Mixed Models for Studying Bacterial Physiology in Mice

Alicia Ter-Cheam

Department of Mathematics and Statistics,
McGill University, Montreal
December, 2021

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of Master of Science

Abstract

The gut microbiome is an ecosystem of microorganisms that are susceptible to changes in the presence of a disturbance, for instance diseases such as intestinal colitis. Through longitudinal studies, we can follow the effect of a perturbation to understand the microbiome's response and how the levels of bacteria and their physiologies change. The bacterial physiologies can be measured as proportions or standardized counts and thus modelled using generalized linear regression models. However, a natural dependence arises in the experimental design wherein several mice are kept in the same cage and repeated measures are taken from each mouse. Therefore, we propose using a generalized linear mixed model with random effects to model correlation between observations within the same cage and observations from the same mouse. We study the time dependence of the bacterial counts and whether there is a relationship between a biological disease state and the abundances of the physiologies. Using various GLMMs, we can show that the gut microbiome is highly dynamic and characterize some of the changes that occur over the course of disease.

Résumé

Le microbiome intestinal est un écosystème de micro-organismes qui est susceptible de changer en présence d'une perturbation, par exemple la maladie inflammatoire intestinale telle que la colite ulcéreuse. Grâce aux études longitudinales, nous pouvons comprendre l'influence de l'effet d'une perturbation sur le microbiome et l'évolution des niveaux de bactéries et leur physiologie. Puisque les physiologies bactériennes peuvent être mesurées comme des proportions ou des comptes standardisés, les modèles linéaires généralisés peuvent être utilisés. Cependant en gardant plusieurs souris dans la même cage et les mesures répétées sont prises pour chaque souris, la conception expérimentale entraîne une dépendance naturelle. Ainsi nous proposons d'utiliser des modèles linéaires mixtes généralisés avec les effets aléatoires pour capturer la corrélation entre les observations provenant de la même cage et les observations de la même souris. Nous examinons la dépendance temporelle des comptes bactériens ainsi que l'existence d'une relation entre l'état de la maladie biologique et les abondances des physiologies. En utilisant les GLMMs, nous pouvons montrer que le microbiome intestinal est très dynamique et caractériser certains des changements qui se prouduisent.

Acknowledgements

I would like to thank my supervisor Professor David Stephens for introducing me to Bayesian analysis and mixed models, as well as many other statistical areas since my undergraduate honours research project. I would also like to thank Mariia Taguer for sharing her data and research with me, and for answering all my questions about microbiology. Finally, I would like to thank my family for supporting me throughout my master's and the unique times during which it took place.

Contribution of the authors

The student was responsible for all of the writing, programming, and analysis contained in this thesis. The student and supervisor contributed equally to the generation of ideas and the editing. Data was provided by Mariia Taguer, PhD candidate, from Corinne Maurice's research lab in the Department of Biology at McGill University. The research from this paper regarding proportions data has been approved for publishing in *mSystems*, and a second paper has been submitted to journals regarding the sequencing application.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background for the application

Constantly inside of an animal's organs, an ecosystem of microorganisms must coexist peacefully at the appropriate levels in order to maintain healthy functioning. These microbes may be symbiotic or pathogenic, but so long as they maintain balance, they may coexist without problems, making up the microbiome [11]. However, diseases such as intestinal colitis, which can cause inflammation or ulcers in the digestive tract [20], are linked to disturbances in the gut microbiome [23].

The data being analysed in this project are provided by PhD candidate Mariia Taguer from Corinne Maurice's lab in the Department of Biology at McGill University. Using mice, they are mimicking intestinal colitis using the drug dextran-sodium sulfate (or DSS) and seeing how this affects the microbiome in the gut [25]. Their interest is in how a chemical perturbation, i.e. chemically altering the function of a biological system, affects bacterial physiology.

By sampling the microbiome, the researchers are able to determine the cell types in which

bacteria from the gut microbiome are abundant. Four different cell types are measured in this experiment: HNA, LNA, PI+, and BONCAT+. HNA and LNA are cells with high and low nucleic acid content, respectively. PI+ is characteristic of dead or highly-damaged bacteria, while BONCAT+ is characteristic of protein-producing bacteria.

2% DSS is introduced on Day 1 of the experiment for 5 days. Beginning on Day 6, the drug is flushed out with water for 5 days until the mice recover. Daily fecal samples were taken from each mouse with the days grouped into 'Disease States' as described in Table 1.1. These disease states were determined through previous biological evidence, and so their groupings won't be changed when trying to observe patterns over time for the different physiologies.

Disease State	Corresponding Days
Baseline	-3-0
Pre-symptoms	1-3
Symptoms	4-6
Recovery	7-12
W1 (Follow-up)	17
W2 (Follow-up)	24

Table 1.1: Disease state groupings, based on previous biological evidence. Follow-up measurements were taken one week after the ten experimental days (W1) and then again another week later (W2, two weeks after ten days).

This experiment was performed 3 times with 5 mice each time. All mice are kept in the same cage during each experiment, allowing us to equate one experiment with one cage. All mice per experiment are of the same sex: all mice are female in experiment 2 and the rest of the mice are male.

Two datasets were collected for this experiment. The first was general cell presence data for each mouse, which were proportions of each cell physiology seen in the samples, and the second was 16S sequencing data, which provided specific cell levels for each type of bacteria. Both of these datasets are explained in further detail in chapters 4 and 5.

1.2 Biological research questions

From these experiments, the researchers wish to understand how the bacteria's physiological makeup changes over time. For instance, if we use the disease state groupings, could we determine whether there is an association between the levels of the different cell types and the disease state? Most relevant might be to focus on the 'symptoms' disease state, and assess if there is a relationship between this period and abundances for each cell type.

Another method of describing a bacteria's physiology is through its dominant cell type. Over time, does a specific bacteria change its dominant cell type? Are there general trends in how bacteria switch between cell types?

We can compare trends at different taxonomic levels. The data provided from the 16S sequencing includes classification of the bacteria up to their genus. Thus, we might wish to aggregate by the phylum or family level to make a more general statement rather than

look at a unique bacteria. At the phylum level, if a bacteria is abundant in HNA in the beginning, would it ever become abundant in LNA or would it stay abundant in HNA the whole time? How is the physiology of *Firmicutes* bacteria changing compared to the physiology of *Bacteroidetes* bacteria?

To try to answer these questions, we will be using generalized linear mixed models, with the distribution chosen according to each dataset. The rest of the paper will be structured as followed. In chapter 2, we begin with a literature review going over existing methods of looking at microbiome data, for which we consider two foci: profiling and modelling. In chapter 3, we introduce generalized linear mixed models (GLMMs) as well as explain Markov chain Monte Carlo sampling (MCMC), which will be used in the Bayesian analysis of our GLMMs. We also perform simulation studies of hierarchical data using GLMMs to explain some of the challenges that come with this model. Chapter 4 covers the application of a logistic GLMM using the first proportions dataset. In chapter 5, we look at the second 16S sequencing dataset in two ways: first with a univariate gamma GLMM to look at each cell type separately, then using a multinomial GLMM to consider the abundant cell type. In chapter 6, we discuss overall findings as they pertain to the research question, as well as possible directions that could still be studied. Conclusions follow in chapter 7.

Chapter 2

Literature review

2.1 The nature of microbiome data

The microbiome, coined in 2001 by Lederberg and McCray, signifies "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" [16]. The study of microbiome alterations looks at the changes in the composition of the microbial taxa whether over time or between a disease group and control group. Researchers are often interested in linking these changes between specific taxa to a phenotype of interest. In the case of the biological research question in the application, we are looking for links between the taxa abundances and the enterotype, which is the classification based on the composition of the gut microbia [1]. For 16S rRNA sequencing, operational taxonomic units (OTUs) are used as the analysis unit to identify sequences [28]; in the research done by Taguer and Maurice, they can be considered equivalent to species. This equivalence is dependent on the taxonomic rank chosen by the researchers; in practice, we will often see OTU used interchangeably with bacteria species or taxon. Figure 2.1 shows the major taxonomic ranks and their ordering from largest to smallest, with the taxonomic classification of

E. coli provided as example. Kingdom, which falls between domain and phylum, is omitted as it was also not in the sequencing data provided.

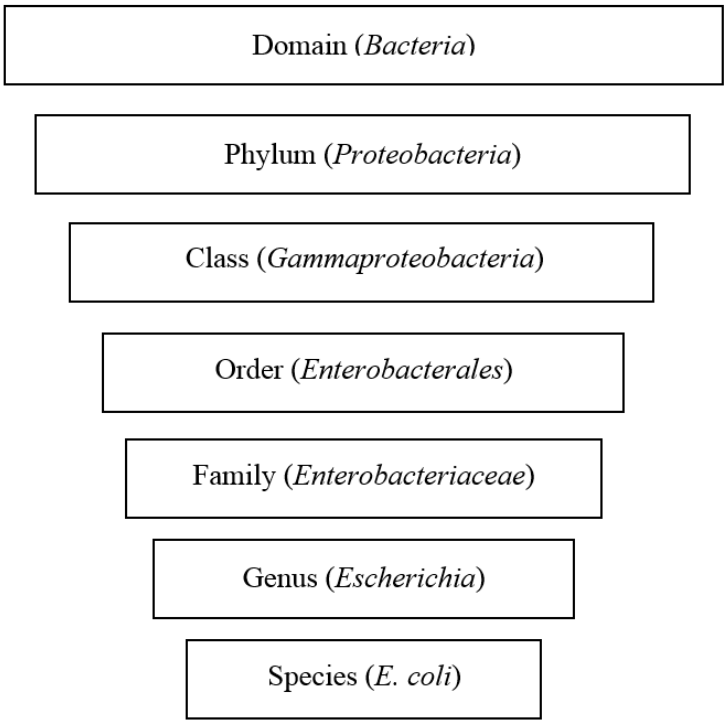


Figure 2.1: Taxonomic classification of *E. coli* bacteria.

Per Xia et al. (2008), there are 4 main features of microbiome data that can provide challenges in analysis [28]. Firstly, next-generation sequencing (NGS), such as the 16S sequencing data seen in Section 5, will often produce uneven total sequence counts of different samples; these counts must be normalized in the overall dataset which in turn generates relative abundances. This creates compositional data such that the increase in abundance of one taxon decreases the relative abundance of the other taxa. This can lead to misinterpretations of the microbiome structure when using traditional statistical methods, such as high false discovery rates and spurious associations [15].

The data are also high-dimensional from high numbers of OTUs and underdetermined, as the number of OTUs is often much larger than the number of samples; this problem is more relevant when we wish to model the data. This can be handled by looking at taxa individually, such as in by Chen and Li (2016) [4], or jointly modelling several taxa of interest [18].

The remaining two features are that the data tend to be over-dispersed and sparse with many zeroes. It is often the case we get sampling or count zeroes when the true population of an OTU is too low to be detected in the sequencing sample [24] or in compositional data as a result of insufficient sequencing depth or sample size [19].

When looking at microbiome data, we consider two foci: profiling and modelling.

2.2 Profiling

Data profiling, i.e. examining data through descriptive statistics and informative summaries [17], can be useful in identifying microbial taxa that are affected by an intervention or a disease state [9].

The technique of Quantitative Microbiome Profiling (QMP) is introduced by Vandeputte et al. (2017) [26] in order to quantify microbial abundances and link variation in the microbial loads to the enterotype associated with Crohn's disease. QMP generates an absolute abundance; this is in contrast to Relative Microbiome Profiling (RMP) which are limited in their ability to reveal possible interaction between microbiota and host health. Since there are currently no statistical methods that exist to deduce absolute abundances from com-

positional data, methods need to be developed to extract these abundances to form a data profile that can be analysed. Thus, in Vandeputte et al., cell counts are used to "transform sequencing data into an absolute microbiome abundance matrix that allowed quantitative microbiome profiling".

QMP is also used by Vieira-Silva et al. (2019) to look at primary sclerosing cholangitis (PSC) and irritable bowel disease (IBD), whereby the quantitative analyses allowed them to identify a pattern between a *Fusobacterium* and the *Veillonella* genera, with *Fusobacterium* being detected only in patients with Crohn's disease [27]. In the QMP, they look at microbiome alterations within inflammation-associated taxa, so that they could be differentiated and reveal markers for similar phenotypes that might present quantitative patterns.

Jian et al. (2020) consider quantitative PCR-based profiling to estimate absolute taxon abundances from the NGS data, as opposed to the cell-based method used by Vandeputte [12]. Galazzo et al. (2020) compared RMP with three different QMP methods and found that the resultant profiles are indeed affected by which microbial quantification technique is used [9].

After a QMP matrix is generated, there are a variety of statistical techniques that can be used to detect microbial alterations. In Vandeputte et al., they use QMP to refer to the quantitative genus abundance. In their analysis comparing Crohn's patients to their control group, they use the Wilcoxon rank-sum test to test differences in microbial load (cell counts per gram of faeces) and Kruskal-Wallis test for the differences in QMP [26]. The Mann-Whitney U-test can also be used to test for differences in bacterial loads between

enterotypes, as was done in Galazzo et al. (2020) [9]. Spearman's correlation coefficient is often used to link taxa abundance to physiological markers and quantify their associations ([26], [27], [12]).

2.3 Modelling

Beyond data profiling and acquiring the abundances, statistical tests are required to understand the interplay between the microbiome and physiology. As abundances are effectively the bacteria counts, the data can be treated as such in statistical modelling.

When modelling, there is the issue of high-dimensionality and the compositional structure that must be considered. There are high numbers of OTUs in the sample and it is often of interest to compare how they relate to each other, especially over time or in the presence of a physiological contrast. The method of jointly modelling the taxa is appealing since it can represent an overall community with a single model. However, we would often want to consider many OTUs in order to best represent the microbial diversity as well as achieve a good model fit [18]. Since the relative abundances are proportions that sum to 1, a Dirichlet-multinomial regression might be used to fit the data such as in Chen and Li (2013) [5]. While hierarchical models can be used to constrain the number of parameters (e.g. La Rosa et al. (2012) [21], Chen and Li (2013) [5], Sankaran and Holmes (2018) [22]), fewer parameters can lead to challenges in modelling the variance structure [22]. Furthermore, it is often difficult to account for the zero-inflated nature of the data in the joint models [18].

A common approach for modelling the relative abundance data is to use the beta distribu-

tion since we are handling a proportion. As mentioned, a challenging feature of microbiome data is how sparse it tends to be. Chen and Li (2016) proposed a zero-inflated Beta regression model with random effects (ZIBR) [4]. As this is a univariate model, they must consider each taxon separately. Martin et al. (2020) modelled microbial abundances using beta-binomial regression [18].

Nonparametric methods have been considered by White et al. (2009) and Segata et al. (2011), also for relative abundances. Morgan et al. (2012) proposed a Gaussian model for transformed observed relative abundances.

However, looking at a single taxon with the beta distribution does not completely cover the scope of the data's compositional nature; recall an increase in the abundance of one OTU will lead to the decrease in abundance of another. Thus, statistical modelling of the absolute abundances, such as those acquired from QMP, is another method. These can also be extended for zero-inflated data often without the problems that befall joint modelling. For absolute abundances, Robinson et al. (2010) use a negative binomial regression. Jiang et al. (2019) [13] propose a Bayesian zero-inflated negative binomial integrative model that "jointly identifies differentially abundant taxa among multiple groups and simultaneously quantifies the taxon covariate associations".

