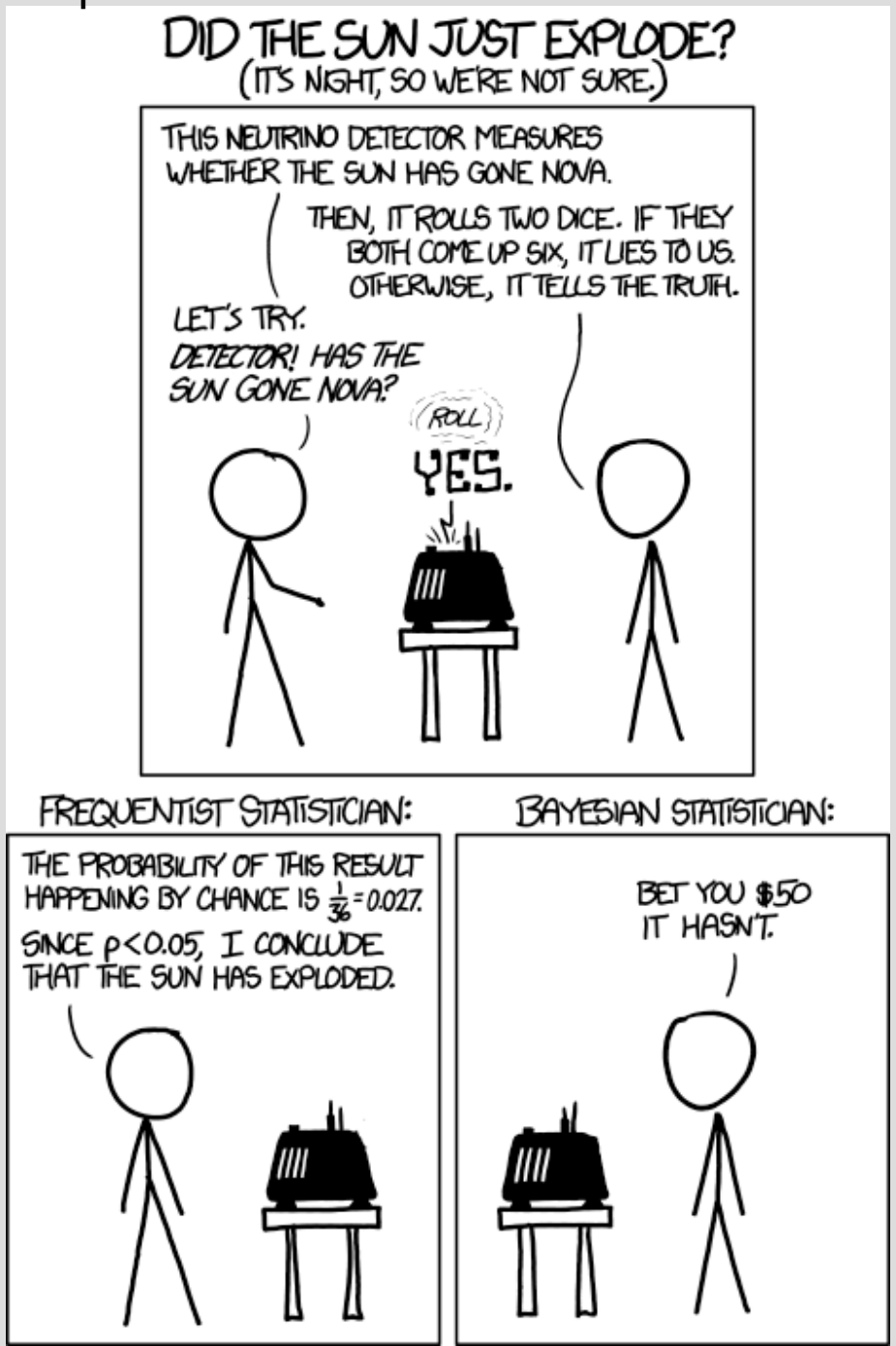
# GWAS 4

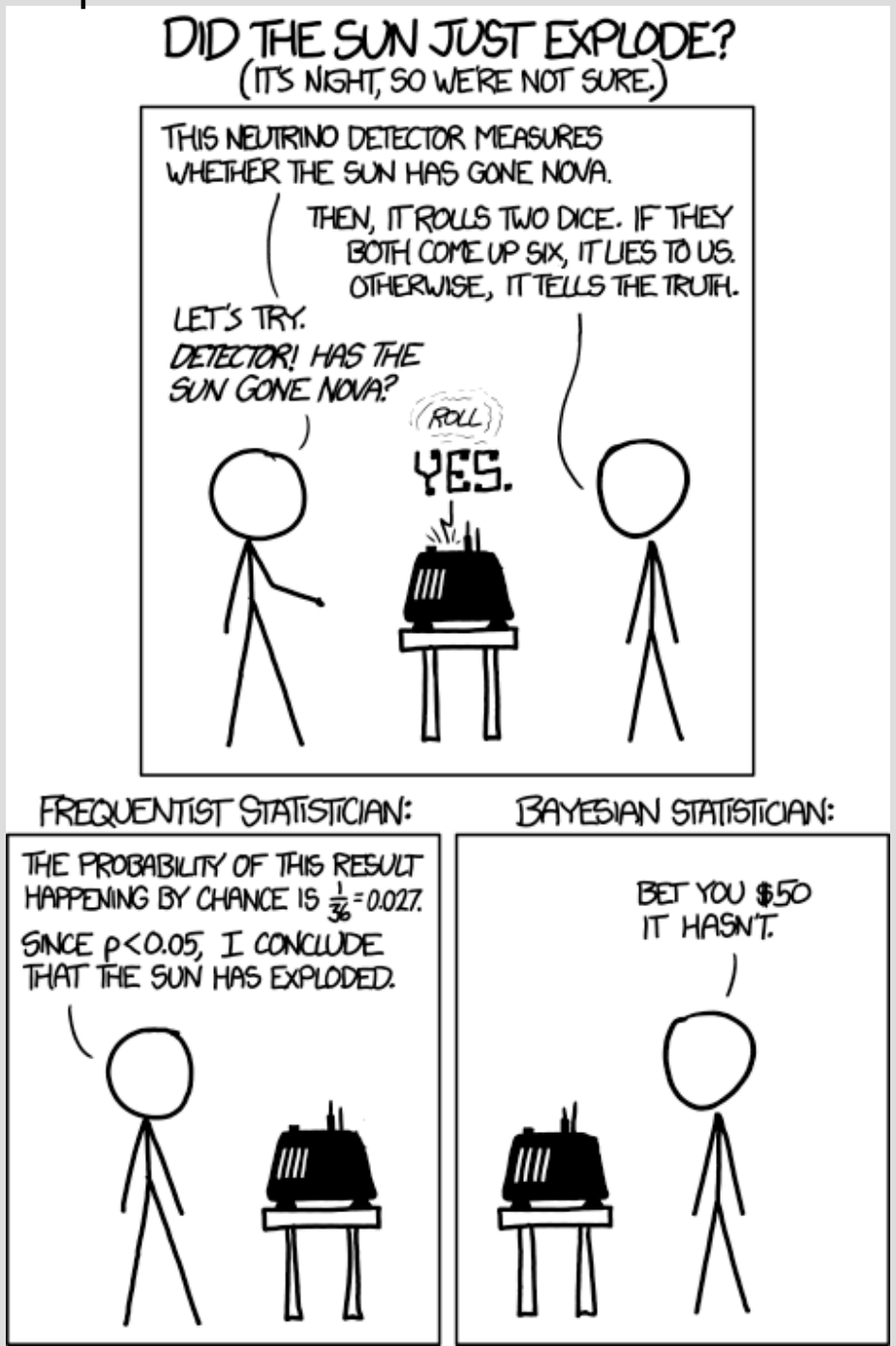
Matti Pirinen
University of Helsinki
4.11.2020

https://xkcd.com/1132/




# BAYES RULE COMBINES PRIOR & OBSERVATION

https://xkcd.com/1132/




# BAYES RULE COMBINES PRIOR & OBSERVATION

X = Sun exploded
Y = Detector says "Yes"

We know P(Y | X) = 0.973 and P(Y | not-X) = 0.027

P(Y) = P(X) * P(Y | X) + P(not-X) * P(Y | not-X)
P(X)*0.973 + (1-P(X)) * 0.027

Bayes rule: $P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$

P(X | Y) = P(X) P(Y | X) / P(Y)
= P(X)*0.973 / (P(X) *0.973 + (1-P(X))*0.027)
= P(X) / (0.0277 + 0.972*P(X))
≤ P(X) / 0.0277 ≤ 40*P(X)

So the observation increases probability of X at most 40 fold compared to prior probability, that is likely very very very small.

