GWAS 2

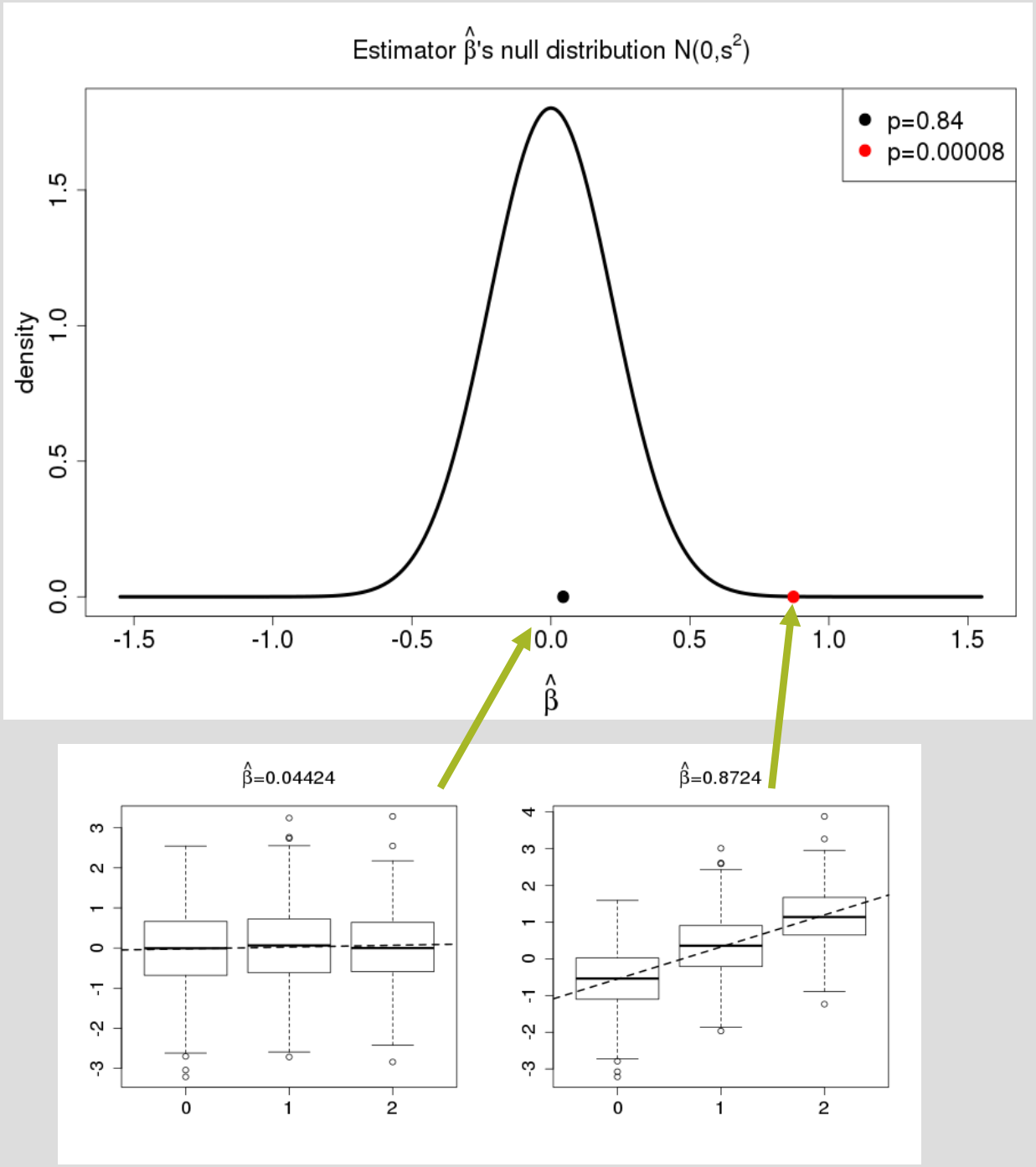
Matti Pirinen
University of Helsinki
15.1.2019

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>1% and N > 100), standard error (SE) of $\hat{\beta}$ describes the uncertainty of the estimate
  - 95% confidence interval for $\hat{\beta}$ is achieved 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

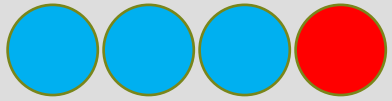
- Is the observed slope plausible if true slope = 0 ?
- P-value: Probability that "by chance" we get as extreme value as we have observed
- P = 0.84: No evidence for deviation from null
- P = 8e-5: Unlikely under the null → maybe not null



MOTIVATION FOR P-VALUE: ARE CASES DIFFERENT FROM CONTROLS?

