Supplemental Statistical Analysis for "Meta-analysis of genotype-phenotype associations in Bardet-Biedl Syndrome uncovers differences among causative genes"

Contents

Part 1: Main Bayesian Analysis **2**
- The data . . . 2
- Handling missingness . . . 4
- The model - an accessible explanation . . . 5
- The model - mathematical formulation . . . 8
- Main results . . . 8
- Summary for functional groups . . . 11
- Summary for BBSome genes . . . 12
- Summary for Chaperonins . . . 13
- Pairwise comparisons of mutations in BBSome genes . . . 13
- Type of mutation - loss of function . . . 16
- Discrepancies between frequentist and Bayesian analysis . . . 17

Part 2: Alternative Models & Model Selection **22**
- Model descriptions . . . 22
- Choosing the model for main analysis . . . 26

Part 3: Conclusions under Multiverse Analysis **42**
- Defining Precise Criteria . . . 42
- Bayesian Comparison . . . 43
- Frequentist results . . . 44
- Combining all results . . . 45

Original computing environment **46**

This document consists of three parts:

- Part 1 describes the Bayesian analysis reported in the main manuscript.
- Part 2 describes all Bayesian models we tried throughout this project and discusses the reasoning behind our choice of model for the main analysis, in particular why between-study variability is crucial and taking into account the type of mutation (i.e. whether it is complete loss of function) is useful while age, and sex can be omitted.
- Part 3 shows how the conclusions of the paper hold under multiple different models.

The complete source code for the analysis can be found at https://github.com/martinmodrak/bbs-metaanalysis-bayes or Zenodo, DOI: 10.5281/zenodo.3243264

Part 1: Main Bayesian Analysis

The data

First let us examine some of the properties of the dataset we are working with - a brief summary follows.

```
## Skim summary statistics
##  n obs: 899
##  n variables: 21
##
## -- Variable type:character ------------------------------------------------------------------------
##        variable missing complete   n min max n_unique
##             age     424      475 899   1   7       92
##         case_no       4      895 899   1  20      800
##    ethnic_group       0      899 899   2   4        9
##       ethnicity       0      899 899   2  34       89
##  mutation_types      70      829 899   3  11        5
##             OBE     150      749 899   1   2        3
##
## -- Variable type:factor ---------------------------------------------------------------------------
##          variable missing complete   n n_unique
##         age_group     424      475 899        7
##         family_id      47      852 899      585
##  functional_group       0      899 899        4
##              gene       0      899 899       20
##  loss_of_function       0      899 899        2
##               sex     403      496 899        2
##            source       0      899 899       85
##
## -- Variable type:numeric --------------------------------------------------------------------------
##  variable missing complete   n mean n_unique
##        CI     234      665 899 0.66        2
##        DD     600      299 899 0.81        2
##     HEART     674      225 899 0.3         2
##       LIV     617      282 899 0.3         2
##        PD     169      730 899 0.79        2
##        RD      65      834 899 0.94        2
##       REN     227      672 899 0.52        2
##       REP     456      443 899 0.59        2
```

Note in particular, that both age and sex are missing in almost half of the records. Also, the data about individual phenotypes (all the numeric columns) is largely incomplete. Some minor clearing is required to use age, as it is stored as character (a combination of age ranges and ages). For some phenotypes we get values that are not 0 or 1 - those correspond to patients that were monitored in multiple studies, but the phenotype data was inconsistent between studies. In our analysis, we treat those patients as exhibiting the phenotype.

`loss_of_function` encodes the fact that the particular mutation certainly leads to the loss of function of the protein (is truncated in both alleles).

For some analyses, we group the genes together according to *functional groups*, those are defined as follows:

functional_group	genes
BBS03	BBS03
BBSome	BBS01,BBS02,BBS04,BBS05,BBS07,BBS09,BBS08,BBS18
Chaperonins	BBS10,BBS12,BBS06
Others	BBS13,BBS14,BBS16,BBS20,BBS21,BBS17,BBS11,BBS19

And here are the counts of individual mutations as observed in the data:

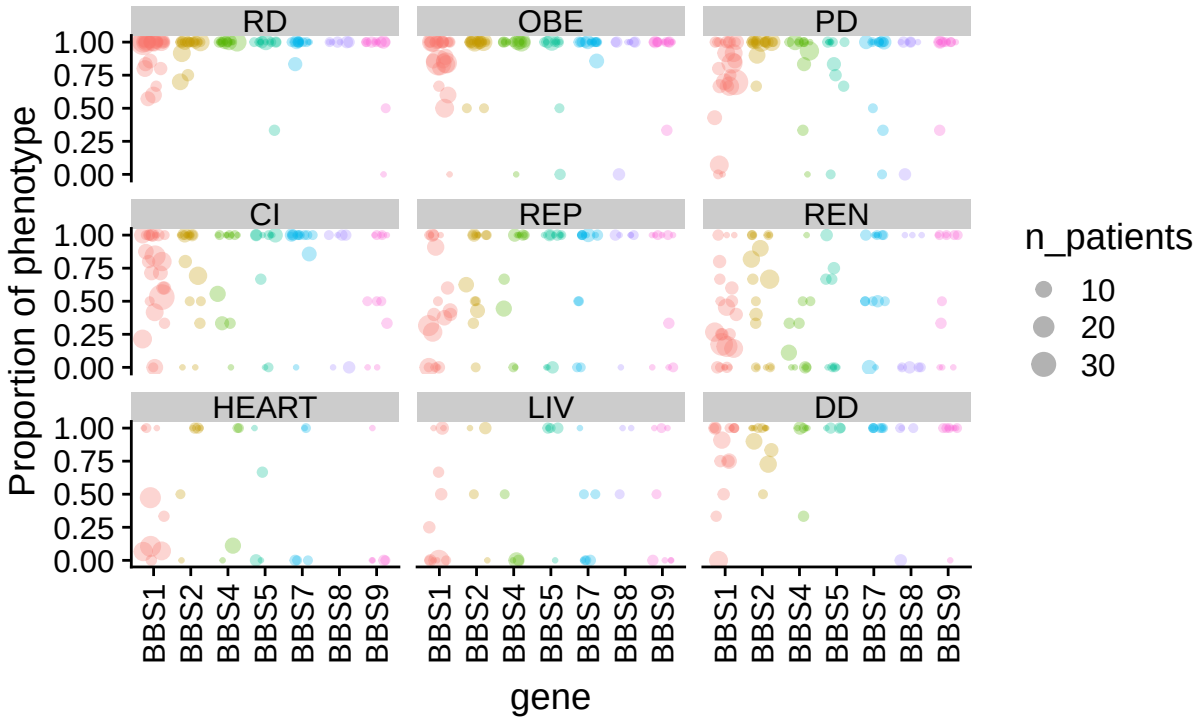
gene	count
BBS01	214
BBS02	85
BBS03	47
BBS04	49
BBS05	34
BBS06	58
BBS07	37
BBS08	17
BBS09	29
BBS10	133
BBS11	1
BBS12	59
BBS13	9
BBS14	58
BBS16	39
BBS17	3
BBS18	1
BBS19	2
BBS20	9
BBS21	15

While we include all of the genes in our computational model, we will mostly show only the most frequent mutations in the results here; those include BBS1 through BBS10 and BBS12.

The phenotypes present in the data are:

```
## [1] "Retinal dystrophy (RD)"    "Obesity (OBE)"
## [3] "Polydactyly (PD)"          "Cognitive impairment (CI)"
## [5] "Reproductive system (REP)" "Renal anomalies (REN)"
## [7] "Heart anomalies (HEART)"   "Liver anomalies (LIV)"
## [9] "Developmental delay (DD)"
```

The data shows considerable between-study (source) variability (showing only the BBSome genes for clarity):

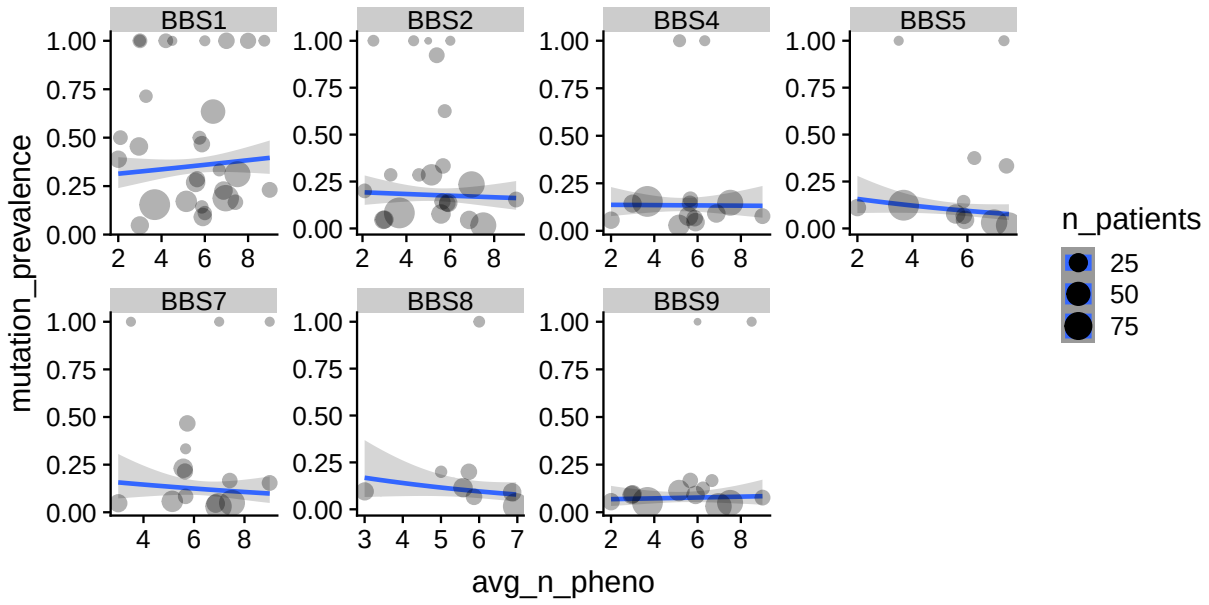


Here, every point represents the proportion of patients with a given mutation that manifested a given phenotype in one study. The point size represents the number of patients with the mutation in the study. We see that even studies with a relatively large number of patients show very different proportions. We are therefore confident, that allowing for between-study heterogeneity is important for analyzing the data correctly. In Part 2, we also show an attempt to model the studies as homogenous and find it to be a bad fit, although most of the qualitative conclusions are supported in both models, as described in Part 3.

Handling missingness

The most problematic missing data problem is the missingness in phenotype data. There are two distinct sources of missingness: a) a study missing the phenotype value for only some patients or b) a study not reporting the status of the phenotype for any patient. Our analysis assumes the phenotype data to be missing at random, i.e. that the decision to not report a given phenotype in a study and missingness for individual patients is independent of the prevalence of the phenotype in the study population. This is probably not true for missingness at the study level, as the investigators are plausibly more likely to report more prevalent phenotypes and more likely to ignore phenotype that was not observed in any patient. Similarly, if data for a specific patient omits a given phenotype (the state of the phenotype is reported as missing data in the original study), it is more likely the phenotype was not present.

However, in our analysis we were unable to find a good way to account for this phenomenon. But since we focus on comparison of the prevalence of individual phenotype across different mutations within a single study and do not compare phenotypes against each other, this should only be a significant issue if the rate of missingness in phenotype values was correlated with the prevalence of individual mutations present in a study (e.g., if studies with high obesity missingness would also tend to have overabundance of mutations in BBS3). Let's check whether this is the case:



In the figure above, each dot is a single study and shows the average number of phenotypes reported per patient vs. the prevalence of mutations in individual genes. Point size corresponds to the number of patients in a given study. Looking at the figure, a strong association of missingness to specific mutations seems implausible, but we can't completely rule out that it biases our results. This is a limitation of our approach and should be taken into account when interpreting our conclusions.

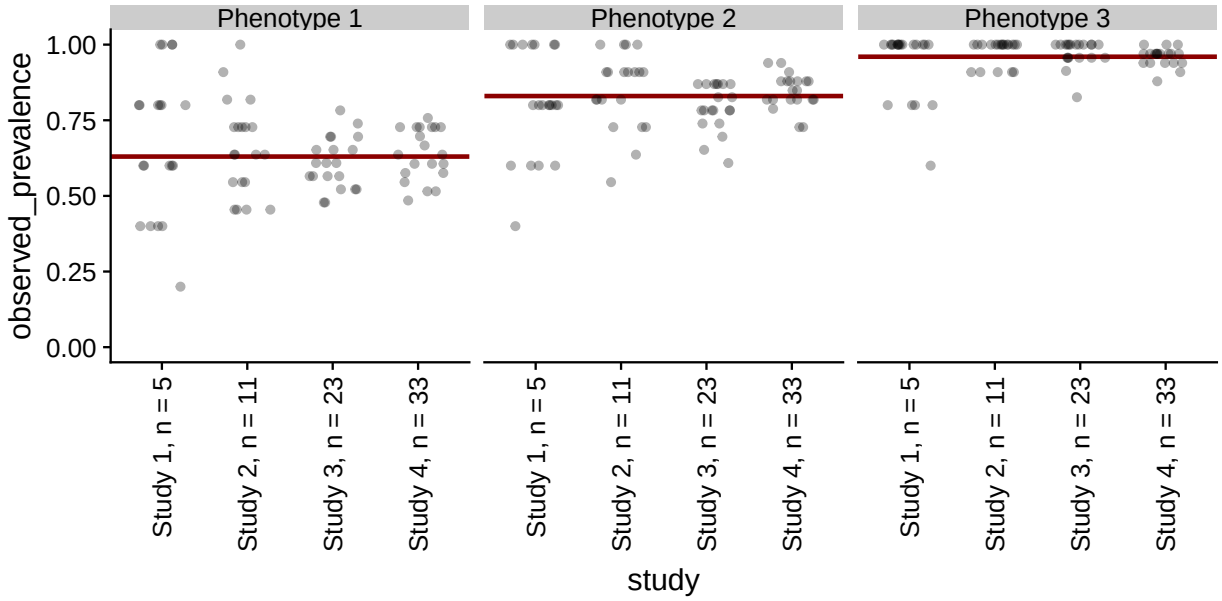
There is large missingness in age and sex data as well. While our main analysis ignores age and sex, Part 2 shows that after accounting for the between-study differences, age and sex differences are already mostly accounted for. We have also tried to impute age and sex data and show that models including imputed age and/or sex provide almost identical results (see Part 3).

The model - an accessible explanation

An exact mathematical formulation is given in the following section. This section may be safely skipped.

As all models, the model we use simplifies and abstracts the medical reality in hope that we can arrive at useful conclusions. Our model is a member of generalized linear model family, using logit link and hierarchical terms in a fully Bayesian treatment. Let's unpack this a little, starting with what a logit link does. In the following, we will describe how we handle a single phenotype, as the estimation for individual phenotypes is mostly independent.

Our model tries to estimate theoretical *true prevalence* of the phenotype in a population - i.e. the probability that a randomly selected patient from the population will exhibit this phenotype. But all we observe is that each individual either exhibits the phenotype or not. Depending on the number of individuals enrolled in a study, the *observed prevalence* will jump more or less around this true prevalence - let's have a look at an example:



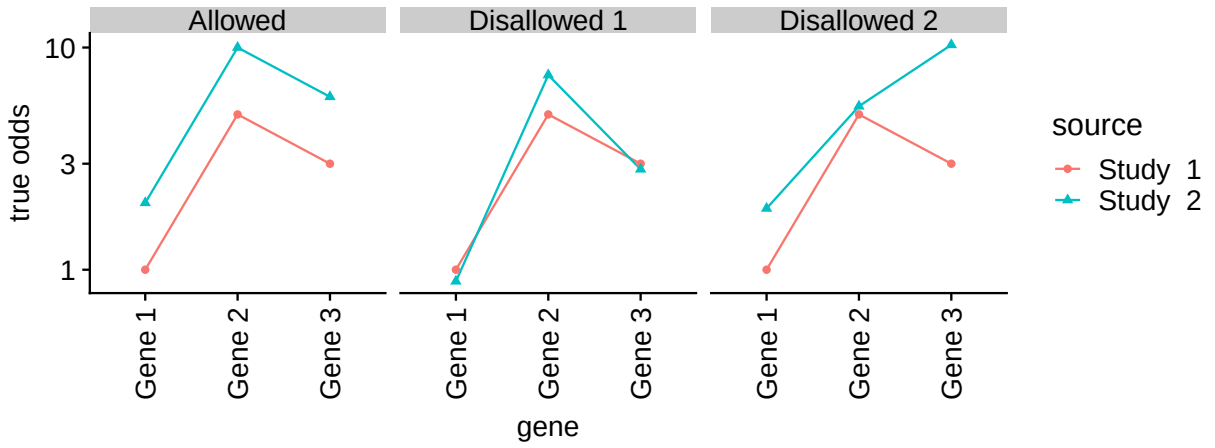
The red line shows the true prevalence of a phenotype (assuming it is the same across multiple studies). Each point shows the observed prevalence of a single possible realization of a study. We see that the observed prevalence can differ substantially from the true value and that the spread decreases with increasing n. In our data, we can't however expect large n, as the biggest n we have is (Castro-Sanchez et al. 2015) with 33 patients having mutation in BBS01 and in 74% of cases there are less than 5 patients with a given mutation in the same study. Also note that the observed values are clustered at discrete "levels" because e.g., among 5 patients you only can have prevalence of 0.2 or 0.4 and nothing in between. Further, you can see that with high prevalence, there is less variation in the observed prevalence.

For mathematical convenience, the model does not work with prevalence directly, but with *odds*. Odds are just another way to express prevalence - for example, when the prevalence is 20%, we expect one patient to exhibit the phenotype for each four patients not exhibiting the phenotype, leading to odds $1:4$ or 0.25 and log odds (base 10 here) of roughly -0.6. Unlike prevalence, which is constrained between 0 and 1, odds can be any non-negative number.

Most of the results of the model are reported as comparisons of odds in different populations. For this, we use the ratio of the corresponding odds. E.g., when the odds ratio is 2, we expect the first population to have twice as much patients exhibiting the phenotype for each healthy patient than the second population.

Technically, the model works with the logarithm of odds. The logit function transforms a probability (prevalence) to log odds, hence the logit "link" used in the model. What is the linear part?

Our model assumes that the true odds of a phenotype is a function of four numbers (*coefficients*): the overall odds of the phenotype in the population, a modifier for the gene the patient has damaged, and a modifier for the study the patient is enrolled in. One additional modifier is added when the mutation is a certain loss-of-function (cLOF). These four numbers are multiplied to arrive at the final odds for the patient. Assuming only cLOF mutations for simplicity, this means that while the odds of a phenotype are allowed to vary between genes and the overall rate of a phenotype may vary between studies, the odds ratio of different genes is the same across all studies. Let's look at an example:



Above, on the leftmost panel, we see that the two studies differ in the odds of the phenotype for each gene, but the ratio of odds for Gene 1 to Gene 2 (and 3) is the same in both studies (since the odds are shown on log scale this manifests as a constant gap between the two lines). This type of between-study variation is allowed. On the other two panels, the odds ratio for Gene 1 to Gene 2 (and 3) differs between studies - this type of variation is not allowed by the model.

Another way to describe the allowed case is that, for both studies a mutation in Gene 2 makes a patient five times more likely to exhibit the phenotype than a patient with a mutation in Gene 1, although the base rate of the phenotype may vary between studies.

The cLOF coefficient is held constant for all genes in a given phenotype, meaning that odds for any given phenotype are multiplied by a small number when the mutation is cLOF. Once again this means that relative odds are the same among cLOF and other mutations, but absolute odds can be higher (or lower) in cLOF mutations.

This is the "linear" part of the model - we multiply odds, which is the same as adding the logarithm of the odds and addition is a neatly linear thing.

Now the "hierarchical" part. This ties the coefficients in the model in two important ways: i) it assumes that small differences in odds across genes and studies are more likely than large differences and updates the estimates accordingly, ii) *partial pooling*: the degree to which odds are allowed to "jump around" across both genes and studies is informed by the data, e.g., if the odds are similar for all genes except one, the model will put higher weight on the possibility that the difference in the last mutation is just noise and shrink its estimate towards the average for other genes. On the other hand, if the odds vary wildly across all genes, the model will assume it is more likely that this is a true variation and it will not shrink the estimate much. The amount of shrinkage also depends on the number of observations as estimates in which there is a larger number of observations are shrunk less. The variability across studies is pooled in a similar way.

Together, those two features result in low risk of overfitting the data, even though we have very little observations for most study - gene - phenotype combinations.

We also allow for a correlation between phenotypes, e.g., that some phenotypes occur frequently together while others rarely manifest in the same patient. Once again, the amount of correlation is estimated from the data.

Finally, the "Bayesian" part: We follow the Bayesian paradigm, so our estimates of the model coefficients are not a single number, but rather a distribution - some values are more likely than others, but the data are insufficient to let us determine the coefficients with high certainty. Therefore, we never report exact numbers, but rather 50% and 95% *credible intervals* of the distribution. Unlike confidence intervals in frequentist analysis, we can directly interpret the 95% credible interval as the interval that contains the true value with 95% probability - assuming our model is correct (which it is not, but we hope it is still a useful abstraction).

