Surrogacy and Biomarker Detection in High Dimensional Data

Rudradev Sengupta

Promotor: Prof. dr. Ziv Shkedy

Acknowledgments

The four year-long epic journey towards the completion of this thesis would have been impossible without the support and guidance of several people and organizations. I therefore would like to offer my sincere gratitudes to all of them. The list here is just a brief and incomplete summary.

My utmost respect and heartfelt gratitude are long due to my promoter Prof. dr. Ziv Shkedy for his continuous guidance, without which this thesis would not have seen the light of the day. Especially, I am very grateful for him being extremely encouraging, patient, enthusiastic and generous to me with his time and ideas - always available to discuss and give feedbacks (not only academic, but also about other aspects of life) throughout last four years. I am very happy to get the chance to continue collaborating with him after finishing my PhD.

I am grateful to Prof. dr. Geert Molenberghs and Prof. dr. Tomasz Burzykowski for reading the thesis thoroughly and providing valuable feedback. I would also like to thank other members of the jury for their time to be in my committee and giving valuable comments which helped to improve this thesis. I feel privileged for collaborating with Janssen Pharmaceutica on several projects and I am thankful to Luc Bijnens for giving me this opportunity by involving me in the "ExaScience Project" and the "Microbiome Project". My visits to Beerse have always been rewarding and productive in one way or the other.

I am really glad to come across so many nice people over the past four years as my colleagues in CenStat - thank you for having me around and being part of JOSS helped me to integrate as well. Thanks to my old (Nolen and Martin) and current (Thao, Alvaro, Evangelina) officemates for making me smile everyday and for our (not so small) chit-chats. Thank you Marijke for all the help - be it about work or reminding me about Doctoral School sessions or about other small, but important things like translating

documents, you always helped me out. To others (my research circle: Theophile, Leacky, Ewoud, Dea, Tadesse, Belay and the "lunch" group: Sammy, Kathy, Tanya, Chella, Yimer, Daniel, Annelies, Thang, Oana, Joris, Mohammed, Trishanta) - it has been a great journey. I will always remember our intellectual and witty conversations during "break"-time and hopefully this is just the beginning of a long lasting friendship. My special thanks are extended to Martine, the most efficient secretariat I have ever come across. Thanks for being the backbone of CenStat and making our lives easier.

I would like to thank all my friends (Sayar, Soudeep, Ananya, Subhabrata, Sayak, Arka, Rounak, Chinmoy, Avijit, Sayantan, Sagnik, Joydeep to name a few), scattered around the world, for always being there irrespective of the challenges of living in different time-zones. Thanks to our vibrant chats/hangouts that helped me survive the last four years and made this journey easier for me. Cheers to ISI and to a decade of friendship. I owe a great debt to Rajarshi, Tamal, Avisek, Saswati, Arkajyoti, Basundhara, Deepan and Sreyasi for being the family away from home. Friends are indeed the family we choose to make.

Lastly, I am eternally indebted to my family for being really supportive throughout my life and it is needless to say that my journey so far would not have been possible without the sacrifices they made for me. Thank you *Maa*, *Baba*, *Didi*, *Rani* & *Angel* for everything.

Rudradev Sengupta
Hasselt, 6 June, 2018

List of Publications

Journal Articles:

- Otava, M., Sengupta, R., Shkedy, Z., Lin, D., Pramana, S., Verbeke, T., Haldermans, P., Hothorn, L.A., Gerhard, D., Kuiper, R.M., Klinglmueller, F. and Kasim, A., (2017). IsoGeneGUI: Multiple Approaches for Dose-Response Analysis of Microarray Data Using R. *The R Journal, 9:1, pages 14-26*

- Sengupta, R., Perualila, N.J., Shkedy, Z., Biecek, P., Molenberghs, G. and Bijnens, L., (2018). High Dimensional Surrogacy: Computational Aspects of an Upscaled Analysis. [***Accepted in Journal of Biopharmaceutical Statistics***]

- Sengupta, R., Molenberghs, G., Alonso, A., Van der Elst, W. and Shkedy, Z., (2018). Single, Multiple and Partial Surrogacy: A Joint Modeling Approach. [***In Preparation***]

- Sengupta, R., Perualila, N.J., Shkedy, Z., Bijnens, L., Ruiz, V.E., Battaglia, T. and Blaser, M., (2018). Development of Microbiome Biomarkers for IgA: A Joint Modeling Approach. [***In Preparation***]

- Sengupta, R., Perualila, N.J., Shkedy, Z., Bijnens, L., Ruiz, V.E., Battaglia, T. and Blaser, M., (2018). Development of High Dimensional Microbiome Biomarkers: A Non Parametric Approach. [***In Preparation***]

- Sengupta, R., Shkedy, Z., Bijnens, L., Ruiz, V.E., Battaglia, T. and Blaser, M., (2018). Development of High Dimensional Microbiome Biomarkers for an Immune Response: Hierarchical Bayesian Approach. [***In Preparation***]

- Sengupta, R., Shkedy, Z., Bijnens, L., (2018). Development of Microbiome Biomarker for Type 1 Diabetes. [***In Preparation***]

Book Chapters:

Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S. and Talloen, W., (Ed.) (2016) Applied Biclustering Methods for Big and High Dimensional Data Using R. CRC.

- Shkedy, Z., Sengupta, R. and Perualila, N.J. (2016)
  Identification of Local Patterns in the NBA Performance Indicators
- Sengupta, R., Trelles, O., Tirado, O.T. and Shkedy, Z. (2016)
  Biclustering for Cloud Computing
- De Troyer, E., Sengupta, R., Otava, M., Zhang, J.D., Kaiser, S., Culhane, A., Gusenleitner, D., Gestraud, P., Csardi, G., Hochreiter, S., Klambauer, G., Clevert, D.A., Perualila, N.J., Kasim, A. and Shkedy, Z. (2016)
  The biclustGUI Shiny App

Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N.J., Shkedy, Z., and Van der Elst, W. (Ed.) (2016) Applied Surrogate Endpoint Evaluation Methods with SAS and R. CRC.