The negative binomial model for absolute abundances could also be further extended to include random effects to account for correlation as a result of repeated measurements or longitudinal correlations within an individual. This is done in Zhang et al. (2017) who used a negative binomial mixed model to detect association between host environmental or clinical

factors and the mouse gut microbiome [29]. Fang et al. (2016) proposed a zero-inflated negative binomial mixed model to account for the sparsity of the data, with application in oesophagitis [7].

For a more comprehensive look into the different models developed for handling microbiome data, see Xia et al. (2018) [28].

Chapter 3

Generalized linear mixed models

3.1 Introduction

In this longitudinal experiment, a natural grouping arises since multiple observations are taken from the same mouse over time. Rather than assuming independence of all observations, we would wish to induce some dependence structure between observations that come from the same individual.

A method of handling the structural dependence is through using a mixed model that includes a random effect for each group level. While fixed effects are common to all units in the population, the random effect will be specific to its group.

The general structure for the models will be to use a mixed effect model with the predictor being the categorical disease state and random effects for mouse and cage, as they are appropriate. This will be done by modelling the response using a generalized linear mixed model that consists of two components: random and systematic.

Our response will be denoted as Y_{cijk} where $c = 1, 2, 3, 4$ represents the different cell types, $i = 1, 2, 3$ represents the different cages/experiments, $j = 1, 2, ..., 5$ represents the

mouse, and $k = 1, 2, ..., K_i$ represents the experimental day. In the random component, Y_{cijk} are conditionally i.i.d. observations from an exponential family distribution, as conditioned on the necessary parameters for the family. In the systematic component,

$$g(\mathbb{E}[Y_{cijk}|X_{ijk}]) = \mu_c + u_i + v_{ij} + X_{ijk}^T\beta$$

$g(x)$ is the link function that will be used for the appropriate generalized linear model (i.e. the Normal in chapter 4, Gamma in 5.2, and multinomial in 5.3). μ_c is a mean value for cell type, u_i is the random effect associated with cage and v_{ij} is the random effect associated with mouse.

Looking at the correlation from this model, we induce a dependence between mice in the same cage, and between observations for a single mouse that are taken over the time span of the experiment. Specific covariance calculations will be stated for each specific model used for each of the applications in chapters 4 and 5. Thus, our sources of variation arise from between the different cages, the different mice, and each daily observation.

In the Bayesian framework, we consider the vector of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ where $\boldsymbol{\beta}$ is the vector of fixed regression parameters and $\boldsymbol{\sigma}$ is the vector of variances/covariances in the correlation model. This provides our distributional model $f(\mathbf{y}|\boldsymbol{\theta})$ for observed data $\mathbf{y}$. Our assumption is that $\boldsymbol{\theta}$ is randomly sampled from a prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ where we will specify our vector of hyperparameters $\boldsymbol{\lambda}$. We can specify our model so that $\boldsymbol{\lambda}$ is known and will generally be chosen to be fairly non-informative, so we thus perform inference for $\boldsymbol{\theta}$ using the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\mathbf{y}|\boldsymbol{\lambda})} = \frac{f(\mathbf{y}|\boldsymbol{\lambda})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int f(\mathbf{y}|\boldsymbol{\lambda})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}}$$

Of course, $\boldsymbol{\lambda}$ is not known in practice; we could specify a hyperprior $h(\boldsymbol{\lambda})$ and integrate over $d\boldsymbol{\lambda}$ as well, which leads to three levels of distributional specification. Alternatively, maximizing the marginal $p(\mathbf{y}|\boldsymbol{\lambda})$ would provide an estimate of $\hat{\boldsymbol{\lambda}}$, which could be used in inference based on the estimated posterior distribution, leading to empirical Bayes analysis [2]. However, in this paper, we will fix $\boldsymbol{\lambda}$ based on prior opinion about the dataset and so that our prior distribution is quite flat.

3.2 Markov chain Monte Carlo sampling

The computational challenge in the Bayesian method is that the posterior distribution is not generally tractable in closed form and hence requires numerical approximation. Markov chain Monte Carlo (MCMC) methods are the most widely used for this context and function by sampling $\boldsymbol{\theta}^{(g)}$, $g = 1, ..., G$ from the posterior distribution. MCMC algorithms create iterative draws from a Markov chain whose stationary distribution is the same as the posterior so that the samples are correlated. We typically use thousands of samples in order for the Markov chain to converge to the correct stationary distribution, as well as for the draws to be reasonably uncorrelated. Since the samples are not themselves i.i.d. draws from the posterior distribution, thinning methods might be used to reduce autocorrelation.

Two of the most popular MCMC algorithms are the Gibbs sampler and the Metropolis-Hastings algorithm.

The Gibbs sampler is advantageous in its clarity and ease of implementation, but it requires readily sampling from each of the full conditional distribution $p(\theta_i, \boldsymbol{\theta}_{j\neq i}, \mathbf{y})$, where we denote the unknown parameters as $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)$. This can be done directly if they are in a familiar form (normal or Gamma distributions) or using a rejection sampling approach. Then, the samples are generated in an iterative fashion, updating each following $\theta_i^{(t)}$ using $p(\theta_i|\theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{t-1}, ..., \theta_r^{(t-1)}, \mathbf{y})$.

However, the full conditionals are often not expressible in closed form, particularly when the prior distribution $p(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ are not conjugate pairs. The Metropolis-Hastings algorithm handles the situation by rejecting or accepting $\theta_i^{(t)}$ based on the ratio of the joint posterior likelihood using the proposed draw θ^* over the joint posterior with $\theta_i^{(t-1)}$.

Under mild conditions, both samplers are expected to converge to the true posterior. These samplers can also be combined, particularly if some of the full conditionals are known in closed form.

For the MCMC, we will be using the software Stan [3], available in the package `rstan` for use with `R`.

3.3 Simulated data

To test the performance of the models, we simulate hierarchical data. We begin with a Normal set-up, generating the first level response

$$Y_{ijk} \sim \text{Normal}(u_i + v_{ij} + X_{ijk}^T\beta, \sigma^2),$$

with the second level $u_i \sim \text{Normal}(0, \sigma_u^2)$ and $v_{ij} \sim \text{Normal}(0, \sigma_v^2)$, and third level $\sigma_u^2 \sim \text{Gamma}(1, 1)$, $\sigma_v^2 \sim \text{Gamma}(1, 2)$ and $\sigma^2 \sim \text{Gamma}(1, 0.5)$. These were chosen since we expect most variation to be between mouse and least within mouse. (We will see that in the real data, the greatest source of variation will turn out to be between days, i.e. within-mouse variation is the highest.)

In the simulations, we run tests with fixed variances $\boldsymbol{\sigma} = (\sigma_u^2, \sigma_v^2, \sigma^2)$ and changing $\boldsymbol{\sigma}$ to test overall performance of the model. When fixing variances, 4 different sets of $\boldsymbol{\sigma}$ were considered as follows:

σ_u (Cage)	σ_v (Mouse)	σ (Individual)
1.23	2.36	0.066
0.26	2.77	0.17
1.13	1.13	0.19
2.36	1.23	0.17

Table 3.1: Fixed $\boldsymbol{\sigma}$ sets used in simulation study.

For simplicity, we considered only a binary treatment variable with a true negative effect of $\beta_1 = -1$.

To compare the effects of experimental design on estimation of the variance parameter, we consider two different set-ups: one with three cages, as in the application, and one with six cages. In both set-ups, we generate $n = 150$ observations with 5 mice in each cage, so that the first set up had 10 'days' of data and the second had 5.

3.3.1 Results

Four chains are run with 2000 iterations, burning the first 1000 as warm-up for a total of 4000 draws.

Results for the fixed $\boldsymbol{\sigma}$ analysis are provided in Table 3.2. We can notice that the mean estimated cage standard deviation is generally lower than the true value, with the exception being when $\sigma_u = 0.26$. Changing the set-up to six cages brings the estimated value closer to the true value in all levels (cage, mouse, and individual). The treatment effect is also quite well estimated, with not much of a difference between the three cage vs. six cage set-up.

In Figure 3.1, we plot all the estimated $\hat{\beta}_1$ from the four different sets of fixed $\boldsymbol{\sigma}$ along with their 95% credible intervals. In Figures 3.2 and 3.3, we plot the random effect estimates and their 95% credible intervals, alongside the true random effect.

Results for the varying $\boldsymbol{\sigma}$ analysis are provided in Tables 3.3 and 3.4. We can note that the mean percent error does not change much for mouse variation when increasing to six cages; however, it drops by at least tenfold in cage and individual variation.

3.3.2 Discussion

Through the simulations, we can see that there is difficulty estimating the variance parameters for the group effect when there are fewer groups, as would be expected. In particular, there is a tendency to underestimate the cage effect. When there are more cages, we are better able to accurately predict the cage effect, which is to be expected as we have more

	3 cages		6 cages	
	Mean	Std. deviation	Mean	Std. Deviation
$\sigma_u = 1.23$	0.91	0.92	1.1	0.7
$\sigma_v = 2.36$	2.3	0.48	2.31	0.32
$\sigma = 0.066$	0.065	0.004	0.065	0.004
$\beta = -1$	-1.00	0.011	- 1.01	0.011
$\sigma_u = 0.26$	0.40	0.62	0.39	0.52
$\sigma_v = 2.77$	2.65	0.52	2.67	0.37
$\sigma = 0.17$	0.17	0.01	0.17	0.011
$\beta = -1$	-1.00	0.031	-0.97	0.027
$\sigma_u = 1.13$	0.91	0.62	1.05	0.43
$\sigma_v = 1.13$	1.09	0.22	1.12	0.16
$\sigma = 0.19$	0.19	0.012	0.19	0.012
$\beta = -1$	-0.96	0.032	-0.96	0.033
$\sigma_v = 1.23$	1.19	0.24	1.22	0.18
$\sigma = 0.17$	0.168	0.01	0.168	0.01
$\beta = -1$	-1.00	0.029	-1.02	0.031

Table 3.2: Simulation results for fixed $\boldsymbol{\sigma}$.

	σ_v (Mouse)	σ_u (Cage)	σ (Individual)	β_1 (Treatment)
Mean	1.98	0.93	0.52	-0.96
Standard deviation	0.60	1.41	0.019	0.051
Mean percent error	0.035	69.6	0.0006	2.4×7.5^{-5}
Mean ESS	591	317	1415	1958

Table 3.3: Results for $\boldsymbol{\sigma}$ estimation in the 3 cages set-up over 500 datasets where $\boldsymbol{\sigma}$ is generated from Gamma distributions described in the third level specifications. ESS stands for 'effective sample size'; note that 4000 samples are generated in each dataset.

available data to estimate σ_u. We see that the standard deviation for the variance parameters decreases as we go from three cages to six, as does our mean percent error. We can also see in Figures 3.6 and 3.7 that the width of the credible intervals for the variance parameters decreased as we considered more cages, hence we are getting more precise results with more cages. Again, this aligns with the intuition that more cages/mice provide more information to estimate the variance.

However, estimating the random effects themselves proved to be rather difficult for the model, as we can see in Figures 3.2 and 3.3. In general, it seems that the model tends to underestimate the random effects. This is more egregious for the mouse random effect, where the true effect tends to be larger because σ_v is also larger in the first set of variance parameters. Often times, the true random effect will fall outside the 95% credible interval.

	σ_v (Mouse)	σ_u (Cage)	σ (Individual)	β_1 (Treatment)
Mean	2	1.05	0.56	-1.01
Standard deviation	0.27	0.58	0.04	0.11
Mean percent error	0.04	5.8	2.4×10^{-5}	2.0×10^{-5}
Mean ESS	965	799	2572	3653

Table 3.4: Results for $\boldsymbol{\sigma}$ estimation in the 6 cages set-up over 500 datasets where $\boldsymbol{\sigma}$ is generated from Gamma distributions described in the third level specifications. ESS stands for 'effective sample size'; note that 4000 samples are generated in each dataset.

The random effect is often not the focus of real data applications, but this pitfall of the model should be kept in mind as we are using it. It is likely that the random effect is being folded into $\hat{\beta}_0$, as the treatment effect does not appear to be greatly affected.

The true fixed effect of $\beta_1 = -1$ was generally very well estimated. Unlike the variance parameters, there isn't much of a difference in estimating the fixed effect in the three cage vs. six cage set-up. From Figure 3.1, the true β_1 value is always captured in the 95% credible interval; this is also true in the varying $\boldsymbol{\sigma}$ simulations. The estimate in the simulations with three cages are slightly closer to the true value, but the differences are too small to be much use in practice.

Our focus in the simulations was the model's accuracy in capturing the true variation, and how that accuracy may be affected by the experimental design. It might be worth going

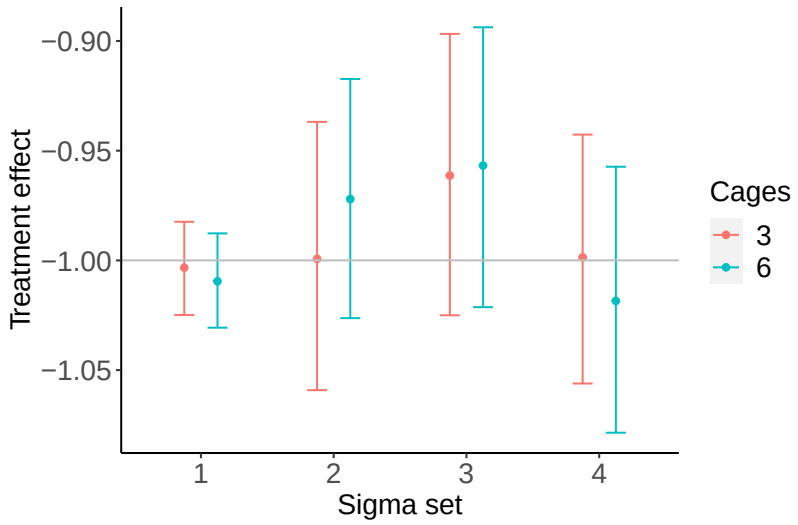


Figure 3.1: Treatment effect (β_1) estimates and 95% credible intervals for each of the fixed $\boldsymbol{\sigma}$ sets described in Table 3.1, in both the three cage and six cage set-up. True value of $\beta_1 = -1$, shown in the grey line.

back and running simulations with varying fixed effects.

All the sample distributions for the estimated parameters seem to have converged; we can see in Figure 3.4 that the chains for the treatment effect β_1 and σ have mixed quite well. There is slightly more autocorrelation seen in the random effect variances σ_u (cage) and σ_v (mouse); in particular, chains 3 and 4 for the cage variance displayed quite a few distinct excursions away from the general mean. However, the overall mixing still looks reasonable considering all chains.

The autocorrelation plots are included in Figure 3.5. In agreement with what we could observe visually in the trace plots, the autocorrelation is the smallest for β_1 followed by σ. Meanwhile, the variance parameters for the random effects have decent amounts of

autocorrelation for at least five lags. Note that the trace and autocorrelation plots were only included for our first $\boldsymbol{\sigma}$ set as described in Table 3.1. These general patterns showed up for the other three sets of $\boldsymbol{\sigma}$.

A final consideration for the convergence of the sampling distribution would be the effective sample size (ESS), i.e. the sample size of an equivalent random sample. Since we had 4 chains with 2000 draws, the first half being warm-up, this gives us a sample of 4000 for each parameter. The ESS is highest for the fixed effect β_1, followed by individual variance σ. It is much lower for the random effect variances. These effective sample sizes all increase when we consider 6 cages, even for β_1. Though lower for σ_u and σ_v, they are still large enough for general inference on their samples.

