Assigning disease liability to genetic variants using prevalence ratios

Supplementary Material

In addition to this file, supplementary data also includes **Supplementary Table 1**, which contains all genetic data (control and disease populations) and disease liability assignments for alleles, and **Supplementary Table 2**, which contains information about variants with BayPR predictions <50% and which have been previously defined as pathogenic or likely pathogenic (P/LP).

Data Sources

All data as described below were de-identified. The CFTR2 project is acknowledged by the Johns Hopkins IRB (NA_00018599) and does not require IRB approval. All other datasets were obtained from publicly available resources and/or from researchers who provided de-identified variant information that meets NIH Exemption 4 (§46.104(d)(ii)) and does not qualify as human subjects research. In some cases, variant names were provided to the analytic team in legacy terminology or protein name only. To obtain the associated nucleotide changes, original laboratory reports or records were reviewed by providing researchers when available. Additional HGVS cDNA and protein changes were obtained using VEP v10135.

"Control" Variant: Data Source: De-identified aggregated genomic data were downloaded from gnomAD v2.1.1, a publicly-available database consisting of 125,748 exome sequences and 15,708 whole genome sequences from unrelated individuals[27]. All genetic data (control and disease populations) and disease liability assignments for alleles can be found in **Supplementary Table 1**.

Cystic Fibrosis (*CFTR*): Data Source: For the CF disease population, de-identified *CFTR* genotype data in the CFTR2 [Clinical and Functional TRanslation of CFTR] (https://cftr2.org) database were provided for 89,052 individuals by national or regional CF patient registries or individual clinics from clinical records[19]. Individuals with no identified *CFTR* variants were excluded. Out of a total of 150,272 potential disease-causing alleles having received some genetic testing, 8,236 were not identified by available genetic testing and a total of 4,492 complex alleles, duplications/deletions greater than 3 base pairs, and variants in regions with poor or no coverage in gnomAD were excluded from analyses to yield a dataset of 137,544 *CFTR* allelic variants. **Curation:** Disease liability for specific alleles was determined by CFTR2 expert panels using clinical characteristics, functional studies, and penetrance analysis.

Phenylketonuria (*PAH*): Data Source: For the PKU disease population, de-identified *PAH* genotype data in the BIOPKU database were provided for 9,953 individuals[28,29]. Out of a total of 19,906 potential disease-causing alleles, 331 either were not identified by available genetic testing or were duplications/deletions greater than 3 base pairs and were excluded from analyses to yield a dataset of 19,575 *PAH* allelic variants. **Curation:** Disease liability of specific alleles was

determined based on frequencies of the metabolic phenotype for genotypes presenting in a functionally hemizygous state and genotypes were compared to blood phenylalanine levels and BH4 responsiveness[20].

Interstitial Lung Disease (*ABCA3*): Data Source: A list of de-identified variants was assembled from candidate gene sequencing was performed on symptomatic infants and children suspected of having genetic surfactant dysfunction in research laboratories at Washington University School of Medicine and Johns Hopkins University School of Medicine and critically reviewed publications[30]. **Data Curation:** Disease liability was determined by expert review of clinical characteristics, supporting imaging, family segregation studies, and/or lung histopathology if available, and mRNA splicing studies.

X-linked Adrenoleukodystrophy (*ABCD1*): Data Source: *ABCD1* variants were identified during clinical testing in the Johns Hopkins Genomics DNA Diagnostic lab from 2016 to 2020. **Curation:** All variants were classified using current ACMG/AMP criteria[6].

Barth Syndrome (TAFAZZIN): Data Source: De-identified *TAFAZZIN* variants were obtained from the Human Tafazzin Gene Variants Database (a sub-database of the Barth Syndrome Registry and Repository), which is publicly available through the Barth Syndrome Foundation (https://www.barthsyndrome.org/research/tafazzindatabase.html). Data were supplied by contributions from clinicians/families[31]. **Curation:** Disease liability was determined by expert review of clinical characteristics, cardiolipin/MLCL levels, and family segregation studies and/or mRNA splicing studies if available.

Marfan Syndrome (*FBN1*): Data Source: Allelic variants were identified during clinical testing at various CLIA-approved commercial labs from 2006 to 2020. **Curation:** Clinical data and variants are reviewed by a single clinical director and genetic counselor. All variants were classified using current ACMG/AMP criteria[6].

Loeys-Dietz Syndrome (*TGFBR1/TGFBR2*): Data Source: Allelic variants were identified during clinical testing at various CLIA-approved commercial labs from 2006 to 2020. *SMAD3, TGFB2,* and *TGFB3* variants were not examined in this study. **Curation:** Clinical data and variants are reviewed by a single clinical director and genetic counselor. All variants were classified using current ACMG/AMP criteria[6].

Supplementary Methods

Summary

An empirical Bayesian approach was used to analyze the counts of allelic variants in the disease-specific population database relative to the control gnomAD population using a two component finite mixture model. In this empirical Bayes model, information across variants is pooled in order

to get better estimates of each individual variant. Theoretically, this allows for more stable estimates, even when there may be few or no variants observed among the control cohort. The amount of information pooled is dependent on the number of variants observed, and the prior estimated from the data. With a richer dataset (the number of variants observed is larger, the total sample size is larger), there is less pooled information, but in a more restricted dataset (the number of variants observed is smaller, the total sample size is smaller), pooling increases. The algorithm is also dependent on the prior; when the prior is has a flatter distribution, reflecting more uncertainty about the distribution of prevalence ratios, the prior will exert less of an effect on the observed prevalence ratio. However, when the prior distribution is more peaked, reflecting less uncertainty about the distribution of prevalence ratios, it will exert more of an effect on the observed prevalence ratio.