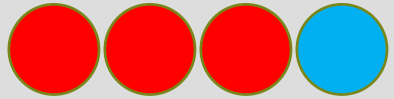
- Assume N1 = N0 = 4
- We want to know: Is the proportion of mutation carriers (red) different between groups?
- We observe: Proportion of carriers in the samples.
- Could the observed difference (75% vs 25%) be just a "chance effect"?

Sample from controls:



1/4 = 25%

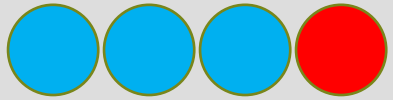
Sample from cases:

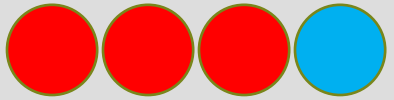


3/4 = 75%

HOW LIKELY IS IT UNDER THE NULL HYPOTHESIS?

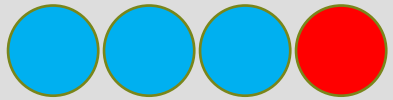
- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations from which these samples are taken?

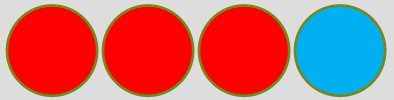


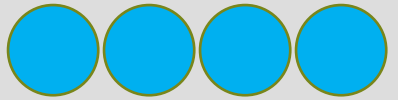


HOW LIKELY IS IT ?

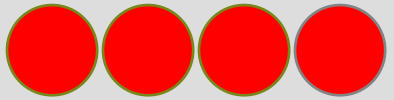
- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations?

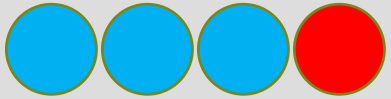




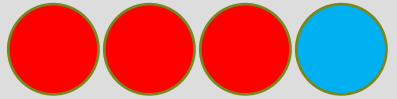


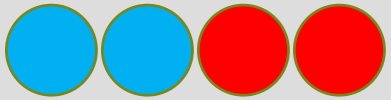
P = 0.014



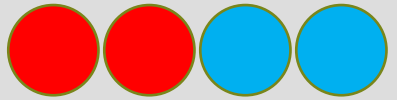


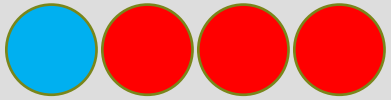
P = 0.229



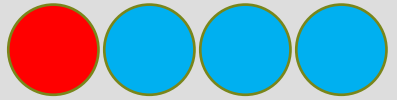


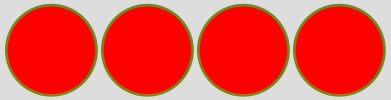
P = 0.514



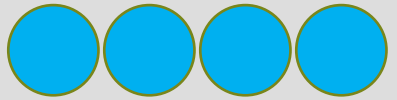


P = 0.229





P = 0.014

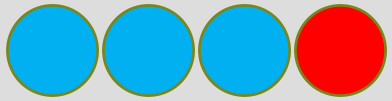


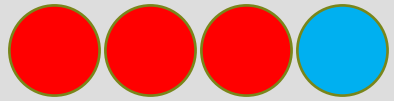
(Computed using combinatorics)

Answer: 0.014 + 0.229 + 0.229 + 0.014 = 0.486

HOW LIKELY IS IT ?