The model - mathematical formulation

We use a generalized linear model with logit link and hierarchical terms. Let us dive into the details. For the model, we expand the data into long form, i.e. each row in the dataset corresponds to a combination of patient and reported phenotype (a patient with reported values for 3 phenotypes would correspond to 3 rows in the long form dataset). The model is specified with the following brms formula, using the Bernoulli family with logit link function:

```
## phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + phenotype:loss_of_function_certain
```

This can be expressed in mathematical notation as:

$$Y_i \sim Bernoulli(\mu_i)$$
$$\text{logit}(\mu_i) = \alpha + \beta^1_{p_i} + \beta^2_{p_i,g_i} + \beta^3_{p_i,s_i} + c_i\beta^4_{p_i}$$

Where $p_i \in \{1, ..., P\}$ is the index of the phenotype for i-th row, $g_i \in \{1, ..., G\}$ is mutated gene for i-th row and $s_i \in \{1, ..., S\}$ is the index of the source study for i-th row. c_i is 1 when the mutation on the i-th row is cLOF and 0 otherwise. α is the intercept and β^1, β^2 and β^3 model the overall phenotype prevalence, phenotype prevalence specific to a given mutation and between-study variability in phenotype prevalence respectively. β^4 models the phenotype-specific effect of cLOF.

Note that this is very similar to running a separate regression for each phenotype, with two exceptions: the overall intercept α is explicitly shared between phenotypes and the structure of the priors introduces some information flow between the other coefficients.

The priors we use for the parameters are:

$$\begin{aligned}
\alpha &\sim N(0, 2) \\
\beta^1 &\sim N(0, \sigma_1) \\
\sigma_1 &\sim N(0, 2) \\
\beta^2 &\sim \mathcal{N}_P(\mathbf{0}, \mathbf{\Sigma}) \\
\mathbf{\Sigma} &= \boldsymbol{\sigma}_2 \bar{\mathbf{\Sigma}} \\
\bar{\mathbf{\Sigma}} &\sim LKJ_P(1) \\
\sigma_{2,p} &\sim N(0, 2) \\
\beta^3_{p,s} &\sim N(0, \sigma_{3,p}) \\
\sigma_{3,p} &\sim N(0, 2) \\
\beta^4_p &\sim N(0, 2)
\end{aligned}$$

Note that the prior on β^1 is P-dimensional multivariete normal $\mathcal{N}_P$, explicitly modelling the correlation $\bar{\mathbf{\Sigma}}$ between the prevalence of individual phenotypes and per-phenotype variance $\sigma_{2,p}$, while the other priors are univariate normal.

The $N(0, 2)$ priors on the various parameters are mildly skeptical in that they exclude that any of the parameter would explain odds ratio larger than ~ 50. In Part 3, we show that the results are almost identical, when the priors are different.

Main results

First, a summary of the model fit as posterior intervals for main model parameters (omitting correlation parameters for brevity):

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 + phenotype) || source)
##    Data: data (Number of observations: 4899)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~gene (Number of levels: 20)
##                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(phenotypeRD)         0.66      0.50     0.03     1.94       1218 1.00
## sd(phenotypeOBE)        0.46      0.33     0.02     1.24       1485 1.00
## sd(phenotypePD)         1.84      0.57     0.88     3.11       1571 1.00
## sd(phenotypeCI)         0.94      0.30     0.46     1.65       2088 1.00
## sd(phenotypeREP)        0.63      0.30     0.13     1.32       2204 1.00
## sd(phenotypeREN)        1.33      0.38     0.73     2.26       2119 1.00
## sd(phenotypeHEART)      0.81      0.59     0.04     2.25       2108 1.00
## sd(phenotypeLIV)        1.41      0.57     0.38     2.66       1720 1.00
## sd(phenotypeDD)         0.80      0.58     0.04     2.20       1715 1.00
##
## ~phenotype (Number of levels: 9)
##               Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(Intercept)     1.59      0.47     0.88     2.72       2280 1.00
##
## ~source (Number of levels: 85)
##                     Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
## sd(phenotypeRD)         2.05      0.46     1.28     3.07       2077 1.00
## sd(phenotypeOBE)        2.39      0.48     1.57     3.43       1364 1.00
## sd(phenotypePD)         1.84      0.38     1.22     2.69       1379 1.00
## sd(phenotypeCI)         1.58      0.31     1.05     2.26       1305 1.00
## sd(phenotypeREP)        2.61      0.57     1.66     3.88       1262 1.00
## sd(phenotypeREN)        1.70      0.34     1.11     2.45       1127 1.00
## sd(phenotypeHEART)      3.37      0.86     1.98     5.30       2077 1.00
## sd(phenotypeLIV)        2.77      0.73     1.59     4.39       2079 1.00
## sd(phenotypeDD)         2.72      0.66     1.60     4.22       1662 1.00
##
## Population-Level Effects:
##                                          Estimate Est.Error l-95% CI
## Intercept                                    1.08      0.57    -0.11
## phenotypeRD:loss_of_function_certain        -0.17      0.48    -1.15
## phenotypeOBE:loss_of_function_certain        0.97      0.37     0.26
## phenotypePD:loss_of_function_certain         0.71      0.33     0.09
## phenotypeCI:loss_of_function_certain         0.53      0.26     0.02
## phenotypeREP:loss_of_function_certain        0.24      0.31    -0.36
## phenotypeREN:loss_of_function_certain        0.77      0.26     0.28
## phenotypeHEART:loss_of_function_certain     -0.93      0.53    -1.99
## phenotypeLIV:loss_of_function_certain       -0.23      0.47    -1.14
## phenotypeDD:loss_of_function_certain         1.29      0.53     0.30
##                                          u-95% CI Eff.Sample Rhat
## Intercept                                    2.20       1481 1.00
## phenotypeRD:loss_of_function_certain         0.76       4494 1.00
## phenotypeOBE:loss_of_function_certain        1.70       5133 1.00
## phenotypePD:loss_of_function_certain         1.36       4223 1.00
## phenotypeCI:loss_of_function_certain         1.04       4933 1.00
```

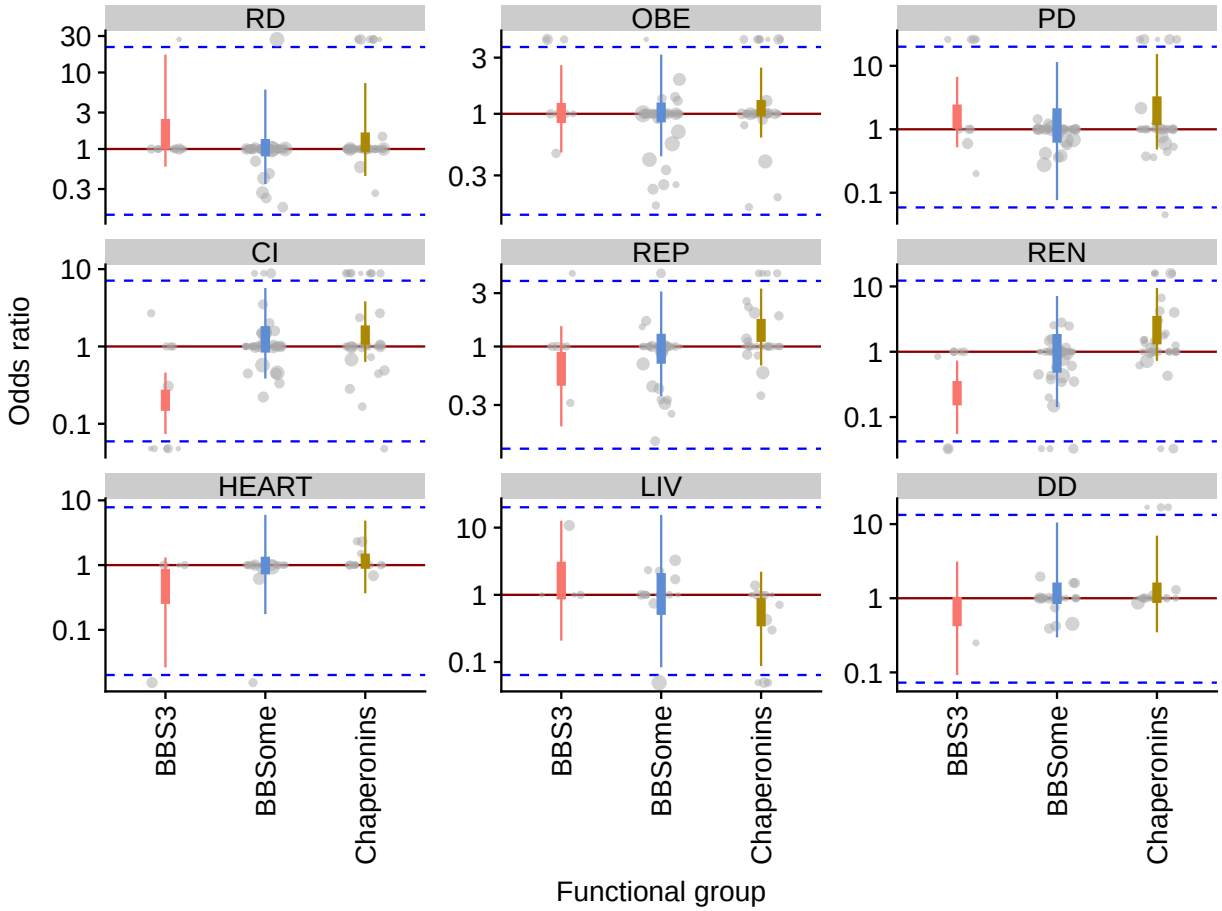
```
## phenotypeREP:loss_of_function_certain       0.84       5202 1.00
## phenotypeREN:loss_of_function_certain       1.28       5933 1.00
## phenotypeHEART:loss_of_function_certain     0.11       5116 1.00
## phenotypeLIV:loss_of_function_certain       0.70       5346 1.00
## phenotypeDD:loss_of_function_certain        2.41       4854 1.00
##
## Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
## is a crude measure of effective sample size, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The fitted model parameters themselves are however hard to interpret, as they operate on log odds scale and there might be notable covariance in the posterior. It is also hard to say how to handle the between-study variability of coefficients. And this variability is substantial - note that the `sd` parameters under `~source` (corresponding to σ_3) admit ranges from ~ 1.1 to ~ 5.1, so the odds of a phenotype, given a mutation can plausibly differ between studies by $1.96 \times \pm 1.1 = \pm 2.16$ to $1.96 \times \pm 5.09 = \pm 9.99$ *on the log scale* (95% of mass of a normal distribution is within $1.96 \times \sigma$ from the mean).

Instead, we will focus on model predictions. In particular, the results we report can be interpreted as if a new study is drawn at random from the same population of studies as we used (i.e. matching all the inclusion criteria) and we directly observe true odds of all phenotypes for all mutations in this study. That is, the predictions do include between-study variability, our uncertainty about the population of studies, our uncertainty about overall prevalence of the individual phenotypes, our uncertainty about the strength of links between mutations and phenotypes and our uncertainty about correlations between the presence of individual phenotypes. The predictions do NOT include the sampling uncertainty of the hypothetical new study. For example when the hypothetical study has the true odds of a phenotype, given a mutation in BBS12, equal to $1 : 2$ (0.5), a study on 20 patients can easily observe odds of $3 : 17$ (0.18) or $11 : 9$ (1.22) simply due to chance, but we will treat the hypothetical study here as having odds of 0.5 and ignore this additional noise in our results.

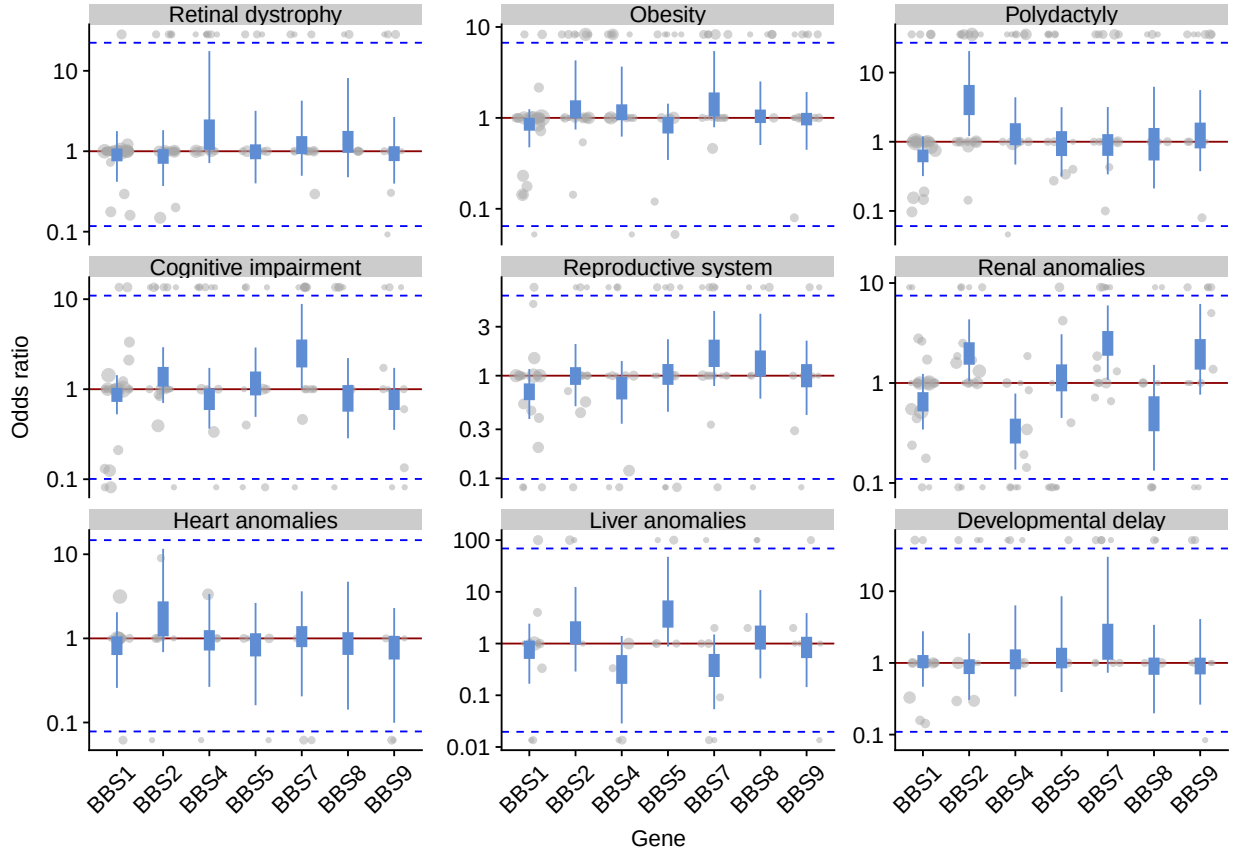
Since those odds can vary wildly between studies, we will focus on various odds ratios (OR) within a single hypothetical study.

Summary for functional groups



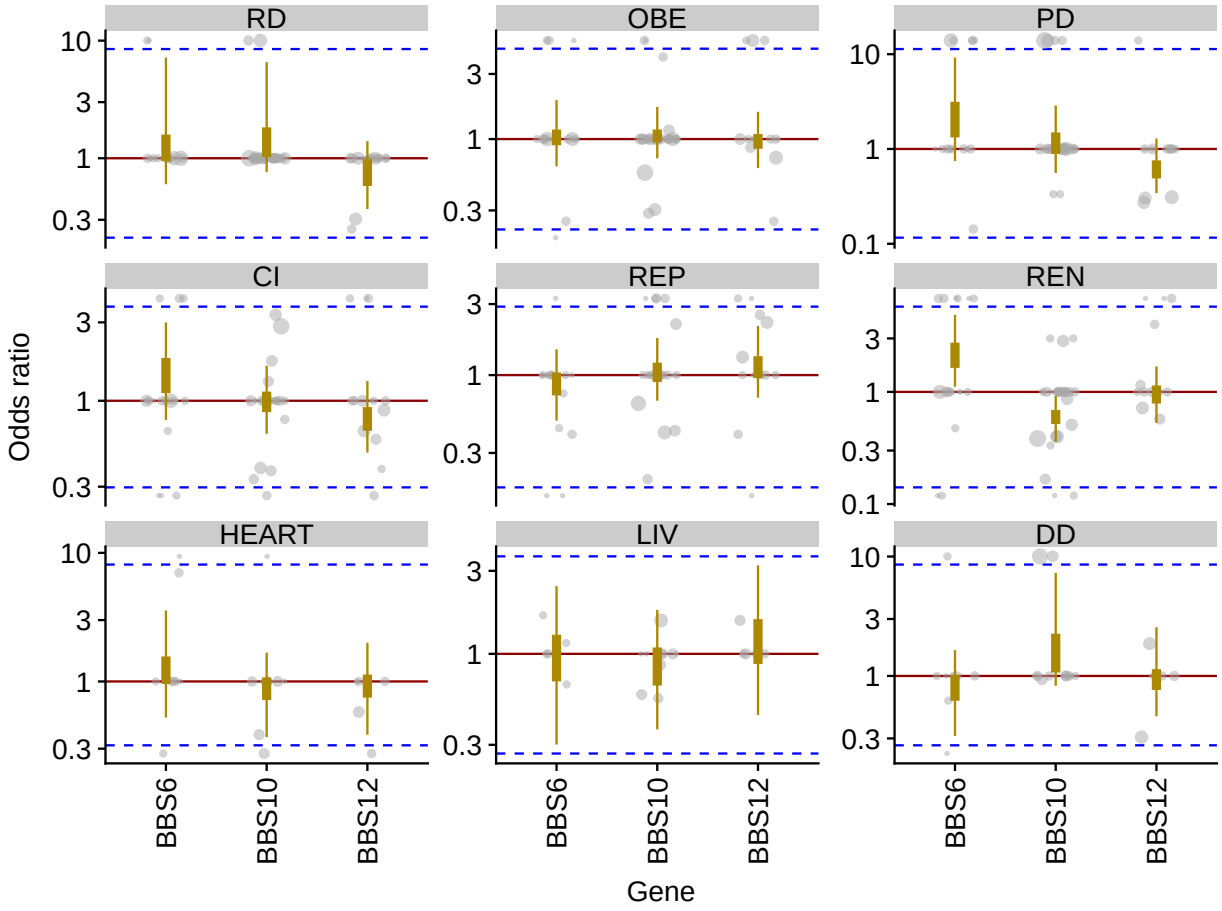
The plot above shows posterior 95% (thin) and 50% (thick) credible intervals for ratio of odds for a phenotype given a random mutation within a functional group to odds for the phenotype given a random mutation across all groups shown. All mutations are assumed to be equally likely - the odds are not weighed by the frequency of the mutations in the dataset. Odds ratios are shown on the log scale and each phenotype has its own scale. Gray dots show the same odds ratio calculated for individual studies included in the meta-analysis. Dots outside the dashed lines correspond to studies, where the empirical odds ratio is 0 or infinity. Dot size represents the number of relevant cases in the study.

Summary for BBSome genes



The plot above shows posterior 95% (thin) and 50% (thick) credible intervals for ratio of odds for a phenotype given a mutation in a gene to odds for the phenotype given a random mutation across all genes shown. All mutations are assumed to be equally likely - the odds are not weighed by the frequency of the mutations in the dataset. Odds ratios are shown on the log scale and each phenotype has its own scale. Gray dots show the same odds ratio calculated for individual studies included in the meta-analysis. Dots below the dashed lines correspond to studies where the empirical odds ratio is 0 or infinity. Dot size represents the number of relevant cases in the study.

