GWAS 4: Bayesian inference

Matti Pirinen, University of Helsinki

Updated: March 13, 2025.

The slide set referred to in this document is "GWAS 4".

A crude statistical inference procedure in the GWAS framework is based on a fixed significance threshold and only determines whether the P-value falls below the genome-wide significance (GWS) threshold (typically 5e-8). However, we typically make more efficient use of the data by reporting effect sizes, their standard errors and exact P-values.

Suppose that there are two variants that both pass the GWS threshold, the first yielding a P-value of 4e-8 and the second of 2e-200. How could we quantify the probability that each of these variants is having a non-zero effect? We have previously learned that the P-value itself is not a probability of the null hypothesis. Let's next look what information and assumptions we would need, in addition to the P-value, in order to be able to talk about a probability of a non-zero effect.

This topic relates to Bayesian methods in GWAS, reviewed by Stephens & Balding.

Below, we will use notation $\Pr(\cdot)$ to denote both probabilities of events and probability densities of continuous random variables. Thus, for example the probability of the null hypothesis is $\Pr(H_0)$, and the probability density of a continous random variable X, evaluated at value x is $\Pr(x)$ (often also denoted by $p_X(x)$).

From significance to the probability of association

In order to talk about a probability of a hypothesis, we need to define the set of all possible hypotheses whose combined probability is 1, i.e., we work under the assumption that one of the hypotheses is true.

Note that for the P-value calculation, we only ever defined the null hypothesis and therefore we are unable to talk about the probability of the null hypothesis because we haven't defined what else might be possible: Even if the data seemed unlikely under the null hypothesis, if the data were even more unlikely under all other possible hypotheses, then the null hypothesis might still be quite probable, and such considerations cannot be done from the P-value alone.

In our case, we consider only two hypotheses:

- H_0 (null hypothesis) : $\beta = 0$ and
- H_1 (alternative hypothesis) : $\beta \neq 0$.

Next, we need to quantify the **prior probabilities** of these hypotheses. These prior probabilities answer to the question: What would I consider as probability of each hypothesis before I have seen the data. The phrase "What would *I*" is there on purpose: prior probabilities are subjective. They are based on whatever knowledge *I* have available. Therefore different persons may have different prior probabilities for the same hypothesis and my prior can (and will!) change as I learn more about the question. For example, $\Pr(H_1) = 10^{-5}$ could be a reasonable prior for a non-zero effect based on what I know about GWAS. (Last

week we saw how a magnitude more stringent assumption $\Pr(H_1) = 10^{-6}$ led us to the common GWS threshold of 5e-8.)

Then, we observe the data $\mathcal{D}$ and our interest is in the probabilities of each hypothesis after we have seen the data. This is the core question of **Bayesian inference**: How does observing the data update our knowledge from our current state of knowledge, described by our prior probabilities $\Pr(H_i)$, into our **posterior probabilities** $\Pr(H_i|\mathcal{D})$? In short: How do we learn from data? Not surprisingly, the answer is the Bayes rule.

4.1 Bayes rule   To write down the Bayes rule (also called Bayes theorem, Bayes formula), we just remember that we are considering joint distributions of two variables. Here the two variables are the hypothesis H_i and the observed data $\mathcal{D}$. We expand their joint probability $\Pr(H_i, \mathcal{D})$ using the conditional probability rule in both ways possible

$$\Pr(H_i \,|\, \mathcal{D})\Pr(\mathcal{D}) = \Pr(H_i, \mathcal{D}) = \Pr(\mathcal{D} \,|\, H_i)\Pr(H_i),$$

from which we can solve the conditional probability that the hypothesis holds given that we have observed the data $\mathcal{D}$:

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

This is the Bayes rule. The conditional probability on the left hand side is also called the **posterior probability** of the hypothesis given the data.