- Perualila, N.J., Shkedy, Z., Sengupta, R., Bigirumurame, T., Bijnens, L., Talloen, W., Verbist, B., Göhlmann, H.W.H., QSTAR Consortium and Kasim, A. (2016)
  High Dimensional Biomarkers in Drug Discovery: The QSTAR Framework

R Packages:

- IntegratedJM: Joint Modeling of the Gene-Expression and Bioassay Data, Taking Care of the Effect Due to a Fingerprint Feature.
  `https://cran.r-project.org/web/packages/IntegratedJM/`
- IsogeneGUI: A Graphical User Interface to Conduct a Dose-Response Analysis of Microarray Data.
  `https://r-forge.r-project.org/R/?group_id=589`

Contents

1 Introduction **1**
1.1 The Single Trial Setting in Surrogacy 1
1.2 High Dimensional Surrogacy 4
1.3 Structure of the Thesis 6

I High Dimensional Surrogacy **9**

2 High Dimensional Surrogacy: An Introduction **11**
2.1 Introduction 11
2.2 The EGFR Drug Discovery Project 12
2.2.1 The EGFR Study 12
2.2.2 Data Structure 13
2.3 Integrated Analysis of Multiple Data Sources: a Joint Modeling Approach for a Single Biomarker 14
2.4 High Dimensional Surrogacy 16
2.4.1 Computational Aspects of High Dimensional Surrogacy 16
2.4.2 Multiple Surrogacy 17

3 High Dimensional Surrogacy: Computational Aspects of an Upscaled Analysis **19**
3.1 Introduction 19
3.2 Computational Aspects of the Joint Model 20
3.2.1 Reduction in Computation Time using Parallelization 21
3.2.2 Benchmark Computation Time for the EGFR Study 22

3.2.3 Parallelization using R packages . . . 23
3.2.4 A User-Specific Parallelization Framework in a Cluster . . . 24
3.3 Application to the Data . . . 27
3.3.1 Joint Modeling for One Fingerprint Feature . . . 28
3.3.2 Analysis for All Fingerprint Features . . . 30
3.4 Discussion . . . 32

4 Single, Multiple, Partial and Orthogonal Surrogacy: A Joint Modeling Approach 37
4.1 Introduction . . . 37
4.2 High Dimensional Surrogacy . . . 39
4.2.1 Modeling Approach for A Single Surrogate . . . 39
4.2.2 Multiple Surrogacy . . . 39
4.2.3 Partial Surrogacy . . . 40
4.2.4 Orthogonal Surrogacy . . . 43
4.3 Application to the Data . . . 44
4.3.1 Single Surrogacy . . . 44
4.3.2 Multiple Surrogacy . . . 44
4.3.3 Partial Surrogacy . . . 46
4.4 Discussion . . . 49

II Detection of Biomarkers in Microbiome Intervention Studies 51

5 Microbiome Intervention Studies: An Introduction 53
5.1 Introduction . . . 53
5.2 Microbiome Measurements at Different Levels of the Phylogenetic Tree . . . 56
5.2.1 Operational Taxonomic Unit (OTU) . . . 56
5.2.2 α-Diversity . . . 57
5.2.3 Family Level Richness . . . 58
5.3 Microbiome Intervention Studies . . . 58
5.3.1 TransPAT Study . . . 58
5.3.2 Type 1 Diabetes (T1D) Dataset . . . 61
5.4 Modeling Approaches for the Analysis of Microbiome Data . . . 65

6 Development of Microbiome Biomarkers for IgA: A Joint Modeling Approach 67
6.1 Introduction . . . 67
6.2 Modeling Approach . . . 68

6.2.1 A Joint Model for Microbiome Measurements and IgA 68
6.3 Application to the TransPAT Data 70
6.3.1 Analysis of α-Diversity 70
6.3.2 Analysis at Family Level 71
6.3.3 Analysis at OTU Level 73
6.4 Discussion 76

7 Development of High Dimensional Microbiome Biomarkers: A Non Parametric Approach 81
7.1 Introduction 81
7.2 A Non Parametric Approach for the Detection of Microbiome Biomarkers 83
7.2.1 Estimation 83
7.2.2 Inference 83
7.3 Application to the Data 87
7.3.1 Differentially Abundant OTUs 87
7.3.2 OTUs Associated with IgA Level at Day 20 91
7.4 Discussion 98

8 Development of High Dimensional Microbiome Biomarkers for an Immune Response: Hierarchical Bayesian Approach 99
8.1 Introduction 99
8.2 Hierarchical Model for Microbiome and IgA: A Path Analysis Approach . 100
8.2.1 Model Formulation 100
8.2.2 Specification of the Prior Distributions 104
8.3 Application to the Data 106
8.3.1 Results for S24-7 Family 106
8.3.2 Results for α-Diversity 108
8.4 Discussion 110

9 Development of Microbiome Biomarkers for Type 1 Diabetes 115
9.1 Introduction 115
9.2 A Poisson / Survival Path Analysis Model 116
9.2.1 A Weibull Model for Time to Develop T1D 118
9.3 Application to the Data 121
9.3.1 Results for α-Diversity 121
9.3.2 Results for S24-7 Family 122
9.4 Discussion 124

