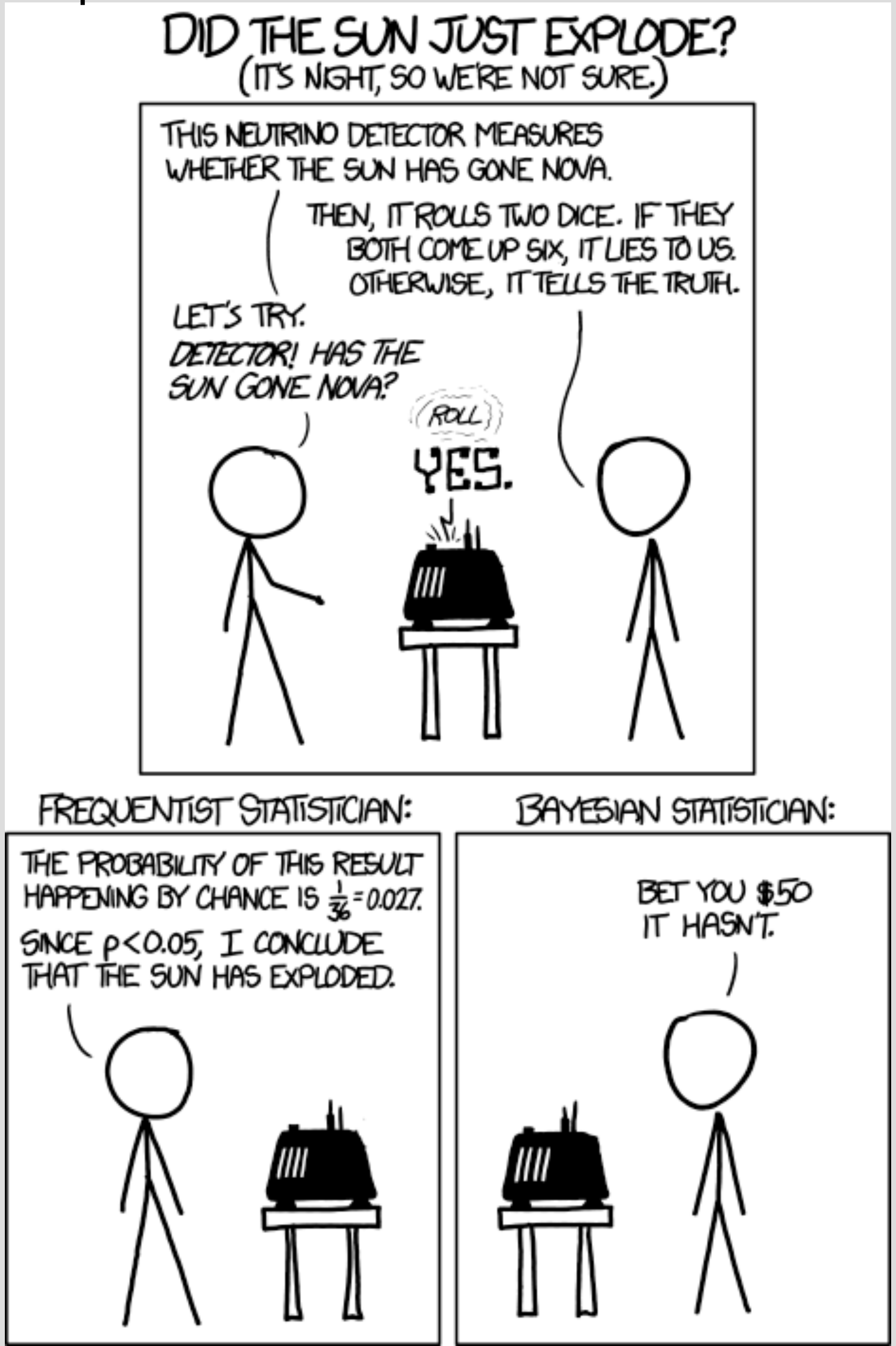
GWAS 4

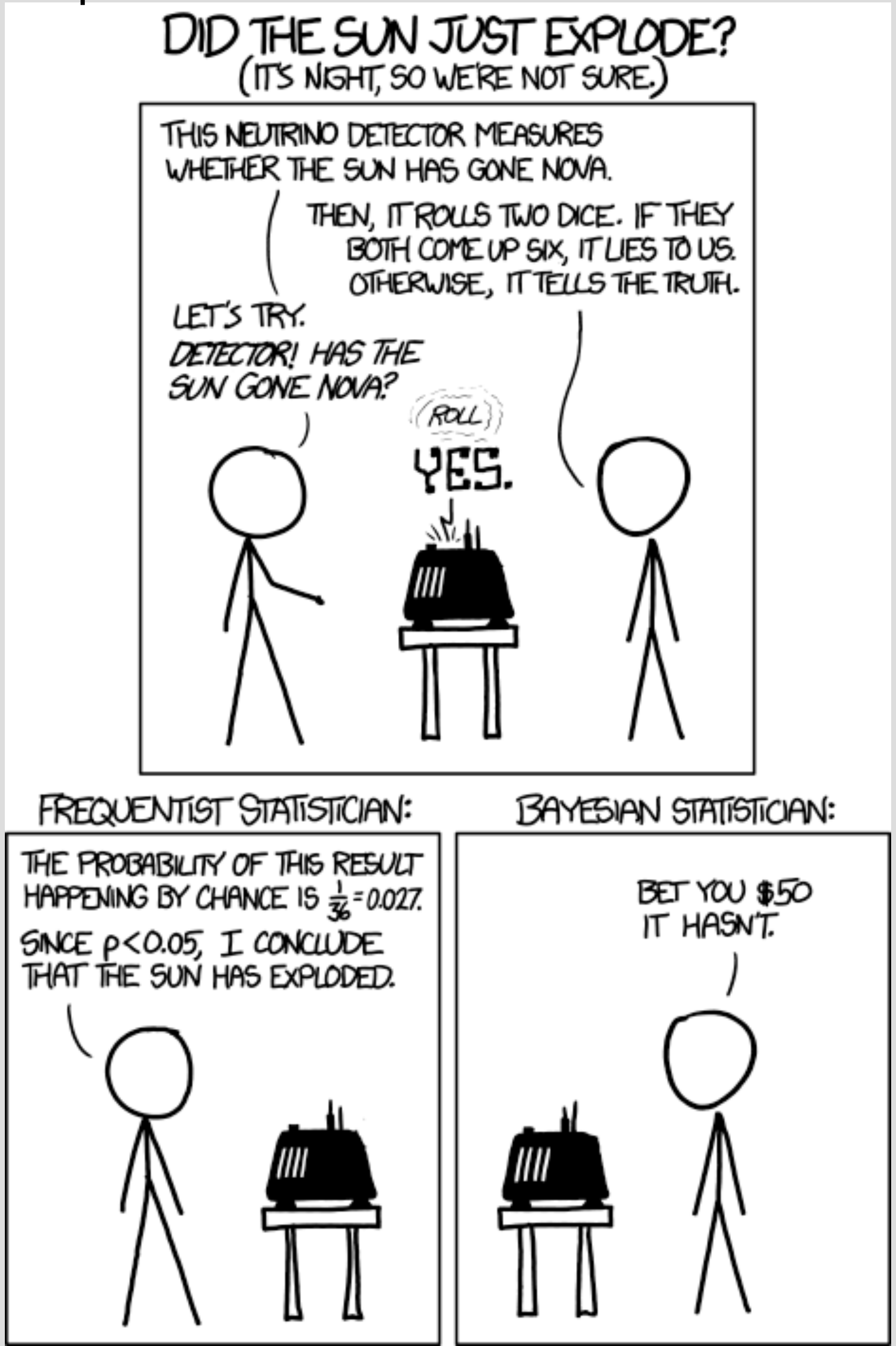
Matti Pirinen
University of Helsinki
Week 2

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"