Since the term $\Pr(\mathcal{D})$ (the marginal probability of the observed data) does not depend on hypothesis H_i, we can get rid of it by taking the ratio of the posteriors of the two hypotheses:

$$\frac{\Pr(H_1 \,|\, \mathcal{D})}{\Pr(H_0 \,|\, \mathcal{D})} = \frac{\Pr(\mathcal{D} \,|\, H_1)\Pr(H_1)}{\Pr(\mathcal{D} \,|\, H_0)\Pr(H_0)}.$$

Hence, in order to compute the posterior probability ratio for the hypotheses, we will still need the terms $\Pr(\mathcal{D} \,|\, H_i)$ in addition to the prior probabilities. $\Pr(\mathcal{D} \,|\, H_i)$ describes what kind of data sets we are likely to see under each hypothesis and with which probability. After these probability densities are specified, the inference is about letting the possible hypotheses compete both in how well they explain the observed data (terms $\Pr(\mathcal{D} \,|\, H_1)$ and $\Pr(\mathcal{D} \,|\, H_0)$) and in how probable they are *a priori* (prior probablity terms $\Pr(H_1)$ and $\Pr(H_0)$).

Example 4.1. The Bayesian inference shows that both the observed data AND the prior knowledge is crucial for a complete inference procedure. Suppose, for example, that a sequencing effort of my genome returns data $\mathcal{D}$ that seem to miss chromosome 6 completely. We have two hypotheses: H_0: "There is a technical error", or H_1: "I don't carry any copies of chromosome 6 (in the cells involved)". The observed result could have resulted from either of these options and hence under both hypotheses $\Pr(\mathcal{D} \,|\, H_i)$ is relatively high, indicating that the data are consistent with both hypotheses. However, the prior odds of $\Pr(H_1)/\Pr(H_0)$ are likely small if we think that it is much more likely that there is a technical error somewhere than that I would be missing chr 6 and still be alive and even pretty healthy. Hence, the posterior conclusion that combines the prior probabilities and the observed data is that it is more probable that we have an error somewhere in the process than that I don't carry chr 6, even though the observation alone couldn't tell the two hypotheses apart.

Example 4.2. Interpretation of a medical test result is the standard example of the use of Bayes rule. Let's apply it to a case where we try to determine whether an individual has a disease given his genotype.

Suppose that each copy of *HLA-DRB1*1501* allele on chromosome 6 increases the risk of *multiple sclerosis* by OR=3. Prevalence of MS-diseases is $K = 0.001$ and population frequency of DRB1*1501 is 0.20. What is probability of ever getting the disease for each genotype (i.e. 0,1 or 2 copies of DRB*1501)?

Answer. Denote by D the event of getting the disease and by X the genotype. Here D has the role of a hypothesis and X the role of the observed data in the formulation of Bayes rule given above. Bayes rule says

that for each genotype $x \in \{0, 1, 2\}$:

$$\Pr(D \mid X = x) = \frac{\Pr(D)\Pr(X = x \mid D)}{\Pr(X = x)}.$$

We know that $\Pr(D) = K = 0.001$ and we can assume that the control frequencies are approximately the population frequencies since the diseases has such a low prevalence. Let's use the function `case.control.freqs()` from GWAS 1 to determine the case frequencies assuming HWE in control population:

```
K = 0.001
or = 3
f = 0.2
cc.f = case.control.freqs(f, or)
cc.f
```

```
##        cases controls
## 0 0.3265306     0.64
## 1 0.4897959     0.32
## 2 0.1836735     0.04
```

The risk to get the disease given the genotype is, according to Bayes rule

```
rbind(genotype = c(0, 1, 2), risk = K * cc.f$cases / cc.f$controls)
```

```
##                    [,1]         [,2]         [,3]
## genotype 0.0000000000 1.000000000 2.000000000
## risk     0.0005102041 0.001530612 0.004591837
```

So even though the disease status has a large effect on the genotype distributions, still even the high risk group has risk $< 0.5\%$, and the genotype doesn't give a practically useful predictive accuracy at the level of an individual from the population because the disease is so rare.

4.2 Probability model for observed GWAS data To use Bayes rule in the GWAS setting, we need to define a probability density function for the observed data under both the null hypothesis and the alternative hypothesis.