10 Discussion and Further Research **125**
10.1 High Dimensional Surrogacy . . . 125
10.1.1 Computational Issues . . . 125
10.1.2 Modeling Issues . . . 126
10.2 Analyzing Microbiome Data . . . 128
10.2.1 Identifying Multiple Microbiome Biomarkers . . . 128
10.2.2 Interaction between OTUs . . . 128
10.2.3 Longitudinal Analysis of Microbiome Data . . . 129

A Single, Multiple, Partial and Orthogonal Surrogacy: A Joint Modeling Approach **139**
A.1 Subclasses of Genes . . . 139

B Microbiome Intervention Studies: An Introduction **141**
B.1 OTU Filtering . . . 141

C Development of Microbiome Biomarkers for IgA: A Joint Modeling Approach **145**
C.1 Family Level Richness . . . 145
C.2 Application to the TransPAT Data . . . 145
C.2.1 Analysis of α-Diversity . . . 145
C.2.2 Analysis at Family Level . . . 148
C.2.3 Analysis at OTU Level . . . 148

D Development of High Dimensional Microbiome Biomarkers: A Non Parametric Approach **153**
D.1 Microbiome Composition . . . 153
D.2 Filtering OTUs . . . 153

E Development of High Dimensional Microbiome Biomarkers for an Immune Response: Hierarchical Bayesian Approach **159**
E.1 Results for Other Active Families . . . 159
E.2 Implementation in `WinBUGS` . . . 166

F Development of Microbiome Biomarkers for Type I Diabetes **167**
F.1 Results for Other Active Families . . . 167
F.2 Implemention in `WinBUGS` . . . 170

Chapter 1

Introduction

1.1 The Single Trial Setting in Surrogacy

In drug discovery experiments, one of the main challenges is a very slow, but costly and inefficient development process. The choice of endpoint(s), to assess the drug efficacy, plays an important role and it influences the duration of the development process. However, measuring the endpoint(s) can become difficult, time consuming and expensive. A "surrogate" endpoint serves as a substitute for the "true" endpoint as it can usually be measured more cheaply and conveniently. However, before using a surrogate as a substitute for the true endpoint it should be validated. Statistical methods for the identification and evaluation of surrogate endpoints in randomized clinical trials have been developed over last three decades (Prentice, 1989, Buyse and Molenberghs, 1998, Burzykowski et al., 2005, Alonso et al., 2016).

A *biomarker* is a surrogate for a certain biological process in a therapeutic intervention experiment. It is defined as an attribute that is objectively measured and evaluated as an indicator of biological or pathogenic processes or pharmacological responses to different types of interventions (Biomarkers Definitions Working Group, 2001). This implies that all surrogate markers are biomarkers, but not all biomarkers can qualify as surrogate markers.

The basic experimental setting in surrogacy is the single trial setting for which, the data structure for a study with n subject, is given by,

$$\mathbf{X}' = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{pmatrix}, \qquad \mathbf{Y}' = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \qquad \mathbf{Z}' = \begin{pmatrix} z_1 \\ z_2 \\ . \\ . \\ . \\ z_n \end{pmatrix}. \quad (1.1)$$

Here, $\mathbf{X}$ denotes the surrogate endpoint (biomarker variable), $\mathbf{Y}$ is the true endpoint and $\mathbf{Z}$ denotes the treatment variable. Note that $\mathbf{X}$ and $\mathbf{Y}$ could be binary, categorical, continuous, counts or time to event endpoints. Figure 1.1 displays the single trial surrogacy setting. Buyse and Molenberghs (1998) suggested a joint model for two Normally distributed endpoints i.e., the Normal / Normal setting, given by,

$$\begin{aligned} X_i &= \mu_X + \alpha Z_i + \epsilon_{X_i}, \\ Y_i &= \mu_Y + \beta Z_i + \epsilon_{Y_i}, \end{aligned} \quad (1.2)$$

where $i = 1, 2, \cdots, n$ indicates the subject and the error terms are assumed to have a joint bivariate Normal distribution with zero mean and variance-covariance matrix,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{X,X} & \sigma_{X,Y} \\ \sigma_{Y,X} & \sigma_{Y,Y} \end{pmatrix}. \quad (1.3)$$

Buyse and Molenberghs (1998) introduced two new measures for the validation of a surrogate endpoint in a single trial setting,

1. The *relative effect*, defined by β/α, is the effect of Z on Y as compared to that of Z on X.

2. The *adjusted association*, captured by ρ, is the correlation between X and Y, after adjusting for the effect of Z on both the variables and can be used as a measure of *individual level surrogacy*. The adjusted association can be estimated by,

$$\rho = \frac{\sigma_{Y,X}}{\sqrt{\sigma_{X,X}\sigma_{Y,Y}}}. \quad (1.4)$$

An example of the Normal / Normal case for a single trial setting is presented in Burzykowski et al. (2005, Age-related Macular Degeneration Study, Chapter 4).

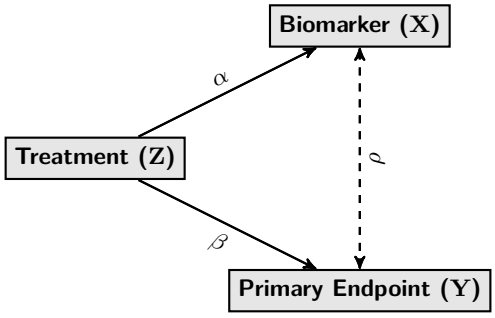


Figure 1.1: The surrogacy framework for two continuous endpoints. $\mathbf{X}$ and $\mathbf{Y}$ are the "surrogate" endpoint and "true" endpoint, respectively. The variable $\mathbf{Z}$ represents a binary treatment group. The parameters α, β and ρ represent the treatment effect on the biomarker, treatment effect on the clinical endpoint and the association between the biomarker and the clinical endpoint after adjusting for the treatment, respectively.