Summary for Chaperonins



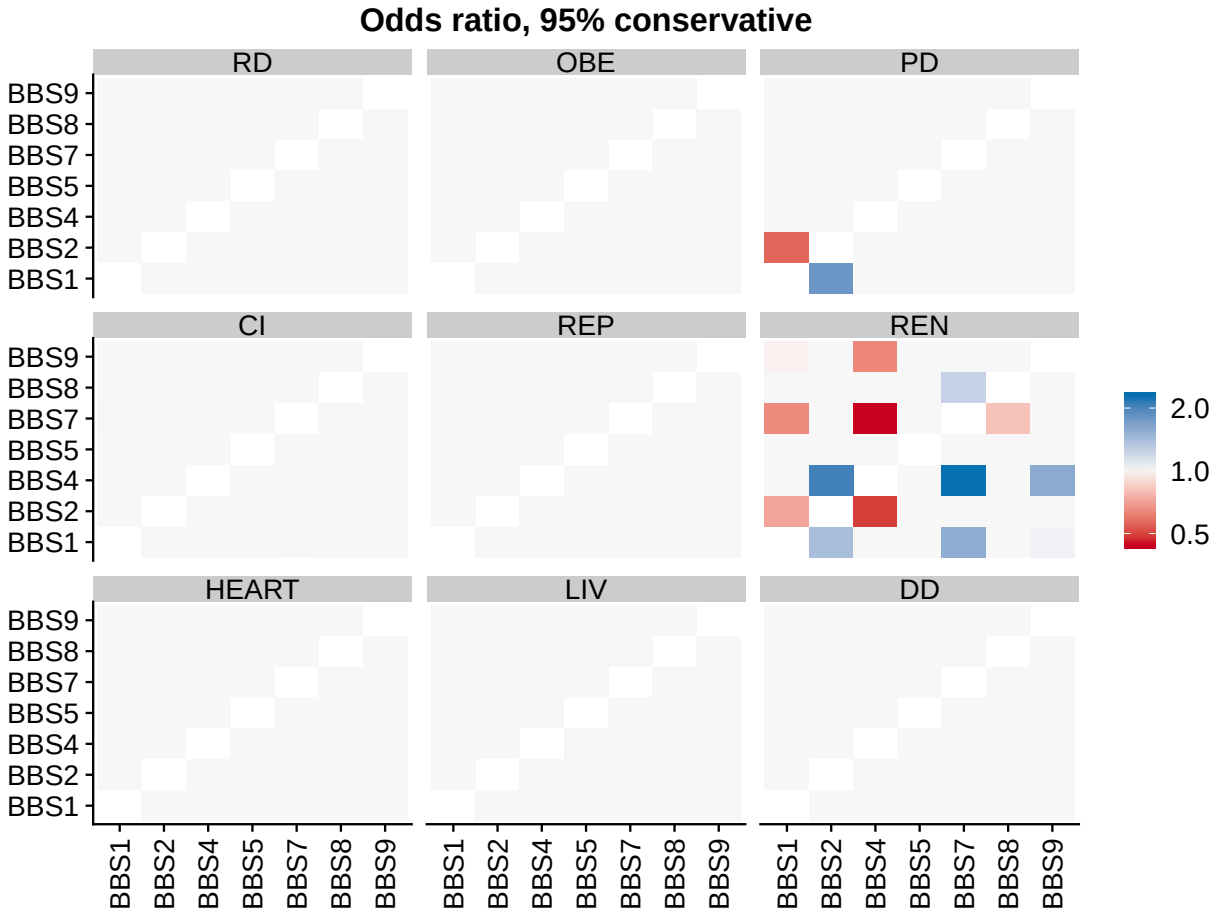
This is the same plot as for the BBSome genes, only showing the members of the chaperonins group.

Pairwise comparisons of mutations in BBSome genes

The summary plots above are not well suited to infer pairwise comparisons, as the estimates for the individual genes / functional groups are not independent. In particular, non-overlapping marginal credible intervals in the summary plot imply that there is a consistent difference, but the converse is not true. If there is a strong positive correlation, there might be a consistent difference even when the above plot would show mostly overlapping marginal posterior intervals.

Pairwise comparisons also have the benefit of better interpretability, as we do not need to rely on odds ratio of the phenotype against some average, which might not be clinically meaningful. In pairwise comparisons, we can directly work with odds ratios for the phenotype given the two mutations in question.

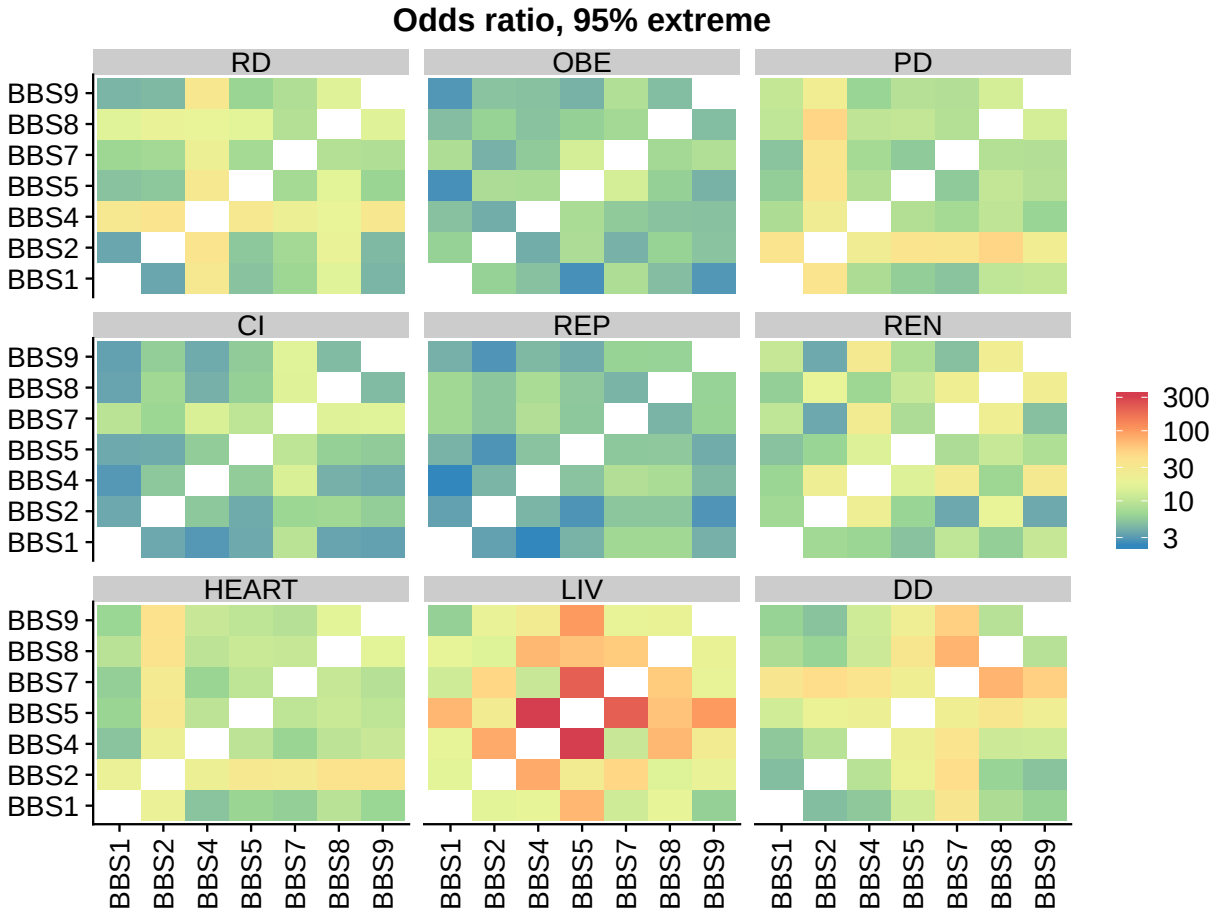
Conservative estimates



The most conservative (closest to one) pairwise odds ratios within 95% posterior credible intervals. The reported odds ratios are for gene on the horizontal axis against the gene on the vertical axis.

This shows pairs, where we are fairly certain there is a systematic difference and the minimal magnitude of this difference consistent with the data.

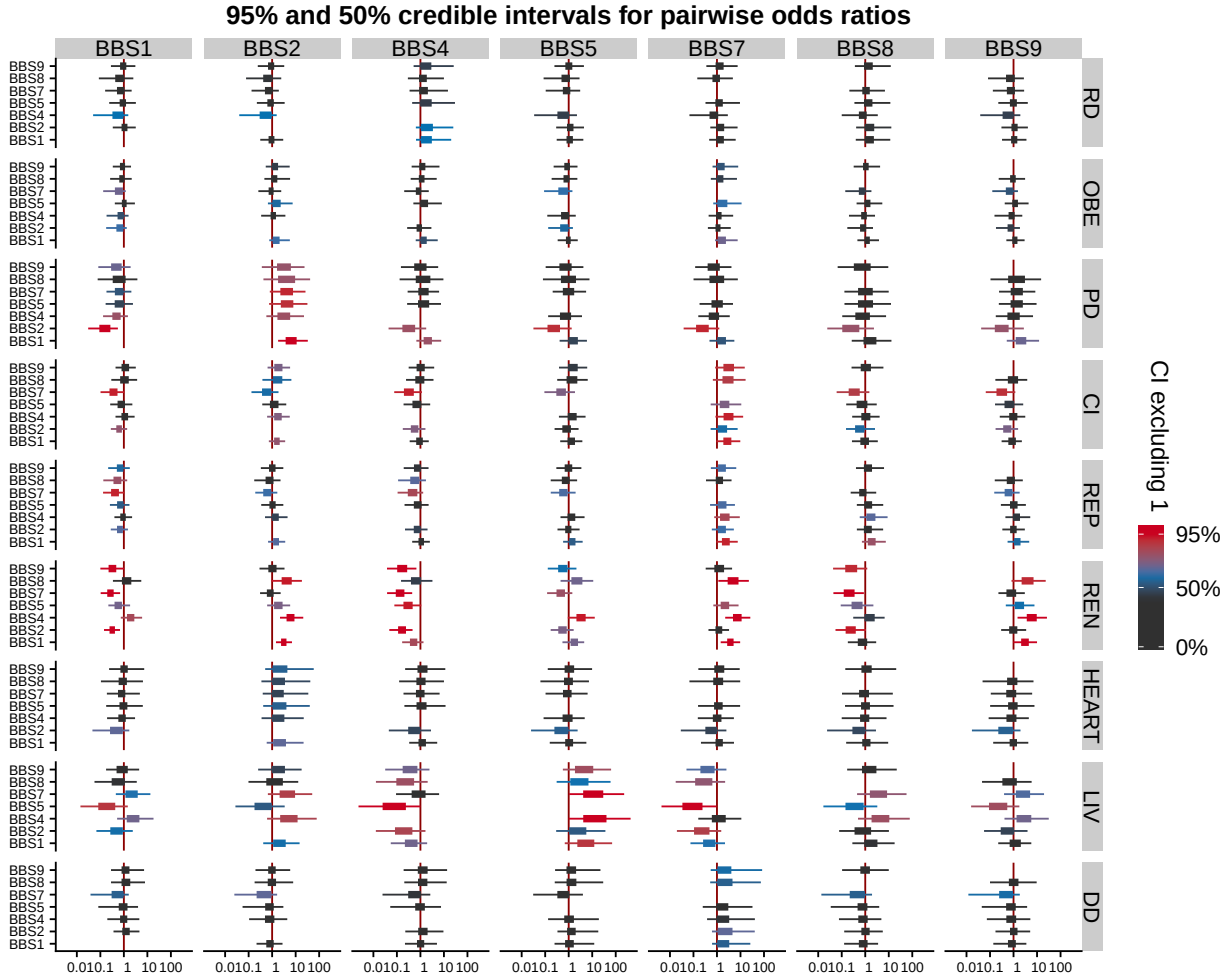
Extreme estimates



The most extreme (furthest from one) pairwise odds ratios within 95% posterior credible intervals. The direction of the effect is not reported as effects in both directions might be similarly plausible - the odds ratios are transformed to be larger than one in all cases.

This shows maximal differences consistent with our model and lets us constrain the differences between mutations for some phenotypes. We see that the data does not let us to put tight constraints on most differences - the tightest we get is OR of 3 - which would still be a very important difference for the clinical prognosis.

Everything at once

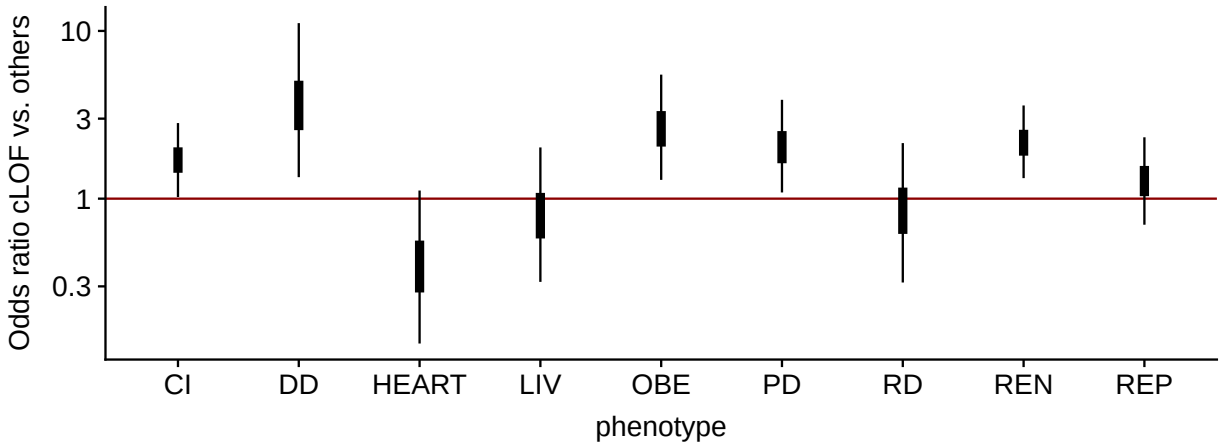


Posterior 95% (thin) and 50% (thick) credible intervals for odds ratio for a phenotype given mutation in the gene on horizontal axis against the gene on the vertical axis. Color indicates the widest central posterior credible interval that does not include one. We deliberately not make any strong cutoff at 95% excluding zero or similar as the tail probabilities have high variance (e.g., where there is less data, one extra positive case could plausibly move this quantity from say 93% to 96%). Odds ratio are shown on the log scale.

This plot integrates the information shown in the plots above, and some more.

Type of mutation - loss of function

What is the effect of whether the mutation certainly leads to the complete loss of function (LOF) of the protein? Since this information (will be furter reffered to simply as LOF) does not vary per gene, let us look at the corresponding odds ratio per phenotype.



For most phenotypes we (expectedly) see that in cLOF mutations the phenotype is more likely. For LIV, RD and REP, the data is not very conclusive. The surprising part is the high posterior probability assigned to cLOF mutations having less severe phenotype in HEART. This might nevertheless be due to biases in reporting or due to higher lethality of HEART phenotype in cLOF mutations (and thus making the patients not included in the dataset). It is however likely not a result of low sample size: there are 225 patients with the HEART phenotype reported, spread roughly equally between cLOF and other mutations.

Discrepancies between frequentist and Bayesian analysis

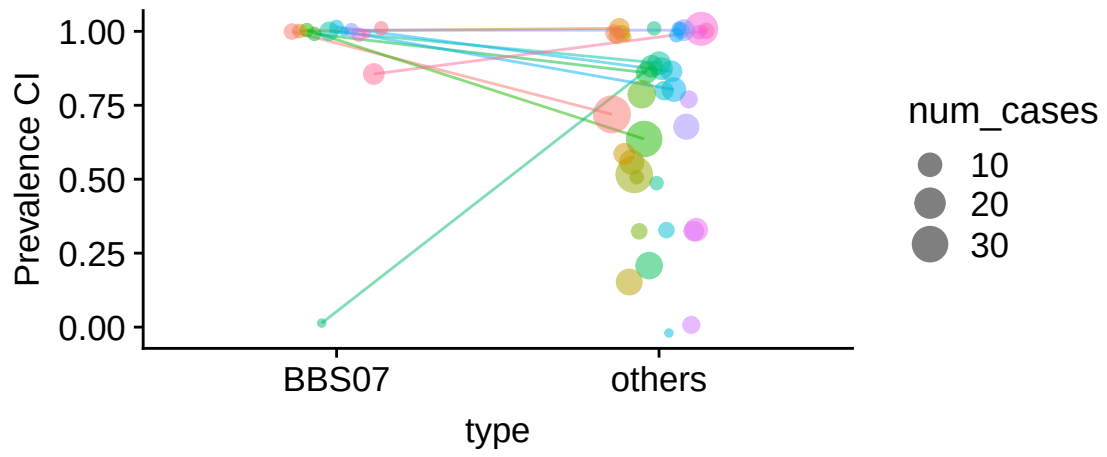
Here we try to understand why the frequentist and Bayesian analyses provided different results for some of the questions we asked. In all cases the answer is that there is a substantial between-study variability, which, when taken into account, prevents us from making firm conclusions about some of the comparisons that are significant in the frequentist analysis (which ignores between-study variability).

CI phenotype and BBS7

Looking at raw proportions in the whole dataset, BBS7 has the highest prevalence of the CI phenotype (and the difference was significant for the frequentist analysis):

gene	prevalence_CI	n
BBS07	0.9375000	32
BBS05	0.8400000	25
BBS02	0.7843137	51
BBS01	0.6464088	181
BBS08	0.6428571	14
BBS04	0.6333333	30
BBS09	0.6111111	18

The picture however gets complicated when we look at individual studies:

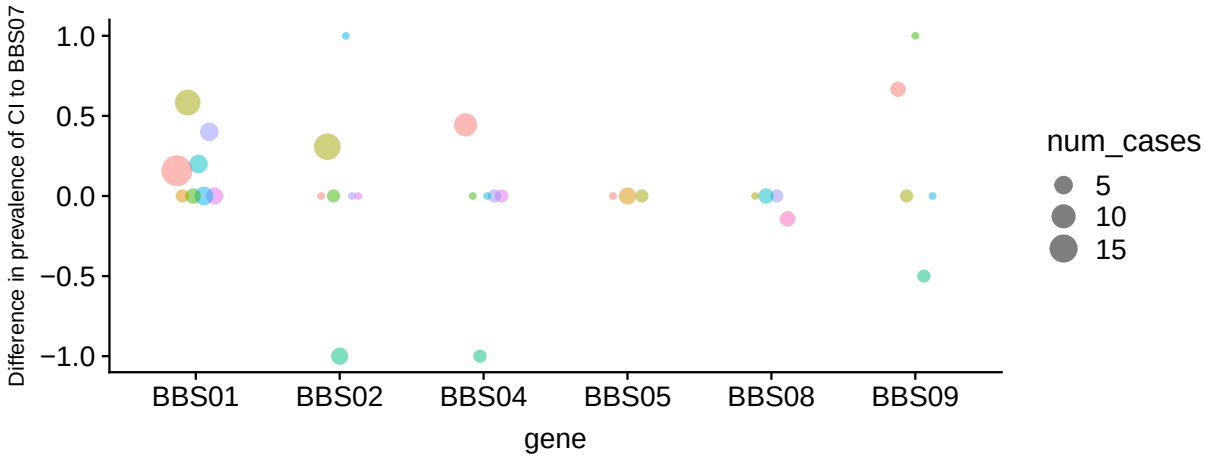


Here, each dot is a study with size indicating the number of cases. We show the prevalence for BBS7 and prevalence across the other mutations. Lines connect estimates from the same study. This shows us that while BBS7 indeed has high CI prevalence (the prevalence is 100% in most studies), there are some studies where the CI prevalence is lower than prevalence in other mutations:

source	prevalence_CI_BBS07	prevalence_CI_others	n_BBS07	n_others
Ullah et al. 2017	0.8571429	1.0000000	7	8
Fattahi et al. 2014	0.0000000	0.6363636	1	11

This includes the Ullah et al. study which has the largest number of BBS7 cases with reported CI phenotype in the whole dataset. So while there is some evidence for BBS7 having unusually high prevalence of CI, we cannot rule out that this is caused by between-study variability.

The above plot showed the total prevalence across all other mutations, we can also look at comparisons of prevalence between individual mutations:



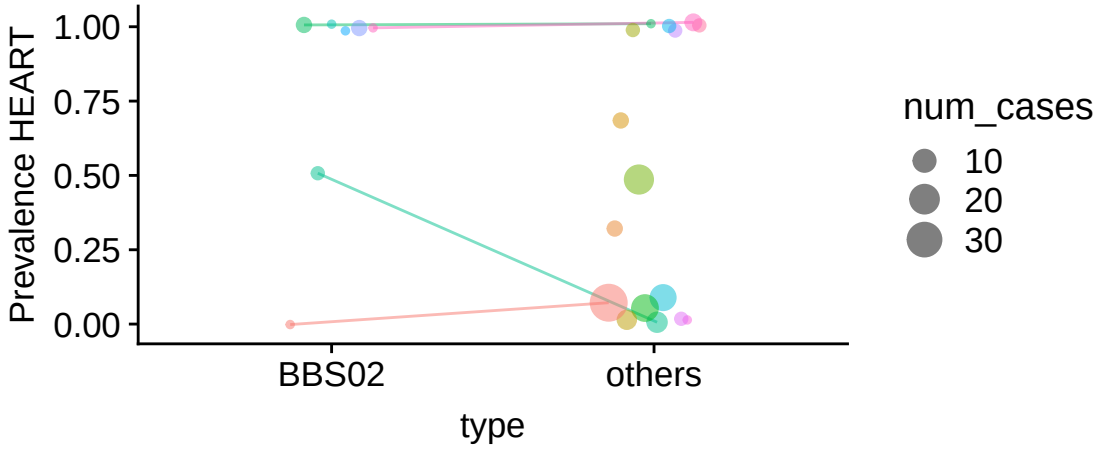
Once again, each dot is a study, the dot size corresponds to the number of patients with the mutation on x-axis and the y-axis shows the difference in prevalence against BBS7 (positive - BBS7 has higher prevalence). Once again, most studies show no or positive difference, but some studies show negative difference and for BBS8 there are actually more studies showing effect in the opposite direction. This means that the model can conclude neither that BBS7 has higher prevalence of CI than each other mutation individually nor (weaker) that it has higher CI prevalence than the average of other mutations.