In a linear regression GWAS model $y = \mu + x\beta + \varepsilon$, the observed data $\mathcal{D}_k = (\boldsymbol{y}, \boldsymbol{x}_k)$ consist of the phenotype vector $\boldsymbol{y}$ and the vector of genotypes $\boldsymbol{x}_k$ at the tested variant k. When we assume Gaussian errors (i.e. each ε_i having a Normal distribution with same variance σ^2), the probability density of data for a fixed value of effect size β and error variance σ^2 is

$$p\left(\mathcal{D}_k \mid \beta, \sigma^2\right) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{x}_k\beta, \sigma^2 I) \propto \exp\left(-(\boldsymbol{y} - \boldsymbol{x}_k\beta)^T(\boldsymbol{y} - \boldsymbol{x}_k\beta)/(2\sigma^2)\right).$$

We have ignored μ here by assuming that y and x_k are mean-centered values; this just simplifies the notation but doesn't affect the inference on parameter β.

Under the null model, we set $\beta = 0$ and in the alternative model, we can set β to some other value b_1. If we do not want to specify our alternative model of non-zero effect by a fixed value b_1, we can use a **prior distribution** for β, for example, by saying that under the alternative model we expect that $\beta \sim \mathcal{N}(b_1, \tau_1^2)$. This means that if the alternative model holds, then the true effect sizes are distributed around the value b_1 with a standard deviation of τ_1. With this prior distribution, the probability of data under H_1 is given by weighting the above likelihood function by the prior probability of each possible value of β:

$$p(\mathcal{D}_k \mid H_1) = \int_{\beta} p\left(\mathcal{D}_k \mid \beta, \sigma^2\right) p(\beta \mid H_1) d\beta = \int_{\beta} \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{x}_k\beta, \sigma^2\right) \mathcal{N}\left(\beta; b_1, \tau_1^2\right) d\beta.$$

(For simplicity, in both models, we may fix σ^2 to its empirical maximum likelihood estimate.)

If we assume that the mean parameter $b_1 = 0$ in the Gaussian prior of β, then the integral can be computed analytically to give

$$p(\mathcal{D}_k \,|\, H_1) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \text{SE}^2\right),$$

where c is a constant and $\widehat{\beta}$ is the MLE of β and SE the corresponding standard error. Note that by replacing τ_1 with 0, we have

$$p(\mathcal{D}_k \,|\, H_0) = c \cdot \mathcal{N}\left(\widehat{\beta};\, 0,\, \text{SE}^2\right).$$

These results show that the quantity $p(\mathcal{D}_k \,|\, H_i)$, that tells how well model H_i explains the data, can be evaluated (up to a constant) by asking how well the model can explain the observed estimate $\widehat{\beta}$. The null model assumes that the estimate is a result of sampling variation alone (only the SE contributes to the deviation from 0) whereas the alternative model allows some additional variance τ_1^2 around 0 due to the truly non-zero effect. Let's demonstrate this.