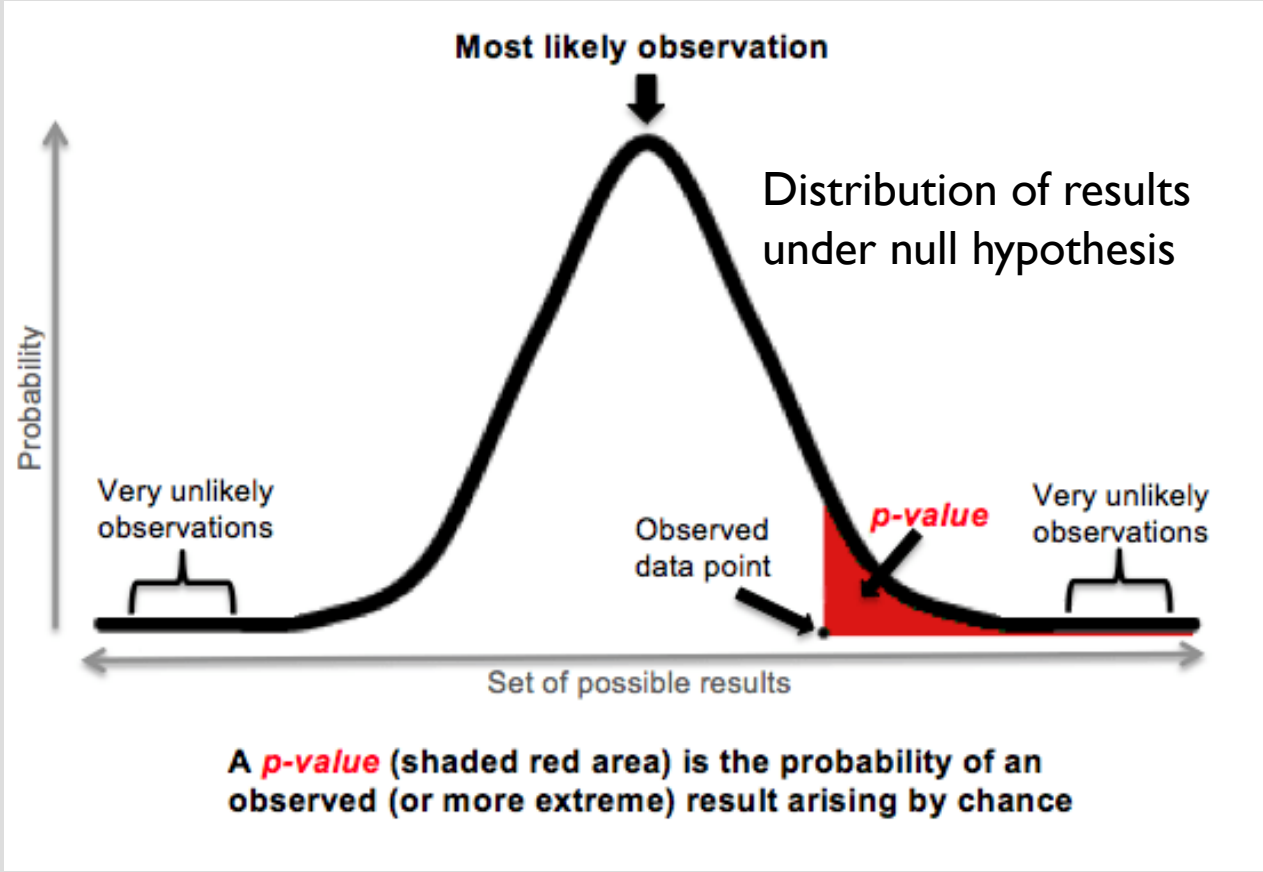
- How likely is it to get at least this large a difference **if** in reality there is **no difference** between the populations?
- Thus in 48.6% of settings where there is no true difference between case and control populations, we would get an observed difference at least as large as 75% / 25%, when we have observed 4 carriers and 4 non-carriers from samples of sizes N1 = N0 = 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? (Because the sample size is so small.)