HEART phenotype and BBS2

The case of BBS2 is similar to BBS7 - looking at proportions across the whole dataset, BBS2 has the highest prevalence of HEART (significant in the frequentist analysis), but the sample sizes are even smaller than in the above case:

gene	prevalence_HEART	n
BBS02	0.8333333	12
BBS04	0.3571429	14
BBS05	0.3333333	9
BBS07	0.3000000	10
BBS01	0.2337662	77
BBS09	0.1250000	8

Let's look at BBS2 compared to prevalence across all other phenotypes:



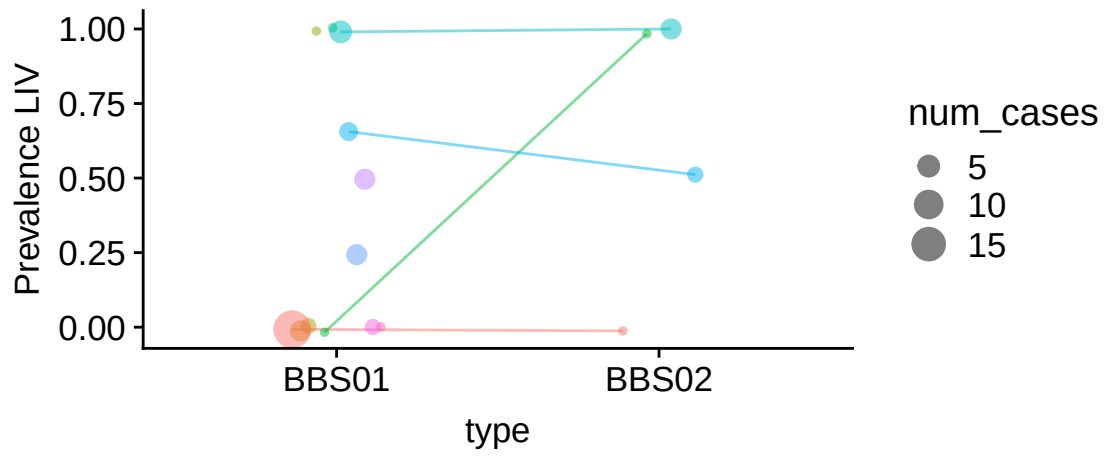
We see there is one study showing higher prevalence, one study showing lower prevalence and two showing equal prevalence. All the other studies either had only BBS2 or had no BBS2 and so cannot directly contribute to a conclusion, despite showing small prevalence of HEART. Once again, we cannot rule out the difference between BBS2 and others in HEART phenotype is simply due to between-study variability.

LIV BBS2 to BBS1 comparison

In the data BBS2 has higher LIV prevalence than BBS1 (again statistically significant, despite the small sample sizes):

gene	prevalence_LIV	num_cases
BBS01	0.2553191	47
BBS02	0.7500000	8

Let's plot the between-study variability:



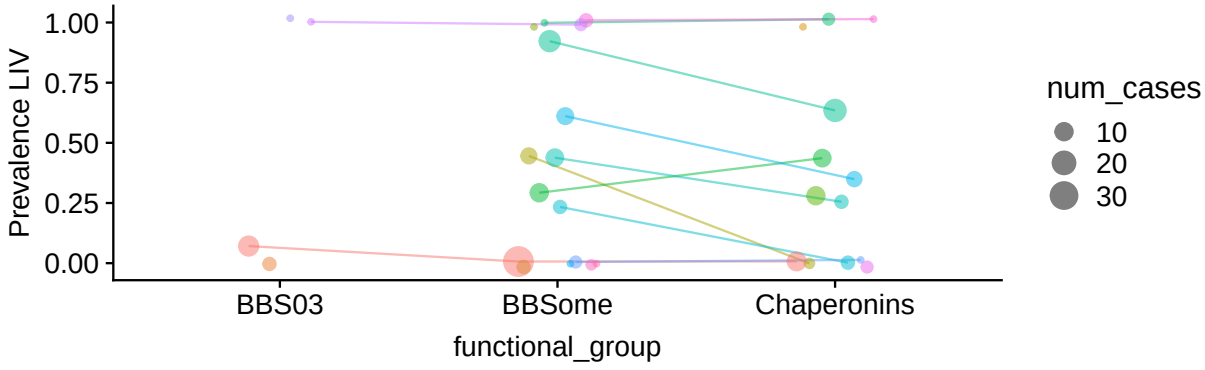
There is one study that shows increase for BBS2 and one that shows decrease for BBS2. And the one showing decrease is actually the study with larger sample size. The difference seems to be driven by studies that have BBS1 patients and no or very few BBS2 patients and can thus be attributed to between-study variability.

LIV phenotype and functional groups

For the LIV phenotype the frequentist analysis disagrees with the Bayesian analysis presented here. In particular, the frequentist analysis shows BBS3 as the least likely to result in LIV and chaperonins as the most likely, while the analysis shown here reports exactly opposite trend. First, let's look at the aggregate frequencies of LIV phenotype:

functional_group	proportion_LIV
BBS03	0.1578947
BBSome	0.3577982
Chaperonins	0.3611111
Others	0.2073171

Indeed, this supports the frequentist conclusions (as this is actually what those are based on). However, this aggregate look ignores the between-study variability. So let's look at what the individual studies show:



In the plot above, each dot is the proportion of patients with LIV given a mutation in a gene from one of the functional groups in a single study. Lines connect values that are from the same study (note that only one study had mutations from all groups and many had mutations in just one group) - those also have the same color. Size of the points represents the number of patients. We added some jitter to let us differentiate the points - notably all the points near zero and one are actually exactly zero and one.

This plot tells a different story: most individual studies, especially those with larger number of cases, show increase in LIV phenotype for BBSome against Chaperonins. There is only one larger study including both BBS3 and BBSome and it has more LIV positive cases for BBS3. The overall opposite trend is driven by a) the only larger studies reporting BBS3 having unusually low proportion of LIV in general and b) studies reporting mutations only for BBS3 or only for BBSome having unusually low proportion of LIV.

However, it seems that the evidence in this direction is not very compelling.

Part 2: Alternative Models & Model Selection

Note: For historical reasons the feature of "certain loss of function" (cLOF) as discussed in the data is called just "lof" in most of the analysis code. This part will thus use "lof" and "cLOF" interchangeably.

Model descriptions

All models are Bayesian varying intercept logistic regressions using the `brms` package. Part 1 of this supplement includes both accessible and complete mathematical description of the model we chose for the main analysis, which will not be repeated here. Generally, all of the terms in the models are varying intercepts, i.e. the model partially pool the estimates for individual groups (genes, sources, ...) towards population mean to achieve more robust inference.

Base models

Base models are those that work with (a subset of) the original dataset, without any imputation. The models are defined in file `models.R`. They differ in the model formula, subsets of the dataset they use and priors for model coefficients. The syntax for formulas in `brms` is described in `brms` manual and will not be explained here. The model may work on the filtered dataset - `lof` means filtered for only the mutations with certain LOF, `family`, `age`, `sex` and `age_sex` corresponds to filtering for patients with reported family, age, sex or both age and sex. Likewise `ethnic_group` and `family` filter only the data that has those values reported. The list of base models follows:

```
## gene_only :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene)
##  data filter:  none
##
## gene_only_filtered_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene)
##  data filter:  lof
##
## gene_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) +
##  phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_lof_per_gene :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype:loss_of_function_certain) || gene) +
##  phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_family :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || family_id)
##  data filter:  family
##
## gene_ethnic_group :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnic_group)
##  data filter:  ethnic_group
##
## gene_ethnicity :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnicity)
```

```
##  data filter:  ethnicity
##
## gene_ethnic_group_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnic_group) + phenotype:loss_of_function_certain
##  data filter:  ethnic_group
##
## gene_ethnicity_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnicity) + phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_ethnic_group_lof_per_gene :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnic_group) + ((0 + phenotype:loss_of_function_certain) ||
##  gene) + phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_ethnicity_lof_per_gene :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || ethnicity) + phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_family_filtered_age_sex_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || family_id) + ((0 + phenotype) || sex) + (0 + age_std_for_model
##  || phenotype)
##  data filter:  family age sex lof
##
## gene_source :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  none
##
## gene_source_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_source_family :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || family_id)
##  data filter:  family
##
## gene_source_lof_family :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || family_id) +
##  phenotype:loss_of_function_certain
##  data filter:  family
##
## gene_source_lof_ethnic_group :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || ethnic_group) +
##  phenotype:loss_of_function_certain
```

```
##  data filter:  ethnic_group
##
## gene_source_lof_wide :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + phenotype:loss_of_function_certain
##  data filter:  none  Note: special priors
##
## gene_source_filtered_lof :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  lof
##
## gene_source_filtered_lof_wide :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  lof   Note: special priors
##
## gene_source_lof_per_gene :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype:loss_of_function_certain) || gene) +
##  phenotype:loss_of_function_certain
##  data filter:  none
##
## gene_source_genecor :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + gene) | phenotype) + ((0 +
##  phenotype) || source)
##  data filter:  none
##
## gene_source_nocor :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) || gene) + ((0 +
##  phenotype) || source)
##  data filter:  none
##
## gene_source_narrow :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  none  Note: special priors
##
## gene_source_very_narrow :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  none  Note: special priors
##
## gene_source_wide :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source)
##  data filter:  none  Note: special priors
##
## gene_source_filtered_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex)
##  data filter:  sex
##
## gene_source_filtered_age :
```

```
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + (0 + age_std_for_model || phenotype)
##  data filter:  age
##
## gene_filtered_age_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex) + (0 + age_std_for_model ||
##  phenotype)
##  data filter:  age sex
##
## gene_lof_filtered_age_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex) + (0 + age_std_for_model ||
##  phenotype) + phenotype:loss_of_function_certain
##  data filter:  age sex
##
## gene_lof_per_gene_filtered_age_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex) + (0 + age_std_for_model ||
##  phenotype) + ((0 + phenotype:loss_of_function_certain) || gene) +
##  phenotype:loss_of_function_certain
##  data filter:  age sex
```

The default priors are $N(0, 2)$ for all model coefficients (half-normal for standard deviations). Very narrow, narrow and wide put $N(0, 0.1)$, $N(0, 1)$ and $N(0, 5)$ respectively for the sd for `gene`. Since `family_id` is a very fine-grained predictor, its sd is given a $N(0, 1)$ prior.

Imputation with the mice package

Including age or sex in the base models is problematic as it involves tossing out 47% or 45% of data, respectively. This results in wide posterior intervals and weak inferences. To try to ameliorate this we also tested running models on datasets with age and sex imputed, using multiple imputation via the `mice` package. We assume that both age and sex can be related to the functional group of the mutation and to each other. Involving further relations (e.g., individual genes) led to warnings from the `mice` package and we thus didn't use those.

The imputed models differ in the formulas used, but all use the default priors and do not filter the dataset in any way.

```
## gene_imputed_age_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex) + (0 + age_std_for_model ||
##  phenotype)
##
## gene_source_imputed_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + ((0 + phenotype) || sex)
##
## gene_source_imputed_age :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + (0 + age_std_for_model || phenotype)
##
## gene_source_imputed_age_sex :
##  formula: phenotype_value ~ (1 || phenotype) + ((0 + phenotype) | gene) + ((0 +
##  phenotype) || source) + (0 + age_std_for_model || phenotype)
```

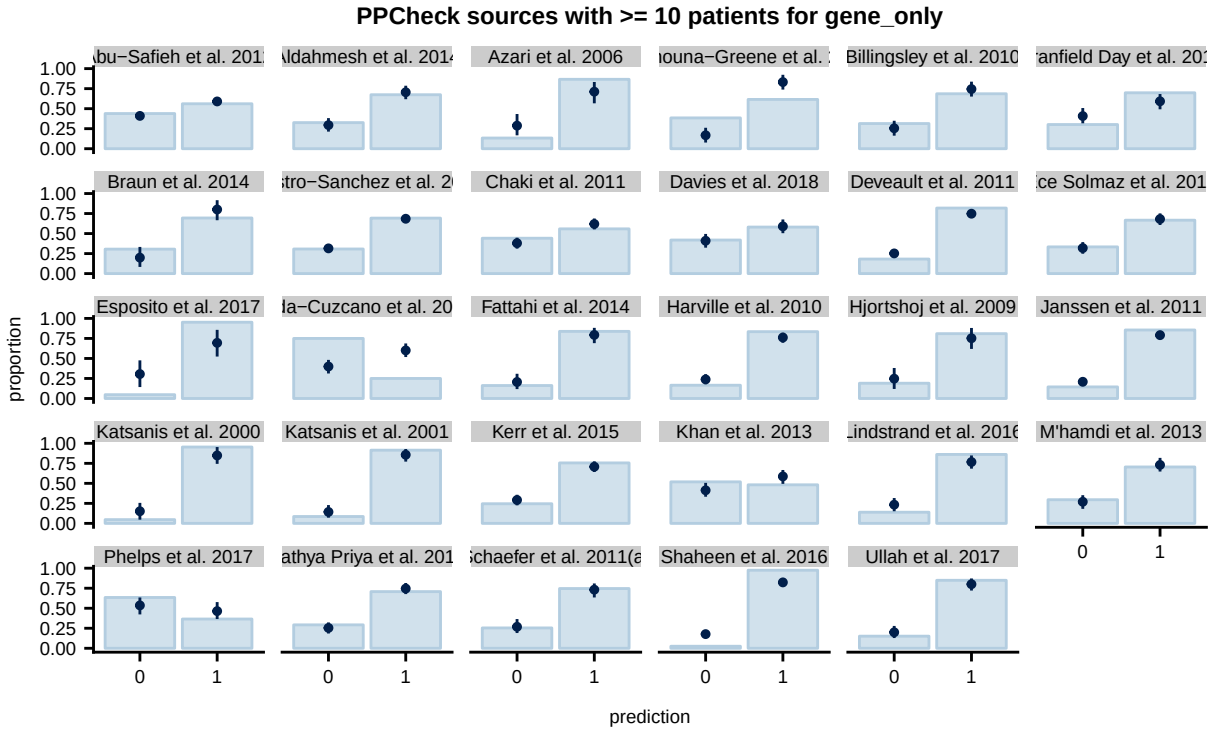
Choosing the model for main analysis

In general we use *posterior predictive checks* (PPCheck) to asses model fit. PPCheck is performed by predicting posterior distribution of possible outcomes implied by the fitted model. This distribution can then be compared to what is actually observed in the data and discrepancies can be noted and used to guide model expansion/selection. In our case, we focus on the prevalence of positive phenotypes across various subdivisions of the data. Of prime interests are subdivisions *not* taken into account by a model - if the model explains groupings that were not included, it is a sign that it works well. If the model consistently misestimates groups it is not aware of, it is an indication that such a group should be involved. We use the `bayesplot` package to perform PPChecks. See Gabry et al. 2018, *Visualisation in Bayesian workflow* for a more thorough discussion of PPChecks. The following discussion is mostly informal and qualitative as we try to balance model fit, model complexity and other considerations. While the choices we make are partly subjective, Part 3 shows that our conclusions are largely robust to defensible variations in model specification.

In most models, we assume phenotype correlations because the data were selected for containing at least two phenotypes and diagnosis criteria is based on having multiple phenotypes. Both of these processes could have introduced correlations. However, the fitted correlations are not conclusive and do not increase the explanatory power of the model (models with different correlation structures are also included).

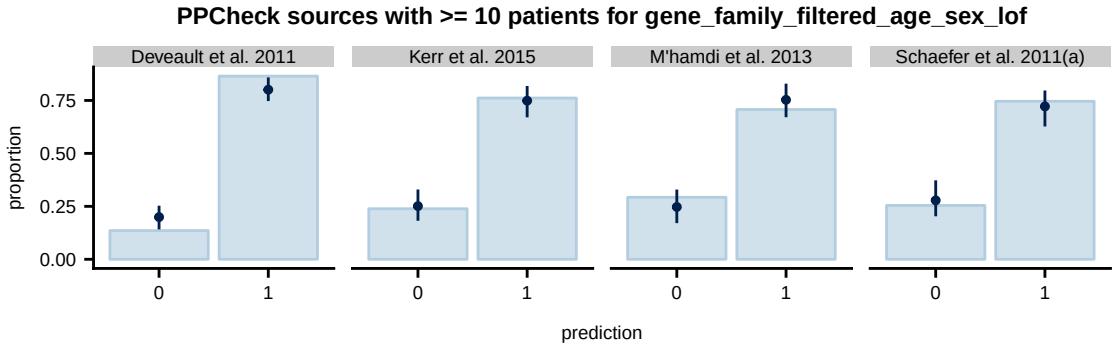
Between-study variability needs to be included

Modeling between-study variability is simply a good practice for any meta-analysis, but PPChecks can convince us that models ignoring it do not fit the data well. Let's start by looking at overall prevalence for studies with at least 10 patients and how the most basic model (`gene_only`, taking only the gene into account) fares:



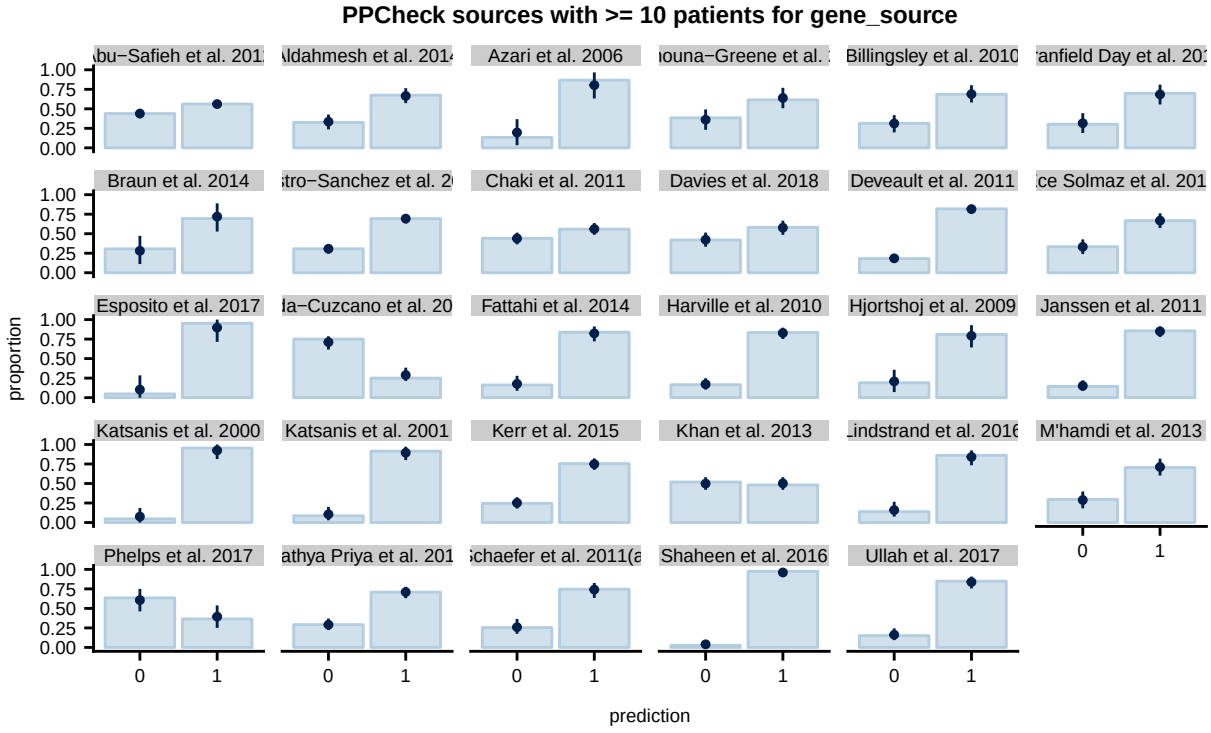
Here, the bars represent actual prevalence in the data, the dots the posterior mean for those subgroups and the lines posterior 95% credible interval. Note that we clump all phenotypes together for simplicity. We see that the model is overly certain in its predictions and the posterior credible intervals frequently miss the actual counts in a study. Almost half of sources have predictions not very consistent with data.

We try to ameliorate this with a complex model without source, but including family age and sex (filtered) and filtered only for loss of function mutations. This means we use very little data and a large number of predictors.



As seen above, even a complex model with little data struggles to fit the Deveault 2011 study. The problems are only larger for less complex models (e.g. using ethnic group instead of family or ignoring some of the predictors, not shown).