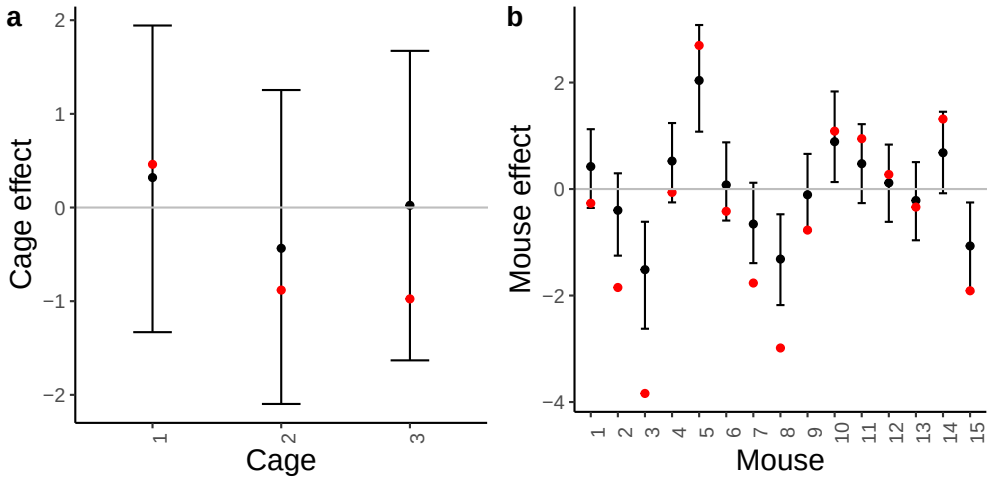


Figure 3.2: Random effect estimates and 95% credible intervals for the three cage set-up when $\sigma_u = 1.234$ (cage) and $\sigma_v = 2.66$ (mouse). The true random effect value is in red.

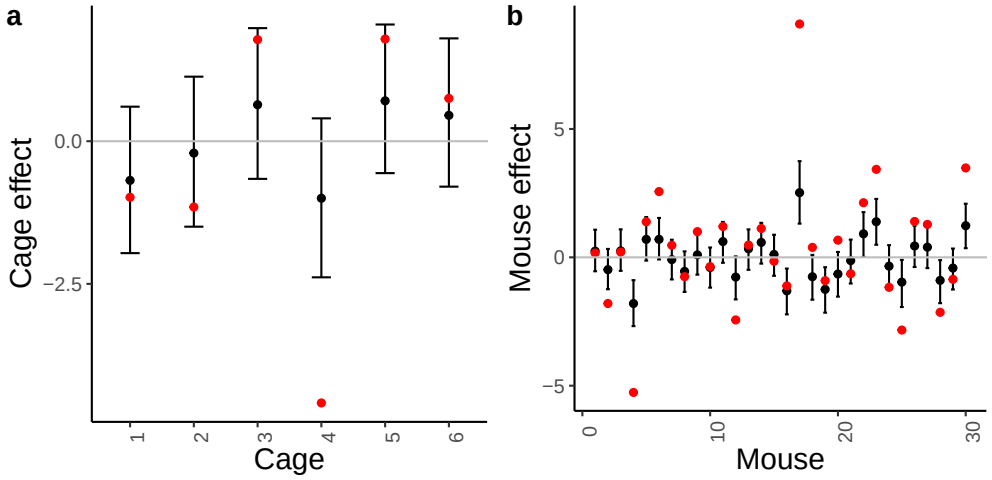


Figure 3.3: Random effect estimates and 95% credible intervals for the six cage set-up when $\sigma_u = 1.234$ (cage) and $\sigma_v = 2.66$ (mouse). The true random effect value is in red.

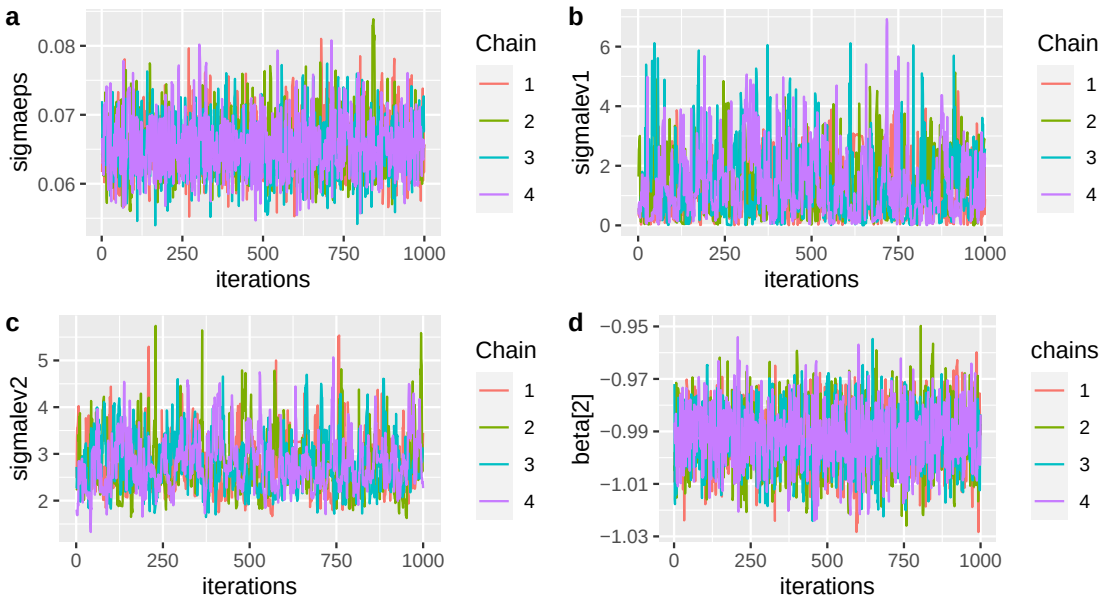


Figure 3.4: Trace plots for (a) $\sigma = 0.066$ individual error, (b) $\sigma_u = 1.234$ cage standard deviation, (c) $\sigma_v = 2.66$ mouse standard deviation, and (d) $\beta_1 = -1$ treatment effect.

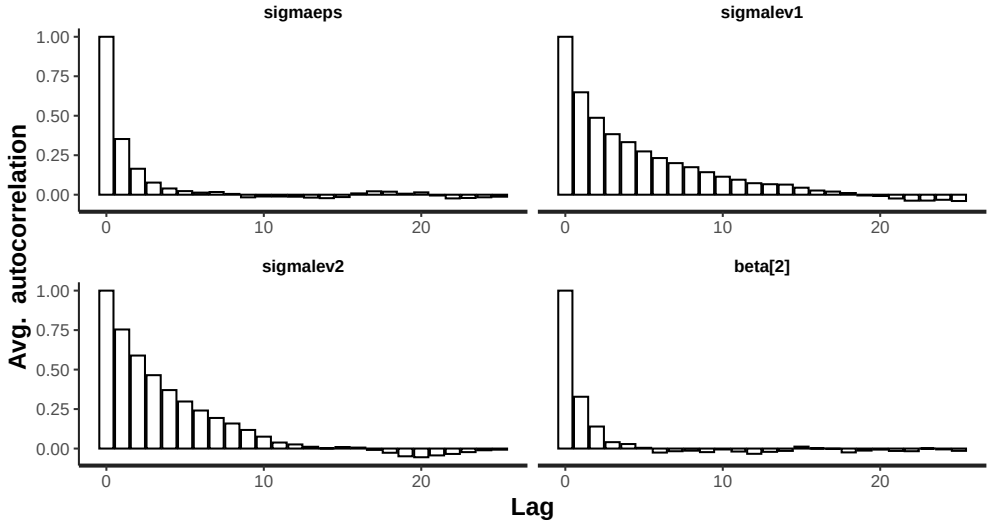
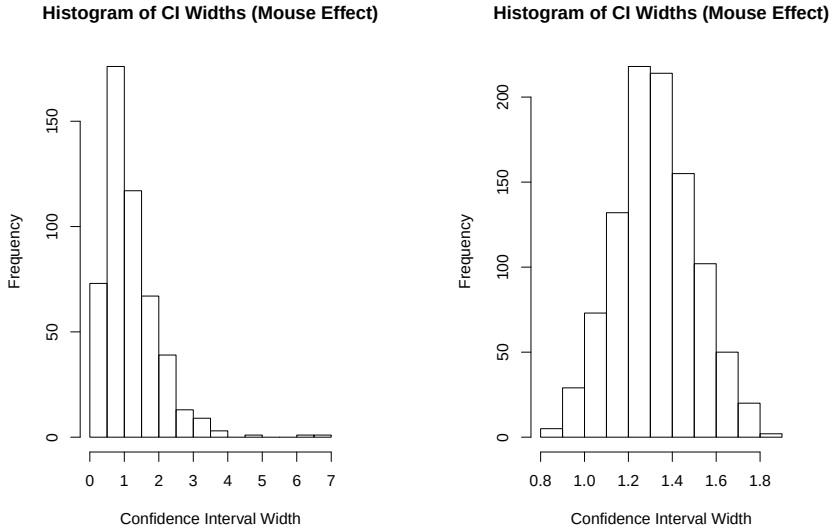
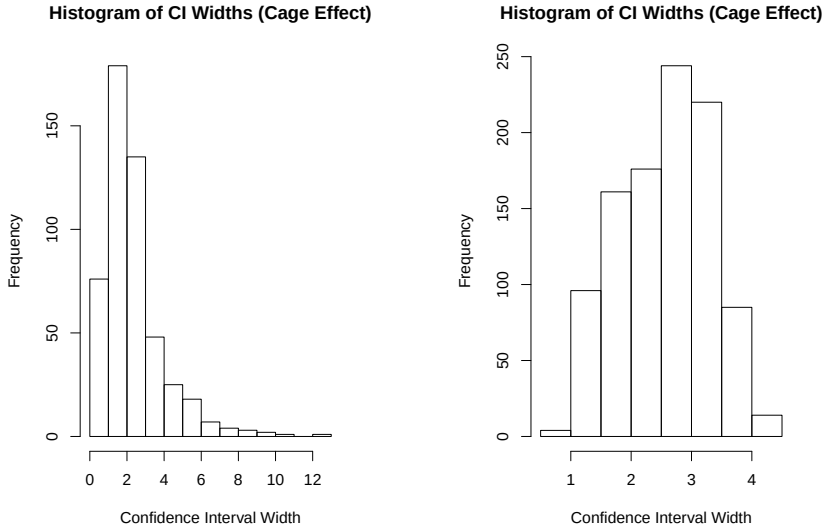


Figure 3.5: Autocorrelation plots for (a) $\sigma = 0.066$ individual error, (b) $\sigma_u = 1.234$ cage standard deviation, (c) $\sigma_v = 2.66$ mouse standard deviation, and (d) $\beta_1 = -1$ treatment effect.



(a) 3 cages (b) 6 cages

Figure 3.6: Confidence intervals for mouse effect; (a) three cages; (b) six cages



(a) 3 cages (b) 6 cages

Figure 3.7: Confidence intervals for cage effect; (a) three cages; (b) six cages

Chapter 4

Application: Cell proportions data

4.1 Introduction

The first set of data was proportion data, which was available only for HNA, BONCAT+, and PI+; LNA proportions were not included in the provided dataset. This data measured the proportion of the sample in which each cell type was present. These measurements are available for all fifteen mice who were split evenly among three experiments (cages).

4.2 Methodology

As the data are in proportions, we use a logistic transformation

$$Y_{cijk} = \log\left(\frac{p_{cijk}}{1 - p_{cijk}}\right)$$

where p_{cijk} is the relative abundance of cell type c for mouse j in experiment i on day k. Therefore $Y_{cikj} \in \mathbb{R}$.

In the model, we include two random effects: u_i, the cage effect and v_{ij}, the mouse effect. X_{ijk} is the covariate; since the perturbation is carried out on all mice, this will be the

day-based disease state. Thus, $\boldsymbol{\beta}$ is an estimated fixed effect. Note that these may also be specific to each cell-type; in Section 4.3, each cell type is modelled separately and thus we would more specifically index these as u_{ci}, v_{cij}, and $\boldsymbol{\beta}_c$.

For the model, we will initially assume that $u_i \sim N(0, \sigma_u^2)$ and $v_{ij} \sim N(0, \sigma_v^2)$. In the analysis of variance, we will also assume incidental error, $\varepsilon_{cijk} \sim N(0, \sigma^2)$. This assumes the variance of one value or one measurement is the sum of the variances of the cage effect, the mouse effect, and the error. Further, the correlation between measurements of the same cell type c taken on the same mouse is

$$\text{Cor}(Y_{cijk}, Y_{cijk'}) = \frac{\text{Cov}(Y_{cijk}, Y_{cijk'})}{\sqrt{\text{Var}(Y_{cijk})}\sqrt{\text{Var}(Y_{cijk'})}} = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma^2}, \ k \neq k'$$

and the correlation between two mice in the same cage is

$$\text{Cor}(Y_{ijk}, Y_{cij'k'}) = \frac{\text{Cov}(Y_{cijk}, Y_{cij'k'})}{\sqrt{\text{Var}(Y_{cijk})}\sqrt{\text{Var}(Y_{cij'k'})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma^2}, \ j \neq j', \ k = k' \text{ or } k \neq k'.$$

We note that cage 1 has fewer observations than cage 2 and 3; there is only one observed day in the pre-symptoms disease state and there are no data points for the baseline disease state. Experiment 1 was performed first, with experiments 2 and 3 running concurrently. Hence why the observations days vary between cage 1 compared to cages 2 and 3.

Using the BD FACSCantoII for flow cytometry, there were three daily HNA and PI+ measurements taken per mouse. BONCAT+ measurements came from using FACS Aria II, and thus there could only be one daily measurement per mouse. There are some equipment discrepancies in the HNA measurements for cage 1, where some were taken using Aria, resulting in only one daily measurement. This overall structure is described in Table 4.1.

	Number of available observations								
	HNA			PI+			BONCAT+		
Disease State\Cage	1	2	3	1	2	3	1	2	3
Baseline	0	60	36	0	30	60	0	14	14
Pre-symptoms	25	43	45	15	45	42	15	5	5
Symptoms	25	44	44	45	45	43	13	15	15
Recovery	50	60	57	75	57	59	25	10	10

Table 4.1: Data structure for cell proportions dataset. 'Baseline' includes experimental days -3 to 0, 'pre-symptoms' includes days 1 to 3, 'symptoms' includes days 4 to 6, and 'recovery' covers days 7 to 12.

We won't be imposing any method for handling these missing data points, as we have no reason to assume a specific non-random missingness structure. As it is on the outcome and missingness depends only on observed data, the missingness model can be ignored in the Bayesian paradigm. Furthermore, we focus on disease state as predictor rather than the specific day, so we ignore missing days so long as there are a reasonable number of data points for each disease state.

In the Bayesian setting, we set priors for the variance parameters according to Gelman

et al (2008) [10], i.e.

$$\sigma_u^2, \sigma_v^2, \sigma^2 \sim \text{Half-Cauchy}(0, 2.5\xi)$$

Here, ξ is the standard deviation of the residuals for the corresponding linear model (usually around 1). For β, we use an improper uniform prior, i.e. the whole real line is considered with equal probability, following Faraway (2016) [8].

We use MCMC sampling to get draws of our posterior samples for the fixed and random effects, as well as the estimated variance parameters. We use the Bayesian model to avoid the problem of 'boundary fits', where the point estimate for mixed effect variance (often σ_u^2, the top level in the hierarchical model) is estimated to be 0. This can often be a problem when i, the number of subjects at that level, is low, as is in our case with $i = 3$.

4.3 Research findings

4.3.1 Exploratory data analysis

For the proportions data, we focus on the first four disease states (baseline, pre-symptoms, symptoms, recovery).

