A Bayesian Meta-analysis of Genetic Background Effects

Nirupama Tamvada
Stat 520A
Term 2, 2022

Abstract

Genetic background effects occur when the same mutations are expressed differentially in terms of phenotype across genetically distinct individuals. Certain allele types have been linked to the exacerbation of genetic background effects, contributing to a wide range of hereditary disorders. Thus, it is imperative to it is crucial to consolidate the existing pool of knowledge on genetic background effects to identify alleletypes that may be linked to such genetic background effects. We thus conduct a Bayesian meta-analysis using a hierarchical Gamma regression model, where we model both the uncertainty of meta-analyses estimates, as well as between-study heterogeneity. Overall, hypermorphic alleles were found to be potentially linked to an increased expression of genetic background effects. While more necessary model fit checks and exploration are pending, this results have significant implications for elucidating the mechanisms of genetic background effects, as well as detecting genes that contribute to this.

1 Introduction

Genetic background effects occur when the same point mutations show different phenotypic effects across genetically distinct individuals [1]. Due to these genetic background effects, certain allele types (variants of genes) contribute to a wide range of hereditary disorders, including, but not limited to, colorectal cancer, hypertension, and phenylketonuria [1]. Additionally, it has been proposed that background effects can influence the emergence of novel traits, as well as the maintenance of deleterious genetic variation within populations [1]. Thus, understanding the causal genetic mechanisms of genetic background effects is a significant topic of interest in biology and medicine. For this reason, it is crucial to consolidate the existing pool of knowledge on genetic background effects through a meta-analysis.

To this end, this study presents a meta-analysis with 44 studies of genetic background effects. The aim of this meta-analysis is to quantify the relationship of different alleletypes, zygosities and phenotypes with the magnitude of genetic background effects. A Bayesian meta-analysis is a particularly attractive endeavour for this problem, as it allows for 1) incorporating prior knowledge 2) modelling the uncertainty of estimating the between-study heterogeneity, which can be problematic in meta-analyses with relatively lower sample sizes and 3) modelling the uncertainty in the estimates effect sizes [2].

In this report, the model construction and prior choices for the Bayesian meta-analysis will be presented. Finally, the report will conclude with a discussion on the results of the meta-analysis model fit, as well as on limitations and potential future additions to the model framework.

Table 1: Metadata and Variable description

Variable	Description
Study	Paper the observation was extracted from
Alleletype	Type of mutation (null, hypermorph, hypomorph, neomorph)
Phenotype Type	Type of phenotype that was measured after mutation (behaviour, disease, metabolic, morphologic and physiologic)
Zygosity	Levels: Similarity of alleles (heterozygotes, homozygotes, hemizygotes)
Sex	Male or Female

2 Methods

2.1 Data Structure

The full dataset used for this analysis consists of data extracted from 40 studies. There are 1146 observations in the data in total, with multiple observations extracted from some studies. The inclusion criteria specified studies done only in species of mice, with one focal mutation in atleast two different genetic backgrounds and with a comparison made to wildtype organisms (controls). The mean of the phenotype measure in the wildtype, in the mutant, the corresponding standard deviations and the sample sizes were extracted from all of the studies. The primary response variable in the final dataset is the absolute value of the log coefficient of variation i.e log(CV). This ratio was calculated as a function of the extracted summary statistics as follows (along with its corresponding standard error, both presented in [3]):

$$log(CV) = log_2\left(\frac{\frac{s_{mutant}}{\overline{X}_{mutant}}}{\frac{s_{wildtype}}{\overline{X}_{wildtype}}}\right) + \frac{1}{2N_{mutant}-1} + \frac{1}{2N_{wildtype}-1} \tag{1}$$

The absolute value of log(CV) is used as we are only interested in the magnitude of the background effects and not the directionality. Thus, we are esentially interested in modelling the magnitude of the ratio of the effect in the mutant against the wildtype controls and its relationship to our predictor variables. The metadata and variable description is presented in Table 1.