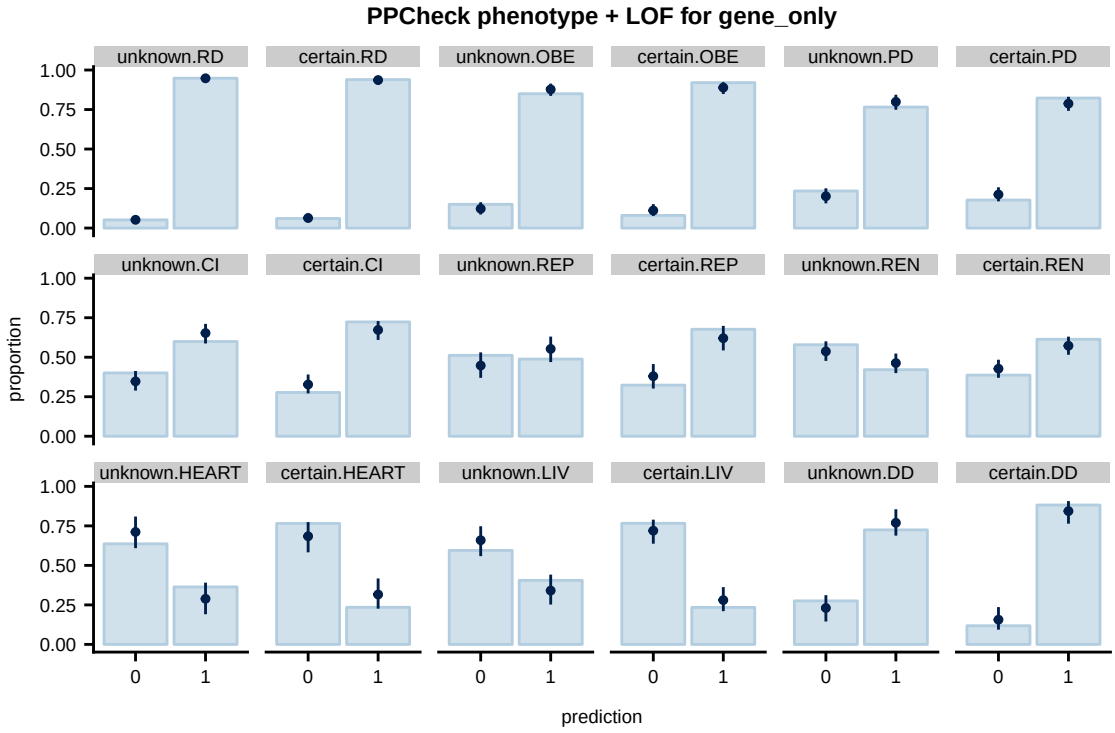
Unsurprisingly, including source explicitly alleviates all of the problems with source, even when other factors are not included:



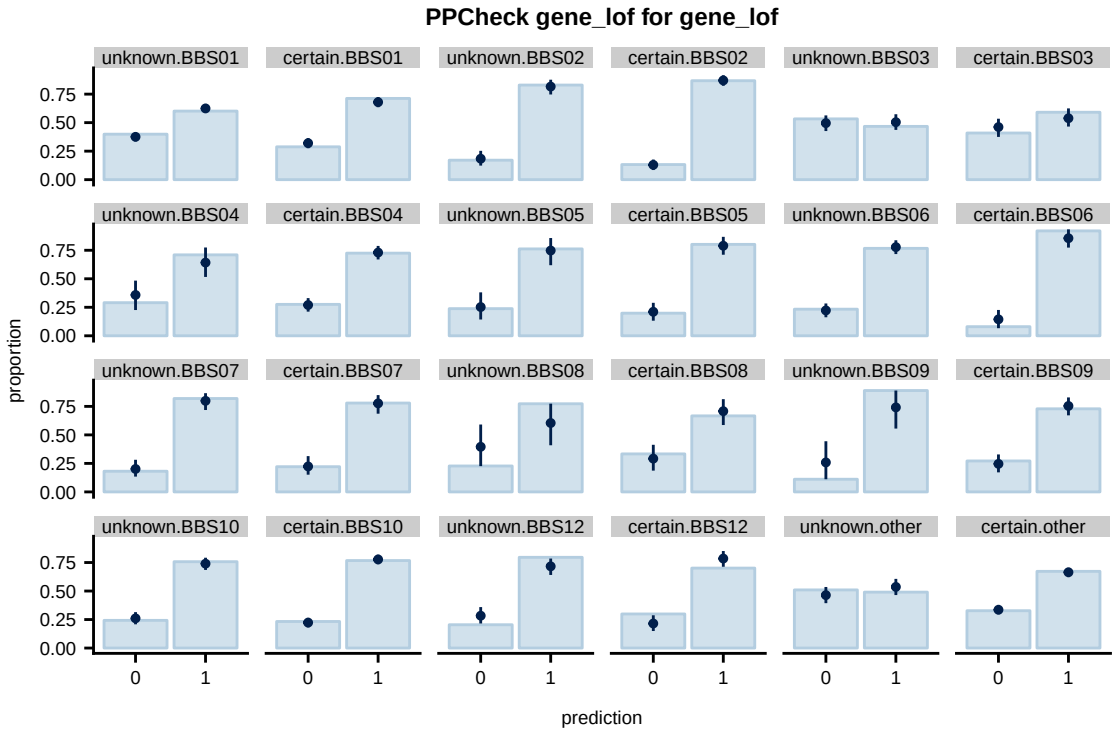
The conclusion is that even the most complex models not including source have problems. Further, we have a good reason to believe there is between-study variability even before looking at the data, as the methodologies for diagnosis are not consistent. Including source is therefore necessary. In light of this, we consider all models not including source as "Problematic fits" to the data.

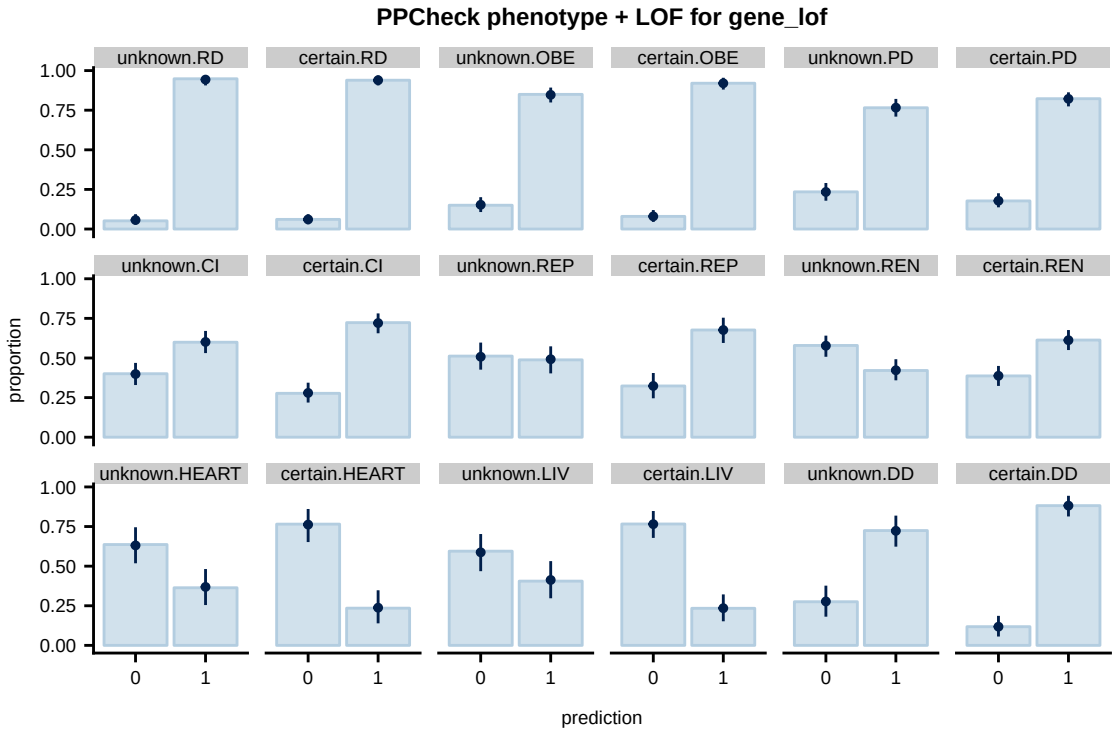
Loss-of-function differences are of a relatively minor importance

Using the simplest model, we see that cLOF differences are slightly problematic (at the edge of model predictions) for many phenotypes, though no gross error is apparent:



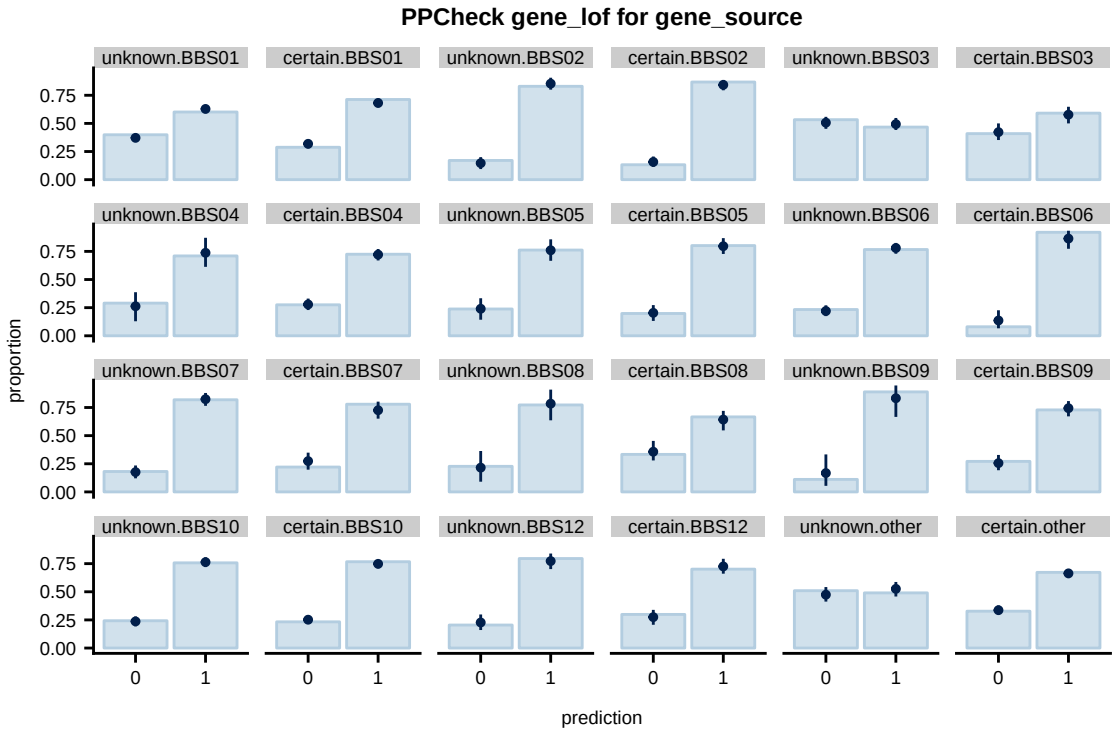
The problems almost disappear when adding a per-phenotype cLOF term to the model (i.e. for a given phenotype the effect of cLOF is assumed equal across all mutations):

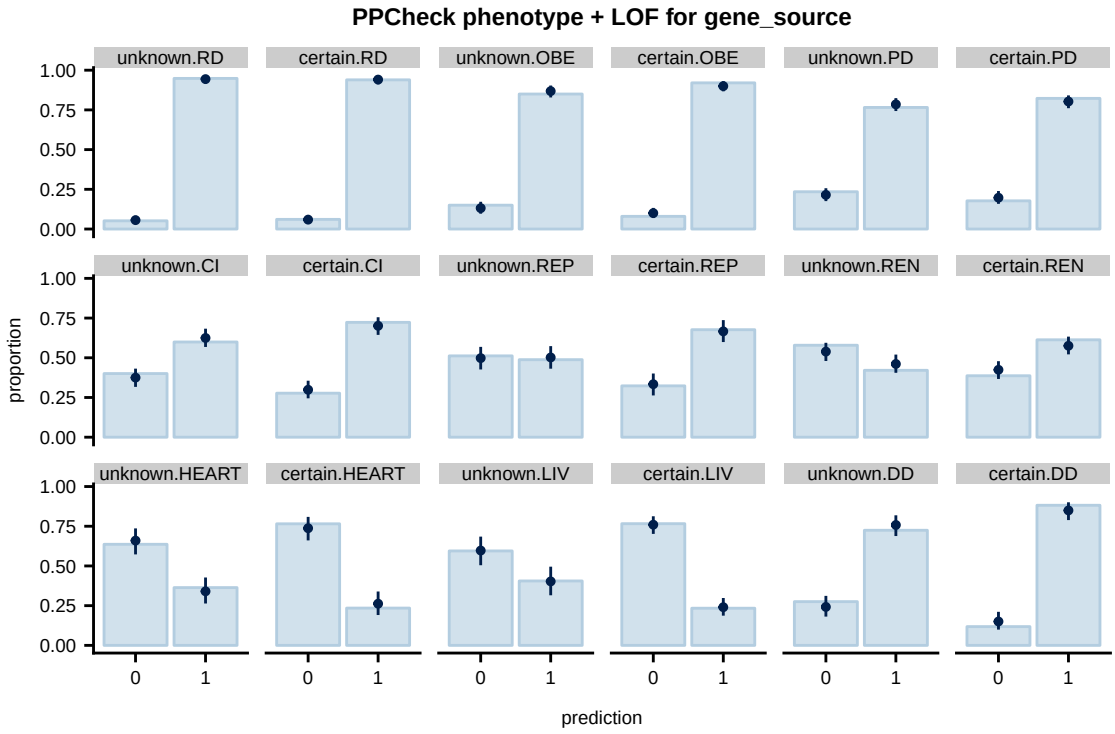




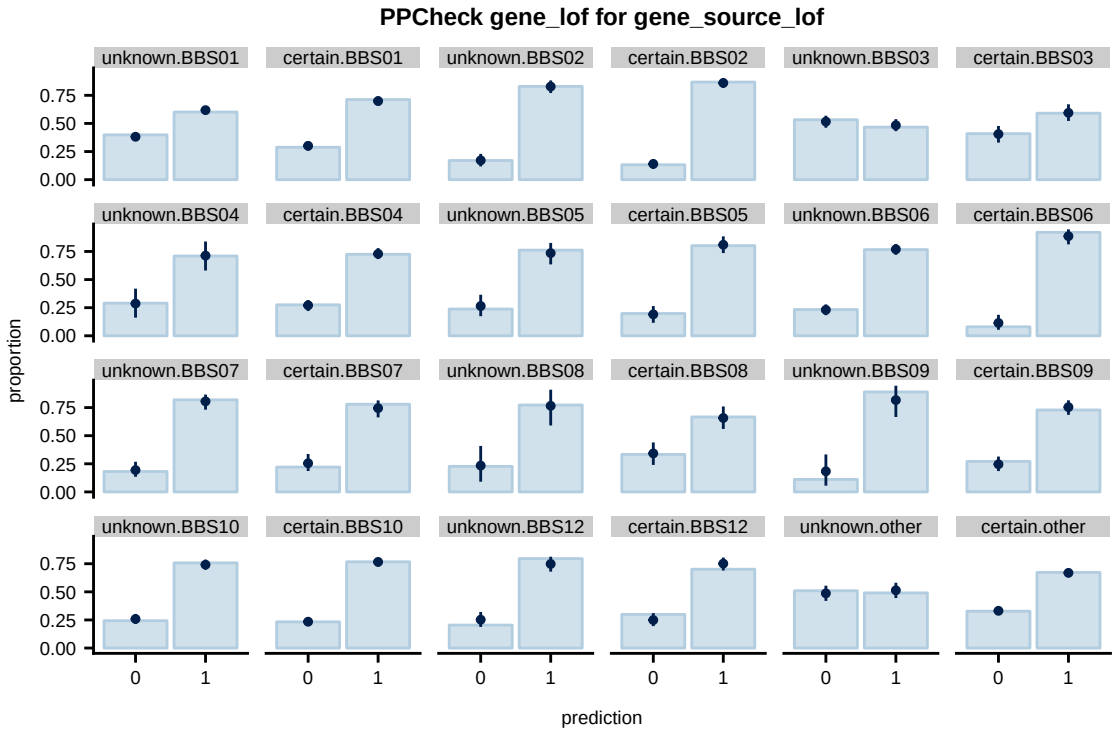
Although the above plot shows that for some genes (most notably BBS8, BBS9 and BBS12) the model has problem explaining differences between cLOF and other mutations.

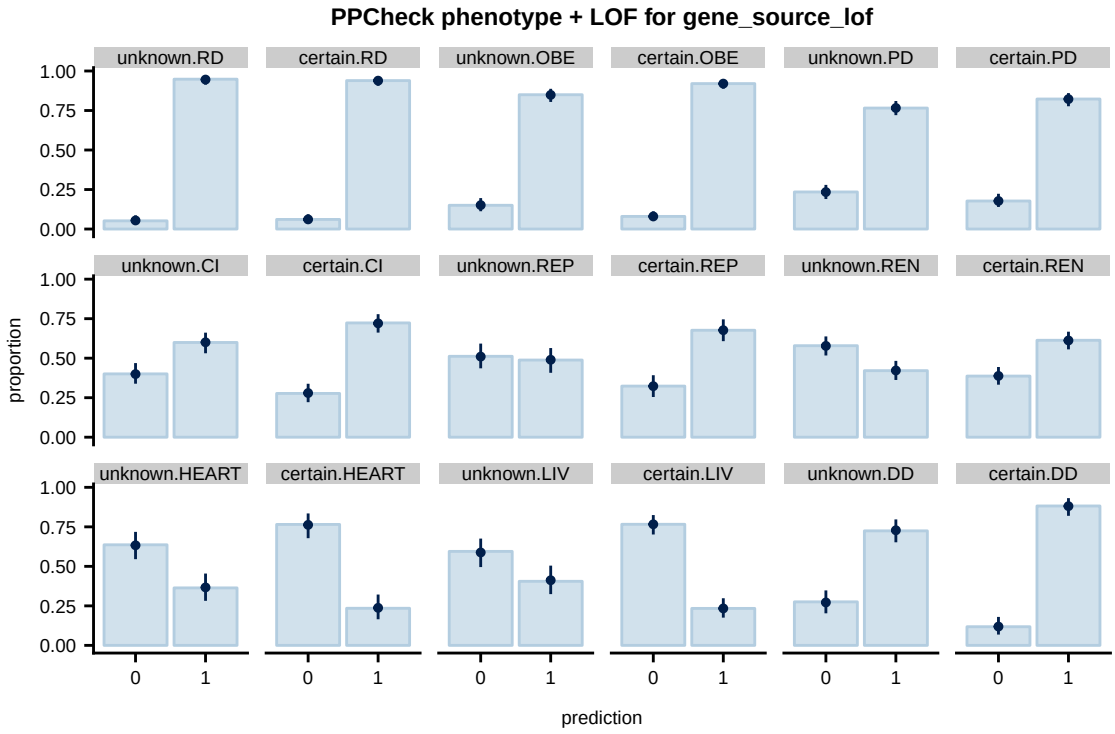
Adding source as a covariate (but ignoring cLOF) ameliorates most of the problems with cLOF looking at both phenotypes and genes:



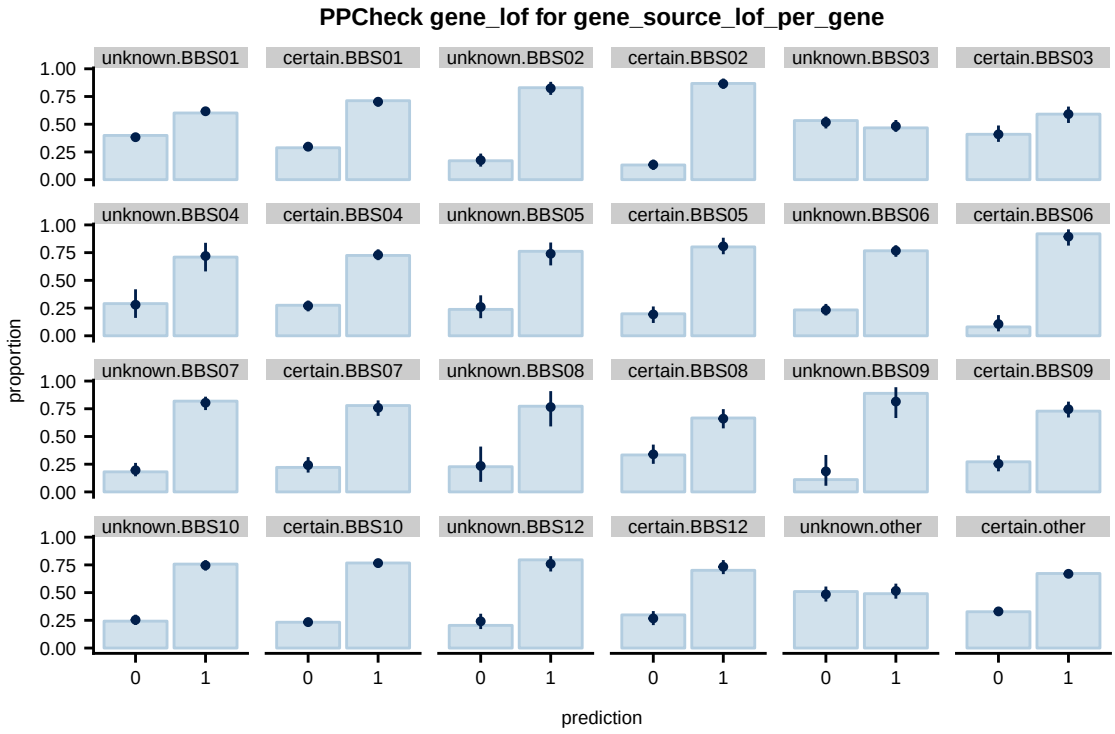


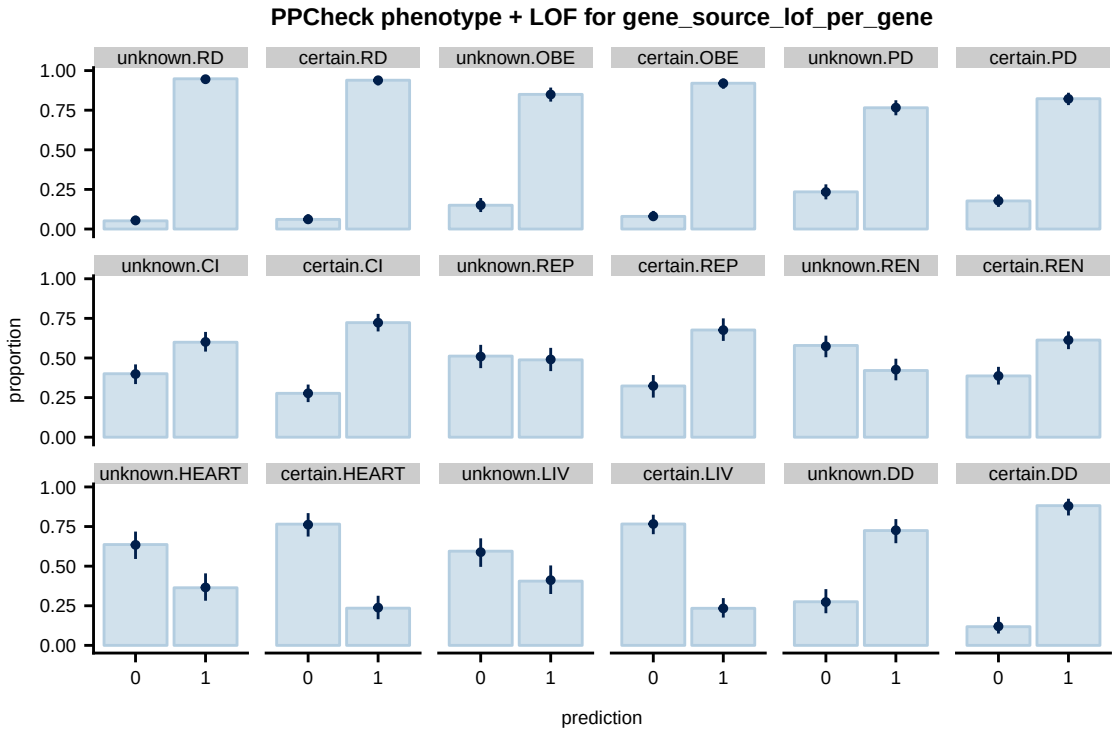
Finally, the problem is mitigated even further when both source and cLOF are included (even when cLOF effect is not allowed to vary with gene).