P-VALUE

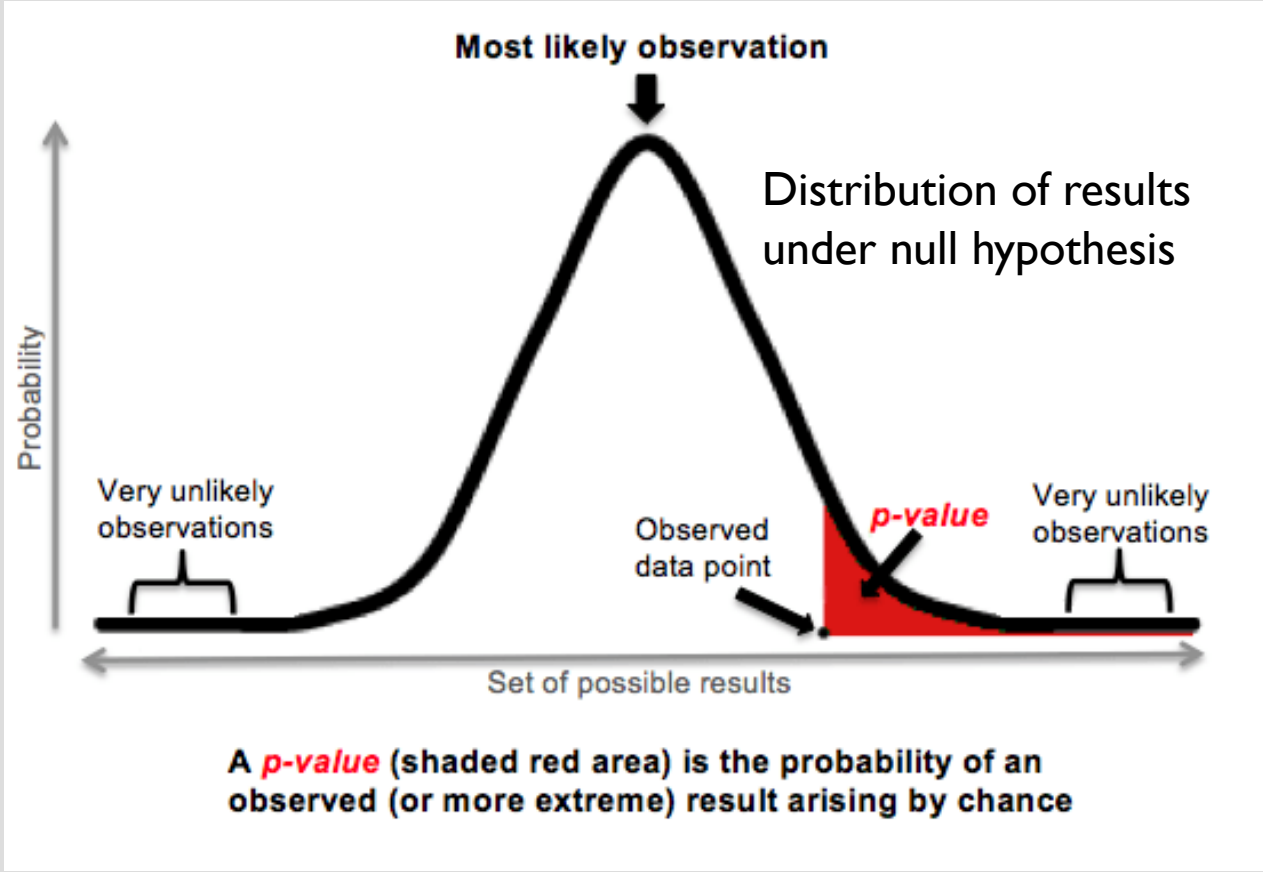
- P-value: Probability of getting at least as extreme data set as the one that has been observed assuming that there is no actual difference between the two groups, i.e., assuming that the observed difference is just a chance effect.
- "At least as extreme" can have different definitions
  - One-tailed (Figure) or two-tailed (default)