Buyse et al. (2000) and Burzykowski et al. (2004) extended the single trial setting to the multiple trials setting in which J trials are used to validate the surrogate endpoint. For the Normal / Normal case the joint model is given by,

$$\begin{aligned} X_{ji} &= \mu_{X_j} + \alpha_j Z_{ji} + \epsilon_{X_{ji}}, \\ Y_{ji} &= \mu_{Y_j} + \beta_j Z_{ji} + \epsilon_{Y_{ji}}. \end{aligned} \tag{1.5}$$

Here, μ_{X_j} and μ_{Y_j}, $j = 1, 2, \cdots, J$, are trial specific intercepts, α_j and β_j are trial specific treatment effects on the endpoints and $\epsilon_{X_{ji}}$ and $\epsilon_{Y_{ji}}$ are assumed to be Normally distributed with zero-mean and variance-covariance matrix given in equation (1.3). The meta-analytic approach introduces a second level of surrogacy, known as the *trial level surrogacy*, based on the joint distribution of α_j and β_j. The trial specific treatment effects can be modeled as,

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} a_j \\ b_j \end{pmatrix}, \tag{1.6}$$

where a_j and b_j are trial specific random effects and are assumed to have a Normal distribution with zero-mean and covariance matrix given by,

$$\mathbf{D} = \begin{pmatrix} d_{a,a} & d_{a,b} \\ d_{b,a} & d_{b,b} \end{pmatrix}.$$

Buyse et al. (2000) proposed a second measure for surrogacy at the trial level, defined by,

$$R^2_{trial} = \frac{d^2_{b,a}}{d_{a,a} d_{b,b}}. \tag{1.7}$$

The information theoretic approach to model the individual level surrogacy was proposed by Alonso and Molenberghs (2007). The mean structure of the true endpoint in equation (1.2) can be written as,

$$E(Y_i) = \mu_Y + \beta Z_i. \tag{1.8}$$

The model specified in equation (1.8) can be compared to a model that includes the surrogate endpoint as a covariate, that is,

$$E(Y_i) = \mu_Y + \beta Z_i + \gamma X_i. \tag{1.9}$$

The parameter γ captures the effect of the biomarker upon the true endpoint. Alonso and Molenberghs (2007) proposed a new measure to evaluate individual level surrogacy,

$$\hat{R}^2_{indiv} = 1 - e^{-\frac{1}{n} G^2}. \tag{1.10}$$

Here, n is the number of individuals and G^2 is the likelihood ratio statistics to compare the models in equation (1.8) and (1.9).

Two measures of individual level surrogacy specified in equation (1.4) and (1.10) were defined in a single trial setting. However, they can be extended for a multiple trial experiment and were discussed in details in Alonso et al. (2016). An example for the evaluation of two Normally distributed endpoints in a single trial setting is presented in Alonso et al. (2016, Age-related Macular Degeneration Study, Chapter 2 and Chapter 10).

1.2 High Dimensional Surrogacy

High dimensional surrogacy is related to the experimental setting in which high number of variables can be used as biomarkers for a response of primary interest. For example, a drug discovery experiment in which transcriptomic data with ℓ genes (features) is available for the ith observation unit (sample) and in addition a response variable Y_i was measured under the condition Z_i for each sample. In this case, there are ℓ triplets (X_{il}, Y_i, Z_i), $l = 1, \ldots \ell$, from which we can select biomarkers. Note that the aim of the analysis is not to find biomarker(s) to replace the response of interest but to identify biomarker(s) in order to gain a better understanding of the underlying biological processes related to Y. For example, the usage of transcriptomic biomarkers to establish the mechanism of action for a set of compounds under development in drug discovery experiments. Note that, the high

dimensional surrogacy framework includes an inference step in which both feature-specific treatment effect α_l and adjusted association ρ_l should be estimated and tested.

The basic data structure we consider in this thesis consists of three data sources, defined in equation (1.11), with a common dimension for which the measurements were observed. Let us consider a study with I samples (i.e., the observation unit, subjects, compounds etc.). Let $\mathbf{X}$ be a $J \times I$ matrix which contains the set of possible biomarkers, $\mathbf{Y}$ be a $B \times I$ matrix of B primary endpoints and $\mathbf{Z}$ be a $K \times I$ matrix that contains K binary variables representing conditions such as treatment group, chemical structure, binary target prediction scores, toxicity scores etc.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1I} \\ x_{21} & x_{22} & \cdots & x_{2I} \\ . & . & . & . \\ x_{j1} & x_{j2} & \cdots & x_{jI} \\ . & . & . & . \\ x_{J1} & x_{J2} & \cdots & x_{JI} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1I} \\ y_{21} & y_{22} & \cdots & y_{2I} \\ . & . & . & . \\ y_{b1} & y_{b2} & \cdots & y_{bI} \\ . & . & . & . \\ y_{B1} & y_{B2} & \cdots & y_{BI} \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1I} \\ z_{21} & z_{22} & \cdots & z_{2I} \\ . & . & . & . \\ z_{k1} & z_{k2} & \cdots & z_{kI} \\ . & . & . & . \\ z_{K1} & z_{K2} & \cdots & z_{KI} \end{pmatrix}. \quad (1.11)$$

The experimental design we consider in this thesis is similar to the single trial setting in surrogacy in the sense that only a singe study was conducted in order to investigate the relationship between $\mathbf{X}$ and $\mathbf{Y}$. For that reason we focus in this thesis on individual level surrogacy.