Allowing the cLOF coefficient to differ per gene does not bring noticeable improvements to fit:

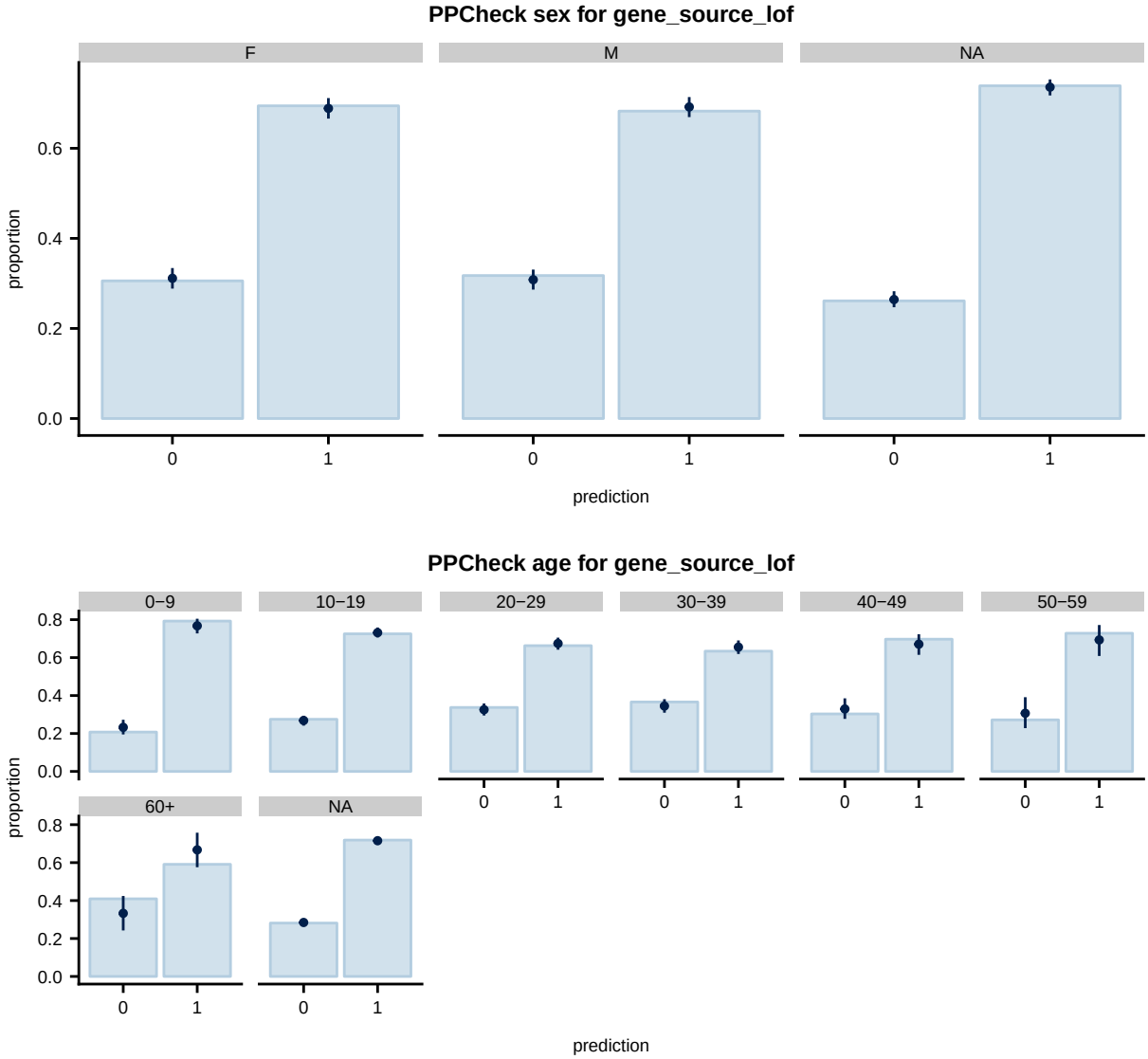




The conclusion is that: a) between source-variability is important as it is able to explain a large portion of cLOF differences even when cLOF is not accounted for b) there is an improvement in including the effect of cLOF per phenotype but further improvement is not observed when the cLOF effect is allowed to vary per gene. Since cLOF is easy to include and does not make the model much more complex, effect of cLOF per phenotype should be considered.

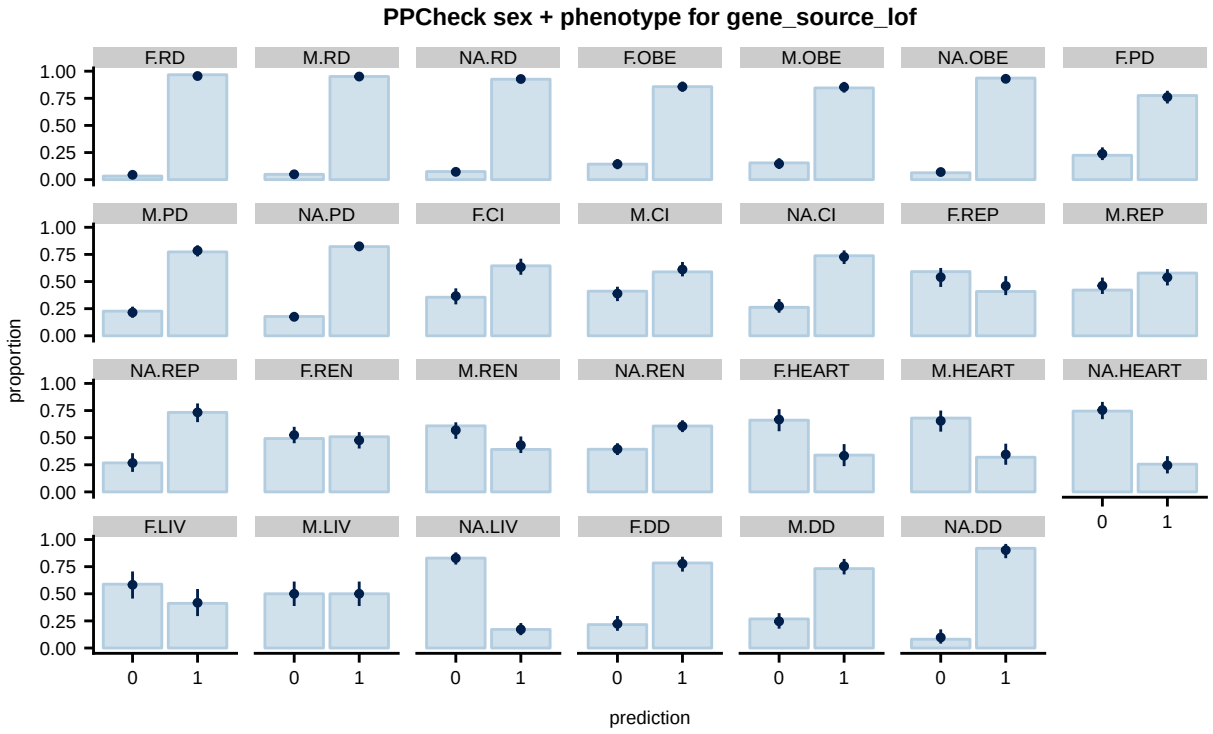
Between-study variability mostly explains age and sex differences

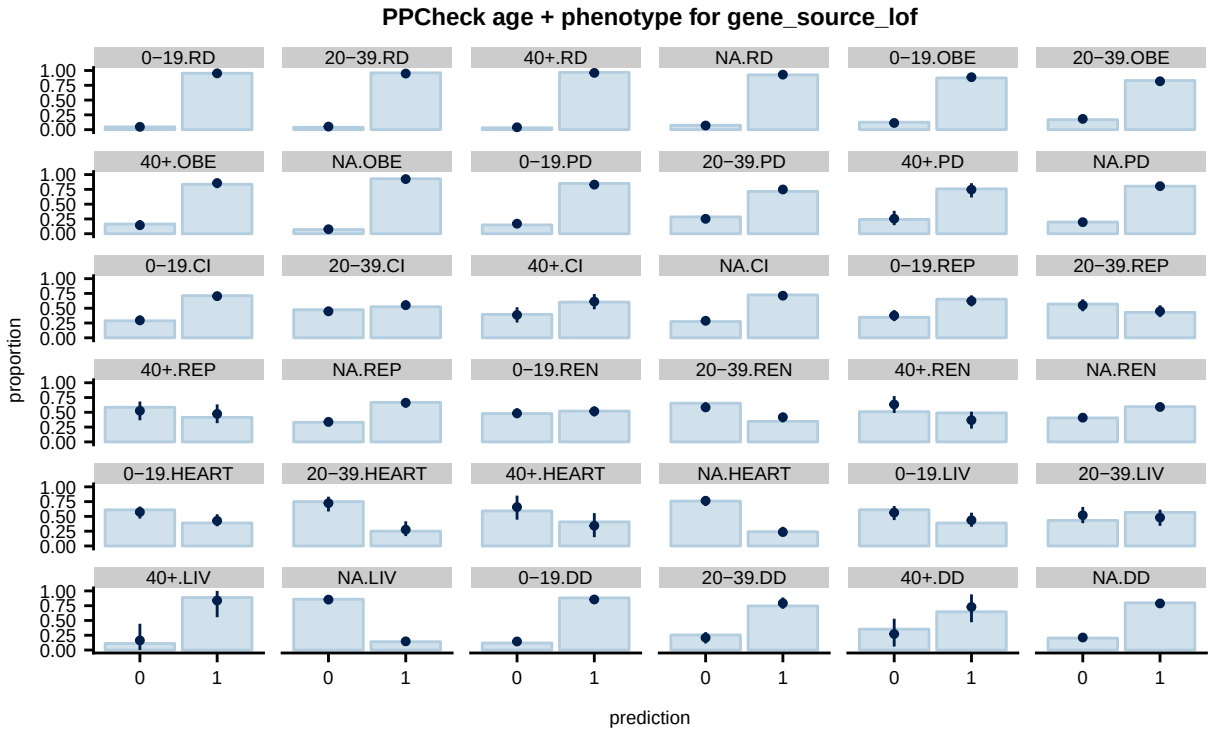
Age and sex differences do not necessarily need to be included, as they are sufficiently well explained by the `gene_source_lof` model. First let us look at overall prevalence by sex and age group:



The `gene_source_lof` model does well, although for some groups the data are on the borders of predicted 95% intervals (e.g. the 0-9 and 60+ groups). Note that missing data are treated as a separate age category.

We can also look at age and sex by individual phenotypes - we collapse some of the age groups to make the age + phenotype plot readable:

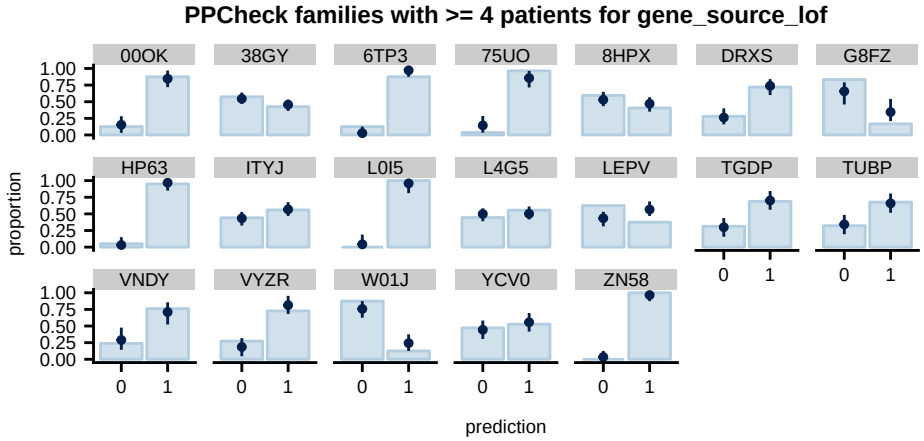


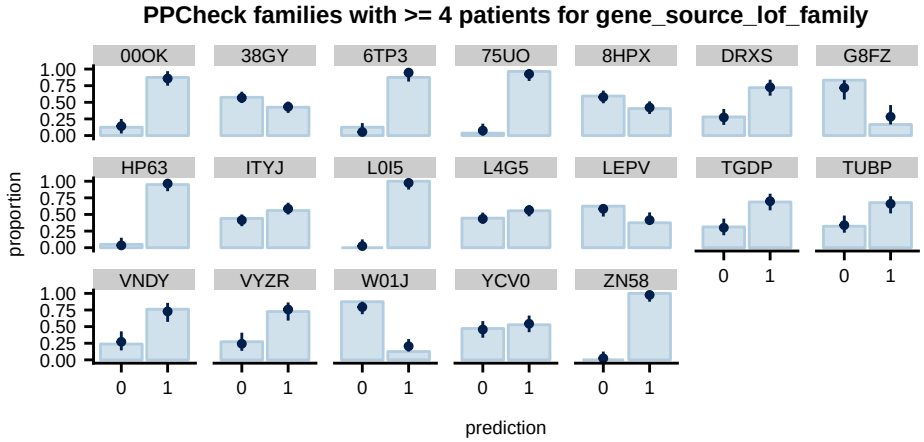


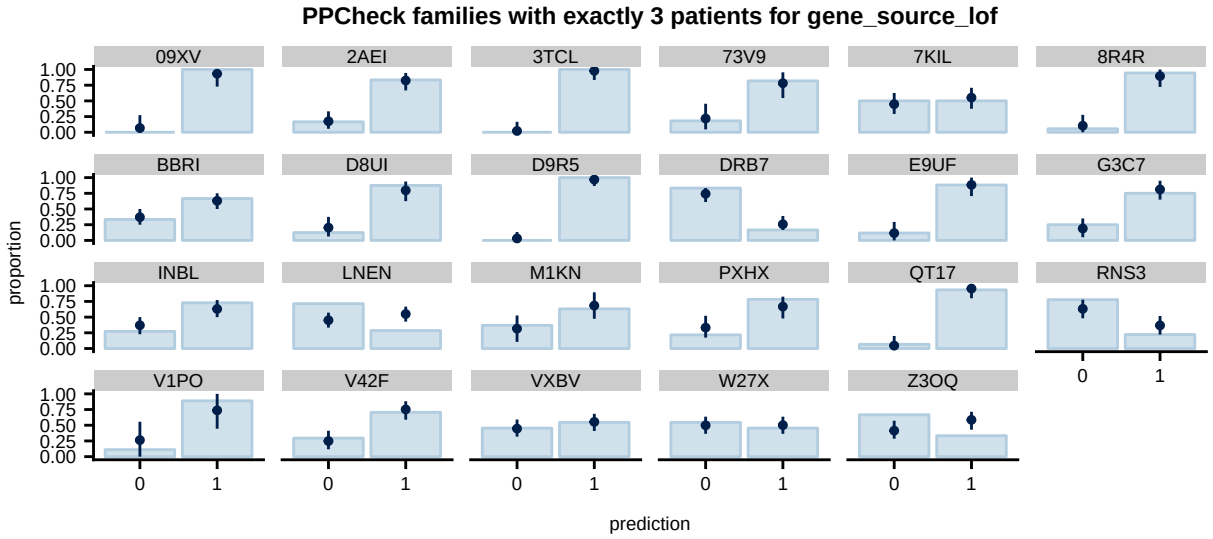
There are problems (e.g., age "40+" for REN phenotype), but no category is completely outside the predicted 95% interval. We chose not to include age and sex in the main model, because they are difficult to handle well due to high missingness - requiring either filtering or imputation, while they are mostly explained by the `gene_source_lof` model.

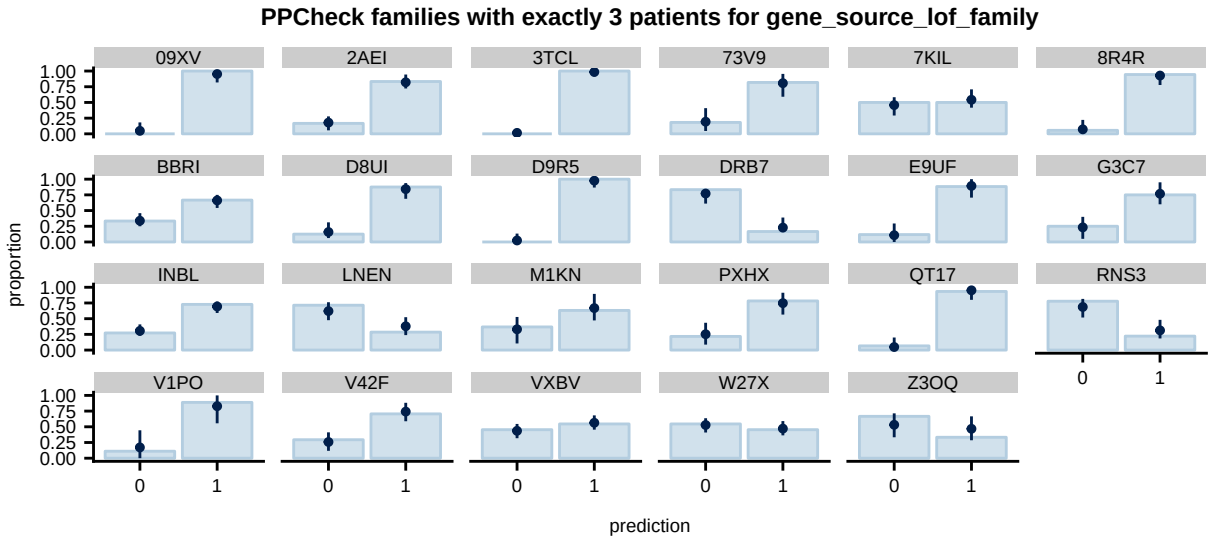
Family information is of relatively minor influence

We see the `gene_source_lof` model already accounts for a huge portion of the variability in prevalence between families, although including family in the model is definitely an improvement for some families.