images.realclear.com

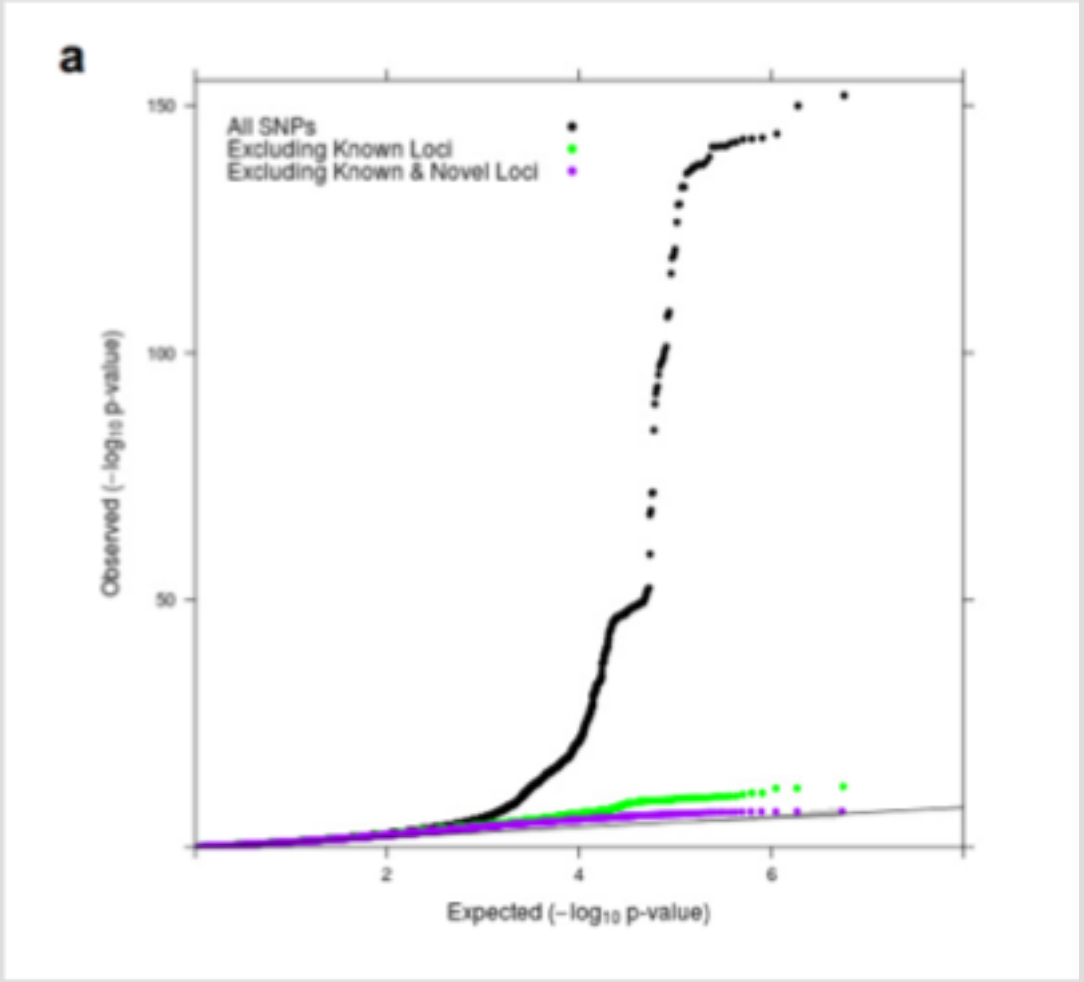
P-VALUE

- Small P-value tells that the observation would have been unlikely if there was no real difference
  - Small P-value can arise because of a real difference ☺
  - **OR** because an unlikely event has happened without a real difference ☹
  - Statistics never claims absolute truth – only informs about appropriate levels of confidence