Early work was focused on transcriptomic data and presented by Lin et al. (2010) and Tilahun et al. (2010). Both authors identified gene-specific biomarkers for continuous outcomes (the distance traveled by the rats under treatment and the HAMD scores for psychiatric patients). Van Sanden et al. (2012) identified gene-specific biomarkers for toxicity data presented as a binary response. Perualila *et al.* (2016a, 2016b) proposed a joint model for the detection of genomic biomarkers as a part of the QSTAR framework for drug discovery data (Verbist et al., 2015).

Note that the triplet $(\mathbf{X}_j, \mathbf{Y}_b, \mathbf{Z}_k)$, consisting of the jth, bth and kth rows from the matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, respectively as specified in equation (1.11) is exactly same as the one in equation (1.1). In Chapter 8, we analyze this single trial setting as a multi trial experiment where each row of the matrix $\mathbf{X}$ is considered as a different trial.

The model, proposed by Perualila *et al.* (2016a, 2016b), can be seen as the joint model for a single trial setting applied in a loop over the dimension of the trascriptomic data. Although, this model allows to identify transcriptomic biomarkers, it did not address two fundamental problems related to high dimensional surrogacy: computation time and joint surrogacy. Both problems are addressed in the first part of the thesis while the second part of the thesis is devoted to a new application of high dimensional surrogacy in the context of microbiome intervention experiments.

1.3 Structure of the Thesis

Figure 1.2 presents the thesis structure. The first part of the thesis is devoted to drug discovery experiments. The observation unit in this type of studies is a compound (i.e., the common dimension in equation (1.11)). The aim of the analysis is to identify transcriptomic biomarkers for a phenotypic variable(s) in order to understand better the mechanism of action (MoA, Ravindranath et al., 2015) of a set of compounds. Each column, in the condition matrix $\mathbf{Z}$, represents the absence or presence of a chemical structure for a specific compound under investigation.

Chapter 2 introduces the joint model, presented in Perualila *et al.* (2016a, 2016b), for transcriptomic data and the EGFR discovery project (Verbist et al., 2015) that is used for illustration in the first part of the thesis. **Chapter 3** is devoted to computational problems. A new parallelization framework that allows to upscale the analysis discussed in Perualila *et al.* (2016a, 2016b) is introduced. The new parallelization framework is compared with a set of `R` packages for parallelization and different configurations for parallelization are investigated in order to find the best solution, in terms of computation time, for the data analysis problem. In **Chapter 4** we discuss different types of surrogacy - single, multiple, partial and orthogonal surrogacy. In this chapter, we extend the setting of single surrogacy $[Y_b, X_j|Z_k]$ to the case with a subset of ℓ candidate biomarkers and discuss the setting of multiple surrogacy (Van der Elst et al., 2018), $[Y_b, X_{j_1}, X_{j_2}, \cdots, X_{j_\ell}|Z_k]$ and partial and orthogonal surrogacy $[Y_b, X_{j_\ell}|X_{j_1}, X_{j_2}, \cdots, X_{j_{\ell-1}}, Z_k]$.

The second part of the thesis is focused on microbiome intervention studies. The two studies analyzed in this part of the thesis are animal experiments that were conducted in order to investigate the association between microbiome variables and clinical variables of interest based on an animal model. In this part, $\mathbf{Y}$ and $\mathbf{Z}$ are both $1 \times n$ vectors, representing a clinical variable of primary interest and the treatment variable, respectively. The biomarker matrix $\mathbf{X}$ is a $m \times n$ matrix containing the information about the microbiome of each subject.

Chapter 5 introduces the microbiome data for two intervention experiments. The first was conducted to investigate the association between the subject's (the host's) immune system and microbiome and the later investigated the association between time to develop type 1 diabetes (T1D) and the host's microbiome.

A parametric joint model and non parametric approaches are discussed in **Chapter 6** and **Chapter 7**, respectively. Bayesian path analysis models, for Poisson and Normal variables and Poisson and time to event variables, are discussed in **Chapter 8** and **Chapter 9**, respectively. An overall discussion and the scope of future work of the research presented in this thesis is given in **Chapter 10**.