Example 4.3.

```r
n = 3000 # sample size for SE calculation
f = 0.3 # MAF for SE calculation
sigma = 1 # error SD
se = sigma / sqrt(2*f*(1-f)*n) # SE for QT GWAS
tau = 0.1 # prior standard deviation for effect size beta under H1

# Let's draw probability densities of "data" under the two models, H0 and H1
# as a function of MLE estimate
x = seq(-0.3, 0.3, by = 0.01)
y1 = dnorm(x, 0, sqrt(tau^2 + se^2) )
y0 = dnorm(x, 0, se)
plot(x, y0, t = "l", col = "cyan", lwd = 2, xlab = expression(hat(beta)),
     ylab = "probability density of data")
lines(x, y1, col = "magenta", lwd = 2)
legend("topright", c("H0", "H1"), col = c("cyan","magenta"), lwd = 2)

# We make a shortcut and don't simulate data at all, but we directly simulate estimates
# Suppose we have two cases, first follows null (beta = 0), second follows alternative (beta = 0.2)
b =c(0, 0.2)
b.est = rnorm(2, b, se) #these are simulated estimates: true means and Gaussian noise determined by SE
points(b.est, c(0, 0), pch = 19, col = c("blue", "red") )
#Next: log of Bayes factor of H1 vs H0, will be explained below
#      use log-scale to avoid inaccuracies.
log.bf.10 = dnorm(b.est, 0, sqrt(tau^2 + se^2), log = T ) - dnorm(b.est, 0, se, log = T)
bf.10 = exp(log.bf.10) #then turn from log-scale to Bayes factor scale
text(b.est[1], 4, signif(bf.10[1], 2), col = "blue")
text(b.est[2], 4, signif(bf.10[2], 2), col = "red")
```
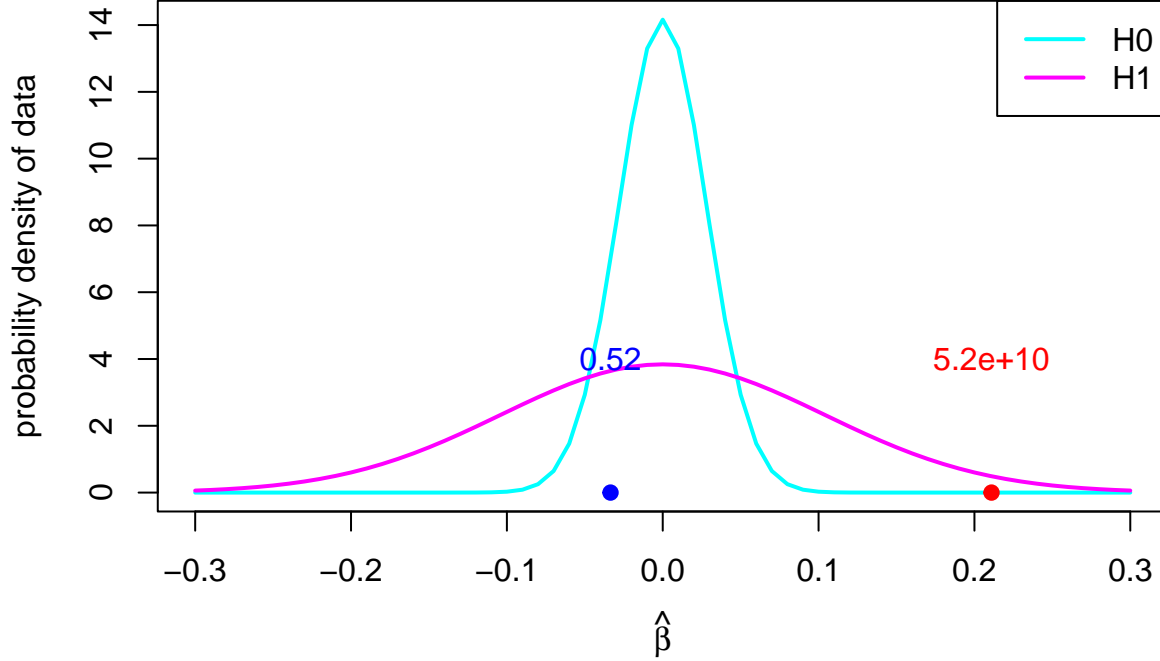
The distribution under H_0 describes what kinds of effect estimates we can get with this sample size and MAF when the true effect is exactly 0. Any deviation from 0 is then by due to statistical sampling variation, as quantified by the SE.

The distribution under H_1 describes what kinds of effect estimates we expect when we have BOTH a non-zero effect size, whose expected range is described by the standard deviation τ_1, and we have ALSO the statistical sampling variation as quantified by the SE.

If the observed estimate $\widehat{\beta}$ is close to 0, then H_0 explains the data better than H_1, whereas the opposite is true when $\widehat{\beta}$ is farther away from 0. With these parameters, H_1 starts to dominate roughly when $|\widehat{\beta}| \geq 0.05$. Values close to zero are relatively more likely under the null than under the alternative. This is because the the null model can only explain well data sets where the effect estimate is near zero whereas the alternative model can also explain observations that are farther away from the zero, and hence the probability mass of the alternative is less concentrated near zero than that of the null. Consequently, the null model yields higher probability densities near 0 than the alternative model that has spread out its probability density more widely away from zero.