Bayes rule: $\Pr(X \mid Y) = \dfrac{\Pr(Y \mid X)\Pr(X)}{\Pr(Y)}$

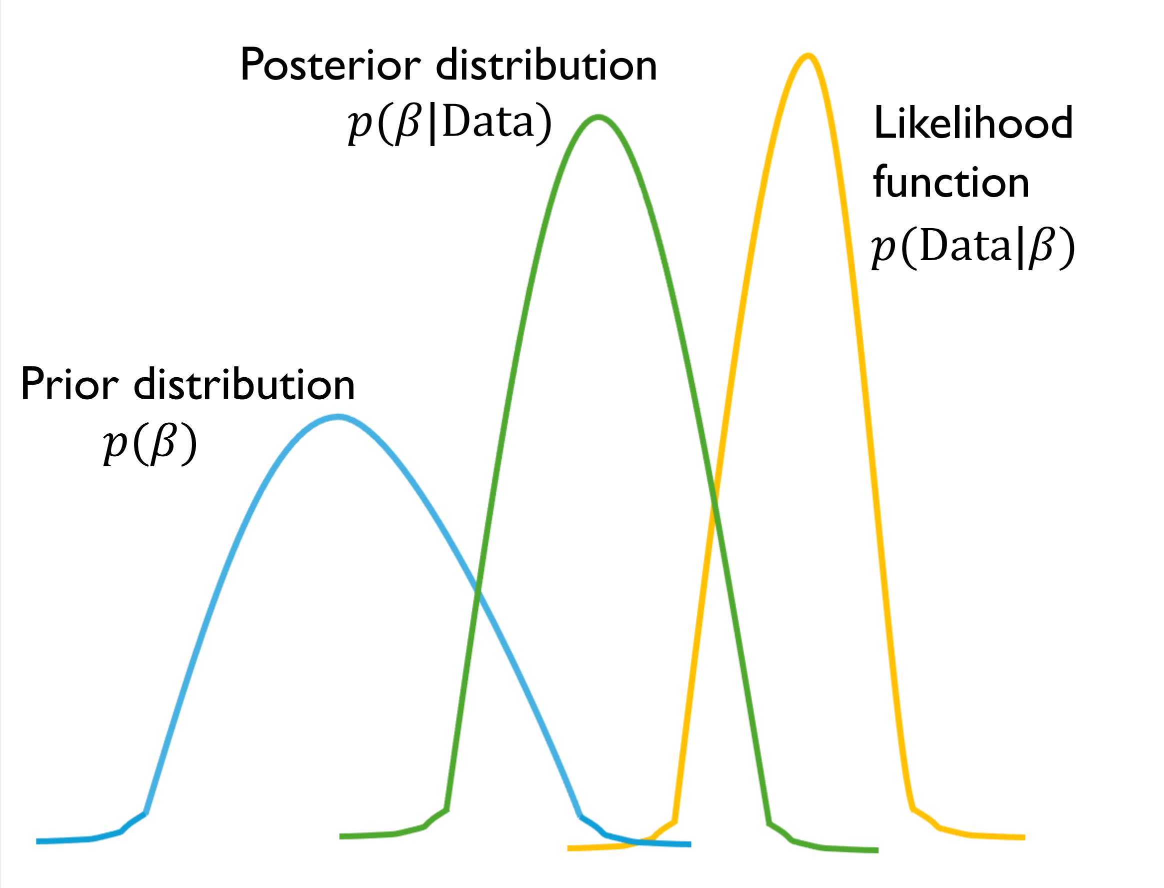
We know $\Pr(Y \mid X) = 0.973$ and $\Pr(Y \mid \mathrm{not}X) = 0.027$.

