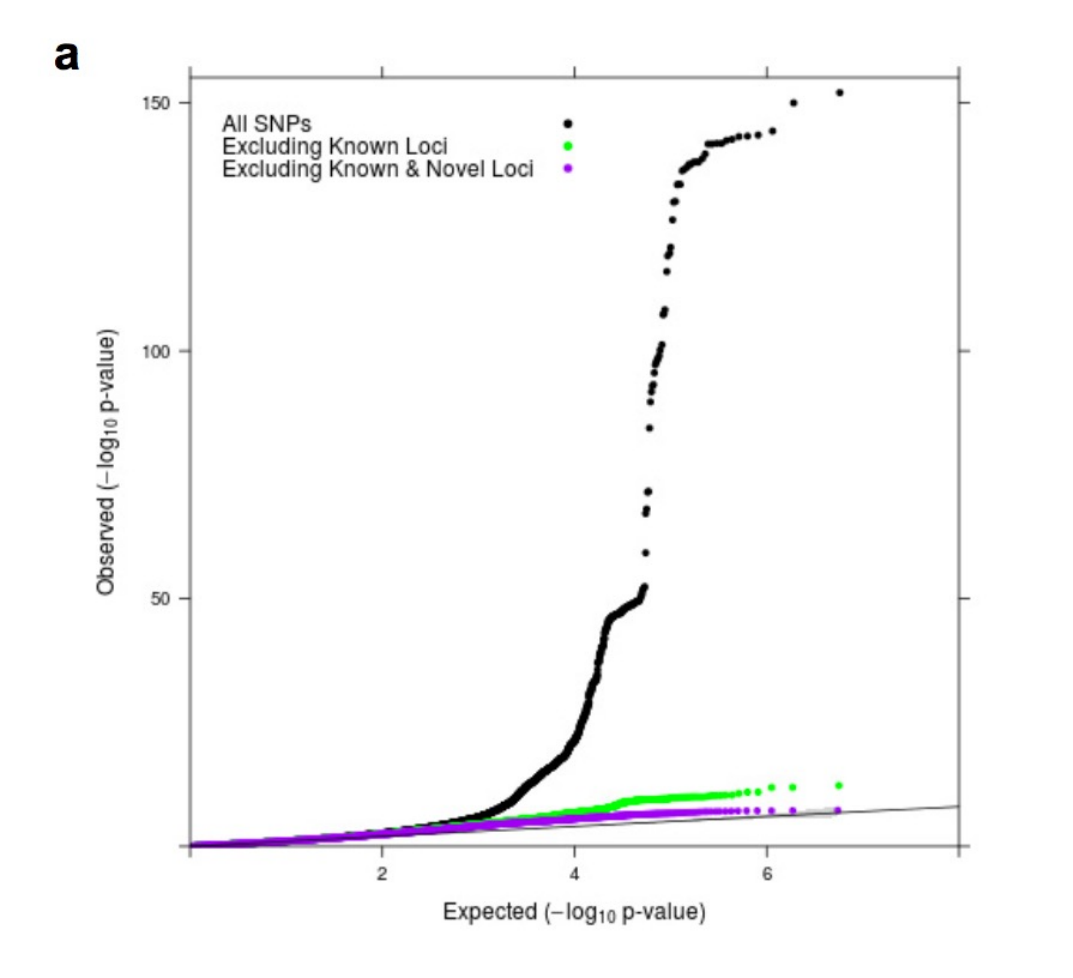
20 indep tests under the null has prob. of 64%
of producing at least 1 "significant" finding