Each component modeled the allele counts of specific allelic variants in the disease-specific database conditional on the total number of observations, using a binomial likelihood, placing a beta prior on a transformation of the prevalence ratio. Estimates were obtained by maximizing the marginal likelihood using the expectation-maximization (EM) algorithm implemented using the optimx package for R version (R Foundation for Statistical Computing, Vienna, Austria). A grid search was used to assess sensitivity to starting values for the EM algorithm, including the parameters of each mixture component and the mixing fraction. The model allows for variation in the size of each database (particularly for true for the control database used in this study), but does assume a constant ratio of the ratio of these database sizes; variation from this ratio would induce a different prior on the prevalence ratio. For assessing the accuracy of our Bayesian algorithm, the probability of a variant belonging to one of the two beta distributions was compared to known disease liability obtained through functional and/or clinical data.

Code availability: The R program used to generate the probabilities can be found on github (https://github.com/melishg/BayPR/).

Introduction

For a particular genetic variant, denoted variant k, the binomial model allows us to infer about the proportion of variants arising from controls, denoted θ_k. This model depends on Y_{1k}, the number of observed variants in cases among the n_{1k} cases assessed for variant k, and Y_{0k}, the number of observed variants in controls among the n_{0k} controls assessed for variant k. The total number of observed variants of type k is denoted T_k, and the relative sample size in cases relative to controls for variant k is denoted $r_k = (n_{1k}/n_{0k})$. We can transform the proportion of variants among the cases to get a prevalence ratio for variant k, denoted γ_k: this tells us how much more prevalent a variant is in cases relative to controls. The transformation from the proportion of variants arising from cases (θ_k) to the prevalence ratio (γ_k) depends on the sample size ratio r_k.

If we have K total variants, instead of viewing each variant in isolation, we can view them as a sample from a population of genetic variants. In this population of variants, the proportion of variants arising from controls follows a beta distribution, whose shape is determined by two parameters, α and β. When viewed in this way, each variant informs us about the distribution of the proportion of variants arising from controls in the larger population of variants, which helps provide more stable estimates for a particular variant when the observed frequency of that variant is low.

In this beta-binomial model, each variant is viewed arising from a population of variants, and in this population, the average proportion of variants arising from controls is $\mu = \alpha/(\alpha + \beta)$, and the variation in these proportions is $\mu(1 - \mu)/(M + 1)$, where $M = (\alpha + \beta)$. From a Bayesian viewpoint, the parameters of the beta distribution, α and β, can be interpreted as observing α prior variants among cases, and β prior variants among controls, equivalent to data from a prior sample size of $M = (\alpha + \beta)$. This information can provide a stabilizing influence when the number of variants of a particular type are small or zero in one group. The stabilizing effect, called shrinkage, depends on the amount of observed data.

Let $s_k = M/(M + T_k)$ denote the ratio of the prior sample size M to the total sample size (prior sample size + number of variants of type k observed). This s_k can be thought of as a 'shrinkage' or 'stabilization' factor for variant k: when the number of observed variants is small relative to the prior sample size, the estimates are stabilized by being 'pulled' or 'shrunken' towards the population mean according to this shrinkage factor. As the number of observed variants becomes larger, this ratio will approach zero, and this shrinkage effect diminishes. We can also view this as a more principled, data-driven way of doing what some may tempted to do in practice: adding some small quantity to a numerator or denominator in order to avoid a numerical singularity.

Instead of viewing all variants as belonging to one homogenous population of variants, we could potentially view a sample variants as arising from a mixture of populations of variants. This is the conceptual basis of the mixture model. The added flexibility of this model allows for more local pooling of information across variants, and determining the sizes of population components and how compatible each variant is with each population component.

There are different methods for obtaining the parameters α and β, which describe the variation between variants in the proportion of variants arising from cases. Here, we used marginal maximum likelihood: finding the values of α and β (equivalently μ and M) that maximize the marginal likelihood of the data, averaging over θ_k. Since we are obtaining α and β from the data itself, not from independent prior information, the 'prior' here can be thought of as a mechanism for stabilizing estimates that approximates a full Bayesian analysis and provides many desirable statistical properties.

Outline

1. Why a Bayesian Approach?

Suppose we have information about the occurrence of K genetic variants from a sample of cases from a population with a given disease and a sample of healthy controls from a population without that disease. One way of estimating the degree of association between a given genetic variant and disease status is to estimate the prevalence ratio: the ratio of genetic variant prevalence in cases relative to that in controls.

If we observe Y_{1k} variants of type k in n_{1k} cases, and Y_{0k} variants of type k in n_{0k} controls, we can estimate the prevalence in cases, denoted π_{1k}, using $\hat{\pi}_{1k} = Y_{1k}/n_{1k}$, and the prevalence in controls, denoted π_{0k}, using $\hat{\pi}_{0k} = Y_{0k}/n_{0k}$. We can estimate the prevalence ratio in cases relative controls, γ_k, using the ratio of the prevalence estimates: $\hat{\gamma}_k = (\hat{\pi}_{1k}/\hat{\pi}_{0k}) = (Y_{1k}/n_{1k})/(Y_{0k}/n_{0k})$. Since it is the ratio of two probabilities, the prevalence ratio could take on any positive value: ratios greater than 1 indicate higher prevalence among cases, and ratios less than 1 indicate higher prevalence among controls.

One issue that may arise is that variants may be extremely rare in one population or another, which may result in unstable prevalence ratio estimates. If few or no variants are observed among controls, the denominator of the ratio becomes very small, resulting in unstable estimates of the prevalence ratio. Bayesian models augment the observed data with a prior distribution, which can provide estimates that are more stable and also have good frequentist statistical properties, such as interval estimate coverage and mean squared error.