$$\frac{\Pr(X \mid Y)}{\Pr(\mathrm{not}X \mid Y)} = \frac{\Pr(Y \mid X)}{\Pr(Y \mid \mathrm{not}X)} \times \frac{\Pr(X)}{\Pr(\mathrm{not}X)} < 36 \times \frac{\Pr(X)}{\Pr(\mathrm{not}X)}$$

posterior-odds Bayes factor prior-odds

So the observation Y increases odds of X at most 36-fold compared to prior odds that are likely very very very small. Thus, the posterior odds of event X remains very very small.

BAYESIAN INFERENCE



$$p(\beta \mid \text{Data}) = \frac{p(\text{Data} \mid \beta)\, p(\beta)}{p(\text{Data})} \propto p(\text{Data} \mid \beta)\, p(\beta)$$

- We are estimating a parameter such as an effect size β in GWAS
- We have some **prior** beliefs about the parameter value but we don't know very accurately
- We gather data and use the **likelihood function** to summarize what the data tells about β
- Bayes rule tells how to combine the prior distribution and the likelihood function into a **posterior** distribution
- If prior is nearly constant across a range of values relative to the amount of info in the data, then the posterior will look like the likelihood function
- If prior of some region is extremely small, then we will need an extremely large likelihood value before posterior will support strongly that region

BAYESIAN MODEL COMPARISON

Posterior probability of hypothesis H_i

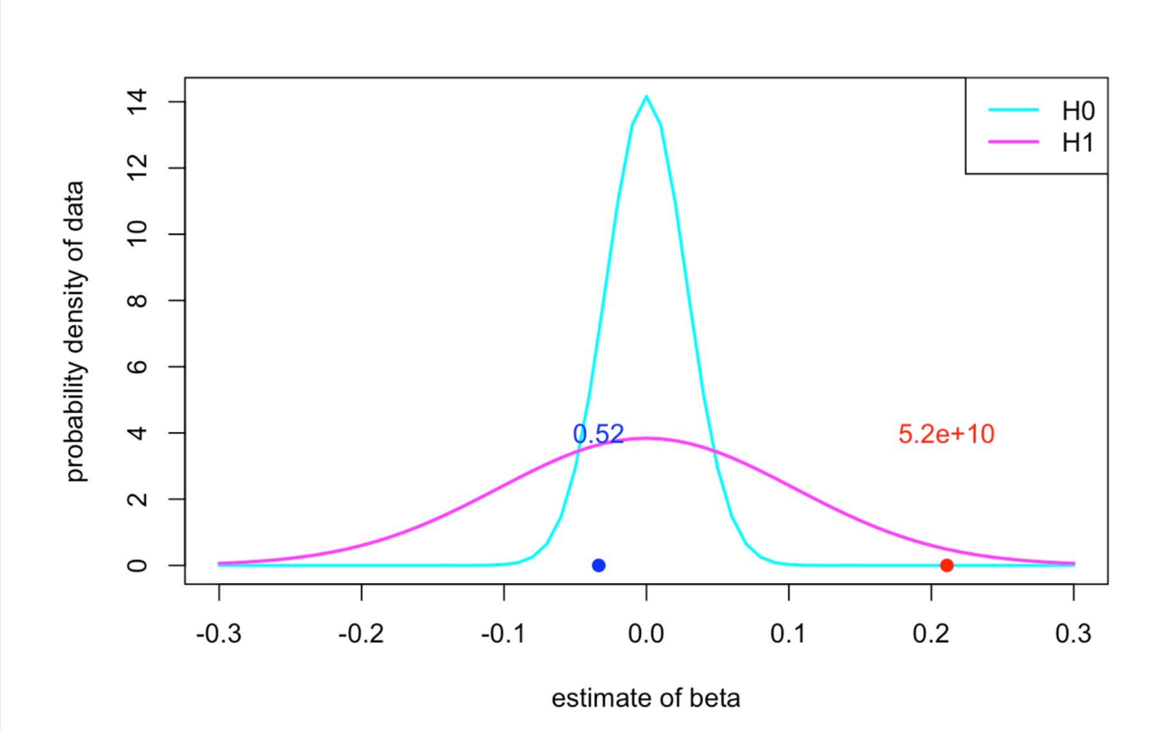
$$P(H_i|\mathcal{D}) = \frac{P(\mathcal{D}|H_i)P(H_i)}{P(\mathcal{D})}, \qquad \text{for } i = 0, 1.$$

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

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?