AMERICAN STATISTICAL ASSOCIATION STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES (THE AMERICAN STATISTICIAN 70, 2016)

https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

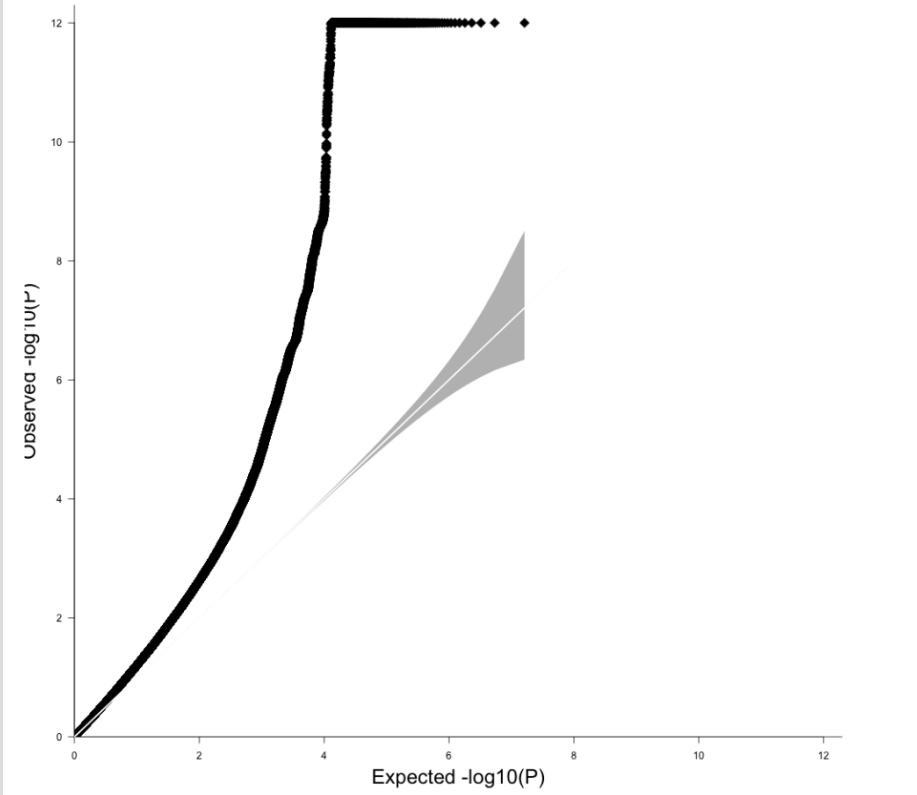
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

QQ-PLOTS IN GWAS



BMI. Locke et al. 2015. Supplementary Figure 1.

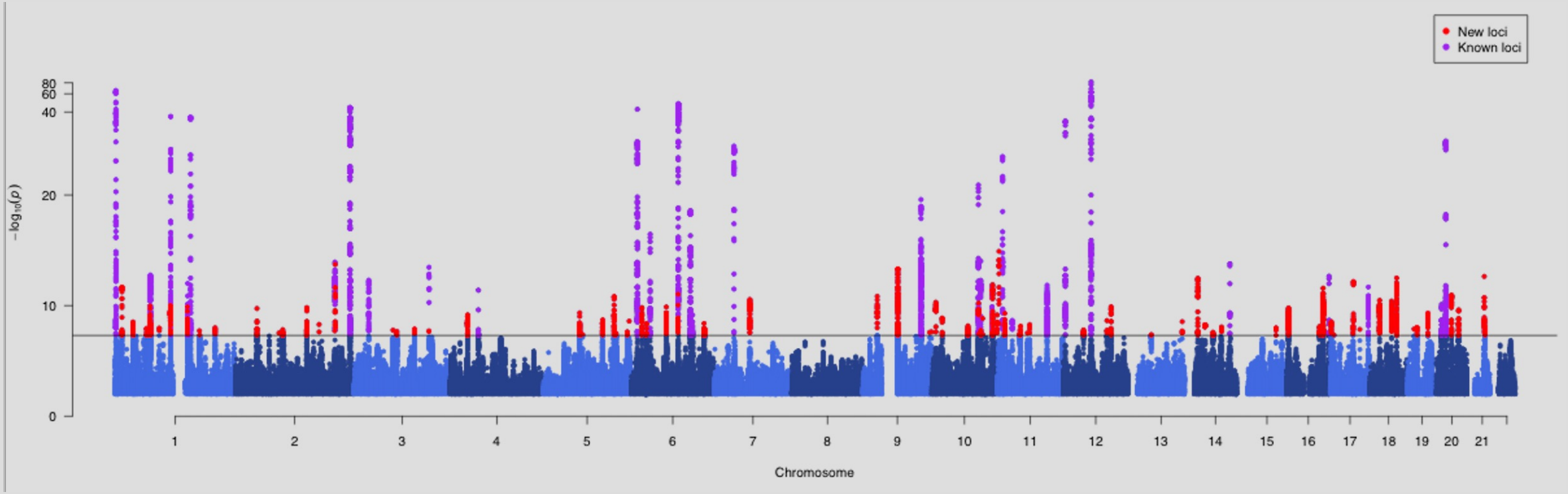
Supplementary Figure 4. QQ-plot of the primary analysis test statistics. In the analysis of all migraine (59,674 cases vs. 316,078 controls), $\lambda_{GC} = 1.24$. For clarity, the observed association *P*-values along the vertical axis have been limited to a minimum value of 1×10^{-12}. The shaded area represents the 95% confidence intervals of expected *P*-values under the null hypothesis.