# BAYESIAN INFERENCE

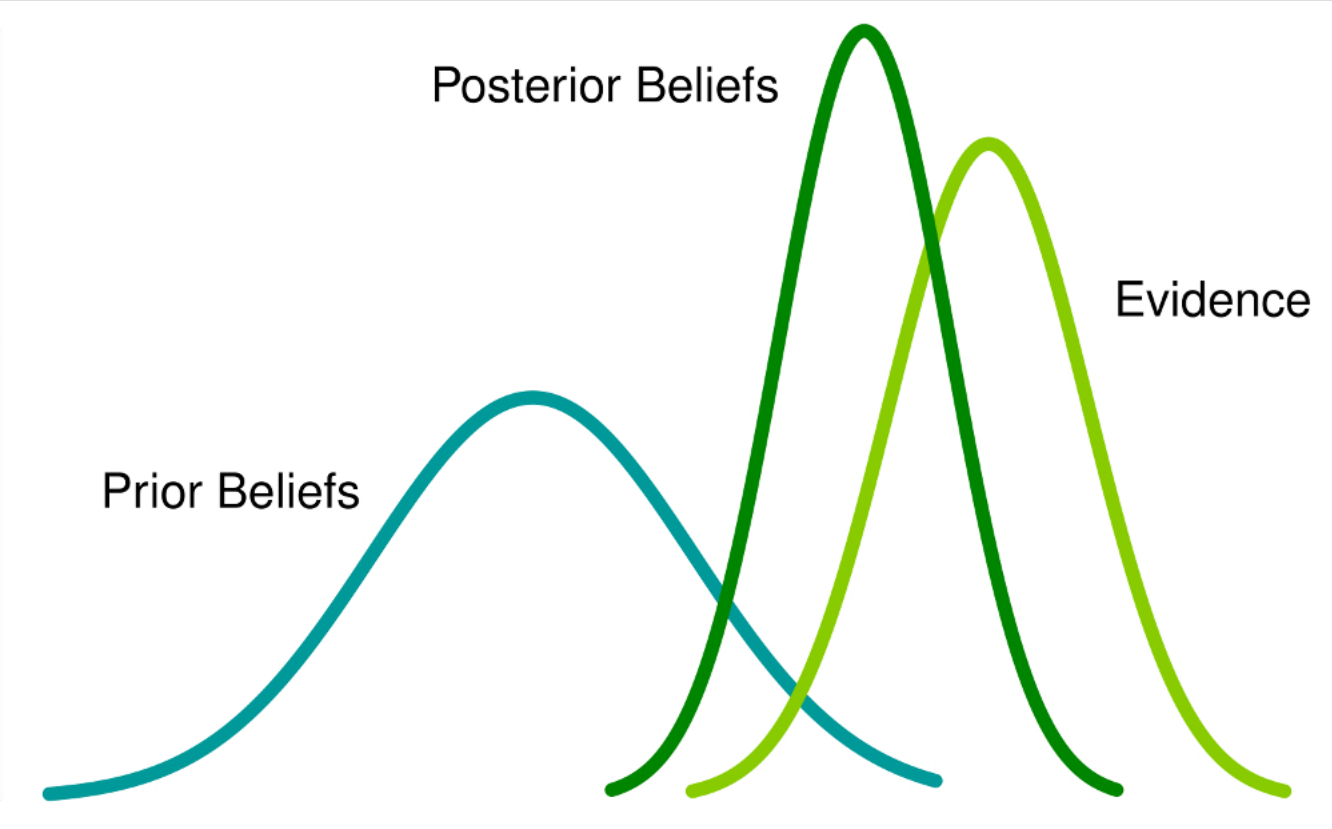



Figure: https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/

- We are estimating a parameter (like an effect size in GWAS)
- We have some prior beliefs where the parameter value is, but we don't know very accurately
- We gather data to learn about the parameter– this gives the evidence based on the gathered data alone
- Bayes rule tells how to consistently combine the prior beliefs and the evidence from the data into a combined posterior belief
- If prior is flat across a range of values (relative to the amount of evidence in data), then posterior will look like evidence in the data
- If prior of some region is extremely small, then we need a lot of evidence before posterior will support strongly that region