We look at the mean abundance over time for each cage in Figures 4.1, 4.2, and 4.3. In the figures, the red line corresponds to cage 1, the green to cage 2 (female mice), and blue to cage 3. The different disease states are split with vertical lines. We also provide these means in Table 4.2.

	Mean value (logit scale)								
	HNA			PI+			BONCAT+		
Disease State\Cage	1	2	3	1	2	3	1	2	3
Baseline	-	-0.17	-0.18	-	1.68	1.11	-	-2.12	-2.47
Pre-Symptoms	0.054	0.091	0.0073	-0.14	2.09	1.71	-1.54	-1.48	-3.39
Symptoms	-0.30	-0.21	-0.34	-0.85	0.62	1.40	-2.28	-2.06	-3.77
Recovery	0.39	-0.0018	0.073	-0.28	2.19	2.10	-2.30	-1.90	-2.69

Table 4.2: Mean proportions for each cell type by disease state, given on the logit scale.

The trends for the abundance do not show much similarity between physiologies. However, there does to appear to be an upward trend from symptoms state to recovery state in most cases. We also see a common downward trend from 'pre-symptoms' to 'symptoms', then a rebound back up to 'recovery'. There is also clear cage variation, where cage 1 (in red) has much lower BONCAT+ values than the other two cages. For PI+, we see that cage 3 has lower values that the other two.

Since there doesn't appear to be a linear trend over time, there will be two main models under consideration: using the disease state as a covariate and dichotomizing the disease states into 'symptoms' and non-'symptoms'. This dichotomization will allow us to see if the 'symptoms' state does correspond to lower levels of HNA bacteria. We refer to the model with all four disease states as the Disease State model and the model with this dichotomization

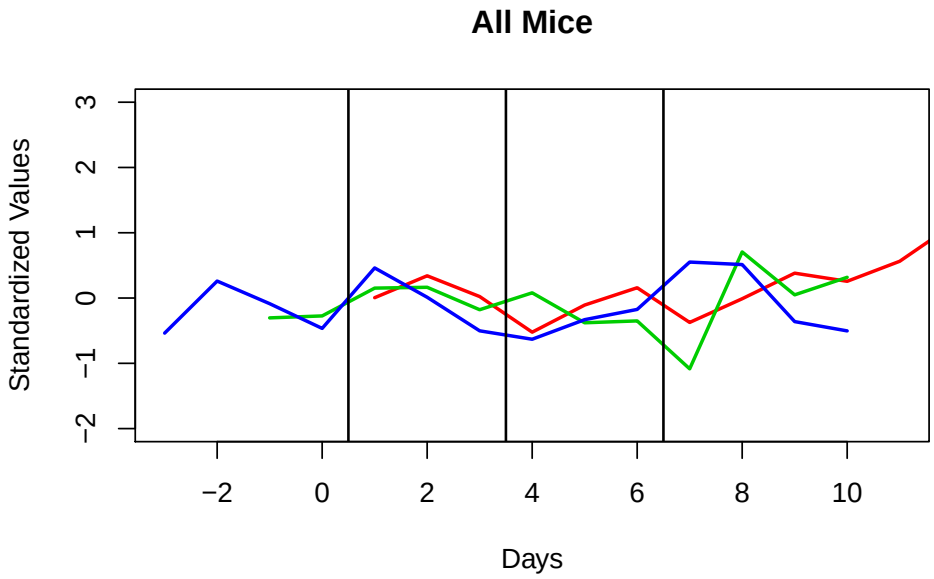


Figure 4.1: Plot of mean HNA abundance for each cage over time.

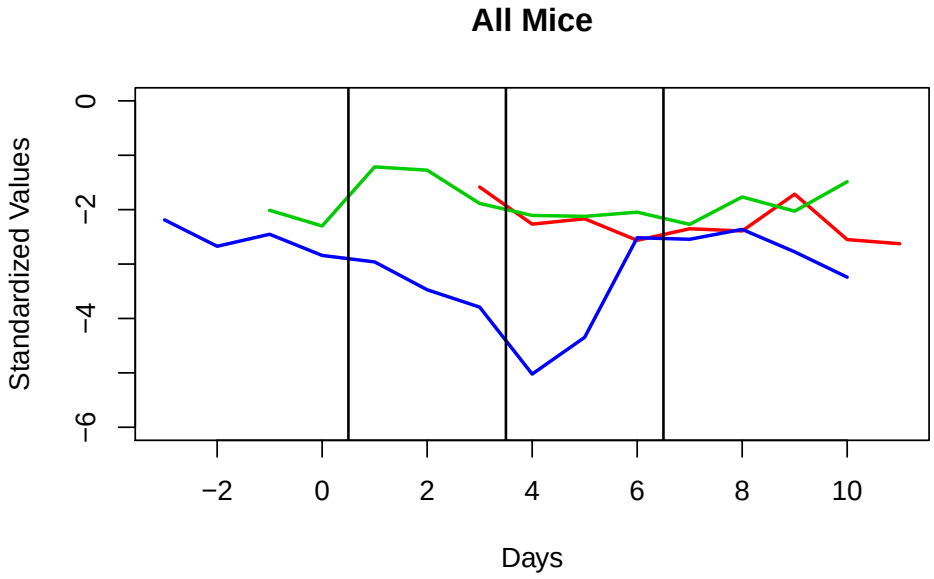


Figure 4.2: Plot of mean PI+ abundance for each cage over time.

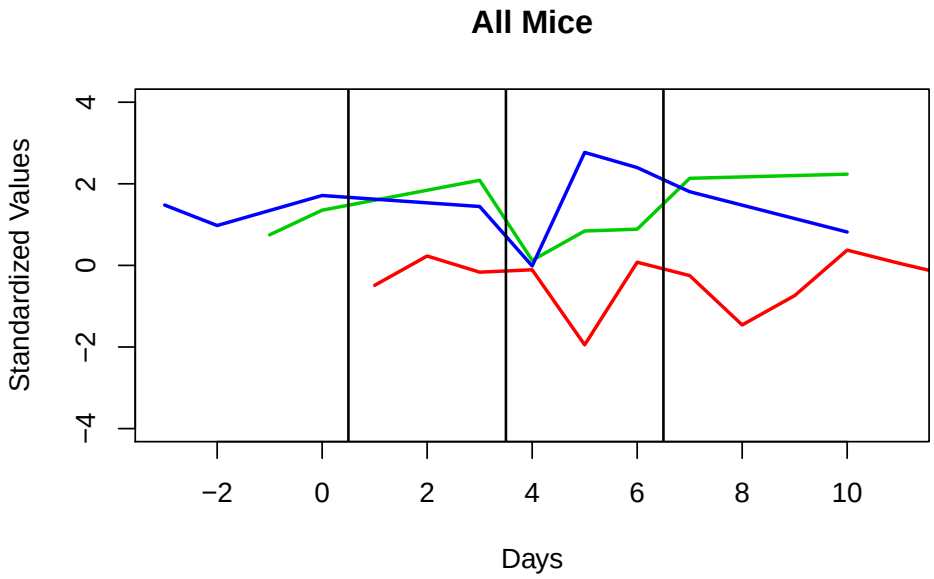


Figure 4.3: Plot of mean BONCAT+ abundance for each cage over time.

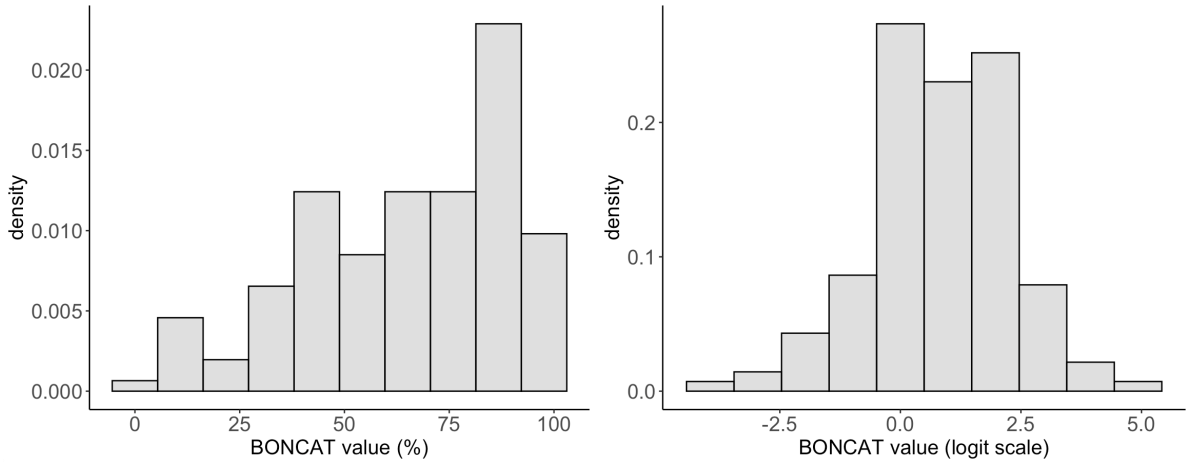


Figure 4.4: Histograms of the outcome (BONCAT+) before and after logit transformation

as the Symptoms model.

Assessing the overall distribution of each cell type using a histogram shows that normality seems to be obtained using the logit transformation, see Figure 4.4 which shows the case for BONCAT+ (similar for the other two cell types).

Variability in the dataset is visualized in box plots of the data. These are available in the Additional Figures section of the Appendix. It is difficult to visually discern whether there exists a difference in the proportions between each disease state, since values within a disease state can be quite varied.

Finally, we might want to assess correlation within the data and how that might inform the type of model we want to use. We've assumed correlation between mice in the same cage, as well as within the measurements taken from the same mouse. Table 4.3 shows the correlations between pre-symptoms, symptoms, and recovery for the three different cell types; baseline was not available for cage 1. In Figure 4.5, these correlations are calculated for all four disease states but only for cages 2 and 3. Because there are uneven observations in each disease state (see Table 4.1), there is no obvious way to measure correlations in the raw data and thus we aggregate by using the mouse/cage means.

For both PI+ and BONCAT+, there is high correlation between the different states but this could feasibly be linked to a within-mouse or within-cage correlation, which are accounted for in the random effects v_j and u_i, respectively. Interestingly, the correlations are much weaker for HNA; this may be because overall variation is smaller for HNA proportions.

	HNA		PI+		BONCAT+	
(Mouse)	Symptoms	Recovery	Symptoms	Recovery	Symptoms	Recovery
Pre-symptoms	0.068	-0.148	0.859	0.822	0.706	0.832
Symptoms	-	0.046	-	0.737	-	0.692
(Cage)	Symptoms	Recovery	Symptoms	Recovery	Symptoms	Recovery
Pre-symptoms	0.961	-0.109	0.996	0.867	0.898	0.990
Symptoms	-	-0.378	-	0.908	-	0.950

Table 4.3: Correlations between mean disease state values, averages aggregated by mouse (top) and then by cage (bottom).

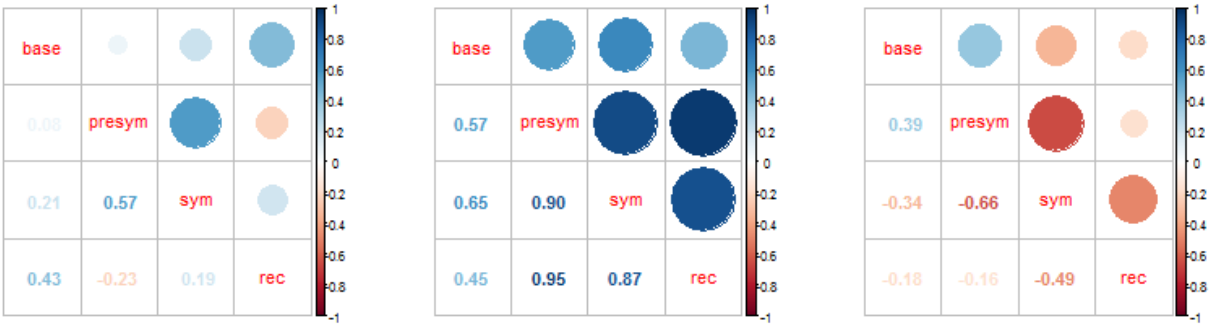


Figure 4.5: Correlations between mean disease state values for the latter two experiments, averages aggregated by mouse. From left to right, correlations are for proportions of HNA, PI+, and BONCAT+.

4.3.2 Results from the model

For all three cell types, a simple linear model was fitted first as reference. For all cell types, there is evidence of a relationship between predictor of 'Disease State' (in both the four category predictor and using the dichotomization). However, the linear model appears to be a poor fit based on the F-statistic and adjusted R-squared values.

Tables 4.4 and 4.5 provide estimates for the two linear models for HNA. Relative to the 'baseline' state, only 'recovery' is significant; this aligns with 'baseline' levels being similar to 'symptoms' and an increase in HNA levels during the 'recovery' state. In the 'Symptoms' model, we see that there is a significant drop during the 'symptoms' state. For PI+ and BONCAT+, we will also see a drop during the symptoms state. However, since we use 'baseline' as our reference in the disease state predictor, there isn't a consistent change between all cell types for the Disease State model.

Posterior results for HNA are provided in Tables 4.6 and 4.7. Again, we see similar values for the fixed effect estimates: when 'baseline' state is treated as the baseline, 'pre-symptom' and recovery states have slightly higher levels of HNA. 'Symptoms' has a slightly lower level, but not entirely non-zero for the 90% credible interval. Meanwhile, 'symptoms' is definitely linked to lower HNA abundance in the 'Symptoms' model.

Looking at the variance estimates, they are quite similar between both models. We find the cage and mouse variation to be very similar in value (around 0.3); however the cage variance itself has much greater variation. Again, day by day variation is still larger than between mouse or between cage variation. Trace plots show reasonable mixing, but we run

	Estimate	Std. Error	Lower C.I.	Upper C.I.	p-value
Intercept	-0.158	0.10	-0.355	0.040	0.117
Pre-symptoms	0.266	0.14	-0.003	0.535	0.053
Recovery	**0.357**	**0.13**	**0.109**	**0.604**	**0.005**
Symptoms	-0.111	0.14	-0.379	0.158	0.418

Table 4.4: Estimates for the Disease State linear model for HNA. Confidence interval estimates correspond to the 95% confidence interval. 'Baseline' is used as the baseline covariate value for disease state. Significant coefficients are bolded.

	Estimate	Std. Error	Lower C.I.	Upper C.I.	p-value
Intercept	0.081	0.05	-0.020	0.181	0.115
Symptoms	**-0.349**	**0.11**	**-0.557**	**-0.140**	**0.001**

Table 4.5: Estimates for the Symptoms linear model for HNA. Confidence interval estimates correspond to the 95% confidence interval. Significant coefficients are bolded.

into more autocorrelation in σ_u, the cage variation, as reflected in its much lower effective sample sizes. We also have non-zero mass at $\sigma_u = 0$, from divergent transitions.

To get better effective sample sizes and reduce autocorrelation, we performed the MCMC again with thinning, i.e. keeping only 1/5 of the draws. Thus, for four chains, we had 10,000 iterations with a burn-in of 5000. This still leaves us with 4000 total MCMC draws. This

thinning increased our effective sample size and did help reduce autocorrelation. That being said, the posterior estimates themselves were largely the same as in the un-thinned sampling.