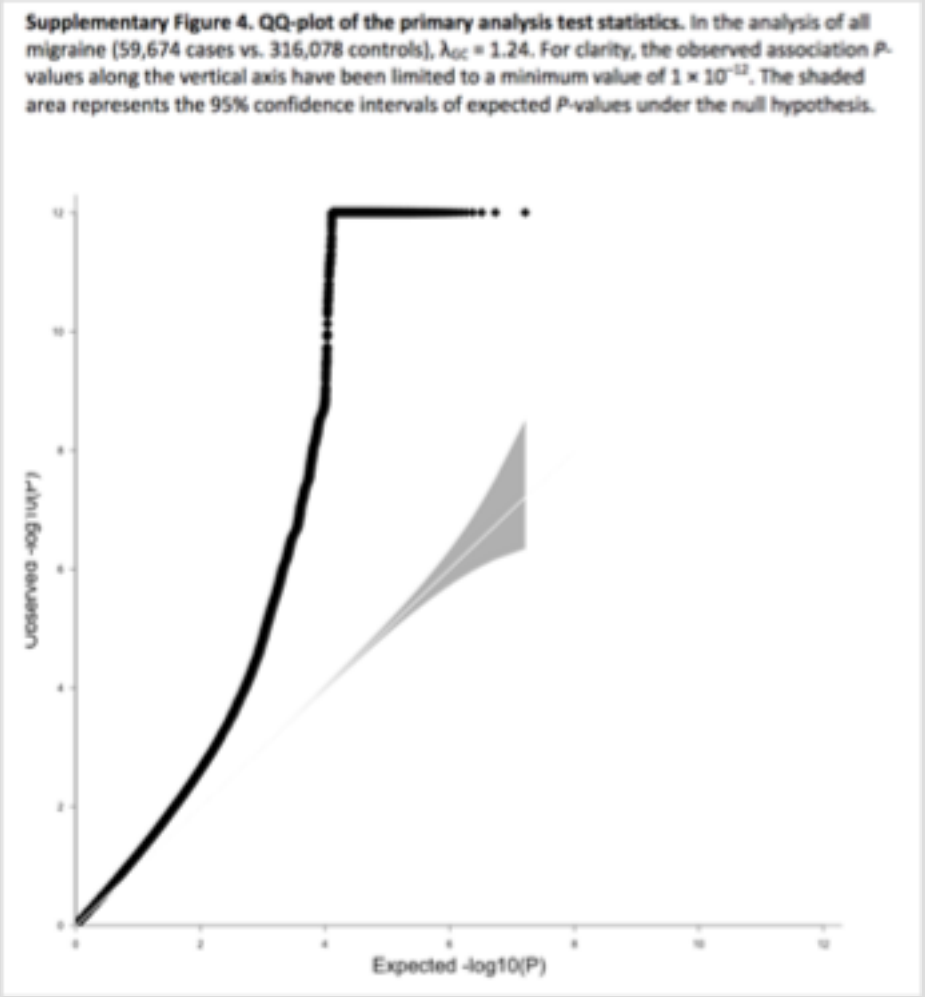


images.realclear.com

QQ-PLOTS IN GWAS



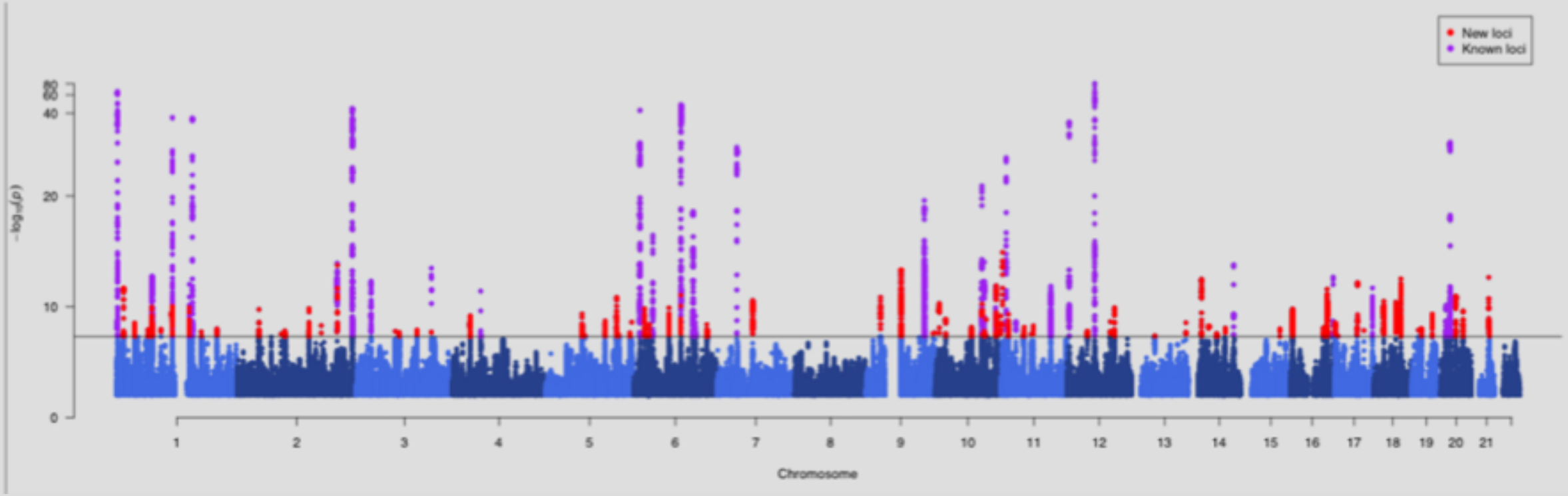
BMI. Locke et al. 2015. Supplementary Figure 1.



Migraine. Gormley et al. 2016.