$p(D \mid H_1)$



BF for blue and red effect size estimates are shown.

- For the NULL hypothesis, true effect size = 0 and hence the observed effect size has distribution N(0, SE^2) – This Normal density evaluated at the observed effect estimate is the evidence term $p(D \mid H_0)$
- For the alternative hypothesis, true effect size is assumed to be sampled 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

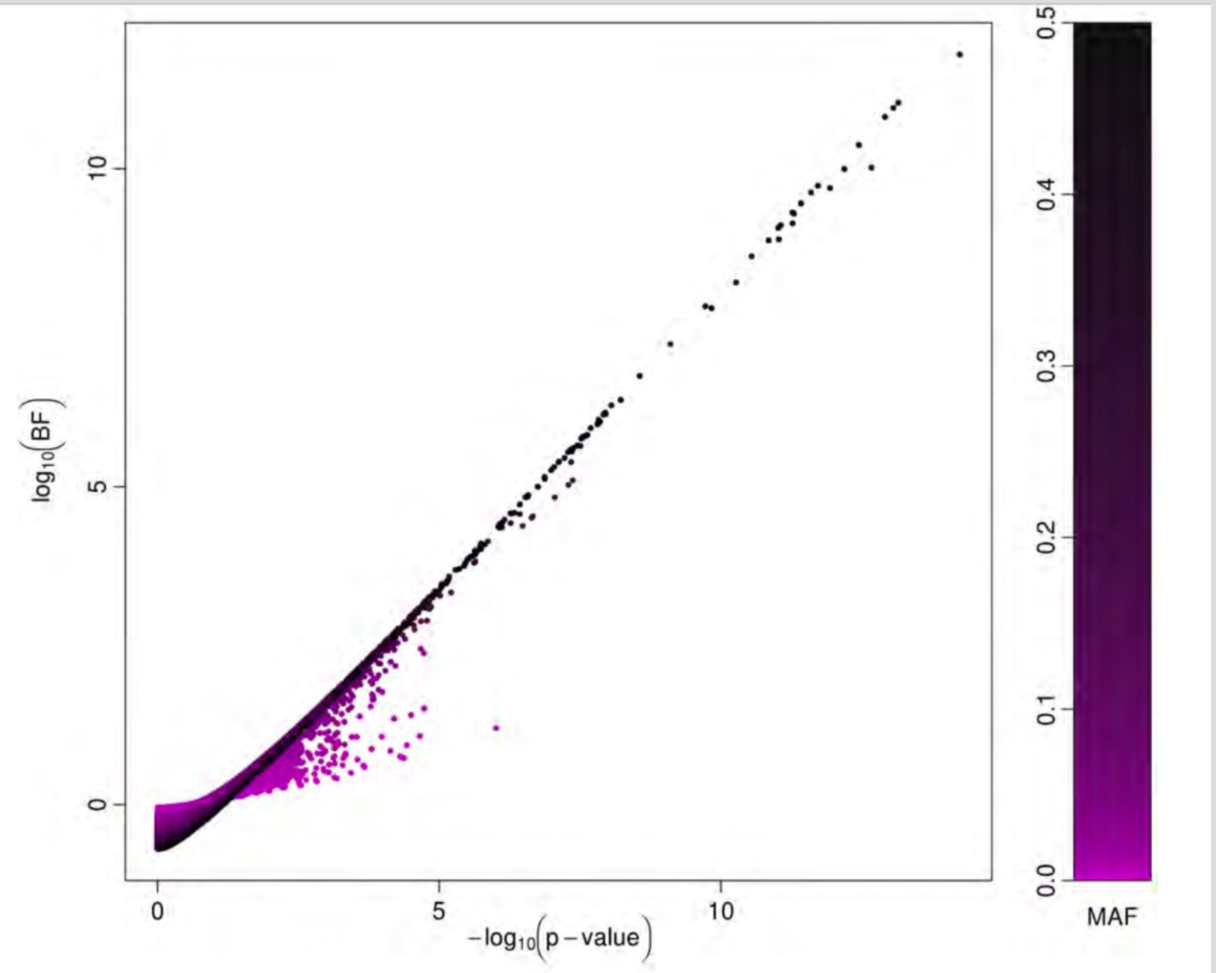