# BAYESIAN MODEL COMPARISON

Posterior probability of hypothesis $H_i$

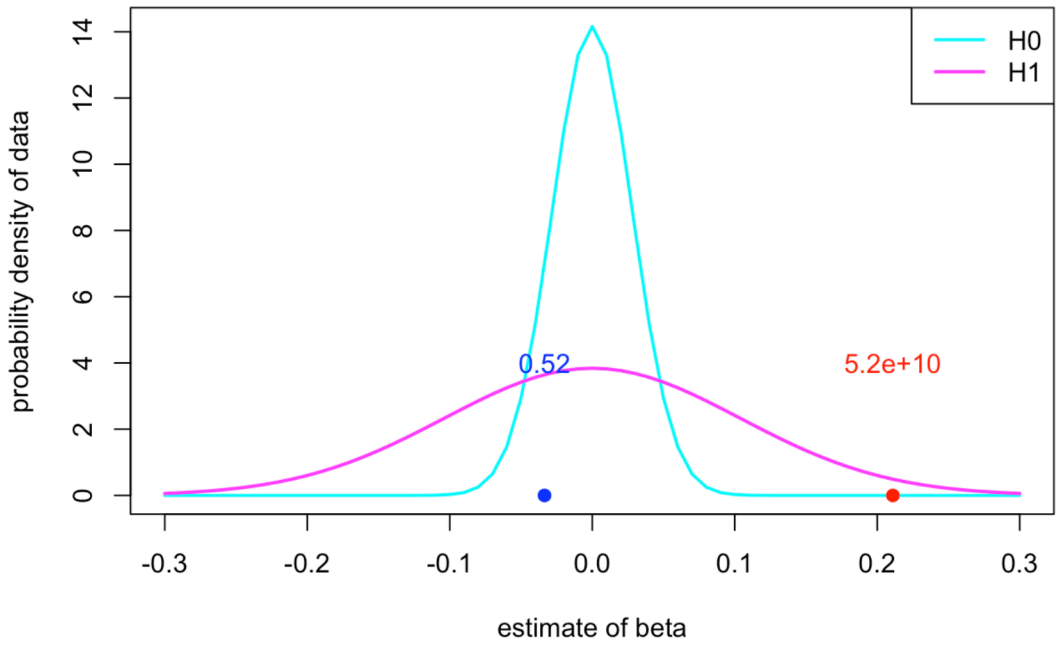
$$P(H_i|\mathcal{D}) = \frac{P(\mathcal{D}|H_i)P(H_i)}{P(\mathcal{D})}, \qquad \text{for } i = 0, 1.$$

$$\underbrace{\frac{P(H_1|\mathcal{D})}{P(H_0|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{prior odds}}$$

To compare the probabilities of two hypotheses we need to define their prior probabilities and the probability distributions how they produce data.

Prior probability of association in GWAS might be in range $10^{-4}$ to $10^{-6}$, but depends on what is known about the variant. What about the Bayes factor?

# P( D | $H_1$)




BF for blue and red effect size estimates are shown.

- For NULL hypothesis, true effect size = 0 and hence the observed effect size has distribution N(0, $SE^2$) – This Normal density avaluated at the observed effect estimate is P ( D | $H_0$)
- For alternative hypothesis, true effect size is assumed to be taken from N(0, $t^2$) and hence the observed effect size has distribution N(0, $t^2$ + $SE^2$)
- Then the Bayes factor is

$$\frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)} \approx \frac{\mathcal{N}\left(\hat{\beta};\, 0,\, \tau_1^2 + \mathrm{SE}^2\right)}{\mathcal{N}\left(\hat{\beta};\, 0,\, \mathrm{SE}^2\right)}$$