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	-0.085	0.23	-0.476	0.278	283	1.02
Pre-symptom	**0.207**	**0.14**	**0.023**	**0.422**	662	1.01
Recovery	**0.265**	**0.12**	**0.058**	**0.471**	626	1.01
Symptoms	-0.177	0.13	-0.401	0.038	857	1
σ_u (Cage)	0.303	0.3	0.019	1.028	271	1.01
σ_v (Mouse)	0.38	0.1	0.246	0.557	1033	1
σ (Residual)	0.93	0.03	0.881	0.982	1346	1

Table 4.6: Posterior estimates for the Disease State linear mixed model for HNA, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

Posterior summaries for PI+ are provided in Tables 4.8 and 4.9. Similar to the HNA data, the 'symptoms' state is associated with lower levels of PI+. Note that the log transformation causes the intercept to be positive, so all other disease states are negative relative to the 'baseline' state. However, symptoms is most negative of all states.

Unlike in HNA, the cage variation is much greater than mouse variation and residual variation. Effective sample size is very low for σ_u with high autocorrelation. Residual/day

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	0.1165	0.23	-0.265	0.51	1062	1.01
Symptoms	**-0.357**	**0.10**	**-0.512**	**-0.194**	3631	1
σ_u (Cage)	0.32	0.32	0.021	0.928	795	1.01
σ_v (Mouse)	0.391	0.1	0.251	0.587	1051	1.01
σ (Residual)	0.932	0.03	0.886	0.982	2657	1

Table 4.7: Posterior estimates for the Symptoms linear mixed model for HNA, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

by day variation remains larger than between mouse variation. Note that the values of the estimates cannot be directly compared since the log was taken of the PI+ proportions. Thinning could be performed to improve chain mixing and reduce autocorrelation, but it still appears convergence was reached for the parameters.

Posterior summaries for BONCAT+ from the Bayesian analysis are provided in Tables 4.10 and 4.11. Similar to what we previously found, the 'symptoms' state is associated with reduced levels of the cell type. Like with the HNA data, 'pre-symptoms' and 'recovery' states have higher levels than the baseline, which is why the credible interval for 'symptoms' is mostly negative, but not entirely nonzero.

As expected from the exploratory analysis is Section 2.1, there is a fair bit of cage

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	2.29	0.88	0.9	3.325	41	1
Pre-symptom	**-0.223**	**0.09**	**-0.371**	**-0.077**	2326	1
Recovery	**-0.247**	**0.09**	**-0.391**	**-0.111**	407	1.01
Symptoms	**-0.620**	**0.09**	**-0.763**	**-0.475**	1914	1
σ_u (Cage)	0.993	0.63	0.322	2.348	129	1.02
σ_v (Mouse)	0.33	0.09	0.215	0.495	970	1
σ (Residual)	0.64	0.02	0.606	0.67	2072	1

Table 4.8: Posterior estimates for the Disease State linear mixed model for PI+, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

variation. As with the PI+ data, it appears that between cage variation is the largest source of variation. Between mouse variation is estimated to be very low for the BONCAT+ data. In the HNA data, σ_u was given a boundary fit in the likelihood-based setting, while we saw that in the Bayesian setting, it was estimated similarly to σ_v but with greater uncertainty. Now, σ_v has a boundary fit, but we now see a Bayesian posterior that is very close to 0.

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	**2.198**	**0.57**	**1.130**	**3.120**	1002	1
Symptoms	**-0.430**	**0.06**	**-0.538**	**-0.328**	1184	1
σ_u (Cage)	0.97	0.65	0.319	0.487	284	1.01
σ_v (Mouse)	0.329	0.09	0.217	0.487	1102	1
σ (Residual)	0.642	0.02	0.609	0.676	3756	1

Table 4.9: Posterior estimates for the Symptoms linear mixed model for PI+, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

4.3.3 Discussion

We can compare the estimated 'Symptoms' effect across the different cell types in Table 4.12. In all cases, the symptoms disease state is associated with lower levels of the cell type. All abundance values are taken on the logistic scale, $\log(\frac{x}{1-x})$, where x is between 0 and 1.

In the 'Disease State' model, the 'symptoms' state is also associated with lower levels of the cell type, but the 'baseline' state in the HNA and BONCAT+ data leads to the credible intervals overlapping slightly with 0.

We must mention the clear time ordering that exists between the disease states. Most importantly, the 'symptoms' state follows the 'pre-symptoms' state and precedes the 'recovery' state. We could consider disease state to be an ordinal covariate; however, it wouldn't

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	-0.121	0.86	-1.569	1.253	1005	1
Pre-symptom	**0.389**	**0.20**	**0.051**	**0.721**	2318	1
Recovery	**0.382**	**0.18**	**0.084**	**0.680**	2041	1
Symptoms	-0.246	0.17	-0.535	0.038	2423	1
σ_u (Cage)	1.445	0.80	0.569	3.15	1064	1
σ_v (Mouse)	0.088	0.07	0.006	0.221	2161	1
σ (Residual)	0.688	0.04	0.62	0.766	4022	1

Table 4.10: Posterior estimates for the Disease State linear mixed model for BONCAT+, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

be fully accurate to describe the states as having any sort of rank despite the time order. The descriptive statistics also didn't support a linear trend in time and therefore neither in disease state. Indeed, we expect cell counts to rebound after the 'symptoms' state, which would contradict any direct linear relation.

The first disease state 'baseline', which covers experimental days -3 to 0, was chosen as the baseline for the categorical predictor. While this seemed fitting, it did make it more difficult to obtain a significant effect for the 'symptoms' state as those cell counts tended to be closer to the ones in the 'baseline' state. If 'pre-symptoms' were to be chosen as a baseline, we

	Mean	SD	5th percentile	95th percentile	ESS	Rhat
Intercept	0.233	0.85	-1.159	1.689	814	1
Symptoms	**-0.502**	**0.13**	**-0.712**	**-0.293**	4783	1
σ_u (Cage)	1.444	0.84	0.552	3.109	1196	1
σ_v (Mouse)	0.089	0.07	0.007	0.215	2302	1
σ (Residual)	0.697	0.04	0.631	0.772	3371	1

Table 4.11: Posterior estimates for the Symptoms linear mixed model for BONCAT+, along with 90% credible interval, effect sample size, and Rhat (to indicate convergence). Non-zero fixed effects estimates are bolded.

	Mean	SD	2.5 percentile	97.5 percentile
HNA	-0.32	0.08	-0.49	-0.16
PI+	-0.46	0.07	-0.60	-0.32
BONCAT+	-0.68	0.19	-1.05	-0.32

Table 4.12: Estimated symptoms effect for the different cell types under the 'Symptoms' model.

would likely lose the significance of the 'recovery' state; the reverse would be true if 'recovery' were chosen. We might see a significant difference to the 'symptoms' state, in contrast. We will see that in future analysis, such as that done in Chapter 5, we select 'symptoms' as our

	Pre-symptoms	Recovery	Symptoms
HNA	0.19	**0.26**	-0.16
PI+	-0.20	-0.28	**-0.66**
BONCAT+	0.56	**0.58**	-0.30

Table 4.13: Estimated disease state effects for the different cell types under the 'Disease State' model, with 'baseline' state as the baseline. Covariates with non-zero credible intervals are bolded.

baseline state because we wish to contrast it with other states, in particular 'pre-symptoms' and 'recovery'. For the purposes of this analysis, using the Symptoms linear mixed model with the binary predictor allowed us to show that there was evidence of a drop in cell counts during the 'symptoms' state.

We also look at the estimated variance effects across the different cell types in Table 4.14 under the 'Symptoms' model. Table 4.15 shows these same proportions standardized by the sample standard deviation for each cell type.

In all cell types, residual variability (day to day) is greater than between mouse variability. However, cage variability is highly dependent on the cell type being measured. As expected from the mean plots in Figure 4.3, the cage effect is rather large in the BONCAT+ data, but small in the HNA data. Mouse variability is also estimated to be very close to 0 in the BONCAT+ data.

	σ_u (Cage)	σ_v (Mouse)	σ (Residual)
HNA	0.27 (0.02, 0.88)	0.33 (0.22, 0.48)	0.78 (0.74, 0.82)
PI+	0.93 (0.27, 2.38)	0.39 (0.24, 0.64)	0.70 (0.66, 0.75)
BONCAT+	1.93 (0.62, 4.85)	0.13 (0.01, 0.36)	1.02 (0.91, 1.15)

Table 4.14: Estimated variance effects for the different cell types under the 'Symptoms' model, along with the 90% credible interval.

	σ_u (Cage)	σ_v (Mouse)	σ (Residual)
HNA	0.32	0.39	0.93
PI+	0.98	0.41	0.74
BONCAT+	1.40	0.095	0.74

Table 4.15: Estimated variance effects for the different cell types under the 'Symptoms' model, standardized by the sample standard deviation of the proportion for each cell type.

Chapter 5

Application: 16S sequencing data

5.1 Introduction

The nature of sequencing data poses a challenge when trying to model it: we often see that it is over-dispersed yet very sparse and is high-dimensional and under-determined [28].

For the larger microbiome dataset, there are counts available for the latter two cages, i.e. a total of 10 mice. There are also 11 days of observations: days -2 to 0, 3 to 7, 10, 17, and 24. We will also consider the same disease state delineation as described in 2.1. Because of the large range of values, the log transformation is taken for all nonzero observations. Any 0 counts are kept at 0.

Note that there is only a single measurement of PI+ for all ten mice whereas there are mouse-specific measurements for the other three cell types. Since it is of interest to determine which cell-type is abundant in the bacteria over time, we take the average cell counts for all mice, see Figure 5.1 for reference.

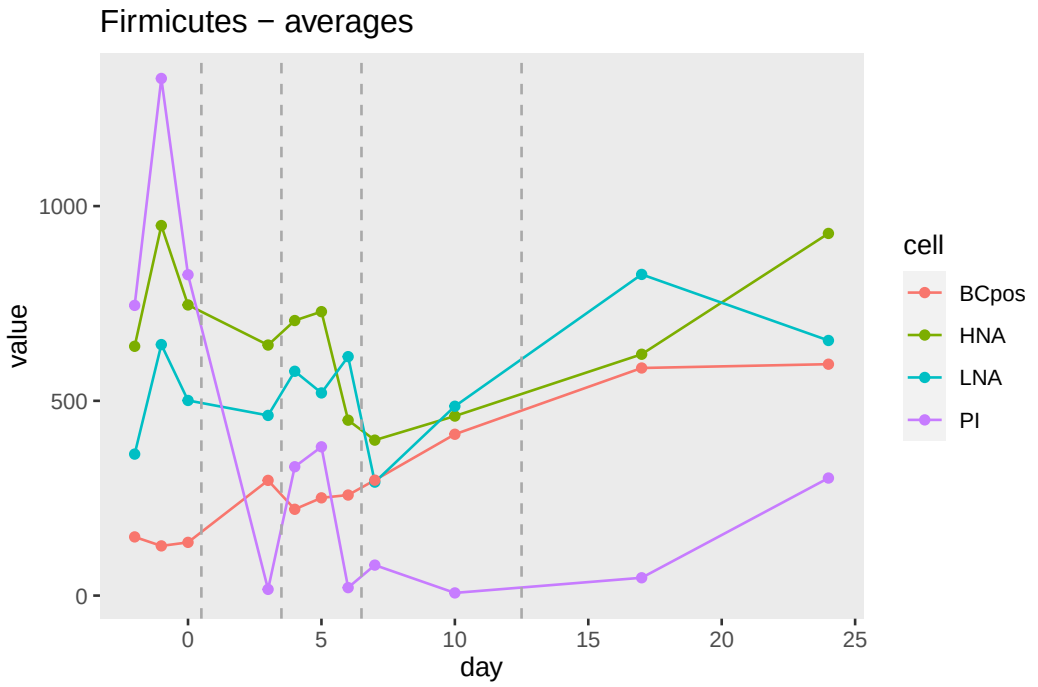


Figure 5.1: Aggregated log cell counts for *Firmicutes* bacteria, averaged over all mice. Grey dashed lines separate the disease states.

5.2 Univariate Gamma GLMM

5.2.1 Methodology

To stabilize some of the overdispersed and underdetermined nature of the sequencing data, we aggregate taxa using the sum of all daily data points for a physiology, still distinguishing by mouse.

We will be modelling the taxa individually, hence fitting the model to each taxon one at a time. While this won't be able to provide a mathematically holistic understanding of how

the cell types change over time, we can still discuss findings relevant to each taxon.

Taking the log of integer cell counts, $Y_{cjk} \in [0, \infty)$. Indexing follows the same notation as 2.2, where $c = 1, 2, 3, 4$ represents the different cell types, $j = 1, 2, ..., 10$ represents the mouse, and $k = 1, 2, ..., K_i$ represents the experimental day.

The model:

$$Y_{cjk} \sim \text{Gamma}(\alpha_{cjk}, \beta_{cjk})$$

$$\log(\mu_{cjk}) = X_{cjk}\boldsymbol{\beta} + v_j$$

$$v_j \sim \text{Normal}(0, \sigma_v^2)$$

$$\alpha_{cjk} = \mu_{cjk}^2/\phi$$

$$\beta_{cjk} = \mu_{cjk}/\phi$$

where v_j is the mouse random effect. With only two cages, we don't include a cage effect, but will include a fixed effect for cage (though there doesn't appear to be any evidence in the descriptive statistics that a cage variable is needed). Thus, there are two fixed effects in the model: β_1 which captures time dependence, and β_2 which is the cage effect.

We consider two methods of handling zero counts, which can still occur with the aggregation over the phyla. First would be to ignore these observations and second would be pad them by a very small value (i.e. 0.01). The second method would allow us to consider all observations, and represents a low count that didn't happen to be captured during measurement.

Because there is only a single PI+ observation for all mice for each OTU, we cannot use a mixed model for PI. Instead, we will have to use a regular gamma GLM.

Then, to consider the relationship between all four different cell types, we consider the Dirichlet distribution. From the four independent Gamma distributions [6], we have that

$$Y_c \sim \text{Gamma}(\alpha_c, \beta_c), \qquad i = 1, ..., 4$$

$$V = \sum_{c=1}^{4} Y_c$$

$$Z_c = Y_c / V$$

$$(Z_1, Z_2, Z_3, Z_4) \sim D(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$$

with $D(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ being the four-dimensional Dirichlet distribution.

In the Bayesian model, we assume the following priors for our model, following [8]:

$$\sigma_v^2, \phi \sim \text{Half-Cauchy}(0, 2.5)$$

and the improper uniform prior for β (i.e. the whole real line is considered with equal probability). For the MCMC draws, we use Stan [3]. We use four chains with 2000 iterations each, throwing out the first 1000 from each chain as the burn-in.

5.2.2 Research findings

For the researchers, there are six OTUs of interest that correspond to: *E. Coli, Akkermansia, B. Thetaiotamicron, Dubiosella Newyorkensis, Turicibacter*, and *Parasuturella*. We will also be looking at the five phyla (*Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria*, and *Verrumicrobia*), the twelve classes, and the ten most populous families.

Still looking at the averages, we sum all cell counts for a given day to determine the proportion of each cell type for a type of bacteria. Continuing with *Firmicutes*, we plot the

proportions in Figure 5.2. We can see that the proportion that PI+ makes up decreases dramatically after day 0. We also see that the BONCAT+ gradually takes up a larger proportion in *Firmicutes* over time. LNA also increases its share of the cell count after day 0, being approximately level with HNA by day 6.