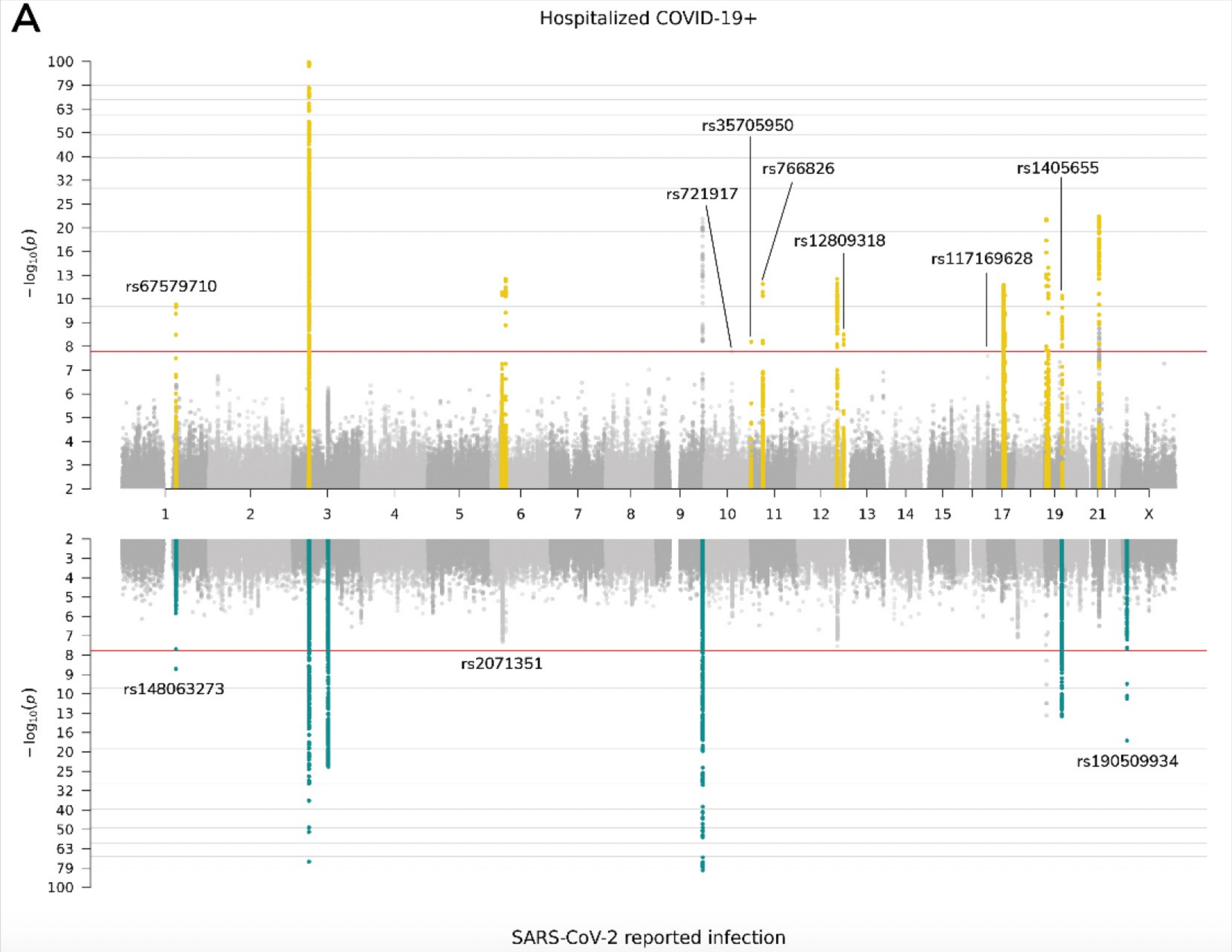
Figure 6.7: **BF versus p-value for Crohn's disease.** Each point represents a SNP from the WTCCC data. BFs are calculated under the conservative prior ($\sigma = 0.2$). Points are coloured according to the MAF, as shown in the legend on the right.

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

WHICH VARIANTS AFFECT COVID INFECTION SUSCEPTIBILITY AND WHICH DISEASE SEVERITY?