# BF VS P-VALUES

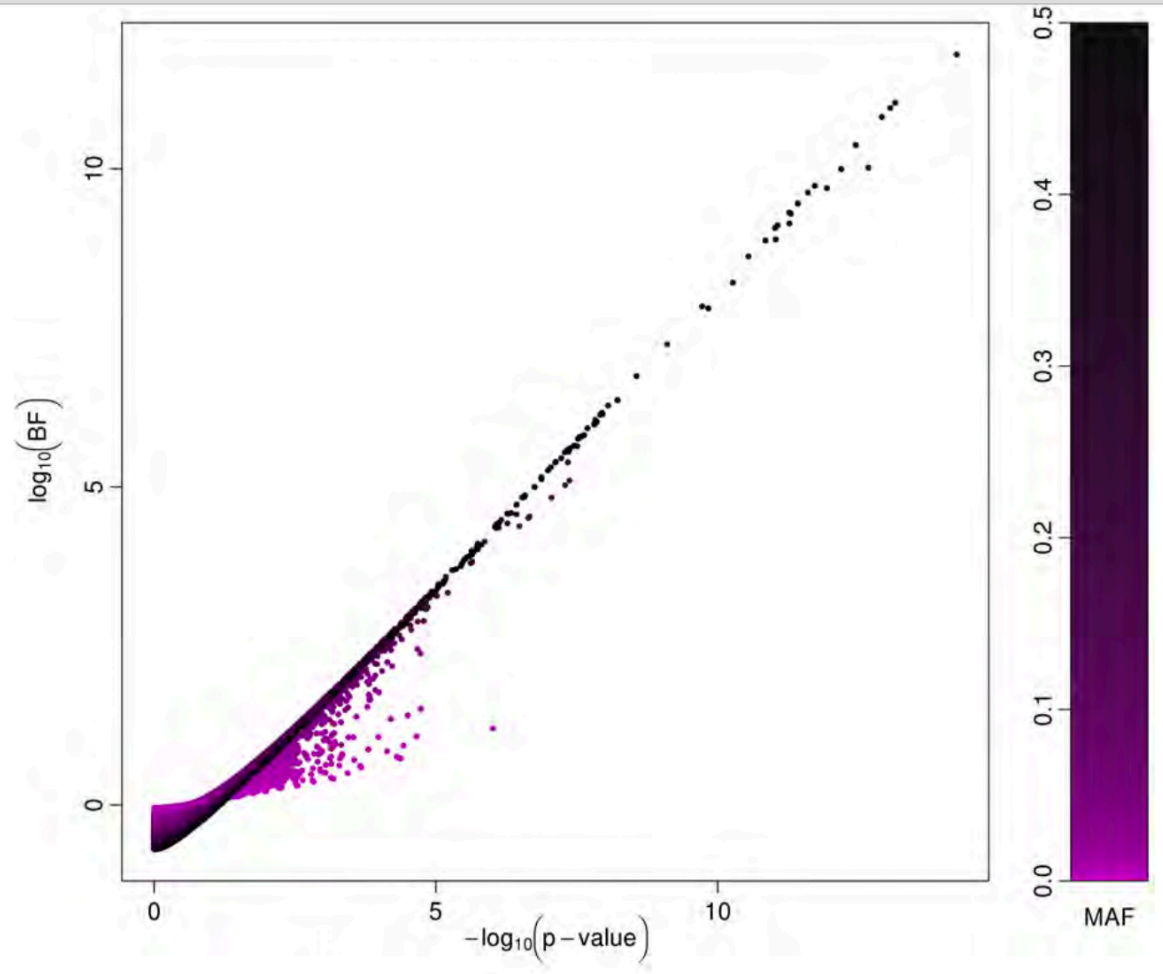


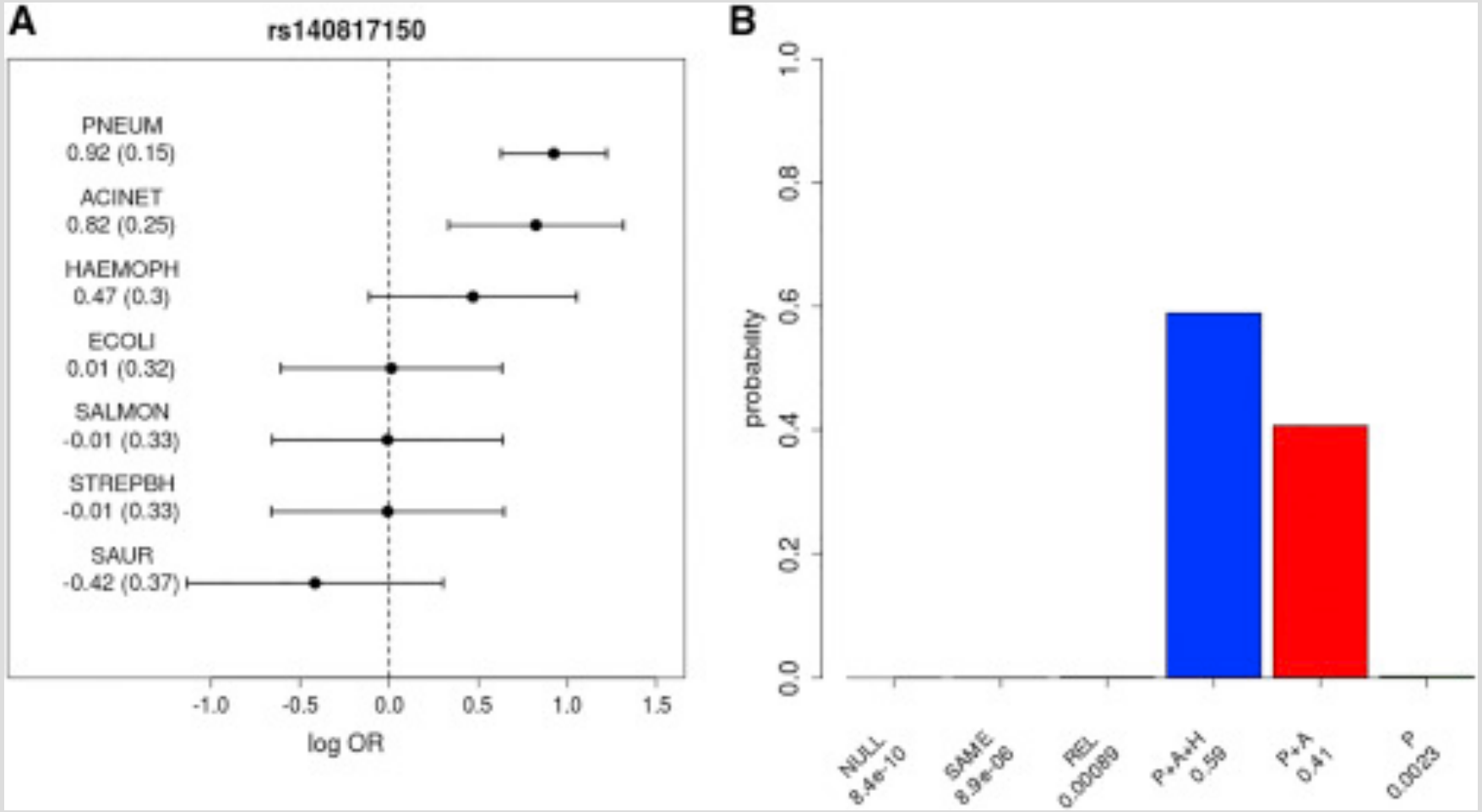

Figure 6.7: **BF versus p-value for Crohn's disease.** Each point represents a SNP from the WTCCC data. BFs are calculated under the conservative prior ($\sigma = 0.2$). Points are coloured according to the MAF, as shown in the legend on the right.

For common variants there is a linear relationship between P-value and BF.

Differences come for rare variants since the standard prior distribution does not allow large effect sizes.