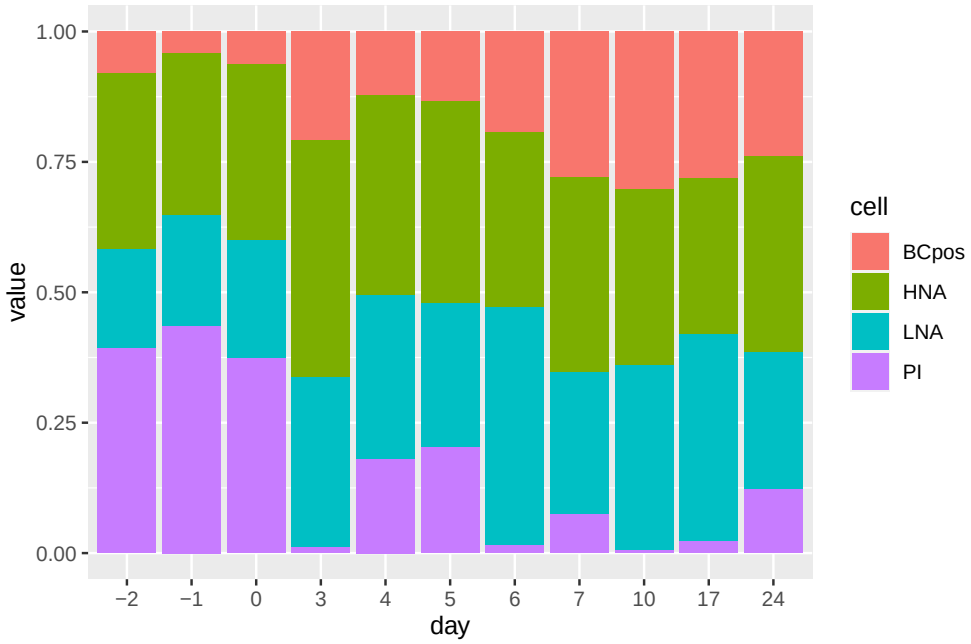


Figure 5.2: Proportions of each cell type for *Firmicutes* bacteria, averaged over all mice.

There were 7 different phyla observed over the 556 OTUs; however, *epsilonbacteraeota* only had 177 counts of which only 2 were nonzero, both on sample day 7. Figures 5.1 and 5.2 showed the patterns in abundance for *Firmicutes* bacteria; however, we do not necessarily see these same patterns in the different phyla. In Figure 5.3, we plot the different proportions of each cell type for the three most populous phyla, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*. In Figure 5.4, we visualize the most abundant cell type, with the proportion of the cell count that it makes up. For ease of reading, this is also done for only these three

most populous phyla.

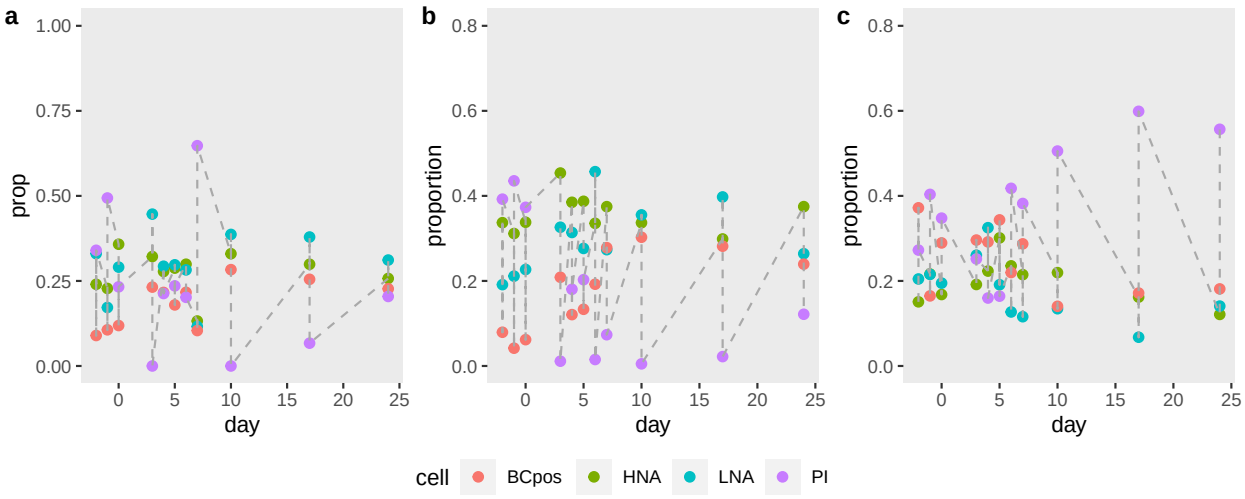


Figure 5.3: Proportions of each cell type for the three most populous phyla, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*, averaged over all mice.

It is difficult to visualize in a single figure the balance between all cell types for all six phyla. We often see levels of HNA, LNA, and BONCAT+ similar to each other, with PI+ levels varying the most for the different phyla. We are again interested in seeing if there is a certain time that helps determine when there are specific changes in these levels.

Without considering the mouse effect, we first see which of the covariates have evidence of a relationship with cell counts.

We find that for time (i.e. day) is a significant predictor for BONCAT+ and PI+. This is expected as we see an increase in BONCAT+ levels over time. For PI+, this is likely because of the high values during the baseline period. It should be noted that treating day as a linear predictor may be inaccurate as we don't expect changes to occur linearly with

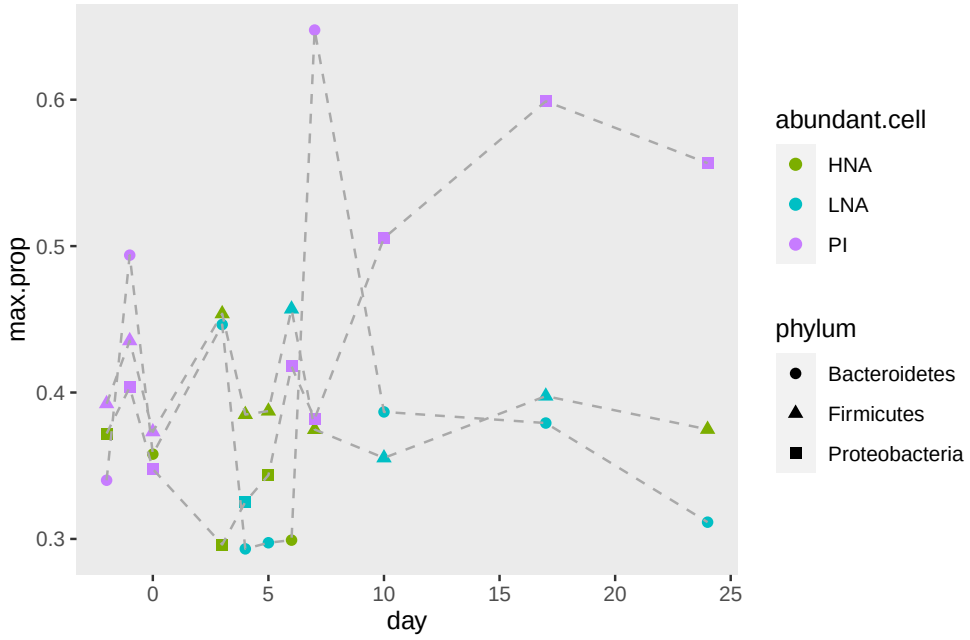


Figure 5.4: Cell type with the largest proportion of cell count for the three most populous phyla, *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*, averaged over all mice.

time, and would be affected by the large values of the follow-up periods, W1 and W2.

For disease state predictor, the 'recovery' state is significantly different than baseline for all cell types. However, this isn't too meaningful in itself, and the 'Symptoms' predictor is only significant in the case of PI+.

The cage has a significant effect for LNA and BONCAT+ in all three models.

We present the results for the *Firmicutes* bacteria with the Gamma GLMM in Tables 5.1, 5.2 and 5.3. Convergence seemed to be very good for these models, with Gelman-Rubin statistics all about 1 and large effective sample sizes.

For PI+, we can only consider the usual gamma GLM since there is a single measurement for all 10 mice. We still use the same Bayesian approach for fitting our model and the same

	Mean	SD	5%	50%	95%
Intercept	6.49	0.12	6.29	6.49	6.67
Day	-0.00	0.01	-0.02	-0.00	0.01
Cage	-0.03	0.19	-0.35	-0.04	0.29
σ_v^2	0.23	0.10	0.11	0.22	0.41
Intercept	6.46	0.16	6.21	6.47	6.70
Symptoms	0.06	0.11	-0.13	0.06	0.23
Cage	-0.03	0.20	-0.34	-0.04	0.27
σ_v^2	0.25	0.11	0.11	0.23	0.45

Table 5.1: Posterior estimates for the gamma GLMM for HNA in *Firmicutes* bacteria, along with 90% credible intervals, with zero counts removed. Non-zero fixed effects estimates are bolded.

priors without σ_v^2. The posterior results are presented in Table 5.4. Convergence was also quite good for this model.

Looking at the estimate mouse effects, we can see that mouse 8 has much lower values than the rest of the mice. This is shown in Figure 5.5, with a similar estimate for HNA and LNA. Otherwise, doesn't appear to be much mouse effect present. We do note that our simulation studies have shown a tendency to underestimate the random effect, though with little impact on the estimation of the fixed effects.

Overall, there doesn't appear to be any consistent trend in amongst all cell types. How-

	Mean	SD	5%	50%	95%
Intercept	6.32	0.15	6.07	6.32	6.55
Day	**0.02**	**0.01**	**0.01**	**0.02**	**0.03**
Cage 2	**-0.40**	**0.19**	**-0.70**	**-0.40**	**-0.10**
σ_v^2	0.26	0.10	0.13	0.24	0.45
Intercept	6.38	0.13	6.17	6.38	6.60
Symptoms	0.14	0.10	-0.03	0.14	0.30
Cage	**-0.33**	**0.18**	**-0.62**	**-0.33**	**-0.05**
σ_v^2	0.22	0.10	0.08	0.20	0.39

Table 5.2: Posterior estimates for the gamma GLMM for LNA *Firmicutes* bacteria, along with 90% credible intervals, with zero counts removed. Non-zero fixed effects estimates are bolded.

ever, this may be expected as we expect different cell types to be more or less abundant over time. We would then proceed by combining the four gamma distributions into a single Dirichlet and seeing if there is a time dependence.

When looking at *Proteobacteria* bacteria, the cage is now only significant for BONCAT+ and time/disease state is only significant for PI+. This aligns with what we can see in the average counts over time for *Proteobacteria* in Figure 5.6.

For the most part, mouse effects tend to be non-significant, except for mouse 8 which always has lower levels for all three cell types (recall that PI does not have unique values for

	Mean	SD	5%	50%	95%
Intercept	6.25	0.38	5.61	6.28	6.84
Day	**0.02**	**0.01**	**0.01**	**0.02**	**0.04**
Cage	**-0.44**	**0.23**	**-0.81**	**-0.44**	**-0.06**
σ_v^2	0.31	0.14	0.13	0.29	0.57
Intercept	5.75	0.14	5.52	5.75	5.98
Symptoms	0.03	0.11	-0.16	0.03	0.22
σ_v^2	0.18	0.10	0.04	0.17	0.37

Table 5.3: Posterior estimates for the gamma GLMM for BONCAT+ *Firmicutes* bacteria, along with 90% credible intervals, with zero counts removed. Non-zero fixed effects estimates are bolded.

each mouse). Mouse 7 has significantly lower levels of HNA, but average levels of the other two cell types.

We use Bayes factors [14] as a means of evaluating whether there is evidence to support a time dependence in the data, particularly as an alternative to the frequentist p-value. The Bayes factor is the ratio of the posterior odds of our hypothesis H_1 to its prior odds. They are provided in Table 5.5. Following Kass and Raftery, we will consider values of 3.2 to 10 to be substantial evidence, 10 to 100 to be strong, and over 100 to be decisive evidence.

Overall, there is plenty of evidence to suggest that there is a time dependence for the different bacterial physiologies in our OTUs of interest. In particular, HNA shows strong to

	Mean	SD	5%	50%	95%
Intercept	6.22	0.20	5.88	6.23	6.55
Day	**-0.03**	**0.01**	**-0.05**	**-0.03**	**-0.01**
Cage	0.00	0.10	-0.17	0.00	0.18
Symptoms	0.09	0.11	-0.10	0.09	0.27
Cage	0.00	0.10	-0.16	0.00	0.18

Table 5.4: Posterior estimates for the gamma GLM for PI+ *Firmicutes* bacteria, along with 90% credible intervals, with zero counts removed. Non-zero fixed effects estimates are bolded.

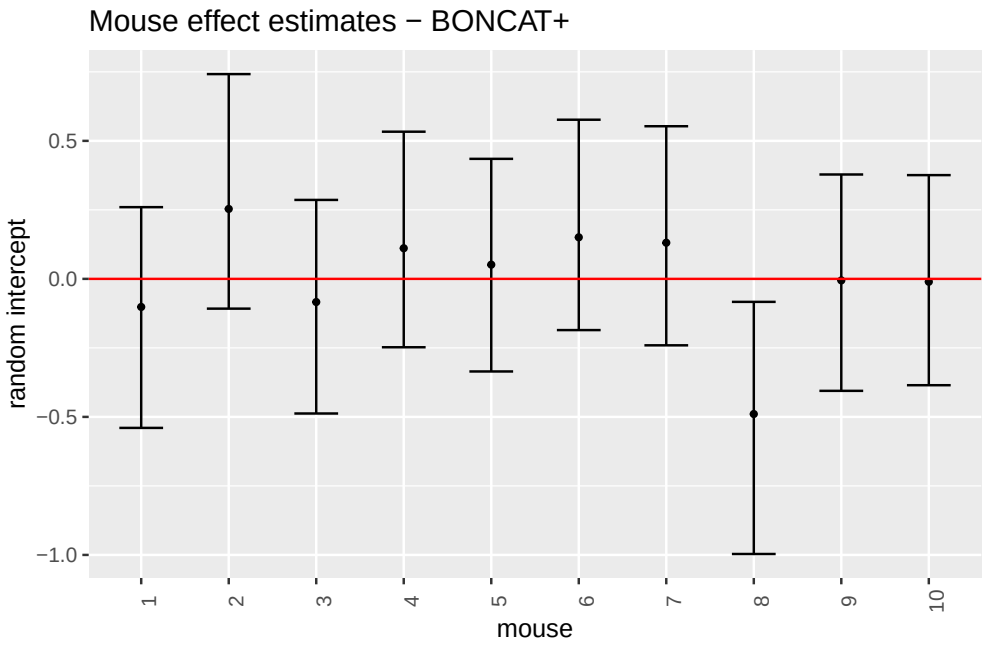


Figure 5.5: 90% confidence intervals for the estimated mouse effects for BONCAT+ *Firmicutes* bacteria.

	HNA	LNA	BONCAT	PI
E. Coli	**4398**	0.05299	62.59	**3.330** $\times 10^{6}$
Akkermansia	**8.309** $\times 10^{6}$	**1.412** $\times 10^{11}$	**1.159** $\times 10^{14}$	0.8112
B thetaiotaomicron	**3.158** $\times 10^{33}$	**1.403** $\times 10^{11}$	**1.187** $\times 10^{39}$	0.7838
Dubosiella newyorkensis	**4.647** $\times 10^{5}$	**1.111** $\times 10^{8}$	**4.245** $\times 10^{6}$	**7.010** $\times 10^{9}$
Turicibacter	0.03308	1.119	**8817**	0.2888
Muribaculaceae	20.21	**2.324** $\times 10^{4}$	0.1292	**1.293** $\times 10^{7}$
Eubacterium	18.82	0.9826	0.1830	6.575
Muribaculaceae	19.57	**2.330** $\times 10^{4}$	0.1385	**1.356** $\times 10^{7}$
Parasutterella	**9265**	**9.409** $\times 10^{7}$	3.193	0.2961

Table 5.5: Bayes factors for the day predictor for the OTUs of interest. Bolded values are those which have shown decisive evidence in favour of a time dependence.

decisive evidence of a time effect for almost all the bacteria except *Turicibacter*. It should be noted that there is evidence of time dependence for at least one physiology for all of the different OTUs; however, *Turicibacter* only has evidence of time dependence in BONCAT.