2. The Probability Model

2.1 The Binomial Likelihood

For each genetic variant $k = 1, 2, \ldots, K$, we assume that we have a random sample of n_{1k} cases and n_{0k} controls who were evaluated for that variant, with Y_{1k} occurrences in cases and Y_{0k} occurrences in controls. The prevalence of variant k in cases is denoted by π_{1k}, and the prevalence in controls is denoted by π_{0k}. The proportion of individuals in a sample of size n of a variant k independently sampled from a population with prevalence π_k can be modeled using a binomial distribution:

$$Y \sim Bin(n, \pi)\colon Pr\{Y = y|n, \pi\} = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

This binomial likelihood can be approximated using a Poisson distribution with rate parameter $\lambda = n\pi$ if n, the number of observations, is large and π_k, the prevalence of the variant, is small.

$$Y \sim Poisson(\lambda = n\pi)\colon Pr\{Y = y|n, \pi\} = \frac{(n\pi)^y e^{-n\pi}}{y!}$$

If we observe Y_{1k} variants of type k in cases, Y_{0k} variants of type k in controls, let T_k denote their sum. We can model the proportion of type k variants that arose from cases out of the the total number of type k variants observed, denoted θ_k, using a binomial distribution:

$$Y_{0k} \sim Poisson(n_{0k}\pi_{0k}),\ Y_{1k} \sim Poisson(n_{1k}\pi_{1k});\ Y_{0k} \perp Y_{1k}\colon (Y_{1k}|Y_{0k} + Y_{1k} = T_k) \sim Bin(T_k, \theta_k),$$

where $\theta_k = (n_{1k}\pi_{1k})/(n_{1k}\pi_{1k} + n_{0k}\pi_{0k})$, the rate of occurrences in cases divided by the sum of the rates in cases and controls. We can divide the numerator and denominator by π_{0k} to parameterize this distribution by the prevalence ratio γ_k:

$$\theta_k = \frac{n_{1k}\pi_{1k}/\pi_{0k}}{n_{1k}\pi_{1k}/\pi_{1k} + n_{0k}\pi_{0k}/\pi_{1k}} = \frac{n_{1k}\gamma_k}{n_{1k}\gamma_k + n_{0k}}$$

Let $r_k = n_{1k}/n_{0k}$ denote the ratio of the sample size of cases relative to controls. Dividing the numerator and denominator again by n_{0k}, the control sample size, gives:

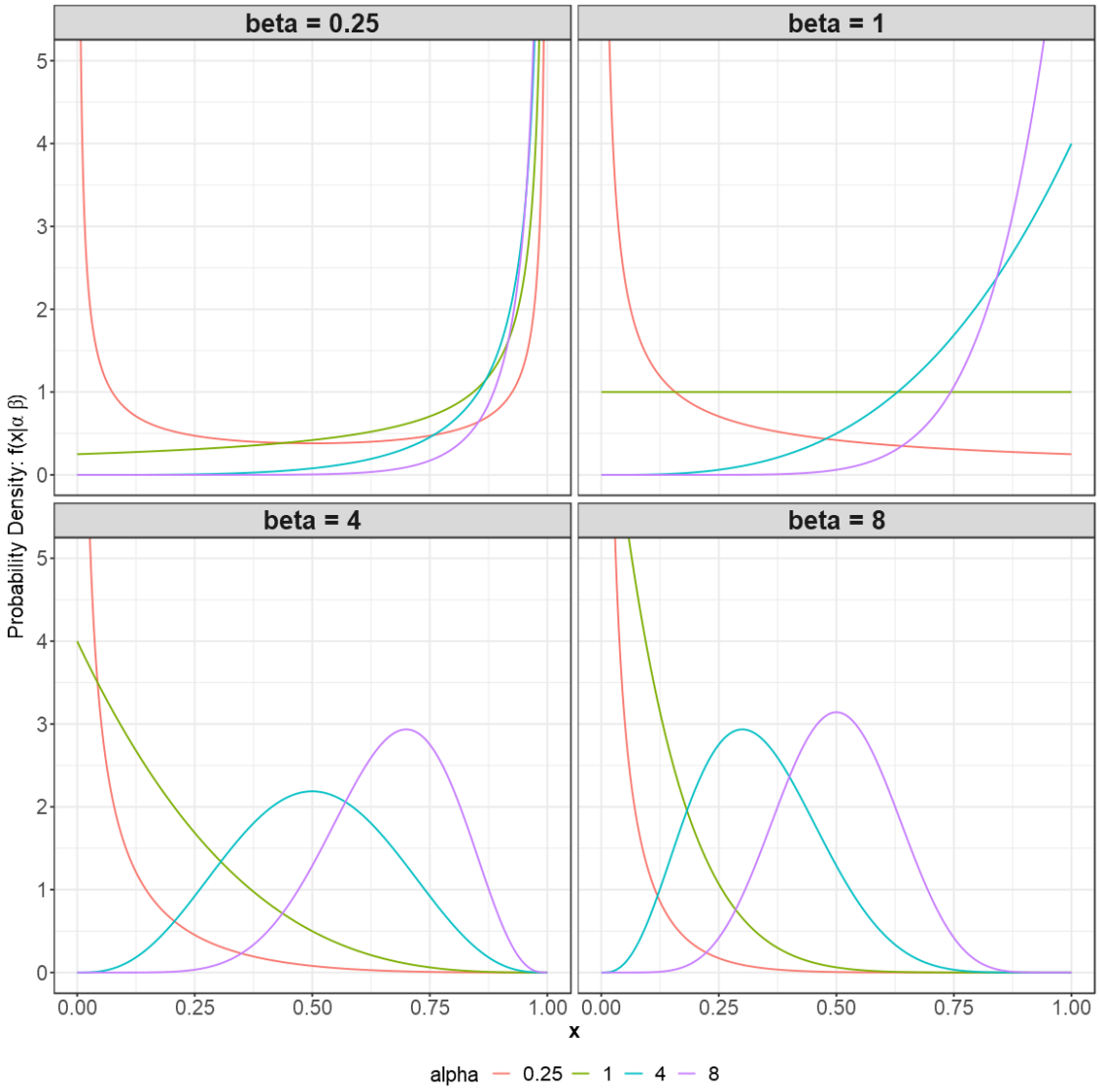
$$\theta_k = \frac{n_{1k}\gamma_k/n_{0k}}{n_{1k}\gamma_k/n_{0k} + n_{0k}/n_{0k}} = \frac{r_k\gamma_k}{r_k\gamma_k + 1}$$

Parameterizing the model this way allows the modeling multiple variants, allowing for variation in sample sizes across different variants, as long as their ratio of sample sizes for any given variant r_k is comparable.