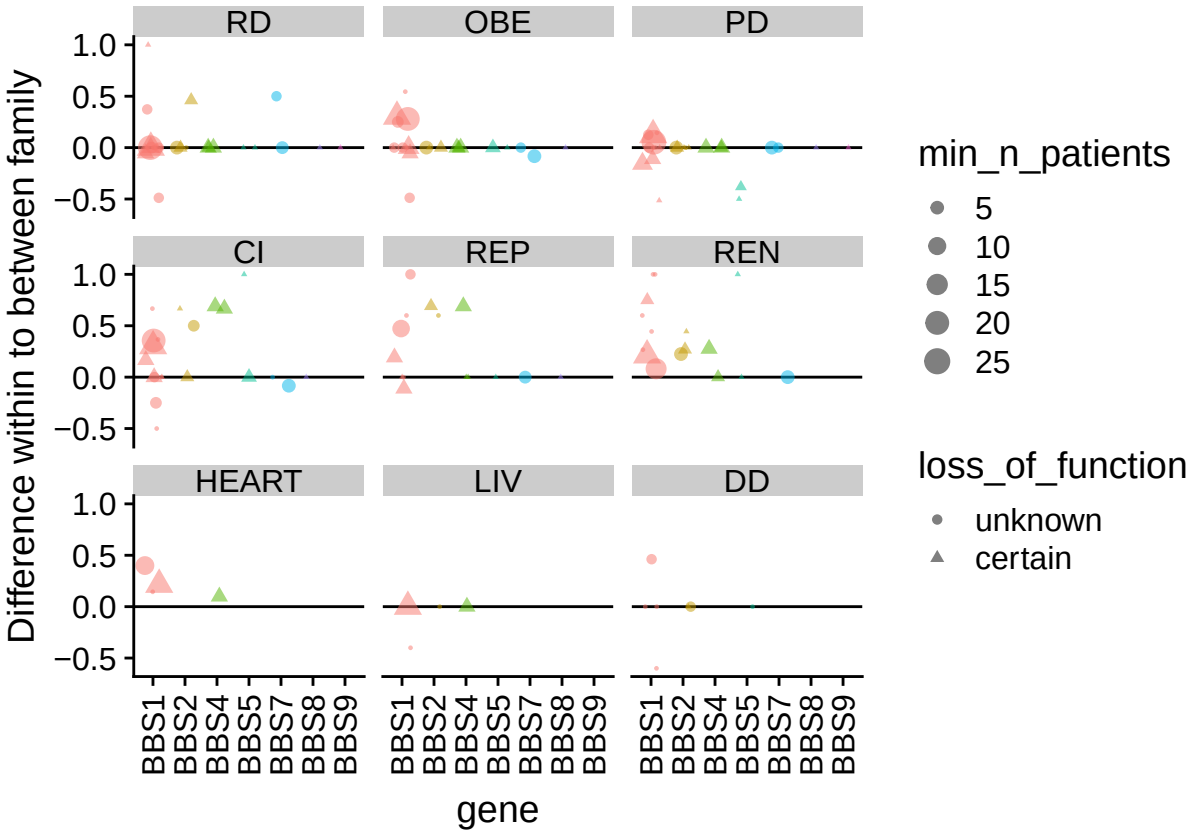


There are few families where the `gene_source_lof` model is fitting poorly (the LEPV and LNEN families) and a small number of those where the actual data is on the border of the predicted 95% interval.

We however consider the improvements due to including family to be small, while adding family greatly increases the uncertainty of the model, as it is a very fine grained predictor. We therefore chose not to include family in the main model. As discussed in Part 3, this does not have notable impact on our conclusions.

Understanding within- and between-family variability

We can also look for some signature of family structure in the data directly:



Here, we take all pairs of patients that are from the same family and all pairs of patients that are from the same source but have no family relationship. The pairs need to have mutation in the same gene and the same cLOF. Over each group of pairs we compute the proportion of pairs that have the same phenotype and then subtract this average for between family pairs from the average of within family pairs. I.e. the higher the number, the more are families with the same mutation homogenous in their phenotypes compared to unrelated individuals. Each point represents one study, the point size represents the size of the smaller group (family or unrelated) in the study.

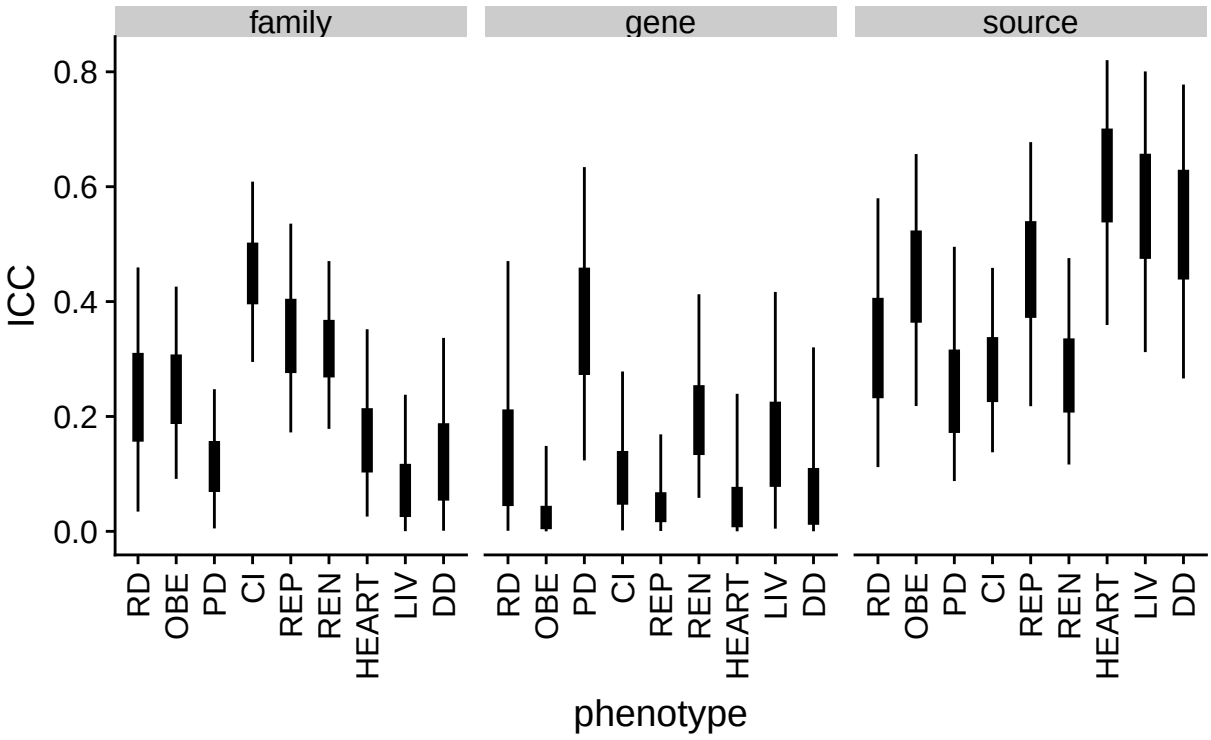
We see that for most phenotypes the results are mostly positive, so family structure (and hence specific type of mutation and/or genetic background) plays some role above just knowing cLOF and the gene where the mutation occurs.

To measure the amount of variability explained by family structure, we can compute the intraclass correlation (ICC) of family for each phenotype following Nakagawa & Schielzeth 2010, 'Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists'. For simplicity, we will use the `gene_source_lof_family` model, i.e. adding family on top of the model chosen for the main analysis. In this setting the ICC can be computed as:

$$ICC = \frac{\sigma^2_{family}}{\sigma^2_{family} + \sigma^2_{source} + \sigma^2_{gene} + \sigma^2_{cLOF} + \frac{1}{3}\pi^2}$$

$$\sigma^2_{cLOF} = Var(\beta_{cLOF} X_{cLOF})$$

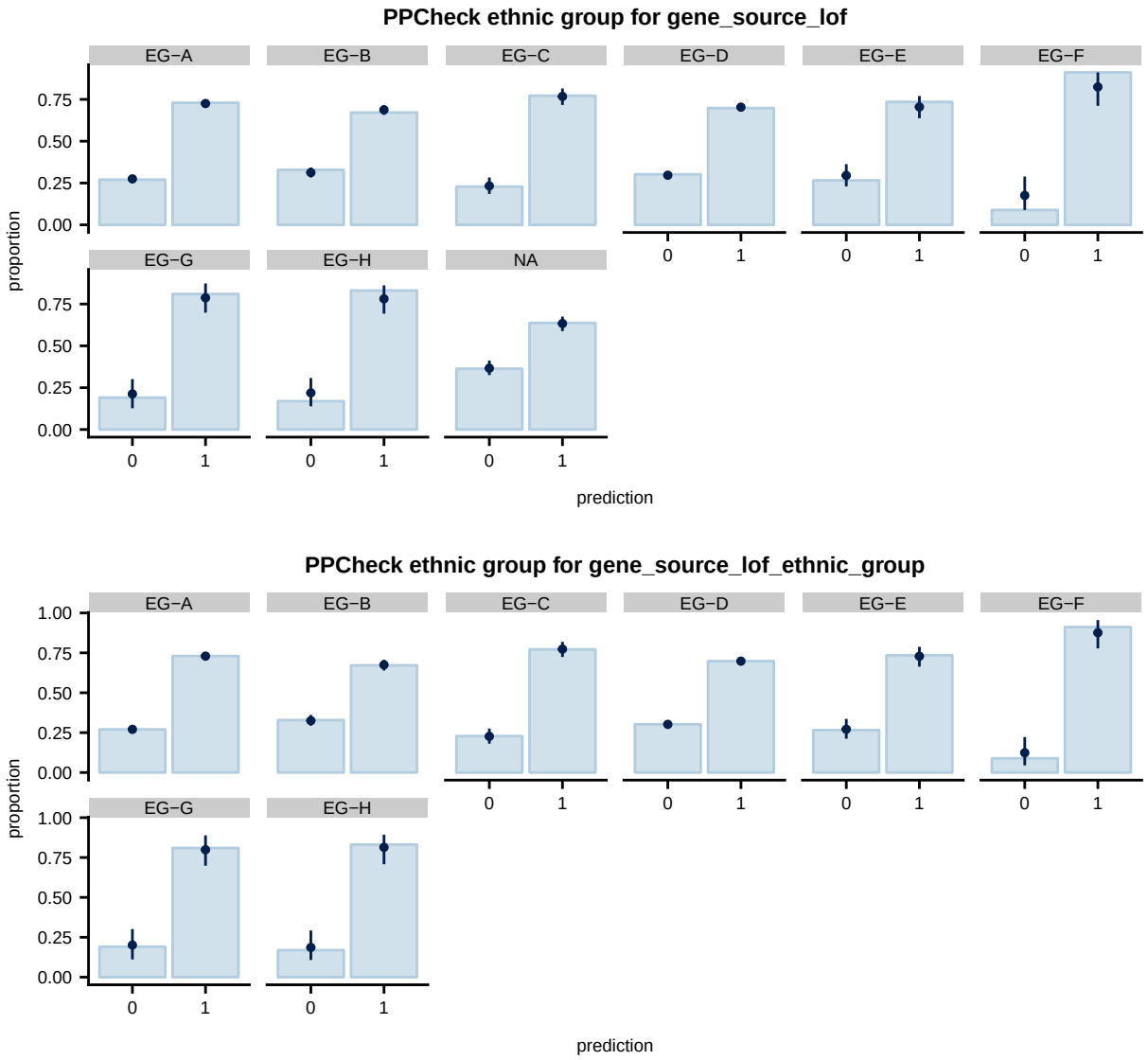
Where σ_{family}, σ_{source} and σ_{gene} are the variances of the respective random effects. Corresponding formulas can be devised for other model terms. The ICC can be very roughly interpreted as the proportion of total variance attributable to the grouping factor (family in this case) on the latent sacale. The estimates of ICC for the main covariates (family, source and gene) for individual phenotypes are:



Thin lines are 95% credible intervals, thick are 50% credible intervals. Note that for family, the ICC roughly corresponds to the plot above, where CI, REP and REN have the most pronounced skew towards family structure playing a notable role. However, for the other phenotypes, very low values of the ICC are consistent with the data. We also see that the ICC for family is likely smaller than for source and likely not much larger than for the gene carrying the mutation.

Due to notable between-source variability, it is hard to make strong conclusions about family structure - it is still quite possible that there is important influence of family for all phenotypes, but it is masked by the between-study variability. We also know that the `gene_source_lof_family` model does not fit much better than `gene_source_lof` and so it is hard to put strong emphasis on those results.

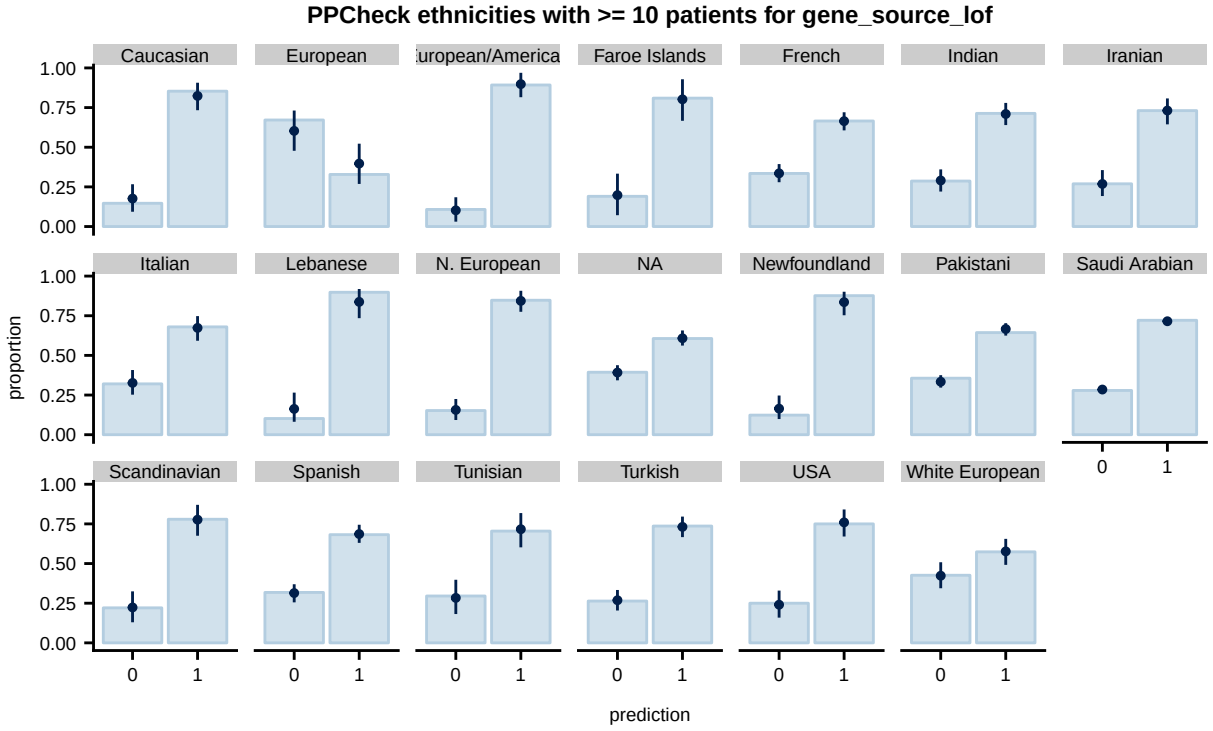
Ethnicity and ethnic groups

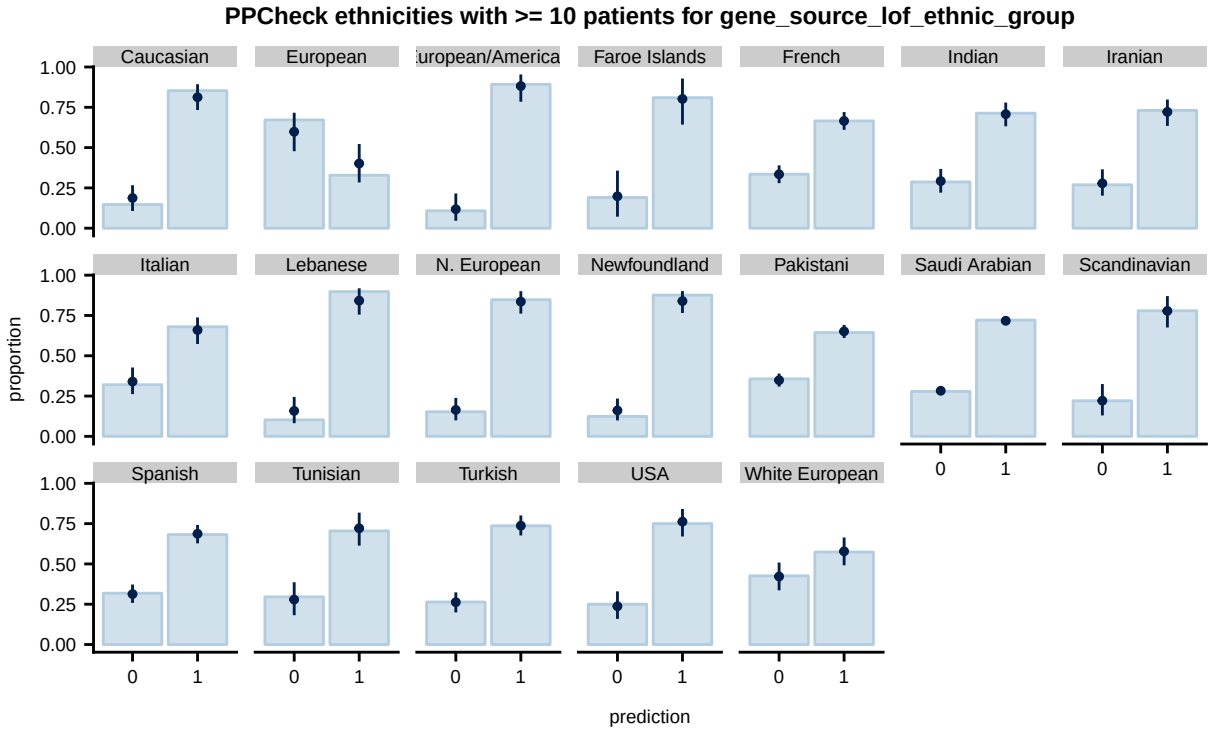


We see that `gene_source_lof` does a good job of fitting most ethnic groups, except for F and H where the observed data are close to the borders of the predicted 95% interval. Fit in those groups is improved upon by including ethnic group in the model. However, the misestimated ethnic groups are exactly those with the fewest patients and so are unlikely to bias the estimates in an important way:

ethnic_group	count
EG-A	273
EG-B	117
EG-C	55
EG-D	330
EG-E	18
EG-F	9
EG-G	8
EG-H	10
NA	79

Finally, we see that `gene_source_lof` fits well even when looking at specific ethnicities and adding ethnic group does not really improve the picture:

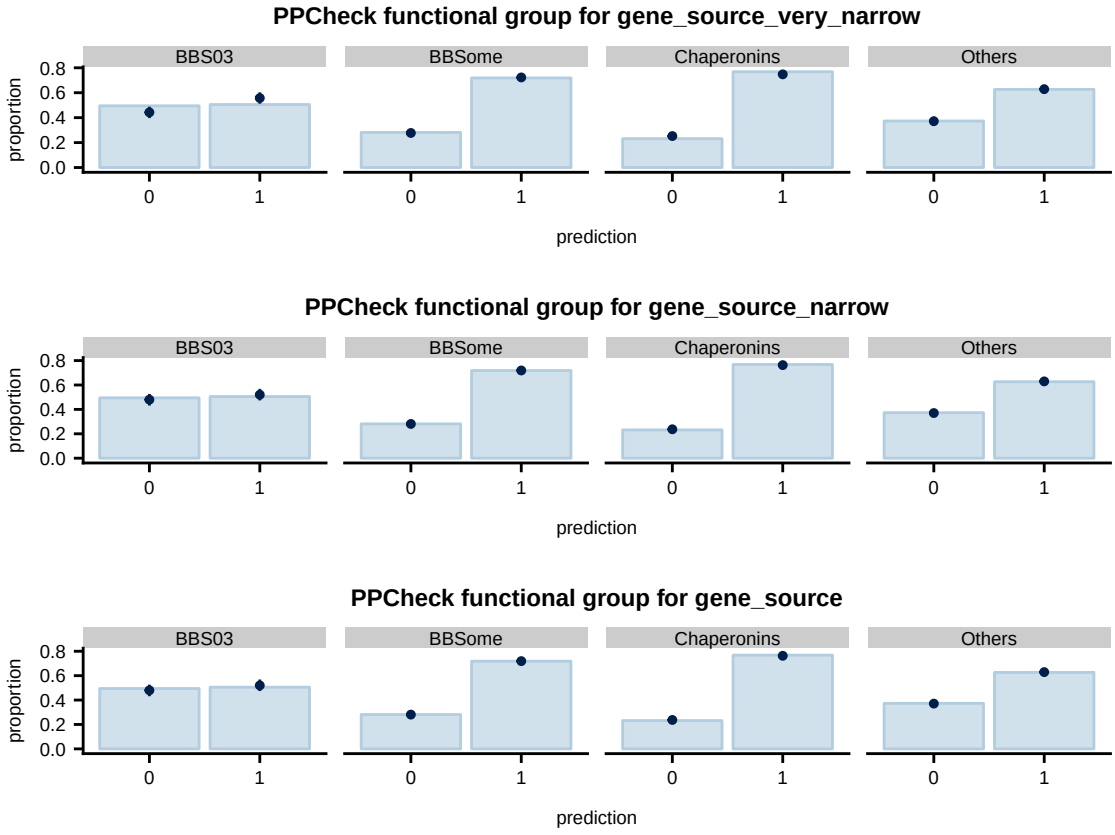




We chose not to include ethnicity or ethnic group in the model, as it is largely explained with just between-study variability.

Prior width

One example that the "very narrow" prior prevents the model from fitting well is that the fit cannot capture the BBS3 and Chaperonins functional groups, while "narrow" (and wider) priors don't have a problem with that:



We therefore consider the `gene_source_very_narrow` model as a problematic fit. Otherwise we didn't find a good reason to prefer either of the "narrow", normal and "wide" priors and this choice seems to be of little consequence for model inferences (as discussed in Part 3).