Damjan Vukcevic 2009, Dphil thesis, Oxford

# BAYESIAN MODEL COMPARISON



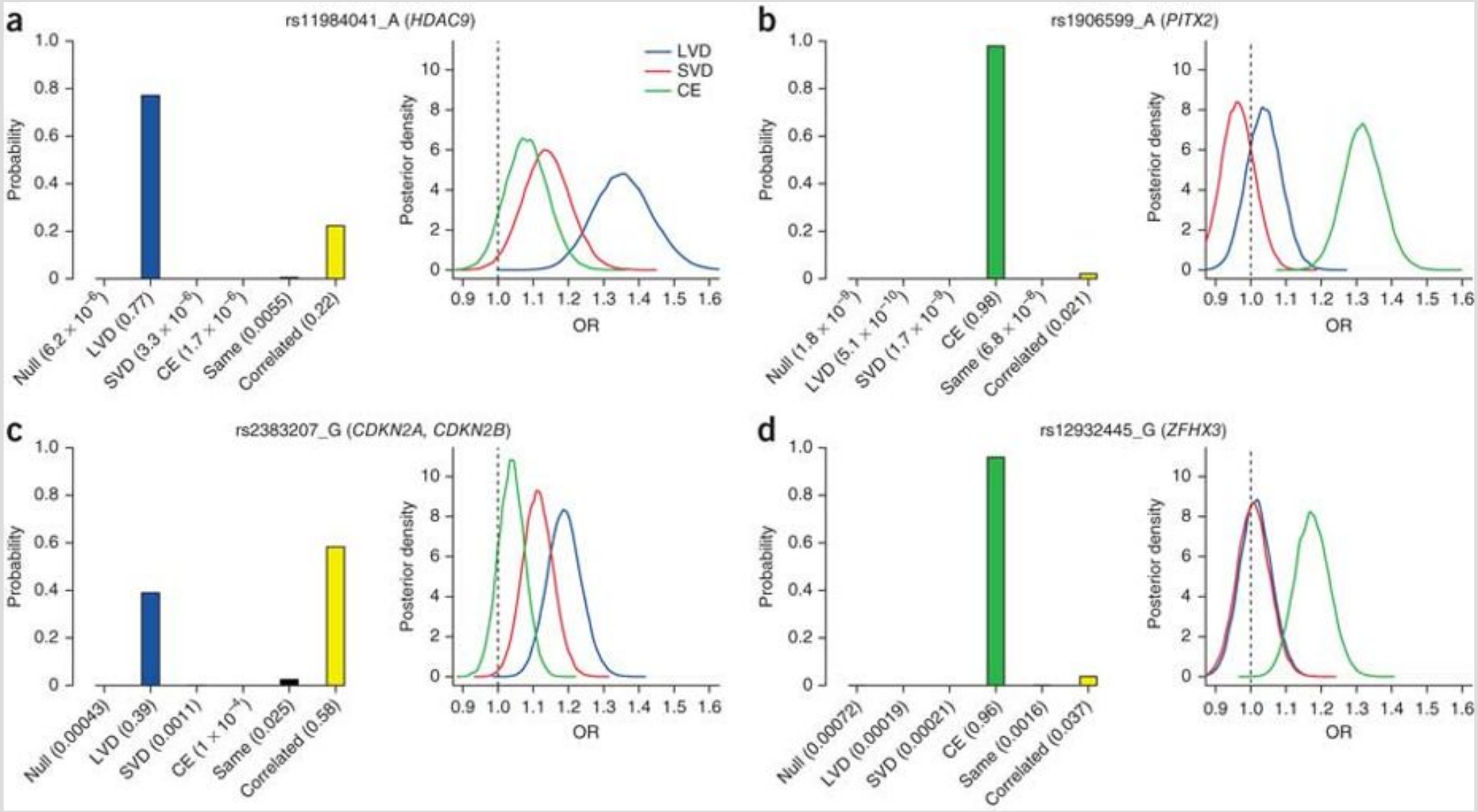

A SNP that associates with Bacetraemia in Kenyan children

The association seems present with several bacetria, but not all.

Bayesian model comparison using ABF framework let us to directly compare different models of association.