Migraine. Gormley et al. 2016.

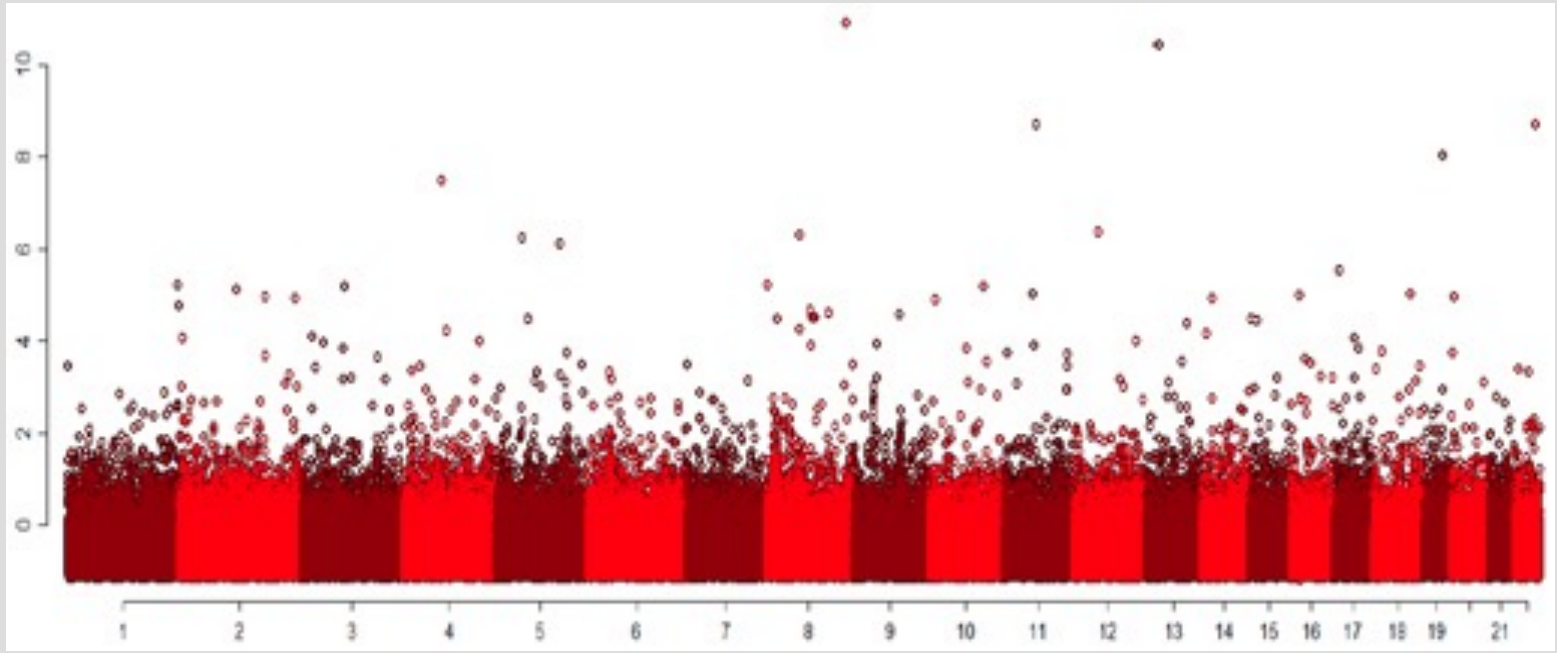
R package 'qqman' can also make a simple qq-plot (but not confidence bounds)