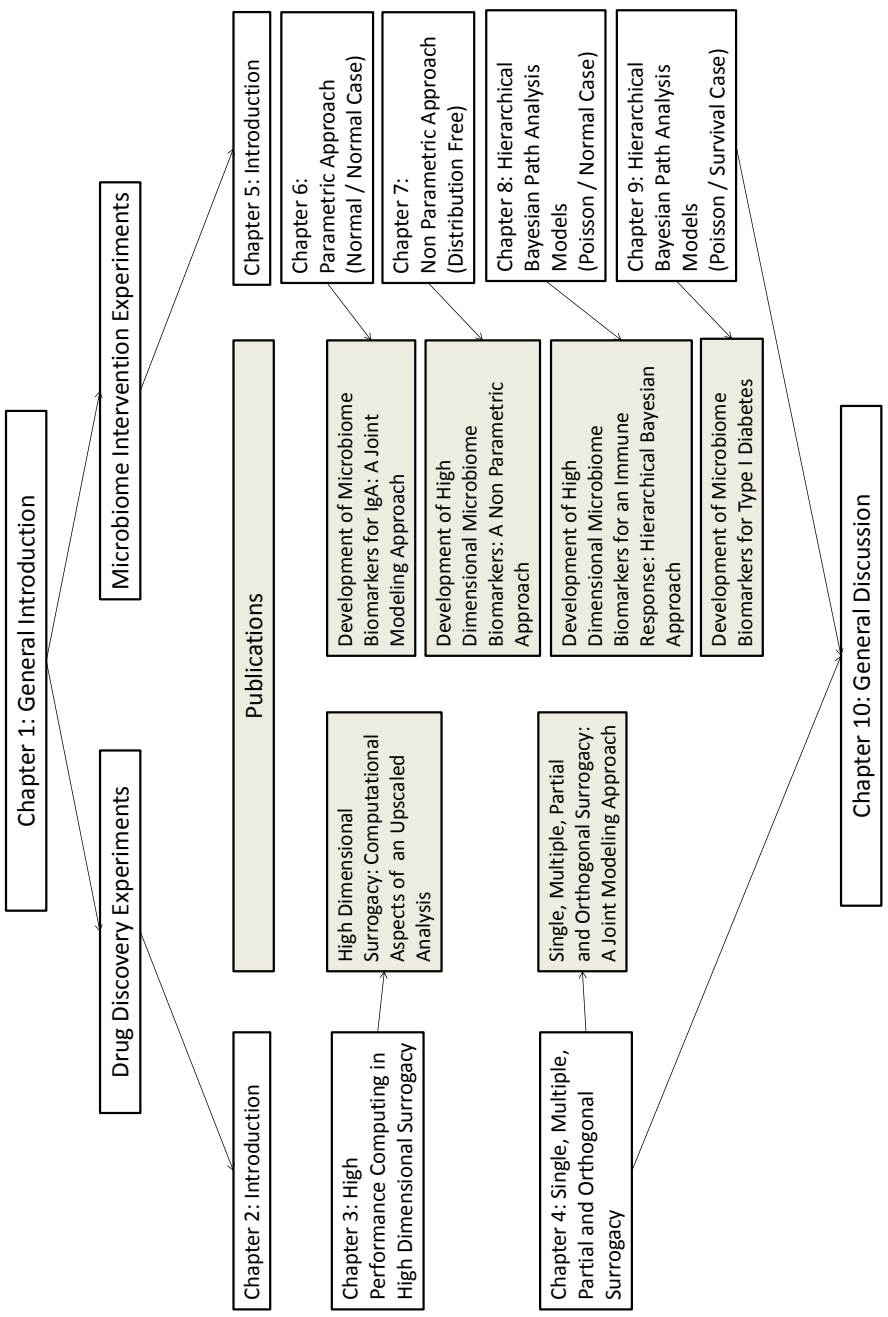


Figure 1.2: Thesis structure and publications.

Part I

High Dimensional Surrogacy

Chapter 2

High Dimensional Surrogacy: An Introduction

2.1 Introduction

The ability and the need to analyze routinely large scale datasets in early drug discovery has become a central issue in the last few years in data analysis workflow of early drug discovery experiments. In a typical drug discovery study, while exploring new compounds, the chemical structure of a specific compound is known, but not all biological processes related to the compound. In order to explore the association between chemical structure of a given compound and biological pathways related to the compound, a new data analysis framework, Quantitative Structure Transcriptomics Activity Relationship (QSTAR, Perualila-Tan et al. 2016a, Verbist et al. 2015), was proposed as an extension of the Quantitative Structure Activity Relationship (QSAR) modeling approach (Bruce et al., 2008, Nantasenamat et al., 2009). The main idea behind the QSTAR approach is to use information about transcriptomics data in order to understand the biological processes related to a new compound and to gain a complete insight about its mechanism of action (MoA).

Figure 2.1 shows the available data sources in a typical high dimensional surrogacy setting. The information available on each compound consists of three sources: chemical structure measured by a set of fingerprint features (FFs, Rogers and Hahn, 2010, Todeschini and Consonni, 2009), bioactivity variables measuered in assays (pIC_{50}, Martin et al., 2002) and transcriptomics data (Amaratunga et al., 2014, Bai et al., 2013, Göhlmann and Talloen, 2009).

Within the QSTAR framework, a possible modeling approach for the identification of transcriptomic biomarkers was based on a feature-specific model, i.e., a model to detect a potential transcriptomic biomarker was fitted in a loop over the transcriptomic dimension for a single chemical structure. In this part of the thesis we investigate two problems which were not addressed before in the context of high dimensional surrogacy: computation time and multiple biomarkers for a single bioactivity variable.
The case study that will be analyzed in the first part of the thesis is presented in Section 2.2. The remainder of the chapter is organized as follows. In Section 2.3 we briefly describe the joint model proposed by Perualila *et al.* (2016a, 2016b) which, in the context of this work discussed in this thesis, can be seen as a *single biomarker* application within the high dimensional surrogacy setting. In Section 2.4 we present the two aspects of high dimensional surrogacy investigated in the first part of the thesis: the computation and modeling aspects related to the high dimensionality of the data.