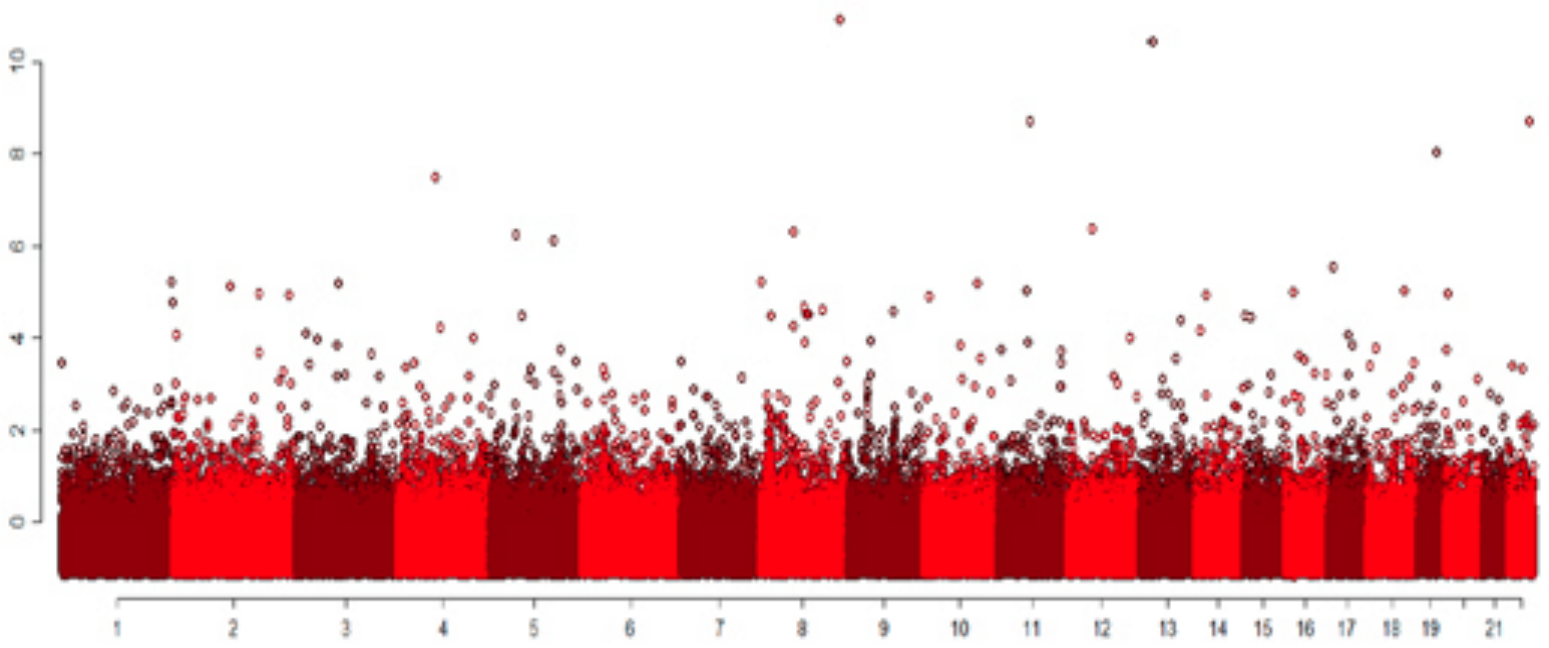
R package 'qqman' can also make a simple qq-plot (but not confidence bounds)

MANHATTAN PLOT



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

MANHATTAN PLOT

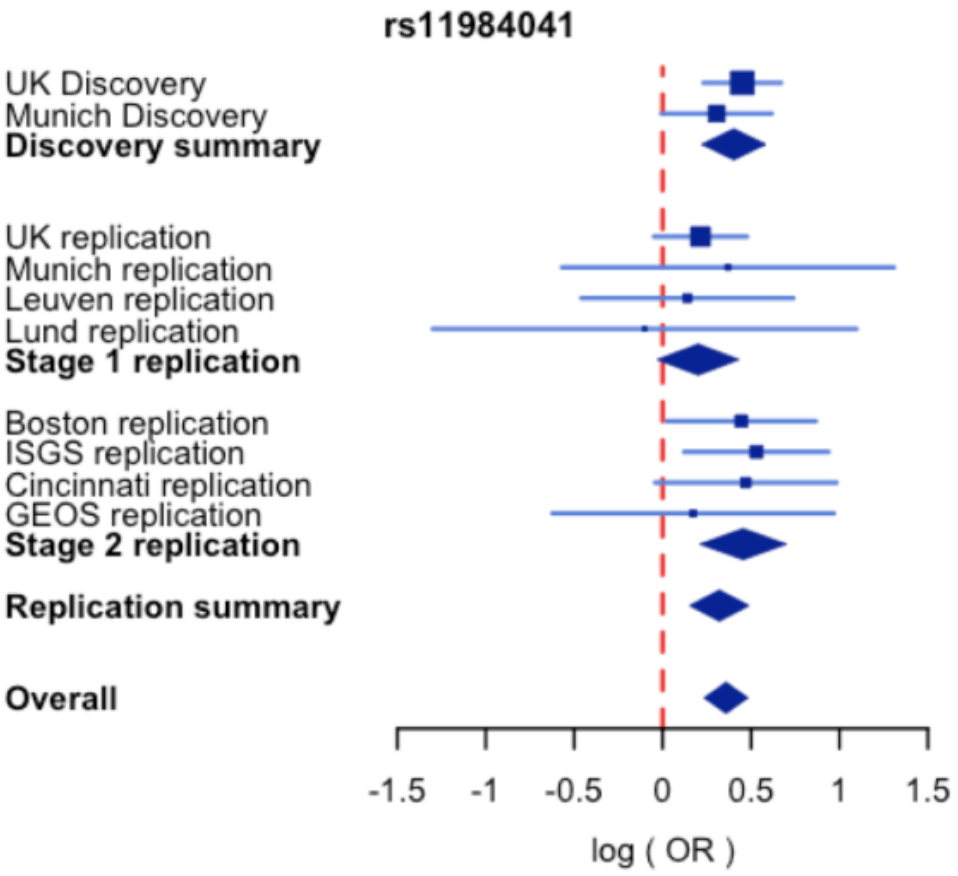


Sebastiani et al.
2010 Science
(retracted 2011 due to QC issues)

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.