Rautanen et al. 2016 AJHG

# BAYESIAN MODEL COMPARISON



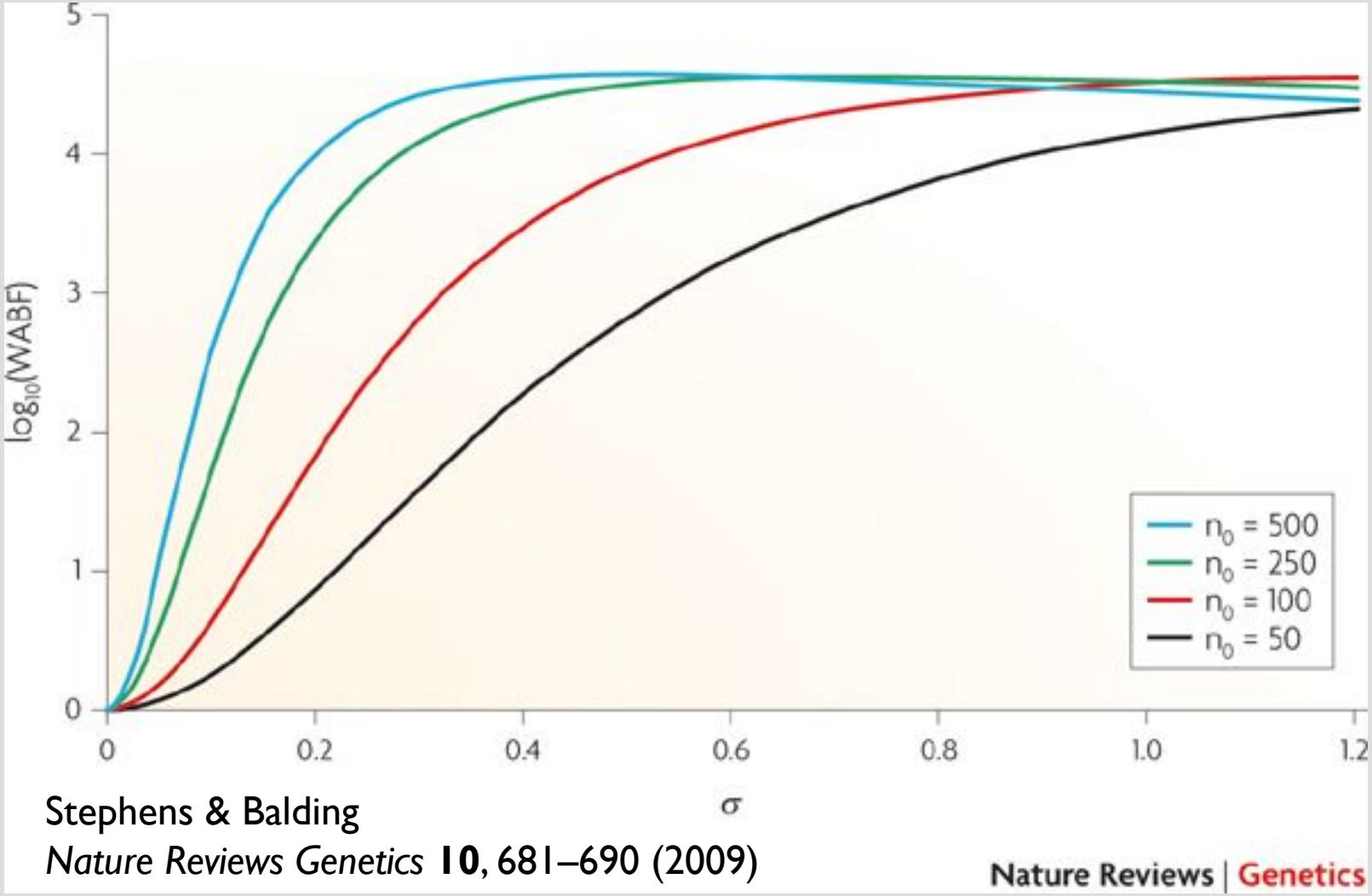

4 SNPs
Associated with
Ischemic stroke

Two SNPs
particularly in
LVD
and 2 in
CE

Subtypes:
LVD large vessel
SVD small vessel
CE cardioembolic

Bellenguez et al. 2012 Nat Gen

# BF FOR A P-VALUE 5E-7




Stephens & Balding
*Nature Reviews Genetics* **10**, 681–690 (2009)

The curves show the Wakefield approximate Bayes factor (ABF) for a SNP with a $p$-value $\approx 5 \times 10^{-7}$ using 4 values of $n_0$, which is the minor allele count among cases and controls combined. There are $n_0$ cases and $n_0$ controls, so the minor allele fraction remains constant at 0.25. As $\sigma$ (the standard deviation of the effect size) increases from 0, the $\log_{10}$(WABF) for each SNP rises from 0 to a maximum value of 4.57 before gradually decreasing as $\sigma$ continues to increase.
If $n_0 \geq 250$, the Bayes factors (BFs) vary by roughly one order of magnitude for $0.2 < \sigma < 1$, but when $n_0 = 50$, the BF varies more markedly, by several orders of magnitude for $\sigma$ in this range. If $\pi = 10^{-4}$, then $\log_{10}(ABF) < 4.57$ implies PPA<0.79. Therefore, under our assumptions, a SNP just reaching the $p$-value threshold of $5 \times 10^{-7}$ still has a substantial chance of being a false discovery.