5.2.3 Discussion

Several models were fit using this dataset: one with only a time predictor (day), one using the disease states as a categorical predictor and one using the same binary 'symptoms' covariate from 4.3. These models were also all fit to five different phyla, each of the classes, the ten

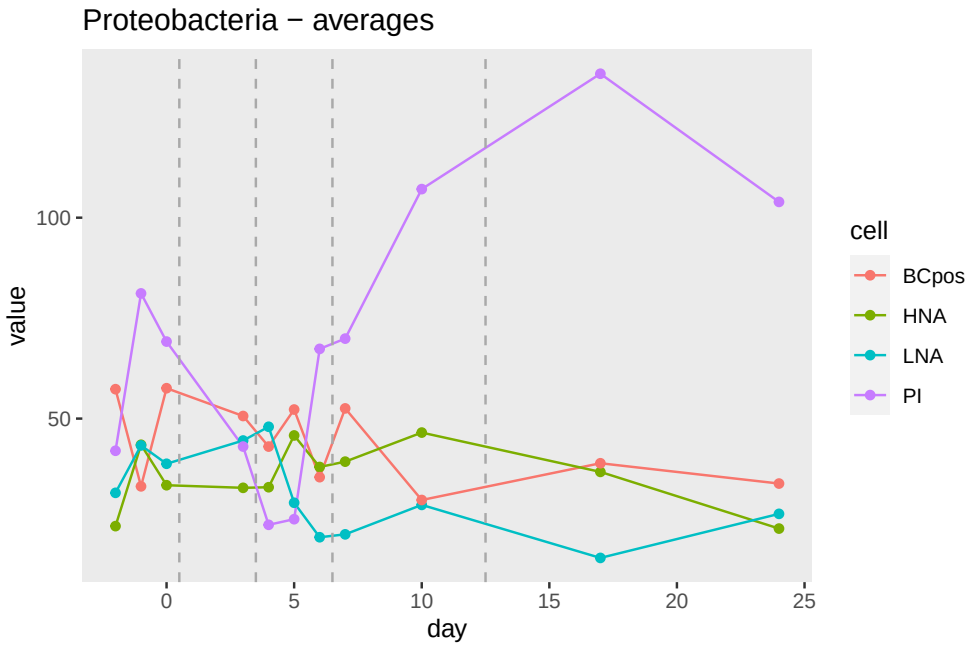


Figure 5.6: Aggregated log cell counts for *Proteobacteria* bacteria, averaged over all mice. Grey dashed lines separate the disease states.

most populous families, and eight OTUs. Hence, it would be tedious to include tables with all the estimated parameters. Furthermore, they likely wouldn't be very informative without a specific taxon we would want to analyze in greater detail.

As mentioned in 5.2.1, a disadvantage to this model is that we cannot easily compare results from the different taxa without direct comparison of the estimated quantities. We also encounter a lot of variability between the different taxa and no clear patterns amongst them. This makes it difficult to conclude much more than the fact that there exists some time-dependence structure in the data.

Ultimately, we are currently satisfied to simply identify the existence of some time dependence as it is sufficient for showing the dynamic nature of the gut microbiome. This also

aligns with what was found in Section 4.3.

5.3 Multinomial GLMM

While the previous methods allowed us to answer questions about how the individual physiologies changed over time, we still wanted a method to answer questions about how the abundances of cell types changed relative to each other. For instance, are certain taxa more likely to be abundant in HNA during the pre-symptoms disease state?

In order to assess these questions, we would require a model that dealt with the multivariate response.

5.3.1 Methodology

We consider a multinomial generalized linear mixed model, where the response is the vector $Y \in \mathbb{Z}^4$ containing the cell counts of HNA, LNA, BONCAT+, and PI+ for the OTUs of interest. Thus, our model is

$$Y_{jk} \sim \text{Multinomial}(\pi_{1jk}, ..., \pi_{4jk})$$

with π_{cjk}, $c = 1, ..., 4$ representing the proportion of cells being of each physiology and such that $\sum_{c=1}^{4} \pi_{cjk} = 1$. Again, $j = 1, 2, ..., 10$ represents each mouse and $k = 1, 2, ..., K$ represents the experimental day. We're going to omit the fixed cage effect used in Section 5.2, as this model already contains many parameters to be estimated and there was no evidence to include it in the previous model.

Following along with the GLMM set-up, we have that

$$\log\left(\frac{\pi_{cjk}}{\pi_{4jk}}\right) = X_{jk}\boldsymbol{\beta}_c + v_j$$

$$(\pi_{1jk}, ..., \pi_{4jk}) \sim D(\alpha_{1jk}, ..., \alpha_{4jk})$$

$$v_j \sim \text{Normal}(0, \sigma_v^2)$$

$$\sigma_v^2 \sim \text{Half-Cauchy}(0, 2.5)$$

where v_j is the mouse random effect and $D(\alpha_{1jk}, ..., \alpha_{4jk})$ is the four-dimensional Dirichlet distribution. Notice that this is similar to when we combined four independent Gamma distributions in Section 5.2.1.

For this data, we consider the disease state predictor, as we had done in the proportions data. In this case, the 'symptoms' state will be our baseline disease state. Figure 5.7 shows the proportions averaged over all mice for each day, as well as averaged for each disease state. We won't be using the time predictor, since there does not appear to be any linear trend. In this analysis, we will focus on just the six OTUs the researchers chose to study: *E. Coli*, *Akkermansia*, *B. thetaiotaomicron*, *Dubiosella newyorkensis*, *Turicibacter*, and *Parasutterella*.

To understand the meaning of the coefficients $\boldsymbol{\beta}$ that are being estimated, we return to the perspective of looking at cell counts. Another way of viewing the multinomial is if we think of the multinomial being built from a conditional Poisson distribution. Let μ_c be the mean count for the c-th cell type. Ignoring indexing on X for simplicity, we could also define:

$$\log \mu_c(X) = \beta_0^c + \beta_1^c x_1 + ... + \beta_1^c x_4 = X\boldsymbol{\beta}_c$$

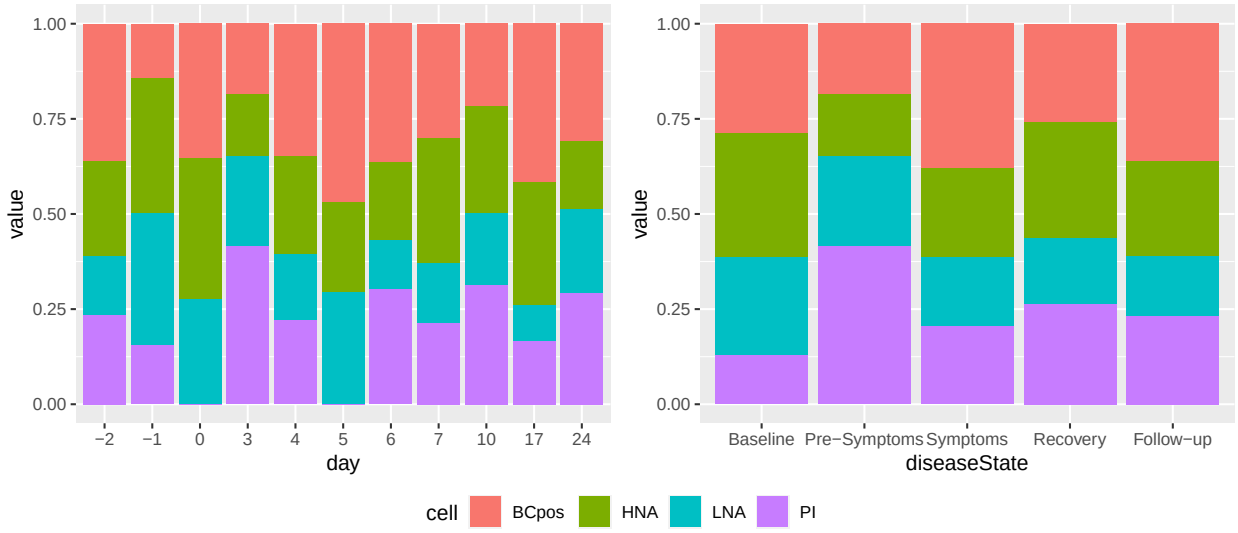


Figure 5.7: Proportions of each cell type for *E. Coli* bacteria, averaged over all mice.

where $x_i = 1, ..., 4$ is the indicator variable for the disease states baseline, pre-symptoms, recovery, and follow-up, respectively.

For identifiability, $\boldsymbol{\beta}_4 = 0$, i.e. we set BONCAT+ as the baseline cell type. The conditional distribution that yields the log-likelihood is a multinomial distribution where we get our probabilities

$$\pi_c(X) = \frac{e^{X\boldsymbol{\beta}_c}}{\sum_{l=1}^{4} e^{X\boldsymbol{\beta}_l}},$$

with indexing by mouse and experimental day left out for ease of reading. Since $\boldsymbol{\beta}_4 = 0$, we have that

$$\pi_4(X) = \frac{1}{1 + \sum_{l=1}^{3} e^{X\boldsymbol{\beta}_l}}, \qquad \pi_c(X) = \frac{e^{X\boldsymbol{\beta}_c}}{\sum_{l=1}^{3} e^{X\boldsymbol{\beta}_l}}, \quad c = 1, 2, 3$$

For the 'symptoms' disease state, $X\boldsymbol{\beta}_c = \beta_0^c$. For all remaining disease states $i = 1, ..., 4$, $X\boldsymbol{\beta}_c = \beta_0^c + \beta_1^c x_i$.

5.3.2 Research findings

Of interest are the odds ratios, i.e. how much more likely a cell is to be a specific physiology over another physiology. In Table 5.6, we include the odds ratios for the different physiologies in *E. coli*. For ease of reading, the rest of these tables for the remaining OTUs are included in the Appendix.

We consider the 'symptoms' state to be our baseline. Thus in the table, we can interpret 0.622 (in the first row, first column of values) to mean that HNA is 0.622 times more likely to be abundant than BONCAT+ during the symptoms period. Meanwhile, we can read 1.632 (in the second row, third column of values) as the odds of an *E. coli* bacteria being LNA is 1.632 times the odds of the bacteria being BONCAT+ during the pre-symptoms state. The bacteria 'being LNA' refers to LNA being the dominant cell type, if we consider a strict classification based on cell abundances.

5.3.3 Discussion

Overall, we can see that there are many significant differences in the odds of a bacteria being more dominant in one physiology over another. That is to say there is evidence that a bacteria is more likely to have proportions of one cell type than another.

There aren't many overall trends that we can grasp just by looking at the six OTUs. We must also consider two baselines concurrently: the disease state baseline as well as the baseline physiology in our odds ratio. It might be useful to find some method that allows us

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	**0.622**	**1.132**	0.897	**1.194**	0.690
LNA	**0.487**	**1.152**	**1.632**	**0.850**	0.557
PI	**0.546**	0.516	**2.556**	**1.141**	0.737
Odds ratio			(relative to HNA)		
LNA	**0.782**	0.796	**1.423**	**0.557**	0.631
PI	0.877	**0.400**	**2.500**	0.838	0.937
Odds ratio			(relative to LNA)		
PI	1.121	**0.502**	1.757	**1.505**	1.485

Table 5.6: Multinomial GLMM odds ratios for *E. coli* bacteria. Bolded values are nonzero at the 90% credible level.

to harmonize all these results, beyond the specific disease state and physiology the research would have to focus on.

This model was very difficult and slow to run using Bayesian inference, especially when including the random effects. None of the random effects had non-zero value, on the 95% credible level. There was also much autocorrelation in the mouse effect.

Chapter 6

Discussion

6.1 On the research questions

The main interest of the researchers was determining whether the bacterial abundances changed over time, particularly between different physiological states throughout the progression of the disease. Through the generalized linear mixed models, we were able to characterize how the bacterial OTU changes between physiological fractions with time. We saw evidence that there were changes in abundances between physiological states. For instance, *Turicibacter* changed in BONCAT+ without changing in HNA or LNA.

We were also able to observe which of the physiologies/cell types was most dominant for the different physiological periods. In the onset of colitis, BONCAT increased moving from PI-dominant bacteria changing to BONCAT-dominant physiology.

Using the proportions data set, we were able to compare the different disease states and whether there was a change in cell proportions between the different states. Of interest was whether this change in proportions was specific to each cell type. Amongst all cell types, we found a drop in proportions during the symptoms state, which we could characterize using

the dichotomized disease state variable.

It should be noted that the sequencing data, which dealt with nine different OTUs, was too variable day-by-day to provide any clear results regarding disease states. The analysis for the sequencing data instead focussed on identifying a time-dependence for each cell type. The highly variable nature of this data made it quite difficult to decide on an appropriate predictor to consider. We saw definite peaks around the symptoms state, i.e. days 4-6 of the experiment; however, the specific day of the peak varied for each cell type and OTU. As we needed to keep it consistent among all the physiologies, we couldn't capture this peak appropriately with any one day. Choosing instead to group days by disease state, as done by the proportions data, also did not provide useful results as the groupings tended to neutralize any of the sharp peaks we saw in the raw data. Thus, just a general time variable was all we could consider.

The logistic regression model from 4.3 and the multinomial model from 5.3 both considered disease state as a categorical predictor. Therefore, one of the states needed to be set as the baseline in the model; this was 'baseline' in 4.3 and 'symptoms' in 5.3. There is more specific discussion in 4.3.3 about the possible impact of these choices. After becoming more familiar with the dataset over time, a more informed decision was made to choose 'symptoms' as the baseline in 5.3. This is because it was of greater interest for the researchers to contrast cell abundances in the different states with the 'symptoms' state. While we might want to revisit the disease state model with different baselines, specifically 'symptoms', we found that the symptoms model with the dichotomized disease state was sufficient to describe the

changes in cell behaviour.

That being said, we found that the diversity is very dynamic over time, as shown by the large within mouse variation relative to the variation between mice or between cage, of which there was no strong evidence. In fact, rarely were the random effects significant when looking at the 90% credible intervals. However, it must be noted that these GLMMs have a tendency to underestimate the true random effect, as shown in our simulation study in Section 3.3. That being said, the simulation study showed that, despite the random effects being underestimated, there was not much impact on the estimation of the fixed effect which is of more interest to the researchers.

6.2 Future work

In order to handle the zero-inflation of the sequencing data, we aggregated different observations to those of the same class, phyla, etc. It might be worth exploring a more direct method of handling the zeros, such as using a Beta-Binomial model.

There is also interest for the researchers to looking into general trends among different bacteria. In our model in Section 5, we tended to look at each OTU (or phylum) individually and make conclusions on an individual basis. However, it would be worth considering different OTUs simultaneously to get a sense of the overall gut microbiome. In particular, this would make it much easier to get a sense of the results, rather than having to parse through each set of estimated parameters that come from each OTU's model.

Chapter 7

Conclusion

The aim of this research was to determine whether there was statistical evidence of changes in physiology during the different biological disease states. We also wanted to see how well a generalized mixed model performed for the data which included both proportions and standardized counts.