2.2 The Beta-Binomial Model

Bayesian statistics augments the likelihood probability model with a prior probability model. Our model for the data is a binomial distribution, which depends on $n = T_k$, the number of variants observed and $\pi = \theta_k$, the proportion of variants arising from the population of cases. Instead of viewing each of our K variants separately, we could view them as a sample from a population of genetic variants, and the proportions of variants due to cases in this population follow a probability distribution called the Beta distribution.

The shape of the beta distribution is controlled by two parameters, α and β. Depending on the values these parameters, the beta distribution can take on many possible shapes, as seen in the figures below. The mean of the beta distribution is given by $E[\theta_k|\alpha,\beta] = \mu = \alpha/(\alpha+\beta)$, and its variance is given by $var[\theta_k|\alpha,\beta] = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.



Supplementary Figure 1: *Examples of beta distributions obtained by varying the values of the parameters alpha, indicated by the color, and beta, indicated by the figure panel.*

We can also re-parameterize the beta distribution in terms of μ, its mean, and $M = \alpha + \beta$:

$$Pr\{\theta_k|\mu, M\} = \frac{\Gamma(M)}{\Gamma(\mu M)\Gamma((1-\mu)M)}\theta^{\mu M-1}(1-\theta)^{(1-\mu)M-1}$$

In this parameterization, $\alpha = \mu M$ and $\beta = (1-\mu)M$. We can think of M as a prior sample size: α prior variants among cases and β prior variants among controls. When parameterized this way, the variance of the beta distribution is given by $var[\theta_k|\mu, M] = \mu(1-\mu)/(M+1)$.

The combination of a beta distribution for the proportion of variants arising from cases, and the binomial likelihood for the number of variants observed in cases, gives a Beta-Binomial Distribution. The parameters of the beta distribution are called hyperparameters, because their values determine the distribution of the proportion parameter in the binomial model.

The probability density function for the prior is:

$$Pr\{\theta_k|\alpha,\ \beta\} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_k^{\alpha-1}(1-\theta_k)^{\beta-1}$$

The likelihood for the observed data is:

$$Pr\{Y_{1k}|T_{1k},\ \theta_k\} = \binom{T_{1k}}{Y_{1k}}\theta_k^{Y_{1k}}(1-\theta_k)^{Y_{1k}-T_{1k}}$$

Note the similarity between these two functions: both are products of the terms θ_k and $(1-\theta_k)$. When combined using Bayes' Rule, they give a posterior distribution is also a beta distribution:

$$Pr\{\theta_k|Y_{1k},\ T_k,\ \alpha,\ \beta\} = \frac{\Gamma(\alpha+\beta+T_k)}{\Gamma(\alpha+Y_{1k})\Gamma(\beta+Y_{0k})}\theta^{\alpha+Y_{1k}-1}(1-\theta)^{\beta+Y_{0k}-1}$$

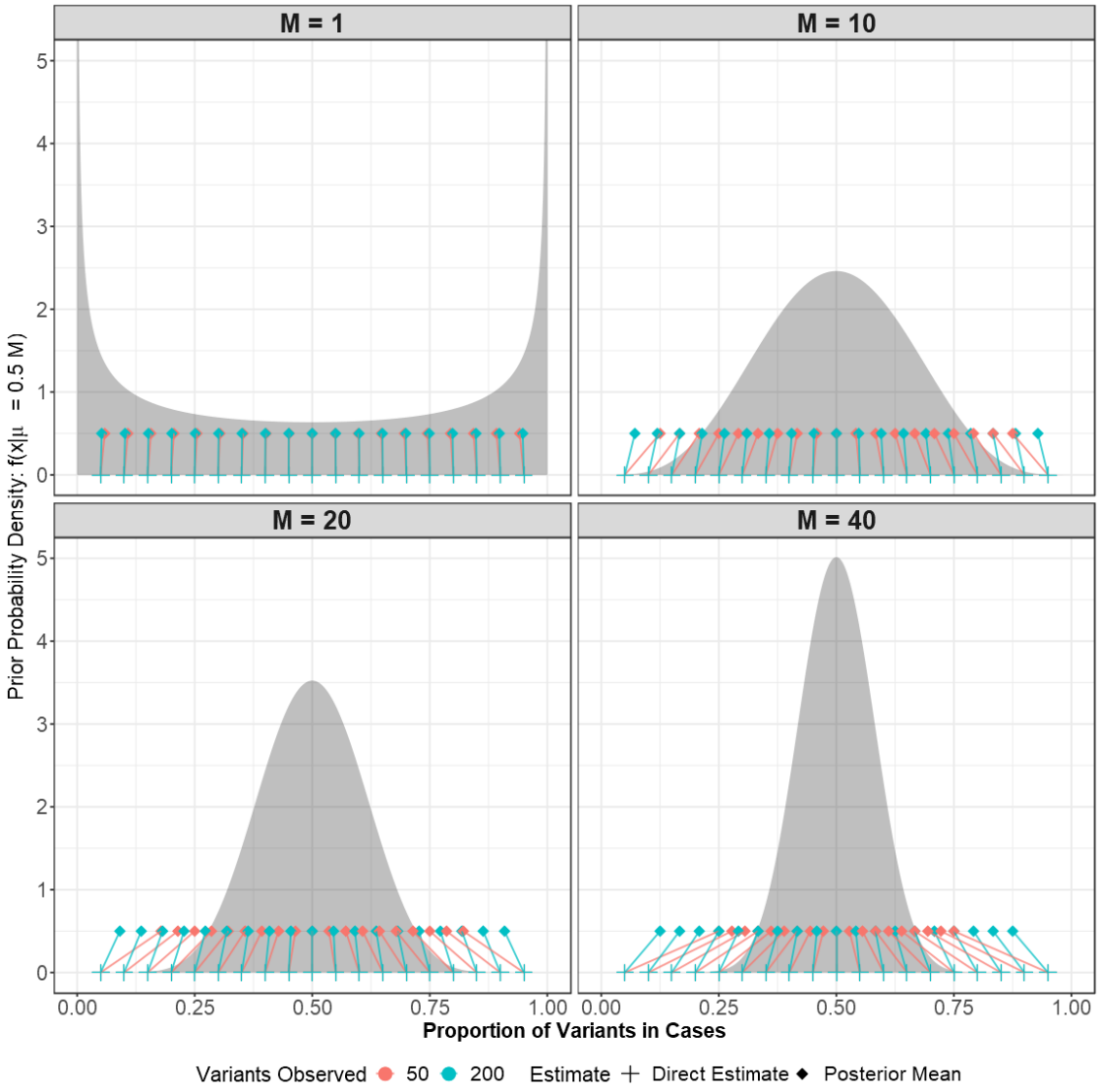
This is just a beta distribution with $\alpha' = \alpha + Y_{1k}$ and $\beta' = \beta + Y_{0k}$: this is why we can interpret α as the prior number of variants observed in cases, and β as the prior number of variants observed among controls. The posterior distribution, and its summaries (such as its mean, variance, and quantiles), allow us to infer about θ_k, the proportion of variants due to cases, for each of our K variants.