2.2 Model Set Up

We begin with the assumption that there is a true (unknown) effect that lies in each of the included studies. We assume that the observed effect, i.e, the absolute `log(CV)` is generated as follows:

$$Y_{ij}|\mu_j, k_j \sim \texttt{Gamma}(\mu_j, k_j + \tau^2) \tag{2}$$

Here, k_j represents the shape parameter of the Gamma distribution and τ represents the between-study heterogeneity. The Gamma distribution is an appropriate assumption here, as our response variable of interest is continuous and only takes values in the positive real space $\mathbb{R}^+$. We then specifically model the mean observed effect size as:

$$log(\mu_i) = \alpha_j + \beta_{ji} x_{ij} \tag{3}$$

We additionally model the shape parameter of the Gamma distribution to account for a relationship with the estimated uncertainties of the observed log(CV) values as:

$$log(k_i) = \theta_j + \gamma_{j1} SE_{1ij} \tag{4}$$

The prior setup for this model is as follows:

$$\alpha_j \sim \texttt{Normal}(0, 5) \tag{5}$$

$$\theta_j \sim \texttt{Normal}(0, 5) \tag{6}$$

$$\tau \sim \texttt{Half-Cauchy}(1, 0.7) \tag{7}$$

$$\beta_i \sim \texttt{Normal}(0, 2) \tag{8}$$

$$SE_{ij} \sim \texttt{Half-Cauchy}(1, 0.7) \tag{9}$$

Half-Cauchy priors were chosen for τ and SE_{ij} (the standard errors of the effect sizes), as this distribution typically has heavy tails, ensuring that larger values of τ and SE_{ij} are possible with some probability, while assuming that lower values are more likely [4]. A relatively narrow normal prior was chosen for the coefficients keeping in mind that the coefficients are on the log scale. This specific Half-Cauchy prior was chosen for the standard errors of the effect sizes to constrain the relationship with the shape to be an increasing one, to the end that a large error lead to a more spread shape of the Gamma distribution. As repeated measures were collected from several very heterogeneous studies, we do expect a larger amount of between-study variability, as well as variation associated with the effect sizes. It should be noted that this model does not account for dependencies between the repeated effect sizes collected from each study. While this model was attempted, it was found to be quite computationally intensive, with a long run-time and thus, this model was not explored. In particular, for this paper, we are primarily interested in the posterior distributions of the coefficients for the contrasts of the `alleletypes`, as well as the study-level effects. The null alleletype is the reference level for the `alleletype` variable.

All the models were run using `brms` which uses `STAN` as a backend [5]. The models were run using `R` (v.4.1.2) on a 2021 M1 MacBook Pro [6]. All the `STAN` code and other `R` code can be found in the following Github repository: `https://github.com/ntmv/bayes-gbe`.

2.3 Model Calibration and Posterior Predictive Checks

In terms of checking model fit, the following checks were examined:

1. Drawing 100 prediction samples from the posterior distribution and overlaying the predicted samples with the observed data

2. LOO probability integral transform (PIT)

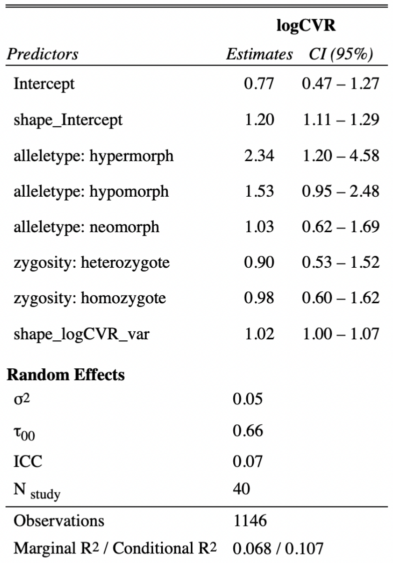
The LOO probability integral integral transform (PIT) combines Leave-One-Out-Cross-Validation and Probability Integral Transform. Essentially, these look to see where each observed point y_i falls in the predictive distribution $p(y_i, y_{-i})$. For a well-calibrated model, values from the kernel density function of all LOO-PIT values should resemble a uniform distribution [7, 8].

The plots for both these checks can be found in the Appendix. Additionally, all estimated posterior distributions and traceplots which show that the MCMC chains have mixed are also available in the Appendix.