With the proportions data, where different cell types were seen as a proportion of the entire gut microbiome bacterial population, we were able to say that there was indeed evidence of change within the bacterial physiology for the different disease states. This data set used a logistic mixed regression model with random effects added for the cage and mouse. It was found that during the symptoms state, there was a significant drop in cell count amongst all cell types, relative to the cell counts in the other states. However, it was more difficult to parse out the other states, such as finding a difference between the recovery and the pre-symptoms state. However, as focus was on the symptoms state, these results were sufficient for the biological analysis.

In the sequencing data, we looked at nine bacterial OTUs and their standardized cell counts, which are more reflect of the raw counts in contrast to the default relative abundances.

The data were highly variable between different OTUs and for each day. Grouping days by disease state, as done with the proportions data, tended to neutralize any of the peaks in cell count that we wanted to characterize. Instead, the analysis focussed on a general time-dependence and we were able to conclude that there was change in abundances over time for all the OTUs.

The generalized mixed models used were usually able to capture the changes in the data, though the random effects themselves were usually not significant. In fact, time-variation was usually much larger than any variation between cages or between mice. In the future, we would be interested in looking at how we might better handle this highly variable dataset, which was also rife with systematic zeros. This might be done using a Beta-Binomial model, for instance.

In our analysis of the data, we were indeed able to find evidence of change in the physiology during the different biological disease states.

Bibliography

[1] M. Arumugam, , J. Raes, E. Pelletier, D. L. Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346): 174–180, 2011. doi: https://doi.org/10.1038/nature09944.

[2] S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Sept. 2014.

[3] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[4] E. Z. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 2016. doi: https://doi.org/10.1093/bioinformatics/btw308.

[5] J. Chen and H. Li. Variable selection for sparse Dirichlet-multinomial regression with

an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1), Mar. 2013. doi: 10.1214/12-aoas592. URL https://doi.org/10.1214/12-aoas592.

[6] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlage, 1986.

[7] R. Fang, B. Wagner, J. Harris, and S. Fillon. Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiology and Infection*, 144(11):2447–2455, Apr. 2016. doi: 10.1017/s0950268816000662. URL https://doi.org/10.1017/s0950268816000662.

[8] J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman &; Hall/CRC, 2016.

[9] G. Galazzo, N. van Best, B. J. Benedikter, K. Janssen, L. Bervoets, C. Driessen, M. Oomen, M. Lucchesi, P. H. van Eijck, H. E. F. Becker, M. W. Hornef, P. H. Savelkoul, F. R. M. Stassen, P. F. Wolffs, and J. Penders. How to count our microbes? The effect of different quantitative microbiome profiling approaches. *Frontiers in Cellular and Infection Microbiology*, 10:403, 2020. ISSN 2235-2988. doi: 10.3389/fcimb.2020.00403. URL https://www.frontiersin.org/article/10.3389/fcimb.2020.00403.

[10] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360 – 1383, 2008. doi: 10.1214/08-AOAS191. URL https://doi.org/10.1214/08-AOAS191.

[11] Harvard T.H. Chan School of Public Health. The microbiome, 2020. https://www.hsph.harvard.edu/nutritionsource/microbiome/, Last accessed on 2020-11-30.

[12] C. Jian, P. Luukkonen, H. Yki-Järvinen, A. Salonen, and K. Korpela. Quantitative pcr provides a simple and accessible method for quantitative microbiota profiling. *PLoS ONE*, 15(1), 2020. doi: https://doi.org/10.1016/j.chom.2018.09.009.

[13] S. Jiang, G. Xiao, A. Y. Koh, J. Kim, Q. Li, and X. Zhan. A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics*, 22(3):522–540, Dec. 2019. doi: 10.1093/biostatistics/kxz050. URL https://doi.org/10.1093/biostatistics/kxz050.

[14] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773 – 795, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291091.

[15] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, May 2018. doi: https://doi.org/10.1038/s41579-018-0029-9.

[16] J. Lederberg and A. McCray. 'ome sweet 'omics - a genealogical treasury of words. *Scientist*, 15(7), 2001.

[17] L. Liu and M. T. Özsu, editors. *Encyclopedia of Database Systems*. Springer Reference. Springer, New York, 2009. ISBN 978-0-387-35544-3. doi: 10.1007/978-0-387-39940-9.

[18] B. D. Martin, D. Witten, and A. D. Willis. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1), 2020. doi: https://doi.org/10.1214/19-aoas1283.

[19] J.-A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, Sept. 2014. doi: 10.1177/1471082x14535524. URL `https://doi.org/10.1177/1471082x14535524`.

[20] National Institute of Diabetes and Digestive and Kidney Diseases. Ulcerative colitis, 2014. `https://www.niddk.nih.gov/health-information/digestive-diseases/ulcerative-colitis`, Last accessed on 2020-11-30.

[21] P. S. L. Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, 7(12):e52078, Dec. 2012. doi: 10.1371/journal.pone.0052078. URL `https://doi.org/10.1371/journal.pone.0052078`.

[22] K. Sankaran and S. P. Holmes. Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614, June 2018. doi: 10.1093/biostatistics/kxy018. URL `https://doi.org/10.1093/biostatistics/kxy018`.

[23] M. Schirmer, L. Denson, H. Vlamakis, E. Franzosa, S. Thomas, N. Gotman, P. Rufo, S. Baker, C. Sauer, J. Markowitz, M. Pfefferkorn, M. Oliva-Hemker, J. Rosh, M. Otley, B. Boyle, D. Mack, R. Baldassano, D. Keljo, N. LeLeiko, and R. Xavier. Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. *Cell Host & Microbe*, 24(4):600 – 610, 2018.

[24] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32):12115–12120, July 2006. doi: https://doi.org/10.1073/pnas.0605127103.

[25] M. Taguer, E. Darbinian, K. Wark, A. Ter-Cheam, D. Stephens, and C. Maurice. Changes in gut bacterial translation occur before symptom onset and dysbiosis in dextran sodium sulfate-induced murine colitis. *mSystems*, 2021. Preprint: accepted for publishing.

[26] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, R. Titol, L. De Commer, Y. Darzi, S. Vermeire, G. Falony, and J. Raes. Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551: 507–511, 2017.

[27] S. Vieira-Silva, J. Sabino, M. Valles-Colomer, G. Falony, G. Kathagen, C. Caenepeel, I. Cleynen, S. van der Merwe, S. Vermeire, and J. Raes. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota

alterations across PSC/IBD diagnoses. *Nature Microbiology*, 4(11):1826–1831, 2019. doi: https://doi.org/10.1038/s41564-019-0483-9.

[28] Y. Xia, J. Sun, and D. Chen. *Statistical analysis of microbiome data with R*. Springer, 2018.

[29] X. Zhang, H. Mallick, Z. Tang, L. Zhang, X. Cui, A. K. Benson, and N. Yi. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1), Jan. 2017. doi: 10.1186/s12859-016-1441-7. URL `https://doi.org/10.1186/s12859-016-1441-7`.

Appendix A

Multinomial tables from Section 5.3.2.

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	0.972	**1.383**	**0.629**	0.976	0.912
LNA	0.998	**1.658**	0.831	1.029	1.118
PI	**0.631**	**2.666**	**0.635**	1.101	**0.453**
Odds ratio			(relative to HNA)		
LNA	1.026	1.230	1.355	1.081	1.258
PI	**0.649**	**1.250**	0.654	0.732	**0.323**
Odds ratio			(relative to LNA)		
PI	**0.632**	**1.016**	**0.483**	0.677	0.256

Table A.1: Multinomial GLMM odds ratios for *Akkermansia* bacteria. Bolded values are nonzero at the 90% credible level.

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	**1.284**	**2.448**	1.072	1.117	1.196
LNA	**1.318**	**2.652**	1.279	1.108	1.248
PI	**0.822**	**3.450**	**0.875**	**0.770**	**0.511**
Odds ratio			(relative to HNA)		
LNA	1.026	1.111	1.224	1.017	1.070
PI	**0.640**	**0.902**	**0.522**	0.441	**0.273**
Odds ratio			(relative to LNA)		
PI	**0.624**	**0.811**	**0.427**	**0.433**	**0.255**

Table A.2: Multinomial GLMM odds ratios for *B. thetaiotaomicron* bacteria. Bolded values are nonzero at the 90% credible level.

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	**1.171**	**1.793**	0.910	1.088	**0.868**
LNA	**1.300**	**1.493**	1.068	1.019	**0.943**
PI	**0.716**	**2.790**	0.854	1.347	1.048
Odds ratio			(relative to HNA)		
LNA	1.111	0.925	1.303	1.040	1.207
PI	**0.612**	**0.952**	0.574	**0.758**	0.738
Odds ratio			(relative to LNA)		
PI	**0.551**	**1.029**	0.440	0.728	**0.612**

Table A.3: Multinomial GLMM odds ratios for *Dubosiella newyorkensis* bacteria. Bolded values are nonzero at the 90% credible level.

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	1.173	**0.791**	**0.595**	**0.794**	**0.742**
LNA	**1.500**	**0.706**	**0.588**	**0.614**	**0.708**
PI	**1.250**	**0.748**	**0.460**	**0.614**	**0.807**
Odds ratio			(relative to HNA)		
LNA	**1.279**	1.143	1.264	0.988	1.220
PI	1.066	1.008	0.824	0.824	1.159
Odds ratio			(relative to LNA)		
PI	**0.833**	0.882	0.652	0.833	0.950

Table A.4: Multinomial GLMM odds ratios for *Turicibacter* bacteria. Bolded values are nonzero at the 90% credible level.

	(Intercept)	Baseline	Pre-Symptoms	Recovery	Follow-up
Odds ratio			(relative to BONCAT+)		
HNA	**2.647**	2.583	1.769	**1.206**	**0.979**
LNA	**2.490**	**6.954**	**8.995**	**1.459**	**1.598**
PI	1.275	**11.684**	**7.190**	**1.269**	**0.944**
Odds ratio			(relative to HNA)		
LNA	0.941	**2.532**	**4.782**	**1.138**	1.536
PI	**0.481**	**2.177**	1.956	**0.507**	0.464
Odds ratio			(relative to LNA)		
PI	**0.512**	**0.860**	**0.409**	0.445	0.302

Table A.5: Multinomial GLMM odds ratios for *Parasutturella* bacteria. Bolded values are nonzero at the 90% credible level.

Appendix B

Additional figures

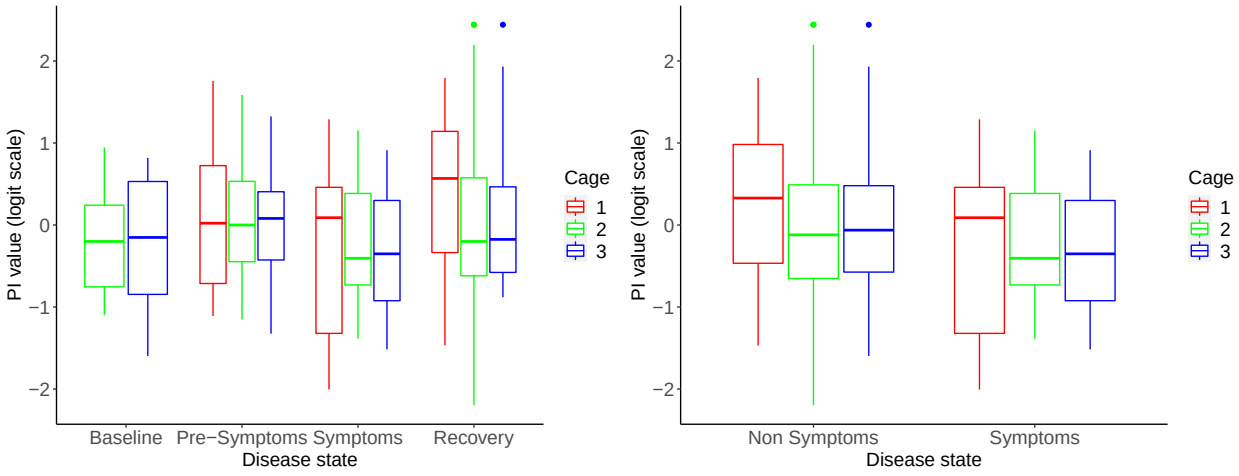


Figure B.1: Additional figure from 4.3.1: Box plot of distributions of the HNA value per cage and per disease state.

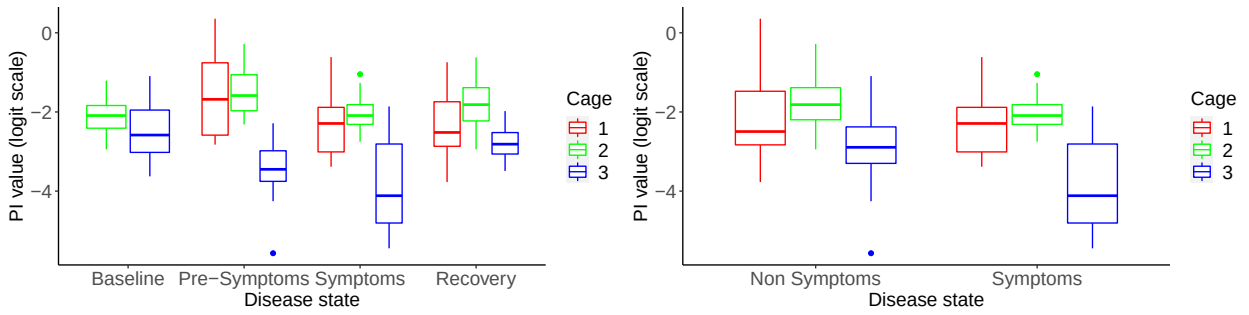


Figure B.2: Additional figure from 4.3.1: Box plot of distributions of the PI+ value per cage and per disease state.

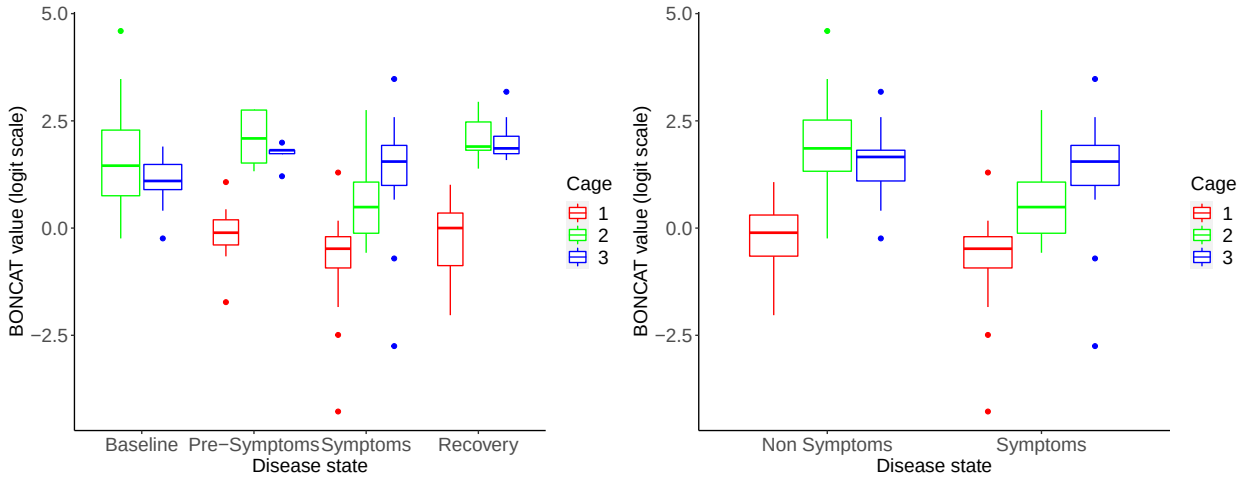


Figure B.3: Additional figure from 4.3.1: Box plot of distributions of the BONCAT+ value per cage and per disease state.