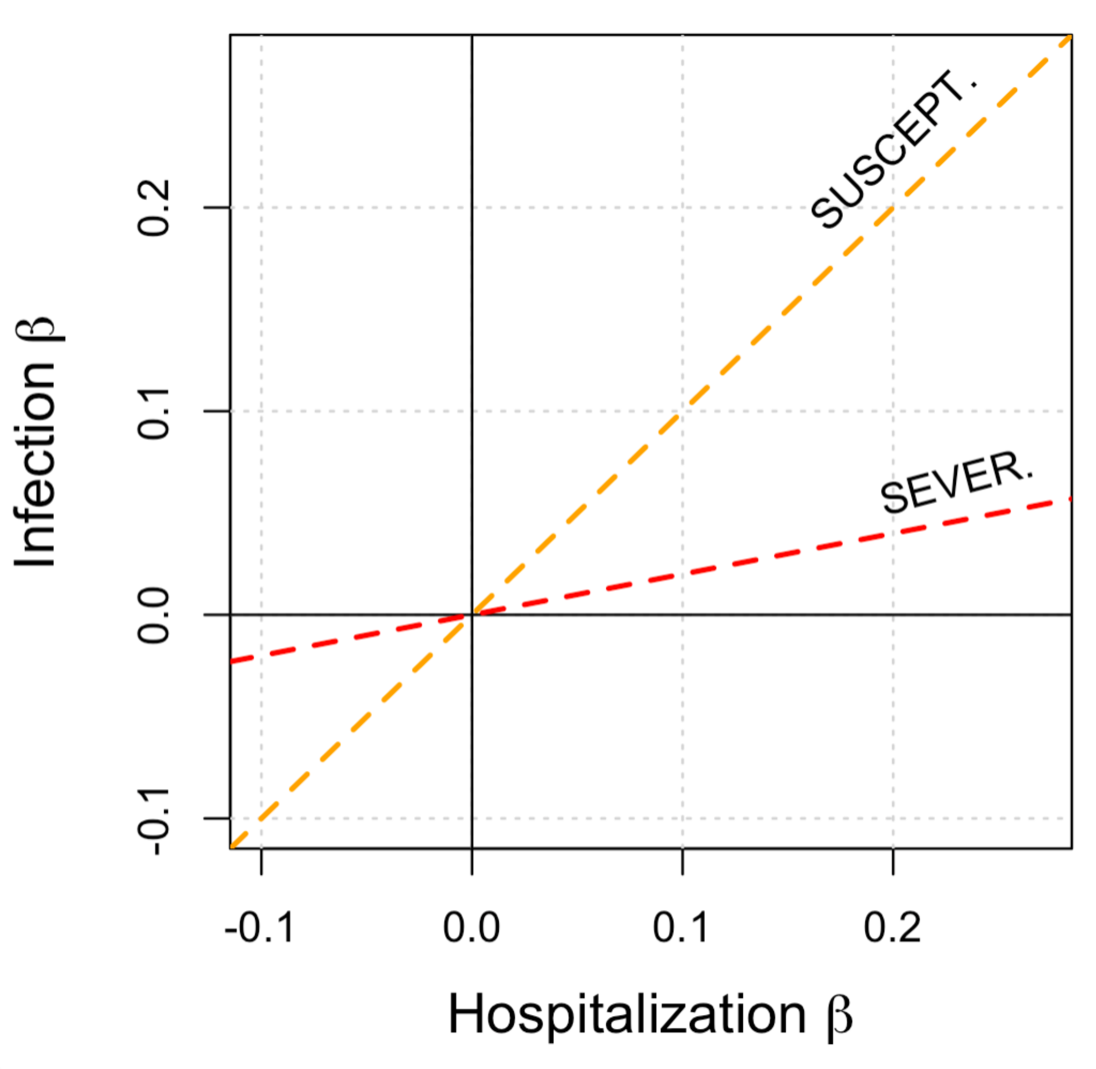
Hospitalization GWAS

Together 23 loci
(*P* < 5e-8)

Infection GWAS

https://doi.org/10.1038/s41586-023-06355-3

EXPECTED TRUE EFFECT SIZES FOR TWO HYPOTHESES



$$\beta_{INF} = \beta_{HOS} > 0$$

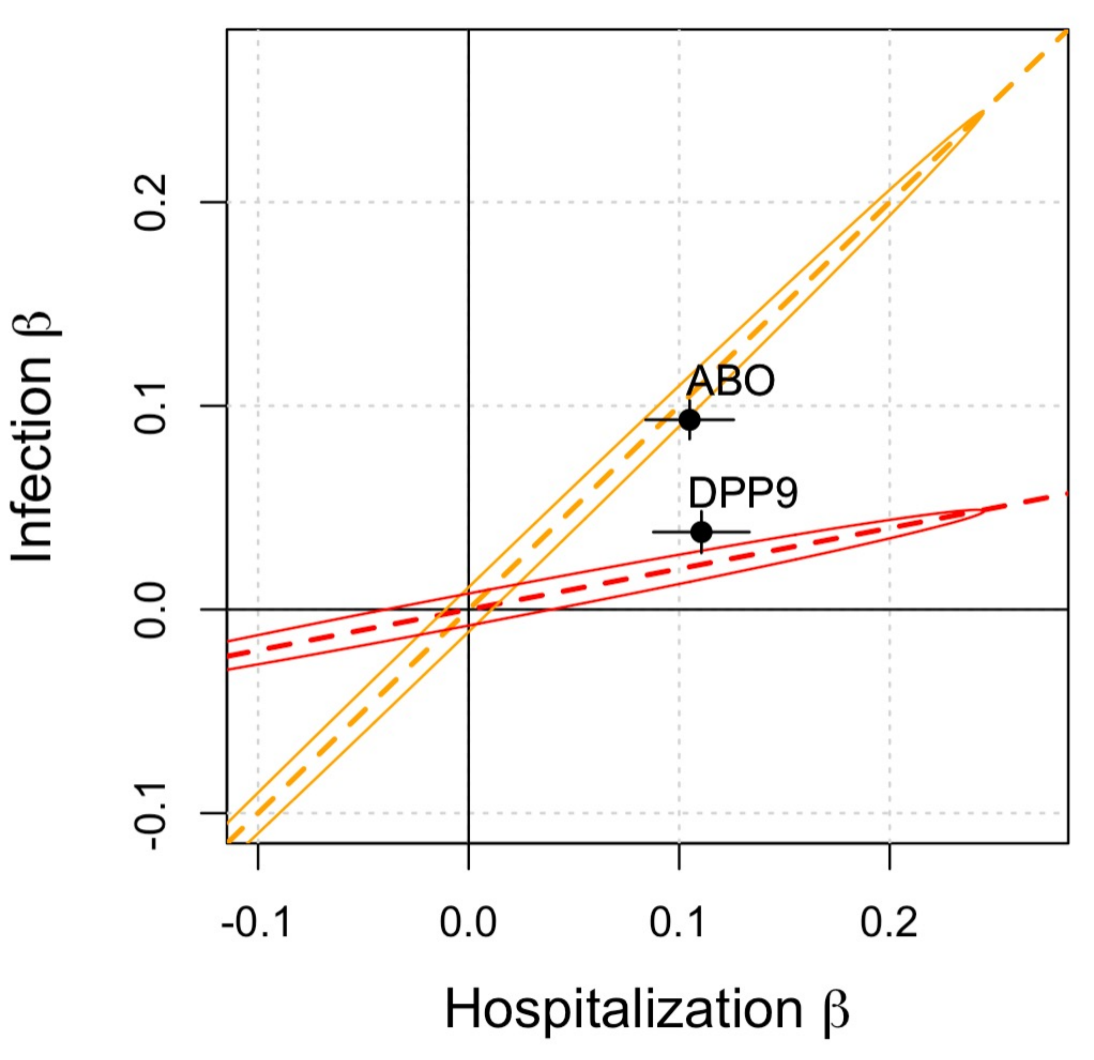
Hypothesis **INF**:
Variant affects infection susceptibility
(but not disease severity)