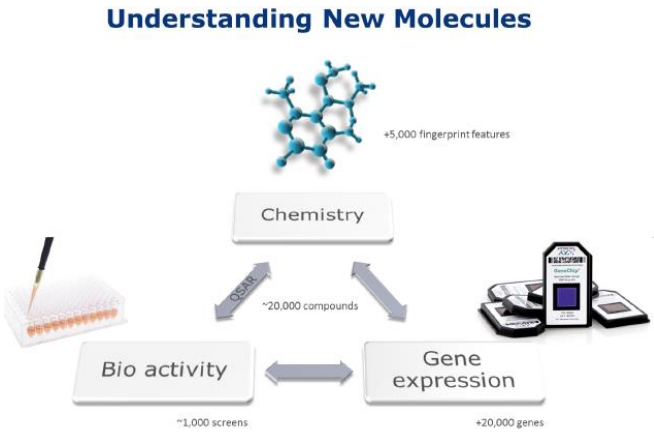


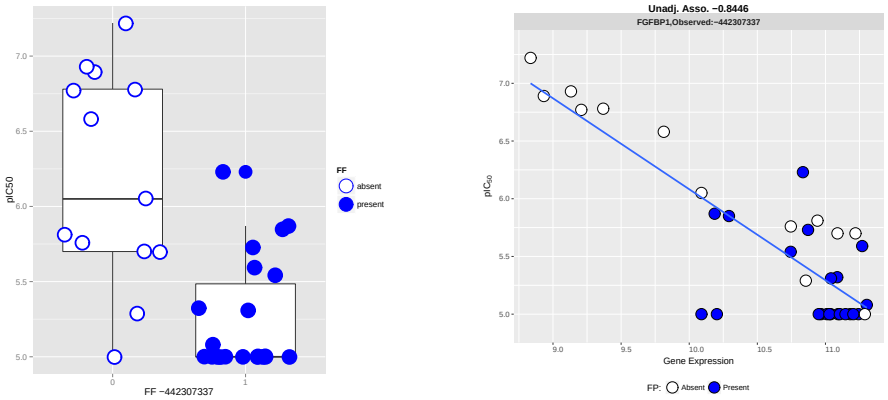
Figure 2.1: Data sources within the QSTAR framework. The observation unit is a compound and three high dimensional data, on both chemical and biological domains, are available.

2.2 The EGFR Drug Discovery Project

2.2.1 The EGFR Study

The EGFR study is a drug discovery project which focuses on inhibition of the epidermal growth factor receptor (Verbist et al., 2015). Thirty-five compounds with a macrocycle

structure were profiled in order to identify compounds with similar biological effects as the current EGFR inhibitors. The compounds, Gefitinib and Erlotinib, served as reference compounds. Gene expression profiles are available for 3595 genes after applying initial filtering steps. For an elaborate discussion about the data production we refer to Verbist et al. (2015). In addition, a total of 138 unique profiles of chemical substructures were identified for this compound set. Figure 2.2a shows the boxplot of the pIC_{50} obtained for the primary bioassay for a specific fingerprint feature: FF-442307337. Figure 2.2b shows a scatterplot for one gene (FGFBP1) versus the pIC_{50}.



(a) **(b)**

Figure 2.2: The EGFR project: FF -442307337. Panel a: boxplot of pIC_{50}. Panel b: example of a particular gene (FGFBP1). Solid blue circles represent compounds for which the FF is present while the empty blue circles represent compounds for which the FF is absent.

2.2.2 Data Structure

Figure 2.3 displays the data structure, for the EGFR project, which is similar to the general data structure presented in Chapter 1. The response variable of primary interest is the bioactivity outcome, measured by the pIC_{50} of 10 different bioassys, $(\mathbf{Y}_{10\times 35})$ which, in the context of surrogate endpoint evaluation, represents the true endpoint. For a specific fingerprint feature, the chemical structure $(\mathbf{Z})$ is a vector of length 35 in which each element is a binary indicator, given by,

$$Z_i = \begin{cases} 1, & \text{if the chemical structure is present in a compound,} \\ 0, & \text{otherwise.} \end{cases}$$

The transcriptomics data consists of a 3595×35 matrix, $\mathbf{X}$, for which the jith entry is the expression level of the jth gene for the ith compound.

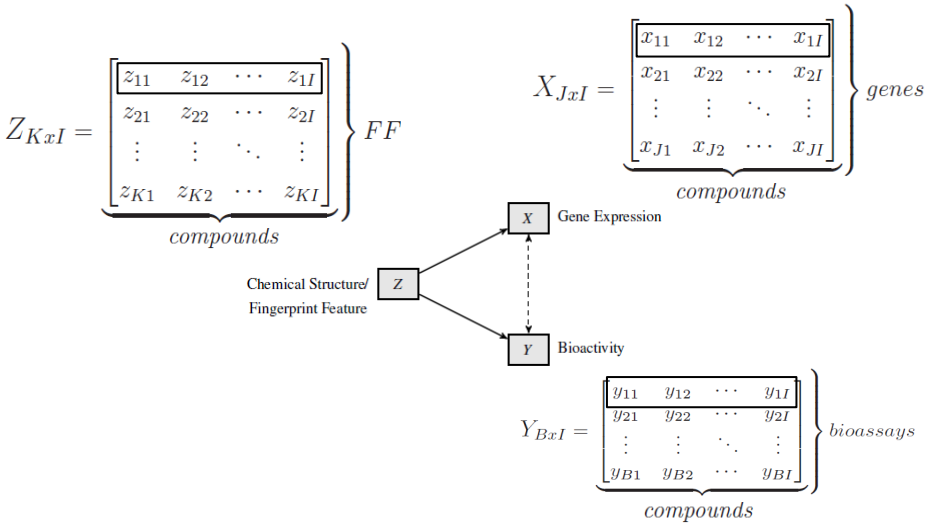


Figure 2.3: Data structure for the EGFR project. The response of primary interest (the "true endpoint" in surrogacy terminology) is pIC_{50} and the potential surrogate matrix is the transcriptomic matrix. The matrix $\mathbf{Z}_{138\times35}$ contains information about the chemical structure of the compounds.

2.3 Integrated Analysis of Multiple Data Sources: a Joint Modeling Approach for a Single Biomarker

In this section, we briefly present the modeling approach for a single biomarker, proposed by Perualila-Tan et al. (2016a) within the setting of high dimensional surrogacy.

Let $\mathbf{X}$ be the transcriptomic matrix where X_{ji} is the jth gene expression of the ith compound, $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let Y_i be the measurement for the bioassay data and Z_i be an indicator variable that takes the value of 1 if a specific fingerprint is present in the chemical structure of the ith compound and zero otherwise.