2.3 Bayes and Shrinkage

The advantage of the Bayesian approach is that the prior distribution acts as a stabilizing influence when the number of observed variants is small, and this influence diminishes as the number of observed variants becomes larger. The effect of the prior can be more easily understood when viewed through the parameterization of μ and M. The mean of the posterior distribution is given by $E[\theta_k|\alpha,\beta] = \mu = \alpha/(\alpha+\beta)$. Re-writing this in terms of μ and M gives:

$$E[\theta_k|\alpha,\beta] = \frac{\alpha+Y_{1k}}{\alpha+\beta+T_k} = \frac{\mu M+Y_{1k}}{M+T_k} = \frac{M}{T_k+M}\mu + \frac{T_k}{T_k+M}\left(\frac{Y_{1k}}{T_k}\right) = s_k\mu + (1-s_k)\left(\frac{Y_{1k}}{T_k}\right)$$

Since M acts as a prior sample size, and T_k is the total number of variants observed, the quantity $s_k = M/(M+T_K)$ represents the proportion of the information coming from the prior, and $(1-s_k) = T_k/(M+T_K)$ represents the proportion of information coming from the observed data. Here we see the posterior mean is a combination of the prior mean μ and the proportion of variants among cases to the total number of variants observed (Y_{1k}/T_k), each weighted according their sample size contribution. When T_k, the number of observed variants is small, the posterior mean is 'pulled' or 'shrunken' towards the prior mean: this gives more stable estimates. As the number of observed variants becomes increasingly larger than the prior sample size, the effect of the prior diminishes.



Supplementary Figure 2: *An example of how the shape of the prior affects the posterior mean, and how this depends on both the number of observed variants, the proportion of variants seen in cases, and the prior sample size. The prior distribution is shown in gray. When the prior sample size (M) is small, the effect of the prior is smallest, and the posterior means are very close to the direct estimates. As the prior sample size increases, direct estimates are 'pulled' or 'shrunken' towards the prior mean (here, 0.5), with greater shrinkage occurring when fewer variants are observed.*

2.4 The Beta Prime Distribution

The beta distribution describes the distribution of a random variable over the interval (0,1), which is convenient for describing a proportion or probability. However, if we want to infer about the prevalence ratio γ_k, we need to transform back to the prevalence ratio scale:

$$\frac{\theta_k}{r_k(1-\theta_k)} = \gamma_k$$

Notice that when $r_k = 1$ (sample sizes are identical between cases and controls), this transformation is from the probability scale to the odds scale. If we have a beta random variable, and transform it to the odds scale, this results in a variable with the beta prime distribution. The shape of this distribution is governed by the same parameters, α and β: If $X \sim Beta(\alpha, \beta)$ then $Y = X/(1-X) \sim Beta\,Prime(\alpha, \beta)$.

If $(Y|\alpha, \beta) \sim Beta\,Prime(\alpha, \beta)$, its mean is given by $E[Y|\alpha, \beta] = \alpha/(\beta - 1)$, and its variance is given by $var[Y|\alpha, \beta] = \alpha(\alpha + \beta - 1)/((\beta - 2)(\beta - 1)^2)$

For a beta-prime posterior distribution of prevalence ratios, we can plug in our posterior values of $\alpha' = \alpha + Y_{1k}$ and $\beta' = \beta + Y_{0k}$. When $r_k \neq 1$ (the sample size differs between cases and controls), the posterior becomes a scaled version of the beta prime distribution. The mean is scaled by $1/r_k$, and the variance is scaled by $1/r_k^2$:

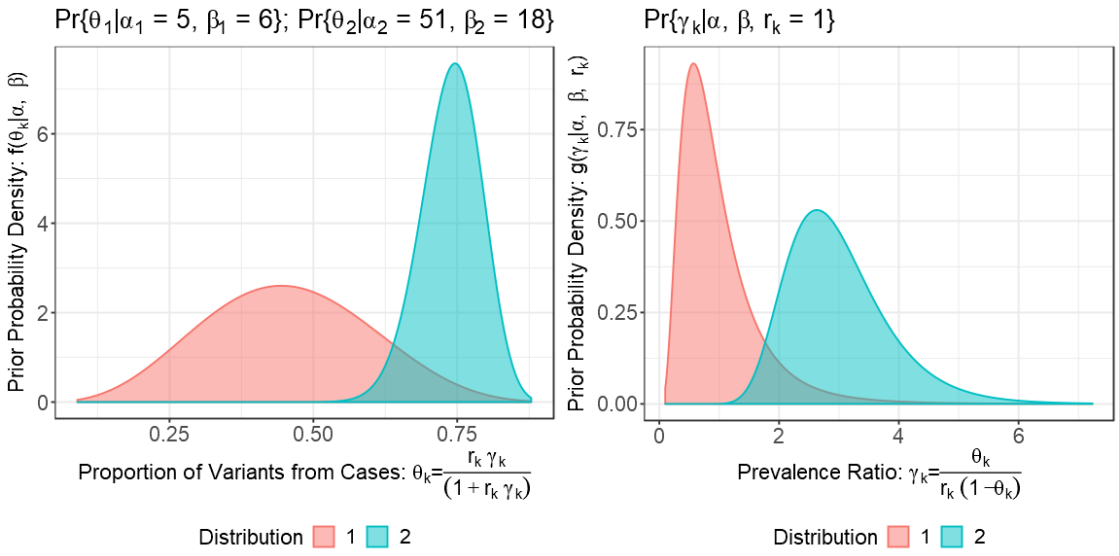
$$E[\gamma_k|Y_{1k}, T_k, \alpha, \beta] = \frac{1}{r_k}\frac{\alpha + Y_{1k}}{(\beta + Y_{0k} - 1)} = \frac{\frac{\alpha}{n_{1k}} + \frac{Y_{1k}}{n_{1k}}}{\frac{\beta - 1}{n_{0k}} + \frac{Y_{0k}}{n_{0k}}}$$