$$\beta_{INF} \approx 0.2\, \beta_{HOS} > 0$$

Hypothesis **SEV**:
Variant affects disease severity (but not infection susceptibility).

Reason why β_{INF} is not 0 under the SEV hypothesis is that the data set has been enriched for hospitalized cases,
and coefficient 0.2 is a specific value computed for this data set.

POSTERIOR PROBABILITIES OF HYPOTHESES



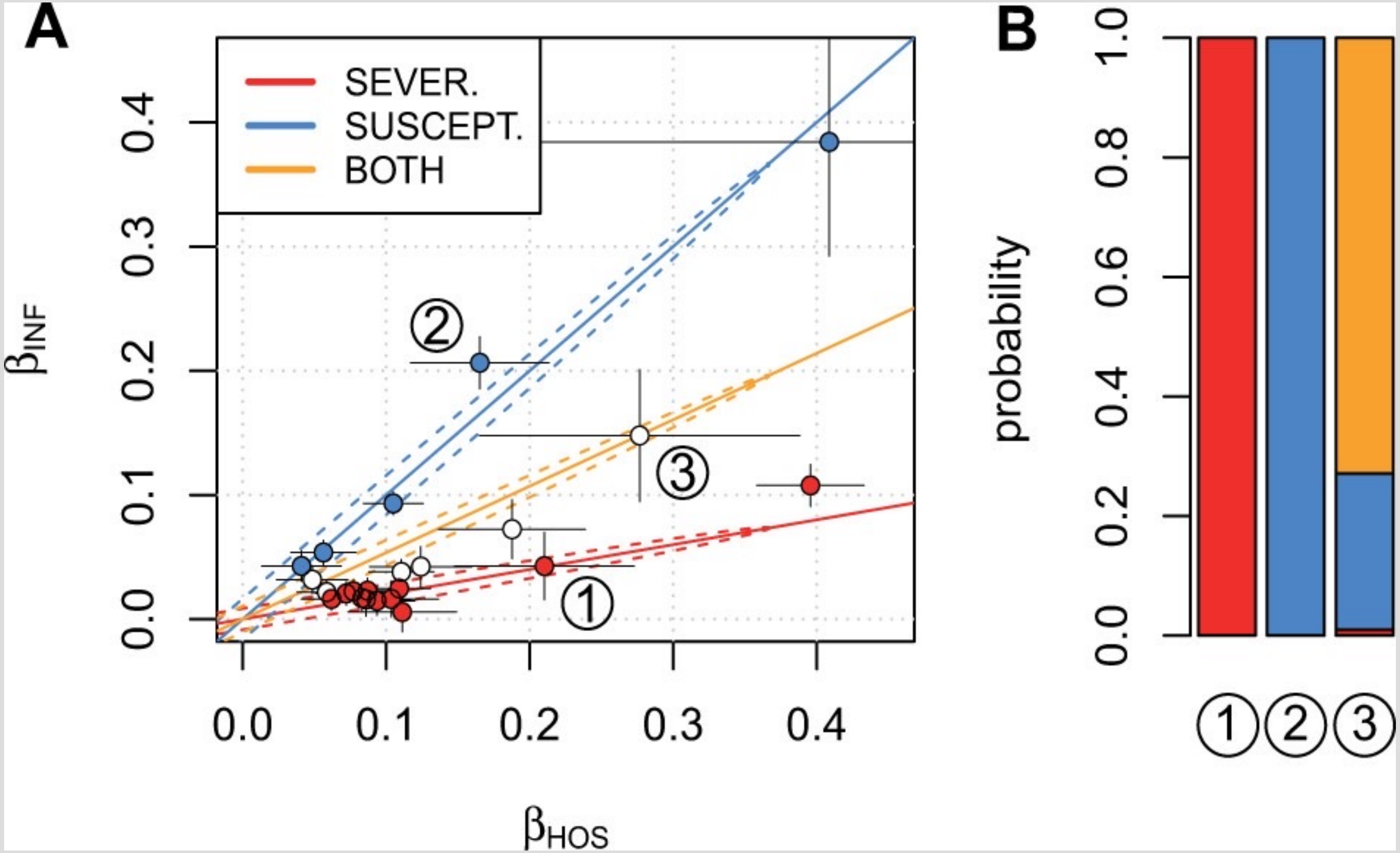
Bayes formula (with equal prior probabilities of models so prior-odds = 1):