4.3 Bayes factor Two points shown in the plot above are examples of possible estimates that could result either under H_0 (blue) or H_1 (red). The values are ratios $\Pr(\mathcal{D}|H_1)/\Pr(\mathcal{D}|H_0)$ of probability densities computed at these two points. When this ratio < 1, the null model H_0 explains the data better than the alternative model and when it is > 1, the opposite is true. This ratio is called **Bayes factor (BF)**, that is the factor that multiplies the prior odds to get to the posterior odds:

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

To interpret the Bayes factor we can think that in order that the posterior probability of the alternative model would be at least 10-fold compared to that of the null, the BF needs to be at least 10 times the inverse of the prior odds. If the prior odds are 1e-5 (1 in 100,000), then we would need a BF of at least 1e+6 (1,000,000) to result in posterior odds > 10.

We are almost there having calculated a proper probability for the null hypothesis. We still need to agree on the prior probability of the null model. Last week we saw that the current genome-wide significance level

can be thought as corresponding to a prior probability of about $\Pr(H_1) = 10^{-6}$. We know that this seems very stringent compared to the true genetic architecture behind complex traits and diseases, but it gives us a conservative reference point. With this prior probability, the posterior odds and posterior probabilities of the alternative model are:

```
post.odds = bf.10 * 1e-6/(1-1e-6) # Pr(H_1|D) / Pr(H_0|D)
post.prob = post.odds / (1 + post.odds) # Pr(H_1|D)
paste(post.prob)
```

```
## [1] "5.22229033295724e-07" "0.999980639798069"
```

For an illustration, let's check the P-values corresponding to these two data sets using the Wald statistic $\widehat{\beta}/\text{SE}$:

```
pchisq( (b.est/se)^2, df = 1, lower = FALSE)
```

```
## [1] 2.342612e-01 6.999659e-14
```

Thus, for the first, the P-value is quite close to 0.2, whereas for the second, it is way below 5e-8. And the Bayesian analysis said that the first data set is almost certain to follow the null whereas the second one is almost certain to have a non-zero effect. Thus, there is no difference between the conclusion of the standard P-value-based GWAS inference and the Bayesian inference, and this is typically the case when there is enough data available. Conceptually, however, there is a large difference between the P-value, which is a probability of observing at least as extreme data under the null as what we have observed, and the posterior probability of the hypothesis itself.

There were several assumptions made in the Bayesian analysis about the effect sizes under H_1 and also about the prior probabilities of the models, and the posterior probabilities of the hypotheses will change when these assumptions change. Instead, the P-value is independent of such assumptions and therefore the P-value remains as a useful simple summary of data that can be computed easily using effect estimate and its SE. The important thing is to know what P-values are and what they are not, and what kind of additional pieces of information would be needed in order to properly quantify the probabilities of the hypotheses.

4.4 Approximate Bayes factor in GWAS   The calculation of the Bayes factor above, that was based on the estimate $\widehat{\beta}$ and its SE, was proposed by Jon Wakefield in 2009. The formula is

$$\text{ABF} \approx \frac{\Pr(\mathcal{D} \,|\, H_1)}{\Pr(\mathcal{D} \,|\, H_0)} \approx \frac{\mathcal{N}\left(\widehat{\beta};\, 0,\, \tau_1^2 + \text{SE}^2\right)}{\mathcal{N}\left(\widehat{\beta};\, 0,\, \text{SE}^2\right)} = \frac{(2\pi)^{-0.5}(\tau_1^2 + \text{SE}^2)^{-0.5} \exp\left(-\frac{1}{2}\frac{\widehat{\beta}^2}{\tau_1^2 + \text{SE}^2}\right)}{(2\pi)^{-0.5}(\text{SE}^2)^{-0.5} \exp\left(-\frac{1}{2}\frac{\widehat{\beta}^2}{\text{SE}^2}\right)}$$

$$= \sqrt{\frac{\text{SE}^2}{\tau_1^2 + \text{SE}^2}} \exp\left(\frac{1}{2}\frac{\widehat{\beta}^2}{\text{SE}^2}\frac{\tau_1^2}{\tau_1^2 + \text{SE}^2}\right),$$

where the alternative model is specified by an effect size prior $H_1 : \beta \sim \mathcal{N}(0, \tau_1^2)$. We have presented the ABF in the form where the alternative is in the numerator and the null in the denominator. Hence, a large ABF means strong evidence in favor of the alternative model. (Wakefield's paper used the inverse of this quantity as Bayes factor, i.e., it computes a Bayes factor comparing the null to the alternative whereas we compare the alternative to the null.)