By adding α/n_{1k} to the numerator, and $(\beta - 1)/n_{0k}$ to the denominator, this stabilizes the prevalence ratio. If we think of α as the prior number of variants observed among cases, and β as the prior number of variants among controls, the numerator is 'pulled' or 'shrunken' towards α, as the denominator is towards β. This pull becomes negligable as the sample size becomes large.

This can also be seen when we paramaterize instead using μ and M:

$$E[\gamma_k|Y_{1k}, T_k, \mu, M] = \frac{\frac{\mu M}{n_{1k}} + \frac{Y_{1k}}{n_{1k}}}{\frac{(1-\mu)M - 1}{n_{0k}} + \frac{Y_{0k}}{n_{0k}}}$$

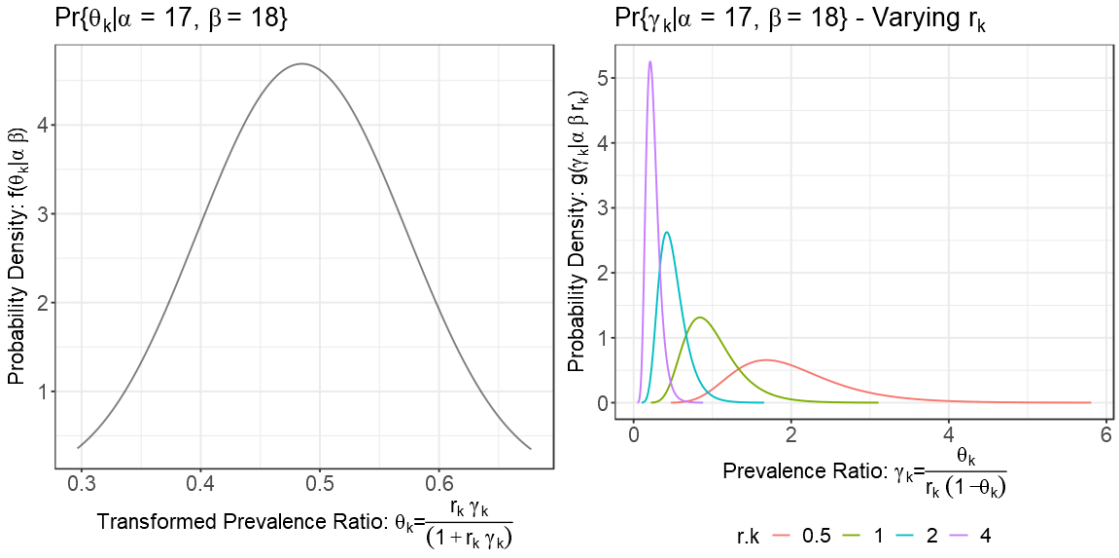
The numerator is 'pulled' or 'shrunken' towards the prior mean μ, and this effect diminishes as n_{1k}, the sample size among cases, becomes large compared to the prior sample size M. The denominator is 'pulled' or 'shrunken' towards $(1 - \mu)$, and this effect diminishes as n_{0k}, the sample size among controls, becomes large compared to the prior sample size M.



Supplementary Figure 3: *Examples of two beta distributions, modeling the proportion of variants observed in cases, and their corresponding distributions of prevalence ratios.*

2.5 The Effect of r_k on the Prior

Note that the transformation to the prevalence ratio scale depends on r_k, the ratio of sample sizes in cases (n_1) relative to controls (n_0). Note how the same beta model could represent populations of variants with lower, equal, or higher prevalence in cases relative to controls, depending on the value of r_k:



Supplementary Figure 4: *Example of how one beta distribution, could represent lower, equal, or higher prevalence between cases and controls, depending on the relative sample size between cases and controls.*

This also illustrates the importance of the assumption about the variation r_k across all K variants: if there is appreciable variation in this ratio across variants, then variants will essentially be 'pulled' or 'shrunken' in different directions.

3. Empirical Bayes Estimation

Up until now, we have been treating the parameters of our beta prior distribution, α and β, as known quantities, and showing how their values 'pull' or 'shrink' the direct estimates towards the prior mean on on the proportion (or θ_k) scale, and 'pull' or 'shrink' the numerator and denominator on the ratio scale. But how do we determine suitable values for these 'hyperparameters?'

Rather than either supplying exact values for these parameters, or specifying a prior distribution over these parameters, we can find the values of these parameters that maximize the *marginal likelihood* of the observed data, averaging over the parameter θ_k: $Pr\{Y_{1k}|T_k, \mu, M\}$

$$Pr\{Y_{1k}|T_k, \mu, M\} = f(Y_{1k}|T_k, \mu, M)$$
$$= \frac{\Gamma(T_k+1)}{\Gamma(Y_{1k}+1)\Gamma(T_k-Y_{1k}+1)} \frac{\Gamma(Y_{1k}+\mu M)\Gamma(T_k-Y_{1k}+(1-\mu)M)}{\Gamma(T_k+M)} \frac{\Gamma(M)}{\Gamma(\mu M)\Gamma((1-\mu)M)}$$

where Γ denotes the gamma function.