$$\frac{\Pr(\text{SEV} \mid \hat{\beta})}{\Pr(\text{INF} \mid \hat{\beta})} = \frac{\Pr(\hat{\beta} \mid \text{SEV})}{\Pr(\hat{\beta} \mid \text{INF})} \times 1$$

posterior-odds Bayes factor x prior-odds

	Pr(SEV | $\hat{\beta}$)	Pr(INF | $\hat{\beta}$)
ABO	1e-38	1
DPP9	0.99999994	6e-8

CATEGORIES OF COVID-19 RELATED VARIANTS

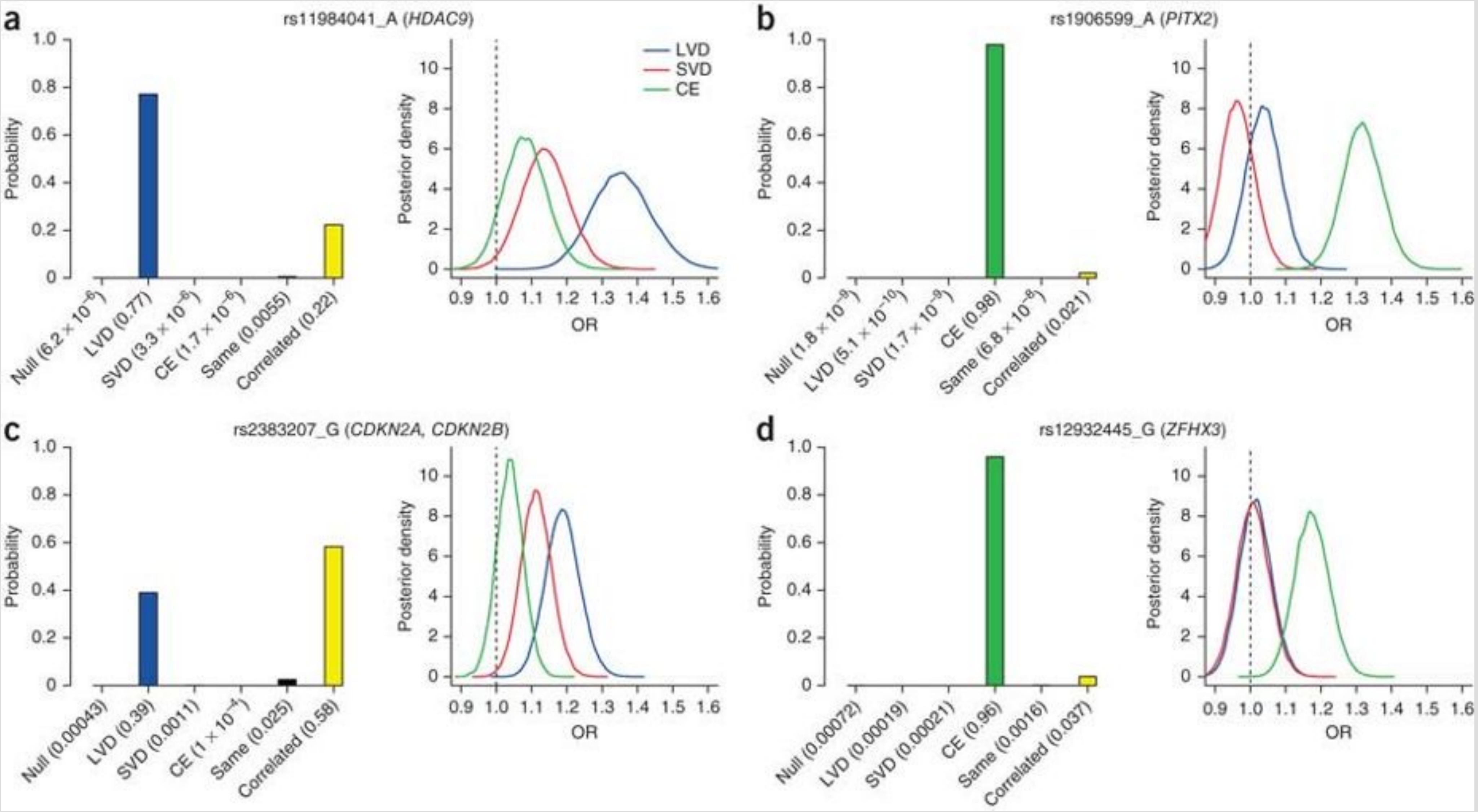


Categorizing variants:
5 infection susceptibility (blue)
12 COVID severity (red)
6 undetermined status (white)

Here we have added an additional model to represent variants that have effect on BOTH the infection and severity.

https://doi.org/10.1093/bioinformatics/btad115

BAYESIAN MODEL COMPARISON



Bellenguez et al. 2012 Nat Gen

4 SNPs
Associated with
ischemic stroke.

3 subtypes:
LVD large vessel
SVD small vessel
CE cardioembolic

Two SNPs
particularly in
LVD
and 2 in
CE