In R, this BF is easy to compute using the `dnorm()` function, as we did above. It is always good to do the ratio of densities on the log-scale to avoid possible numerical underflows/overflows. In `dnorm()`, this happens by adding `log = TRUE` to the command. Then the ratio of densities becomes a difference between the log-densities:

```
log.bf = dnorm(b.est, 0, sqrt(tau^2 + se^2), log = TRUE) - dnorm(b.est, 0, se, log = TRUE)
bf = exp(log.bf) #turn from log-scale to Bayes factor scale
```

The same formula can be used when $\widehat{\beta}$ and its SE originate from a disease study analyzed by logistic regression. In that case, the formula is an approximation based on the assumption that the logistic regression likelihood has a shape of a Gaussian density function. Therefore, this approach is generally called an **Approximate Bayes Factor (ABF)**.

ABF can be computed from the observed GWAS estimates $(\widehat{\beta}, \text{SE})$ once we have chosen the variance parameter τ_1^2 of the effect size distribution under the alternative. How should we do that?

Example 4.4. Let's assume that the non-zero effects have a distribution $\mathcal{N}(0, \tau_1^2)$ and we want to determine τ_1 in such a way that with 95% probability the effect (of a SNP with MAF = 0.25) explains less than a proportion p of the total phenotypic variance $v = \text{Var}(y)$. We will first compute the effect size β_p that explains exactly the phenotypic variance of $p \cdot v$, and then we will find the SD parameter τ_1 for which 95% of the probability mass of $\mathcal{N}(0, \tau_1^2)$ is included in the interval $(-\beta_p, \beta_p)$.

```
v = 1 # Set this to the phenotypic variance
p = 0.01 # Effect explains less than 1% of the trait variance,
target.prob = 0.95 # with this probability
maf = 0.25 # minor allele frequency
# 2*maf*(1 - maf)*b^2 = p * v --> b = +/- sqrt(p*v/(2*maf*(1-maf)))
b = sqrt(p * v / ( 2 * maf * (1 - maf) ) )
tau.seq = seq(0, 1, 0.001) #grid to evaluate tau
x = pnorm(b, 0, tau.seq, lower = FALSE) #what is the upper tail prob. at b for each value of tau?
tau.1 = tau.seq[which.min( abs(x - (1 - target.prob)/2) )] #which is closest to target prob?
#Check that the probability mass in (-b,b) is indeed close to target
print(paste0("tau.1=",tau.1," has mass ",signif(1 - 2*pnorm(b, 0, tau.1, lower = FALSE),3),
             " in (-",signif(b,4),", ",signif(b,4),")."))
```

```
## [1] "tau.1=0.083 has mass 0.951 in (-0.1633, 0.1633)."
```

Example 4.5. For a case-control GWAS, we want to find such τ_1 that with 95% probability a variant can increase odds by at most 1.30-fold (OR $\leq$ 1.30). Now we get the critical point β_p as equal to $\log(1.30)$ and then proceed as above.

```
or = 1.30 # Effect is at most this large
target.prob = 0.95 # with this probability
b = log(or)
tau.seq = seq(0, 1, 0.001) # grid to evaluate tau
x = pnorm(b, 0, tau.seq, lower = FALSE) # what is the upper tail prob. at b for each value of tau?
tau.1 = tau.seq[which.min( abs(x - (1 - target.prob)/2) )] # which is closest to target prob?
# Check that the probability mass in (-b,b) is indeed close to target
print(paste0("tau.1=",tau.1," has mass ",signif(1 - 2*pnorm(b, 0, tau.1, lower = FALSE),3),
             " in (-",signif(b,4),", ",signif(b,4),")."))
```

```
## [1] "tau.1=0.134 has mass 0.95 in (-0.2624, 0.2624)."
```

Note that the above choices of τ_1 do not model very large effect sizes. There are variants that explain, say, over 10% of the variation of a quantitative trait or that have an OR $\geq$ 3.0 for some disease. To properly model those variants in the Bayesian framework, we would need to use several prior distributions and average the results across them (Bayesian model averaging).