4. Using a Mixture Model

Instead of assuming that all variants are represented by one beta distribution that describes the proportion of variants among cases, we can relax this assumption by assuming that the observed data were generated from a mixture of different beta distributions.

Instead of our model having only two parameters, μ and M, our mixture model will have 5 parameters: μ_1 and M_1, which control the shape of one beta distribution, μ_2 and M_2, which control the shape of the second beta distribution, and ϵ, which is the proportion of variants arising from the second beta distribution:

$$f(Y_{1k}|T_k, \mu_1, M_1, \mu_2, M_2, \epsilon) = \epsilon f(Y_{1k}|T_k, \mu_1, M_1) + (1-\epsilon) f(Y_{1k}|T_k, \mu_2, M_2)$$

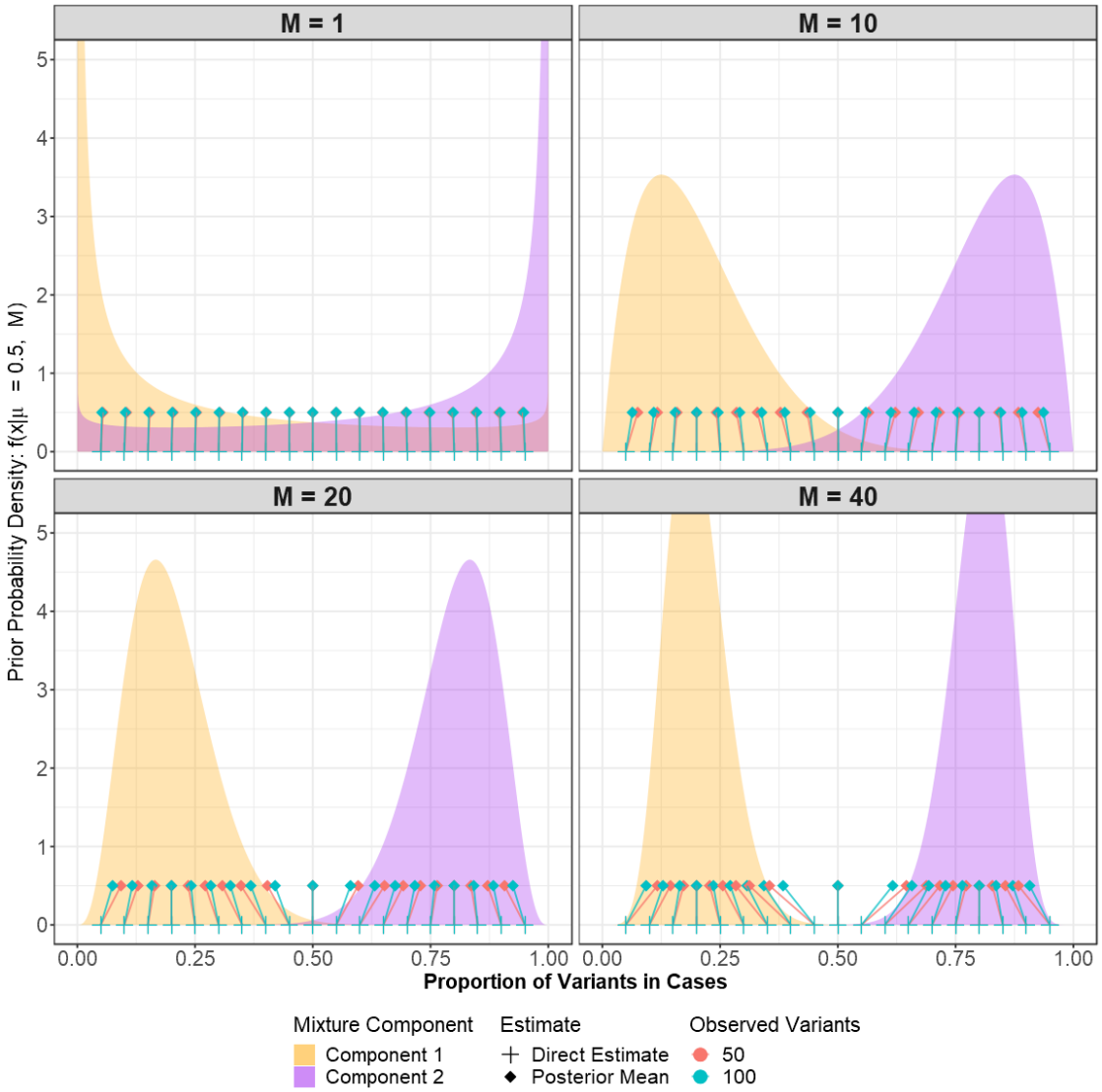
Note that two different parameter vectors ($\mu_1 = a, M_1 = b, \mu_2 = c, M_2 = d, \epsilon = e$) and ($\mu_1 = c, M_1 = d, \mu_2 = a, M_2 = b, \epsilon = 1-e$) result in identical values of the mixture model: for this reason, the constraint $\epsilon < 0.5$ is imposed to identify a unique solution. Estimates of these parameters are obtained by maximizing the marginal likelihood of the mixture.

Instead of each direct estimate being 'pulled' or 'shrunken' in the same direction, each distribution will exert a different 'pull' on the data, with the 'pull' being related to the relative compatibility between each model component and the data.