3 Results and Discussion

3.1 Mutations in hypermorphic alleles largely contribute to genetic background effects

Table 2: Standard reporting of meta-analysis model coefficients. All coefficients are on the exponent scale.



	logCVR	
Predictors	*Estimates*	*CI (95%)*
Intercept	0.77	0.47 – 1.27
shape_Intercept	1.20	1.11 – 1.29
alleletype: hypermorph	2.34	1.20 – 4.58
alleletype: hypomorph	1.53	0.95 – 2.48
alleletype: neomorph	1.03	0.62 – 1.69
zygosity: heterozygote	0.90	0.53 – 1.52
zygosity: homozygote	0.98	0.60 – 1.62
shape_logCVR_var	1.02	1.00 – 1.07
Random Effects		
σ^2	0.05	
τ_{00}	0.66	
ICC	0.07	
N $_{study}$	40	
Observations	1146	
Marginal R^2 / Conditional R^2	0.068 / 0.107	

Table 2 above presents the point estimates (the posterior mean estimates) of the estimated model coefficients. In particular, from this table and from the posterior distributions presented in the Appendix, we see that in the pooled effects, `zygosity` does not seem to have a significant relationship with a change in genetic background effects with mutations, as the posterior distributions of the `zygosity` contrasts are centered around zero. However, when looking at `alleletype`, overall, we see that the contrast of hypermorphs compared to null mutations seems to be centered around 2 on the exponent scale. This indicates that overall, in mutations in hypermorphic alleles seem to have a relationship with differential genetic background effects. This is an important result for biology as it provides evidence that diseases primarily linked to such hypermorphic alleles can lead to comparatively more adverse phenotypes in different genetic backgrounds, such as ABCA4 disease, Alzheimer's and a host of other autoimmune diseases [9, 10]. This result also has the potential to speed up the detection of putative casual genes which are differentially expressed among humans, as a greater focus can be established on such hypermorphic alleles [9, 10].

3.2 Model fit and calibration

Looking at the posterior predictive checks in the Appendix, we see that predictions from the fitted model do align reasonably well with the observed data, indicating that the model is likely a good fit for the data and that the model is likely correctly specified for the data. Additionally, the the samples from the Probability Integral Transformed estimated kernel density do seem to fall into a reasonable range of variables drawn from a uniform distribution. This is potentially a good indication that our model is well calibrated that that the MCMC sampling has generated enough effective samples to approximate the posterior well. It indicates that our model likely approximates the true generative distribution of the data well, which therefore leads to the probability transformed values approximating the uniform distribution closely.

3.3 Limitations and future steps

While we did assess some posterior predictive checks and calibration measures for our model, it is additionally important to include a simulation-based calibration of the parameters estimates to our workflow. Additionally, while our priors were chosen to be weakly informative based on best practices, a prior predictive check would have been ideal to assess the validity of our prior choices. More posterior predictive checks such as comparing simulated posterior summary statistics to the observed data statistics would also be useful.

A large limitation in our analysis which was mentioned previously is that we did not account for the dependence between the repeated measures in this study, which future iterations of this project should attempt to do in a computationally efficient manner. Additionally, most studies in the literature perform meta-analyses assuming a normal distribution. This meta-analysis is likely one of very few which both had a response effect size modelled by the Gamma distribution, and additionally modelled the shape parameter to account for the uncertainty in the calculated log(CV) estimates (i.e did not assume that the standard errors were completely realized, as other studies did). Thus, a useful step is to additionally compare this model with its frequentist counterpart as a sensitivity analyses, in order to check that our approach is a valid one. However, overall our approach here is a novel one in the meta-analysis field. While our model can additionally benefit from more checks, it makes an important contribution by highlighting the important role that hypermorphic alleles play in contributing to genetic background effects.

References

[1] Mullis, M. N., Matsui, T., Schell, R., Foree, R., amp; Ehrenreich, I. M. (2018). The complex underpinnings of genetic background effects. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-06023-5

[2] Rhodes, K. M., Turner, R. M., amp; Higgins, J. P. T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. Journal of Clinical Epidemiology, 68(1), 52–60. https://doi.org/10.1016/j.jclinepi.2014.08.012

[3] Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., amp; Senior, A. M. (2014). Meta-analysis of variation: Ecological and evolutionary applications and beyond. Methods in Ecology and Evolution, 6(2), 143–152. https://doi.org/10.1111/2041-210x.12309