MANHATTAN PLOT



A good quality Manhattan plot of common variants
shows clusters of similar P-values: neighboring variants support each other.

MANHATTAN PLOT

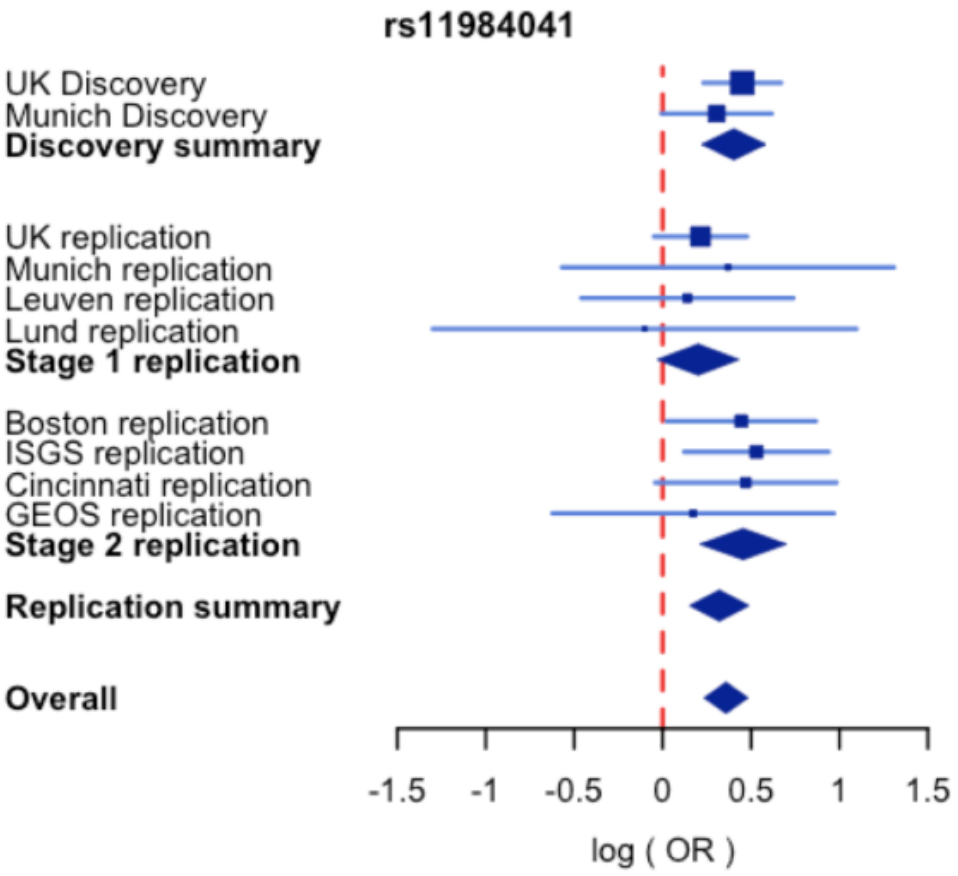


Sebastiani et al.
2010 Science
(retracted 2011 due to QC issues)

Manhattan plot like this suggests that there may be quality control (QC)
problems with individual variants that are not supported by their neighbors.

Especially in case-control analyses, where cases and controls are genotyped separately,
strict QC must be iterated until Manhattan plot looks clean.