From this mixture distribution, in addition to obtaining posterior means, variances, and quantiles, we can additionally obtain:

- The marginal likelihood ratio: $MLR_{2/1} = Pr\{Y_{1k}|T_k, \mu_2, M_2\}/Pr\{Y_{1k}|T_k, \mu_1, M_1\}$
- The posterior odds of belonging to component 2 vs. 1: $PO_{2k} = (\epsilon/(1-\epsilon))MLR_{2/1}$

- The posterior probability of belonging to component 2: $P_{2k} = PO_{2k}/(1 + PO_{2k})$



Supplementary Figure 5: *An example of a mixture prior, and the effect of each mixture component on the direct estimate. When the prior sample size (M) is small, the effect of the prior is smallest, and the posterior means are very close to the direct estimates. As the prior sample size increases, direct estimates are 'pulled' or 'shrunken' towards each of the priors. The 'pull' of each prior depends on the marginal likelihood ratio: the relative compatibility between the direct estimate and each prior probability component.*

Supplementary Results

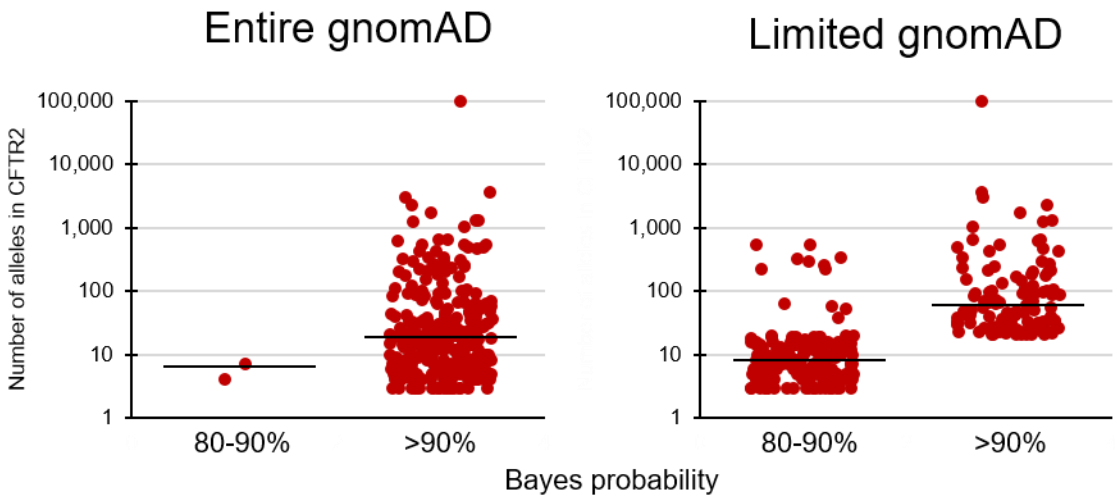


Figure S1. Data is jittered for display purposes only in this Figure S1; precise allele counts and BayPR probabilities are provided in the Data Supplement. A total of 286 P/LP *CFTR* variants with Bayes scores >80% as determined using the entire gnomAD dataset (exomes and genomes; ~140,000 individuals) are grouped by their scores (80-90% and >90%) plotted by allele count within CFTR2. There is no significant difference in allele counts between these two groups (Wilcoxon rank sum; p=0.1275). A total of 280 P/LP *CFTR* variants with Bayes scores >80%, as determined by the limited gnomAD dataset (genomes only; ~15,000 individuals), are grouped by score and plotted against allele count within CFTR2. The variants in the 80-90% group are less frequent in CFTR2 than the variants in the >90% group (Wilcoxon rank sum; p<0.001).

Table S1. Sensitivity and specificity analysis using a limited gnomAD dataset (genomes only)

	Variant interpretation	n variants scored	BayPP score threshold*	Positive n (%)	Negative n (%)	Sensitivity (95% CI)	Specificity (95% CI)	n variants scored	BayPP score threshold*	Positive n (%)	Negative n (%)	Sensitivity (95% CI)	Specificity (95% CI)
CFTR	Path / Likely path	296	90%	110 (37.2)	186 (62.8)	37.2 (31.6, 42.9)	94.1 (71.3, 99.9)	296	80%	280 (94.6)	16 (5.4)	94.6 (91.4, 96.9)	94.1 (71.3, 99.9)
	Benign / Likely benign	17	10%	1 (5.9)	16 (94.1)			17	20%	1 (5.9)	16 (94.1)		
PAH	Path / Likely path	393	90%	376 (95.7)	17 (4.3)	-	-	393	80%	390 (99.2)	3 (0.8)	-	-
ABCA3	Path / Likely path	225	90%	218 (96.9)	7 (3.1)	96.9 (93.7, 98.7)	100.0 (47.8, 100.0)	225	80%	218 (96.9)	7 (3.1)	96.9 (93.7, 98.7)	100.0 (47.8, 100.0)
	Benign / Likely benign	5	10%	0 (0.0)	5 (100.0)			5	20%	0 (0.0)	5 (100.0)		
FBN1	Path / Likely path	178	90%	178 (100.0)	0 (0.0)	-	-	178	80%	178 (100.0)	0 (0.0)	-	-
TGFBR1	Path / Likely path	23	90%	23 (100.0)	0 (0.0)			23	80%	23 (100.0)	0 (0.0)		
TGFBR2	Path / Likely path	55	90%	2 (3.6)	53 (96.4)	-	-	55	80%	54 (98.2)	1 (1.8)	-	-
ABCD1	Path / Likely path	61	90%	61 (100.0)	0 (0.0)	100.0 (94.1, 100.0)	100.0 (2.5, 100.0)	61	80%	61 (100.0)	0 (0.0)	100.0 (94.1, 100.0)	100.0 (2.5, 100.0)
	Benign / Likely benign	1	10%	0 (100.0)	1 (0.0)			1	20%	0 (100.0)	1 (0.0)		
TAFAZZIN	Path / Likely path	143	90%	143 (100.0)	0 (0.0)	-	-	143	90%	143 (100.0)	0 (0.0)	-	-

*Score thresholds were set to determine whether a variant had a positive or negative result from the Bayesian XYZ algorithm for the purposes of sensitivity and specificity analysis. For P/LP variants, Bayesian XYZ scores >90% or >80% were considered true positives. For B/LB variants, Bayesian scores <10% or <20% were considered true negatives. P/LP and B/LB determinations were made by expert review. Only genes with both P/LP and B/LB underwent sensitivity and specificity analysis.