[4] Harrer, M., Cuijpers, P., Furukawa, T. A., amp; Ebert, D. D. (n.d.). Chapter 13 Bayesian meta-analysis: Doing meta-analysis in R. Chapter 13 Bayesian Meta-Analysis | Doing Meta-Analysis in R. Retrieved March 24, 2022, from https://bookdown.org/MathiasHarrer/Doing$_{M}eta_{A}nalysis_{i}n_{R}/bayesian - ma.html$

[5] Paul-Christian Bürkner (2021). Bayesian Item Response Modeling in R with brms and Stan. Journal of Statistical Software, 100(5), 1-54. doi:10.18637/jss.v100.i05

[6] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[7] Schad, D. J., Betancourt, M., amp; Vasishth, S. (2021). Toward a principled bayesian workflow in cognitive science. Psychological Methods, 26(1), 103–126. https://doi.org/10.1037/met0000275

[8] Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., amp; Gelman, A. (2019). Visualization in bayesian workflow. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(2), 389–402. https://doi.org/10.1111/rssa.12378

[9] Magno, L., Lessard, C. B., Martins, M., Lang, V., Cruz, P., Asi, Y., Katan, M., Bilsland, J., Lashley, T., Chakrabarty, P., Golde, T. E., amp; Whiting, P. J. (2019). Alzheimer's disease phospholipase C-gamma-2 (PLCG2) protective variant is a functional hypermorph. Alzheimer's Research amp; Therapy, 11(1). https://doi.org/10.1186/s13195-019-0469-0

[10] Zernant, J., Lee, W., Collison, F. T., Fishman, G. A., Sergeev, Y. V., Schuerch, K., Sparrow, J. R., Tsang, S. H., amp; Allikmets, R. (2017). Frequent hypomorphic alleles account for a significant fraction of ABCA4 disease and distinguish it from age-related macular degeneration. Journal of Medical Genetics, 54(6), 404–412. https://doi.org/10.1136/jmedgenet-2017-104540

Appendix

3.4 Posterior Predictive Checks

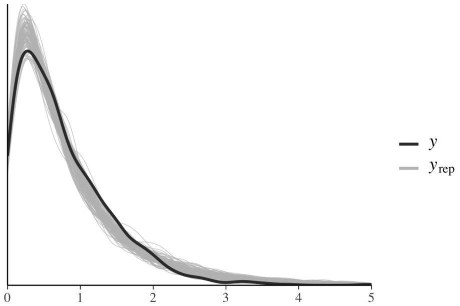


Figure 1: Posterior Predictive Check

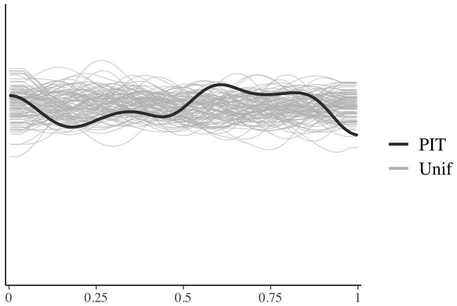


Figure 2: LOO Probability Integral Transform

3.5 Posterior Distributions and Traceplots

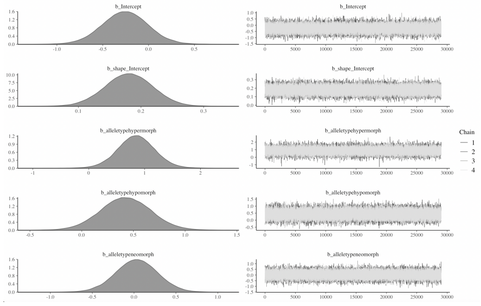


Figure 3: Posterior Distributions and MCMC Traceplots

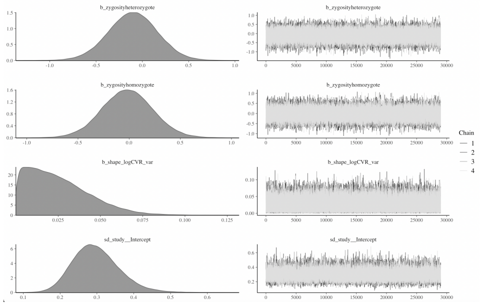


Figure 4: Posterior Distributions and MCMC Traceplots