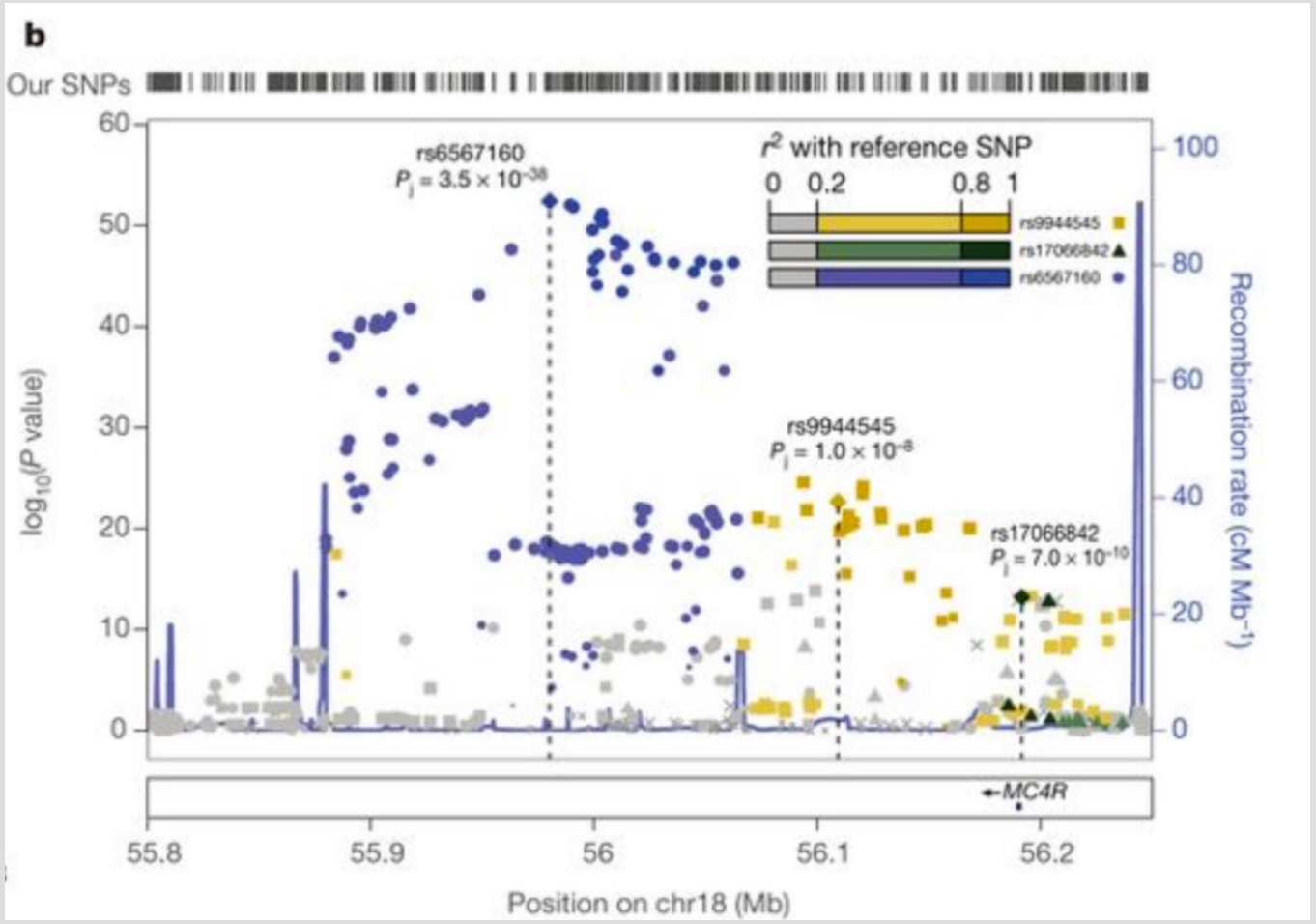
REPLICATION



- Forest plot shows beta and 95% CI for different studies
- We want many cohorts to support the association
- We combine all results into one meta-analyzed result

WTCCC2 & ISGC:
Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke.
Nat Genet. 2012 44(3):328-33

GWAS LOCUS WITH MANY CORRELATED VARIANTS

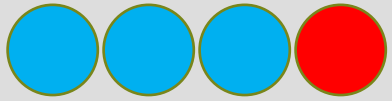


Locke et al. 2015
Nature

MOTIVATION FOR P-VALUE IN CASES-CONTROL SETTING?