Model selection verdict

We have shown that source (between-study variability) has to be included and there is an advantage in including cLOF per phenotype, but not much improvement when including cLOF per phenotype and gene. Since the `gene_source_lof_per_gene` model is too flexible for the limited amount of data (results in very wide posterior intervals, spanning odds ratio up to 10000), we think `gene_source_lof` is a better choice. We have further shown that adding family or ethnicity information improves the fit only a little while making the model more complex and will therefore not be included for the main analysis.

Part 3: Conclusions under Multiverse Analysis

In this part we show how conclusions we discuss in the main paper hold under different model choice. A wide variety of models was tested. Those are briefly described in the comparison. Their exact formulations can be found in Part 2.

Note: For historical reasons the feature of "certain loss of function" (cLOF) as discussed in the data is called just "lof" in most analysis code. This part will thus use "lof" and "cLOF" interchangeably.

Defining Precise Criteria

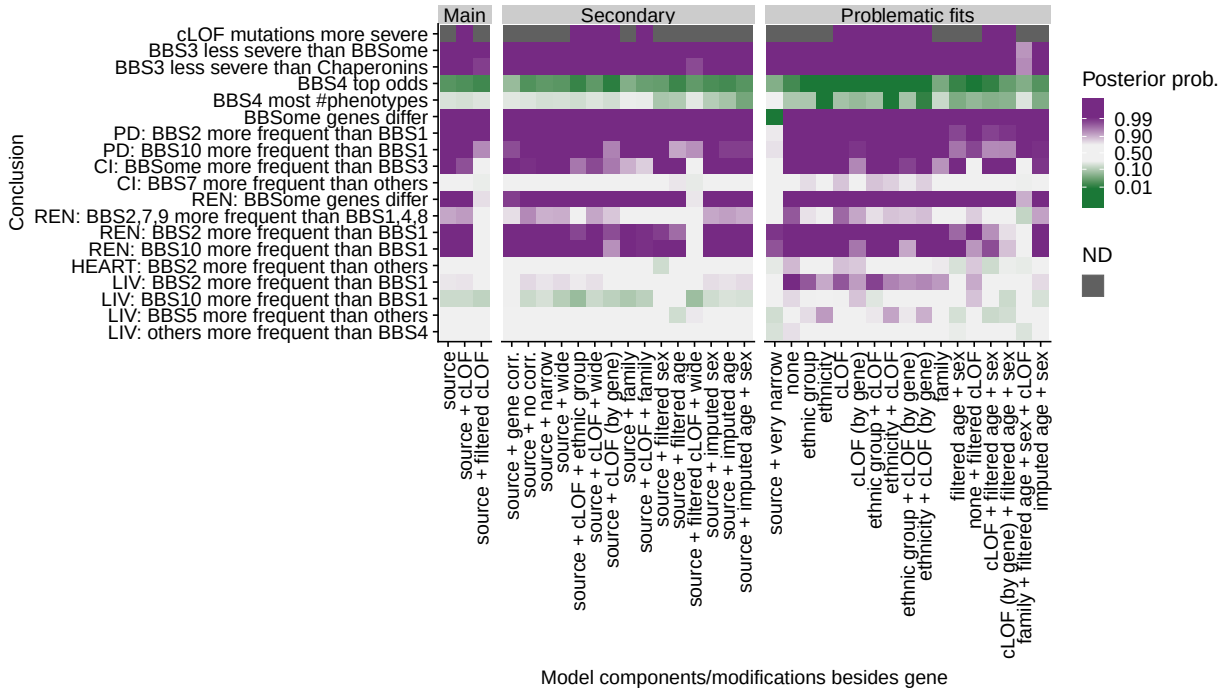
There is some flexibility in defining the exact model configurations consistent with a given conclusion. In general, when discussing whether there is or isn't a difference, the Bayesian model will always say "yes there is a (possibly small) difference" - the posterior probability of a difference of exactly 0 is 0. Instead we choose a threshold of a "clinically relevant effect", which for us is odds ratio outside of (0.5,2) and any effect outside of these bounds is counted as "different".

Similarly, other plain-English statements need to be transformed to an exact predicate that can be considered true or false for a given assignment of numerical values to model coefficients, to let us evaluate its posterior probability. The exact definition of the individual tested statements follows.

- The severity of BBS is worse in patients with LOF mutations than in patients with other mutations.
  - Ignored for models that do not include LOF.
  - Measured as posterior probability, that the LOF effect is positive (OR > 1) for at least 5 BBSome genes in 3 phenotypes.
- The data suggest that mutations in BBS3 have lower severity than mutations in different functional groups of genes.
  - Measured as posterior probability, that odds ratio is < 1 for at least 3 phenotypes for at least 1/2 of pairwise gene comparisons.
- The data suggest a difference between the severity of BBS in patients with mutations in different BBSome subunits.
  - Measured as posterior probability, that there is clinically relevant effect for at least 3 phenotypes for at least 5 pairwise comparisons.
- BBS4 phenotype is the most severe of all BBSome-encoding genes.
  - Probability that odds for BBS4 are among the top 3 odds for at least 5 phenotypes.
  - Probability that a patient with BBS4 has the highest total number of phenotypes present.
- Mutations in different BBSome subunits predispose to different renal phenotype.
  - Measured as posterior probability, that there is clinically relevant effect for at least 5 pairwise comparisons.
- Differences between small groups of genes: probability that all pairwise odds ratios are greater/less than 1 (depending on the direction of the comparison).
  - Cognitive impairment is less frequent in BBS3 patients compared to other patients (all canonical BBS genes).
  - Cognitive impairment is more frequent in BBS7 patients compared to other patients with mutations in BBSome-encoding genes.
  - Renal involvement is less frequent in BBS1, BBS4 and BBS8 patients compared to BBS2, BBS7 and BBS9 patients.
  - Patients with BBS2 mutations are more likely to have heart anomalies compared to patients with other mutations in BBSome-encoding genes.
  - Patients with BBS5 mutations are more likely to have liver anomalies compared to patients with other mutations in BBSome-encoding genes.
  - Patients with BBS4 mutations are less likely to have liver anomalies compared to patients with other BBSome mutations.
- Individual differences in phenotype between two genes: directly the probability that the OR > 1 for patients with cLOF mutation.
  - Polydactyly is more frequent in BBS2 patients compared to BBS1 patients.

- Polydactyly is more frequent in BBS10 patients compared to BBS1 patients.
- Renal involvement is less frequent in BBS1 patients compared to BBS2 patients.
- Renal involvement is less frequent in BBS1 patients compared to BBS10 patients.
- Liver involvement is less frequent in BBS1 patients compared to BBS2 patients.
- Liver involvement is less frequent in BBS1 patients compared to BBS10 patients.

Bayesian Comparison



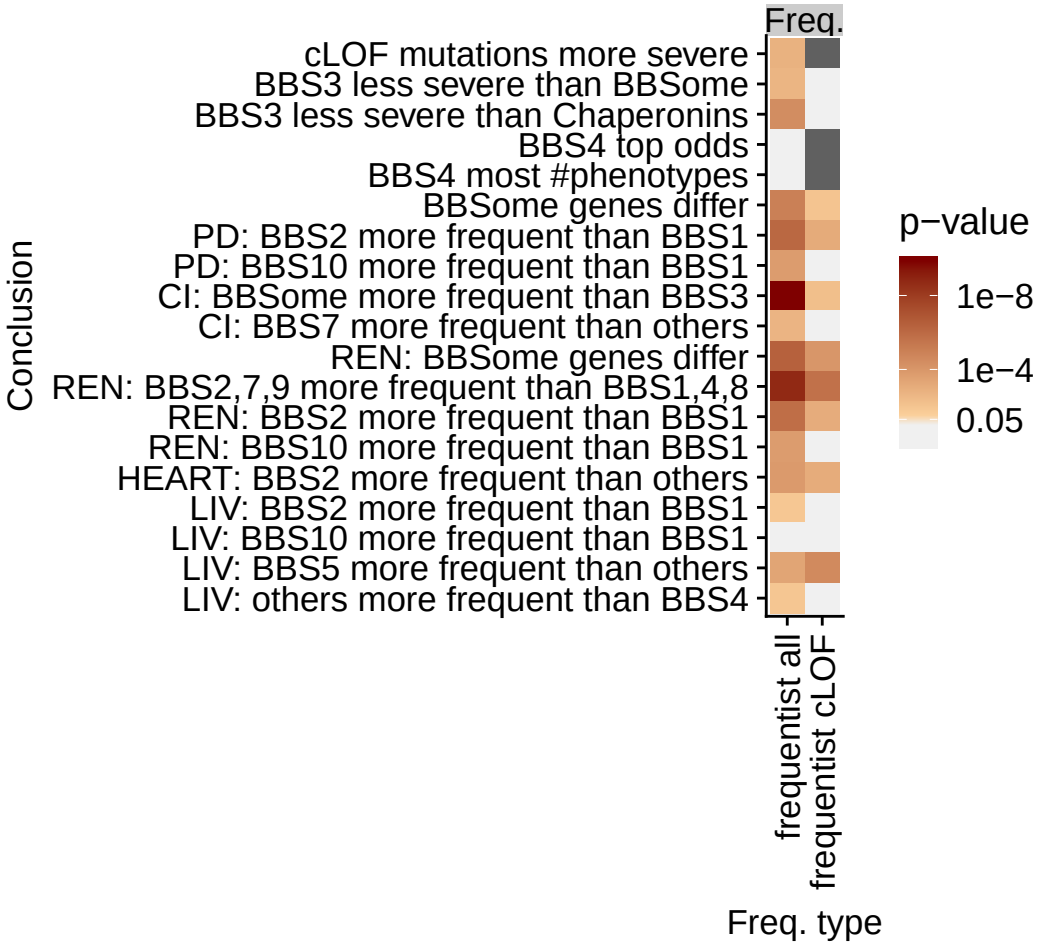
A heatmap of posterior probability of statements characterizing individual conclusions. Note that the probability is on logit scale. All Bayesian models include gene as covariate, but may also include additional covariates: source, age, sex, ethnicity or ethnic group, family and certain loss of function (cLOF) - either as a global covariate or by gene. Since age and sex are not available for all data, we can either fit the model only to patients where those are reported (filtered) or impute missing data (imputed). Instead of using cLOF as a covariate, we can fit the model using only patients with cLOF mutations (filtered cLOF). For most models we include a correlation structure across phenotypes (e.g., that two phenotypes occur frequently together across all genes), but this structure may be absent (no corr.) or replaced with a correlation structure across genes (gene corr. - e.g., that two genes have similar pattern of effects across all phenotypes). We also tried modifying the width of prior distributions (wide, narrow, very narrow). See Part 2 of this supplement for a detailed description of all models and the imputation procedure. Dark grey indicates that the question could not be evaluated for the given model (currently only asking for cLOF differences in models that exclude cLOF). The "Problematic fits" category is reserved for models we know do not capture some important variability in the dataset, as discussed in Part 2.

Note: posterior probabilities are clamped to be at least 0.001 and at most 0.999 as the sampling scheme used does not let us be very confident in the tails of the distribution. It is possible that with more computational resources some of the posterior probabilities will be more extreme.

Some patterns to notice:

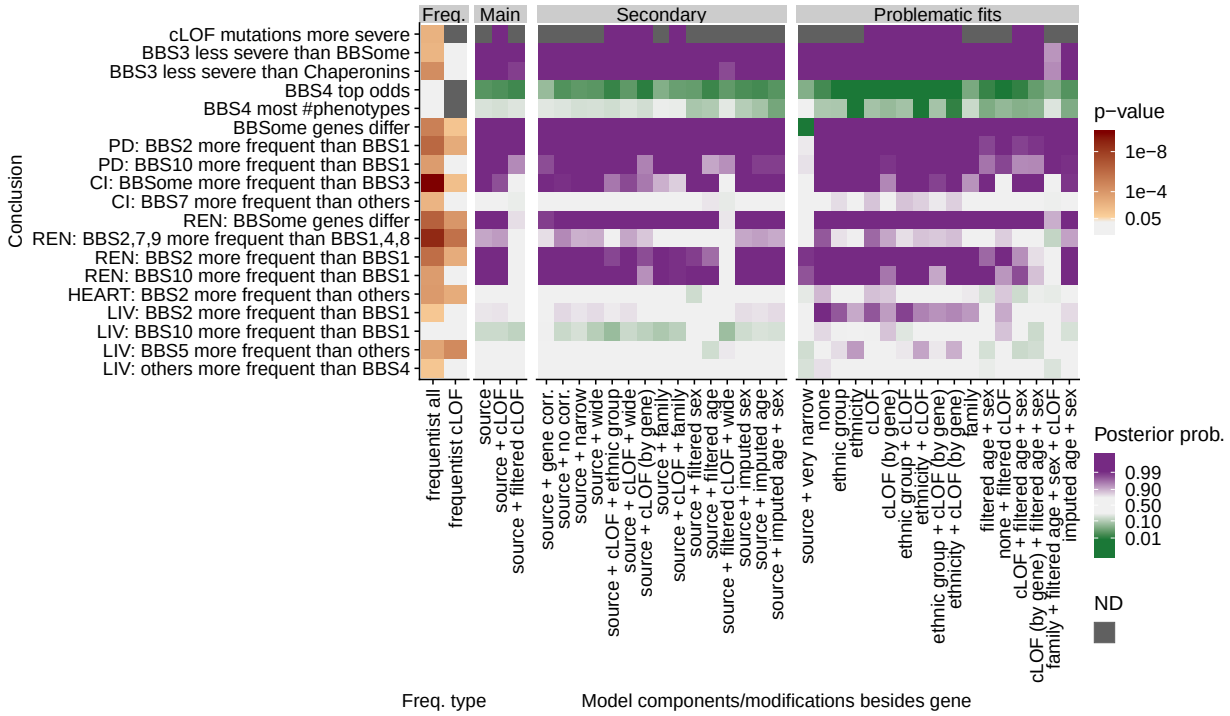
- When fitting filtered datasets, there is less certainty implying less strong evidence in both directions.
- Using very narrow priors on gene coefficients ($N(0, 0.1)$, i.e., that almost all odds ratios should be less than ~ 1.2) unsurprisingly results in little evidence for directional differences between genes.
- Other than noted above, the conclusions are not sensitive to model choice.

Frequentist results



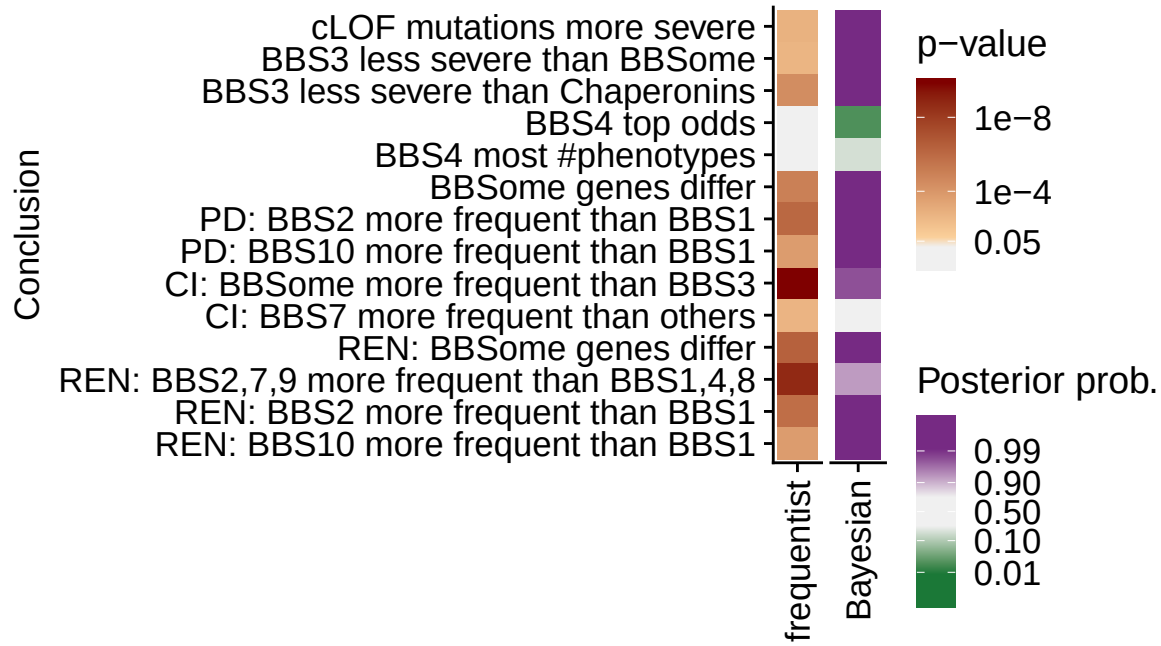
Heatmap of p-values from frequentist analysis for the same conclusions (exact computations described in the main body of the paper and in the analysis code and report on Zenodo, DOI: 10.5281/zenodo.3243400).

Combining all results



This is a combined plot for both frequentist and Bayesian analysis, once again showing mostly consistent results even when frequentist analyses are taken into account. The only big difference is that some of the HEART and LIV conclusions are not supported by most Bayesian analyses, or only those ignoring between-study variability. Also, Part 1 discusses some of those conclusions and shows how between-study variability is likely important for them.

Finally, a brief summary of the main analyses as shown in the main paper.



Original computing environment

This report was built from Git revision `9d4d48b9f6bcb07b0ce74a6ca7bd01d48bb250f0` on 14 June, 2019

```
## R version 3.5.3 (2019-03-11)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] mice_3.3.0         lattice_0.20-38    svglite_1.2.1
##  [4] knitr_1.23         cowplot_0.9.3      bayesplot_1.6.0
##  [7] tidybayes_1.0.1    forcats_0.3.0      stringr_1.3.1
## [10] dplyr_0.8.1        purrr_0.2.5        readr_1.1.1
## [13] tidyr_0.8.1        tibble_2.1.1       tidyverse_1.2.1
## [16] here_0.1           readxl_1.1.0       skimr_1.0.3
## [19] brms_2.8.0         Rcpp_1.0.1         rstan_2.18.2
## [22] StanHeaders_2.18.0 ggplot2_3.1.0
##
## loaded via a namespace (and not attached):
##   [1] minqa_1.2.4              colorspace_1.3-2
##   [3] ggridges_0.5.1           rsconnect_0.8.8
##   [5] rprojroot_1.3-2          ggstance_0.3.1
##   [7] markdown_0.8             base64enc_0.1-3
##   [9] rstudioapi_0.8           svUnit_0.7-12
##  [11] DT_0.4                   mvtnorm_1.0-8
##  [13] lubridate_1.7.4          xml2_1.2.0
##  [15] splines_3.5.3            bridgesampling_0.5-2
##  [17] codetools_0.2-16         shinythemes_1.1.1
##  [19] jsonlite_1.5             nloptr_1.2.0
##  [21] LaplacesDemon_16.1.1     broom_0.5.0
##  [23] shiny_1.1.0              compiler_3.5.3
##  [25] httr_1.3.1               backports_1.1.2
##  [27] assertthat_0.2.1         Matrix_1.2-15
##  [29] lazyeval_0.2.1           cli_1.1.0
##  [31] later_0.7.5              htmltools_0.3.6
##  [33] prettyunits_1.0.2        tools_3.5.3
##  [35] igraph_1.2.2             coda_0.19-1
##  [37] gtable_0.2.0             glue_1.3.1
##  [39] reshape2_1.4.3           cellranger_1.1.0
##  [41] nlme_3.1-137             crosstalk_1.0.0
##  [43] xfun_0.3                 ps_1.3.0
##  [45] lme4_1.1-18-1            rvest_0.3.2
##  [47] mime_0.5                 miniUI_0.1.1.1
```

```
##  [49] gtools_3.8.1             pan_1.6
##  [51] MASS_7.3-51.1            zoo_1.8-4
##  [53] scales_1.0.0             colourpicker_1.0
##  [55] hms_0.4.2                promises_1.0.1
##  [57] Brobdingnag_1.2-6        parallel_3.5.3
##  [59] inline_0.3.15            shinystan_2.5.0
##  [61] RColorBrewer_1.1-2       yaml_2.2.0
##  [63] gridExtra_2.3            gdtools_0.1.7
##  [65] loo_2.1.0                rpart_4.1-13
##  [67] stringi_1.2.4            highr_0.7
##  [69] dygraphs_1.1.1.6         pkgbuild_1.0.3
##  [71] rlang_0.3.4              pkgconfig_2.0.2
##  [73] matrixStats_0.54.0       evaluate_0.11
##  [75] rstantools_1.5.1         htmlwidgets_1.3
##  [77] labeling_0.3             processx_3.3.1
##  [79] tidyselect_0.2.5         plyr_1.8.4
##  [81] magrittr_1.5             R6_2.4.0
##  [83] mitml_0.3-6              pillar_1.4.1
##  [85] haven_1.1.2              withr_2.1.2
##  [87] xts_0.11-1               nnet_7.3-12
##  [89] survival_2.43-3          abind_1.4-5
##  [91] modelr_0.1.2             crayon_1.3.4
##  [93] jomo_2.6-4               arrayhelpers_1.0-20160527
##  [95] rmarkdown_1.10           grid_3.5.3
##  [97] callr_3.2.0              threejs_0.3.1
##  [99] digest_0.6.17            xtable_1.8-3
## [101] httpuv_1.4.5             stats4_3.5.3
## [103] munsell_0.5.0            shinyjs_1.0
```