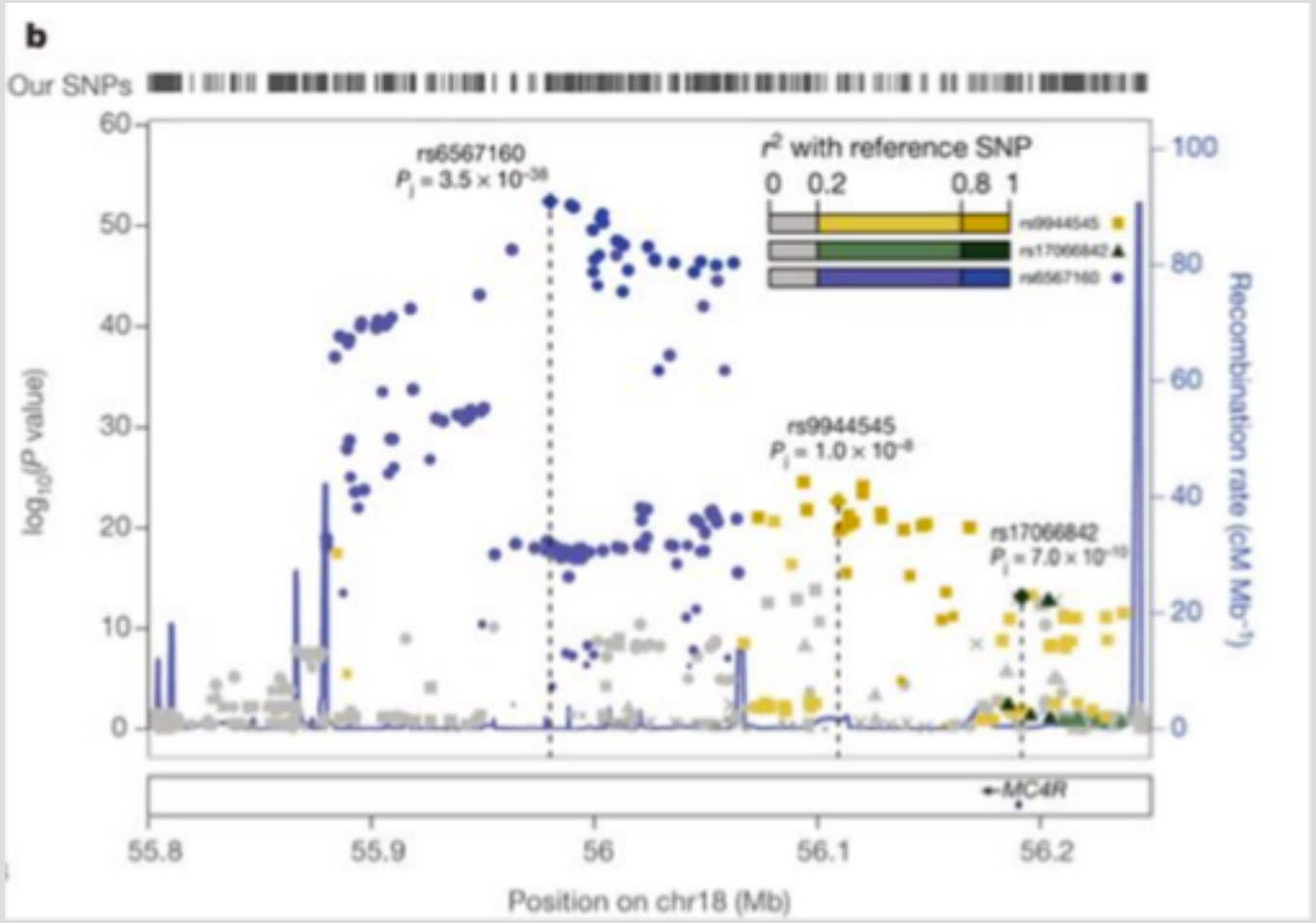
REPLICATION



- 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