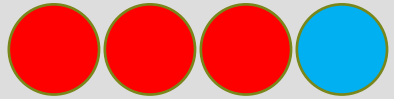
- Assume N1 = N0 = 4
- We want to know: Is the proportion of mutation carriers (red) different between groups?
- We observe: Proportion of carriers in the samples.
- Could the observed difference (75% vs 25%) be just a "chance effect"?

Sample from controls:



1/4 = 25%

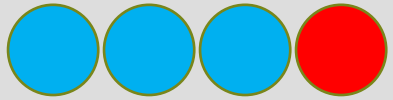
Sample from cases:

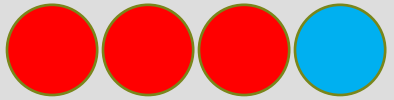


3/4 = 75%

HOW LIKELY IS IT UNDER THE NULL HYPOTHESIS?

- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations from which these samples are taken?

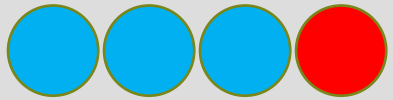


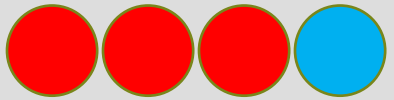


HOW LIKELY IS IT ?

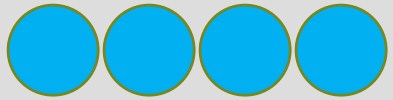
- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations?

Observation:

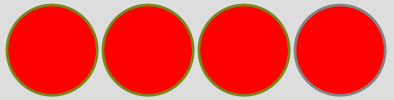




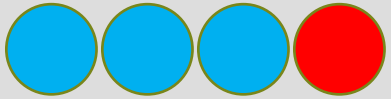
All Possibilities and their probabilities under the null of no difference between the groups



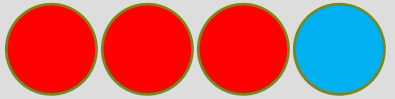
P = 0.014

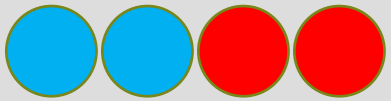


(Computed using combinatorics)

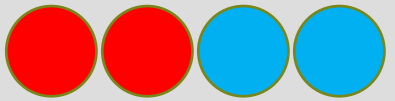


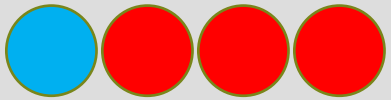
P = 0.229



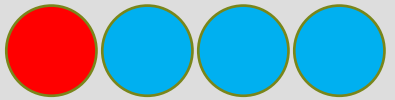


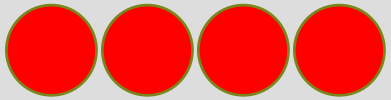
P = 0.514



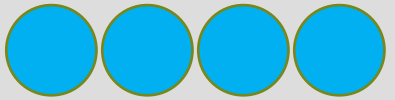


P = 0.229





P = 0.014



Answer: 0.014 + 0.229 + 0.229 + 0.014 = 0.486

HOW LIKELY IS IT ?

- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations?
- Thus in 48.6% of settings where there is no true difference between case and control populations, we would get an observed difference at least as large as 75% / 25%, when we have observed 4 carriers and 4 non-carriers from samples of sizes $N_1 = N_0 = 4$.
  - This observation is not at all convincing evidence for a true difference, even though 75% vs 25% may sound large!
  - Why is this the case? (Answer: Because the sample size is so small.)