Following Perualila *et al.* (2016a, 2016b) we formulate a gene-specific joint model that allows to detect which gene is differentially expressed and which gene is correlated with the response, taking into account that the fingerprint feature can influence both endpoints:

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{X_j} + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \mathbf{\Sigma}_j\right], \tag{2.1}$$

where $\mathbf{\Sigma}_j$ is a gene-specific covariance matrix, given by,

$$\mathbf{\Sigma}_j = \begin{pmatrix} \sigma_{X_j,X_j} & \sigma_{Y,X_j} \\ \sigma_{Y,X_j} & \sigma_{Y,Y} \end{pmatrix}. \tag{2.2}$$

The parameters α_j and β represent the fingerprint feature effects for the jth gene and the response, respectively, while μ_{X_j} and μ_Y are gene-specific and the response-related intercepts, respectively. Gene-specific association with the response, adjusted for a possible fingerprint effect, can be estimated using adjusted association, defined by Buyse and Molenberghs (1998) as,

$$\rho_{Y,X_j} = \rho_j = \frac{\sigma_{Y,X_j}}{\sqrt{\sigma_{X_j,X_j}\sigma_{Y,Y}}}. \tag{2.3}$$

The case with, $\rho_{Y,X_j} = 1$ indicates a deterministic relationship between the gene expression and the response after accounting for the effect of the fingerprint feature (Alonso et al., 2016).

As pointed out by Perualila *et al.* (2016a, 2016b) the joint model, formulated in equation (2.1), allows testing for differentially expressed genes. Hence, for each gene, we test the hypotheses

$$\begin{aligned} H_{0_j} &: \alpha_j = 0, \\ H_{1_j} &: \alpha_j \neq 0. \end{aligned} \tag{2.4}$$

In order to test whether the expression level of a gene and the bioassay readout are correlated, we specify the hypotheses

$$\begin{aligned} H_{0_j} &: \rho_{Y,X_j} = 0, \\ H_{1_j} &: \rho_{Y,X_j} \neq 0, \end{aligned} \quad \text{or equivalently} \quad \begin{aligned} H_{0_j} &: \sigma_{Y,X_j} = 0, \\ H_{1_j} &: \sigma_{Y,X_j} \neq 0. \end{aligned} \tag{2.5}$$

Note that under the null hypothesis in equation (2.5), the covariance matrix specified in equation (2.2) is reduced to

$$\mathbf{\Sigma}_j = \begin{pmatrix} \sigma_{X_j,X_j} & 0 \\ 0 & \sigma_{Y,Y} \end{pmatrix}. \tag{2.6}$$

For a microarray with J genes, $2 \times J$ null hypotheses should be tested, which implies that an adjustment for multiple testing should be applied. Throughout the analysis, presented in the first part of the thesis, we apply the False Discovery Rate (FDR) approach proposed by Benjamini and Hochberg (1995).

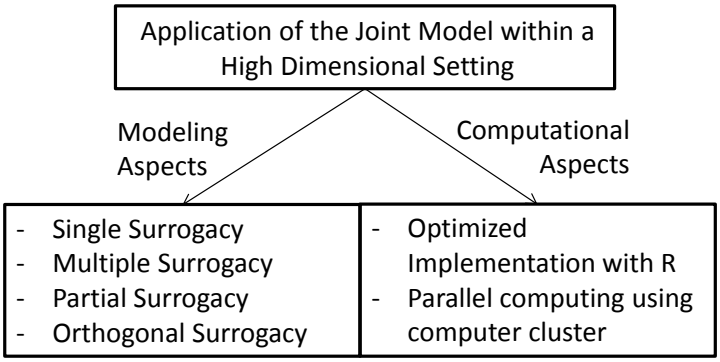


Figure 2.4: Modeling and computational aspects of high dimensional surrogacy.

2.4 High Dimensional Surrogacy

The term high dimensional surrogacy was used by (Perualila et al., 2016b) to describe the setting in which the joint model formulated in equation (2.1) was used to detect transcriptomics biomarkers among a set of J candidates. As pointed out by (Perualila et al., 2016b) this setting is closely related to the single trial surrogacy setting described in Burzykowski et al. (2004) as it is focused on the identification of a single biomarker at the time i.e., a single biomarker per model. In Perualila *et al.* (2016b), the model was implemented in a loop which allows detecting multiple biomarkers for the response of interest. In this part of the thesis, we focus on two different aspects of high dimensional surrogacy, shown in Figure 2.4: computational problems and multiple surrogacy. The first is related to computational solutions for an upscaled analysis using the joint model presented in Section 2.3 for the setting in which all fingerprints are analyzed while the later is related to a surrogacy setting in which there are multiple candidates that can serve as biomarkers for the response of primary interest. Note that similar to Perualila *et al.* (2016b), we focus on a single trial setting and therefore, in surrogacy terminology, we focus on individual level surrogacy measures.

2.4.1 Computational Aspects of High Dimensional Surrogacy

Within the QSTAR setting, the analysis was conducted for a specific fingerprint feature of interest. In practice, a gene-specific joint model was fitted using only one `for loop`, in R. For the EGFR Study, an application of the joint model within a single a loop over the genes' dimension requires 377 seconds (in a laptop with i5 processor) to complete

the entire analysis for all the genes, for a particular fingerprint feature. However, for an analysis consists of all fingerprint features and all genes, two `for loops`, one over the gene expression dimension and another over the fingerprint feature dimension should be implemented. The double `for loop` provides a simple solution, but it requires 14.5 hours which makes the analysis over the dimension of the fingerprint feature not practical. In **Chapter 3**, we discuss several computational solutions that can be used to reduce the total computation time. We focus on parallel programming and present the results for both `R` packages for parallel programing and a user specific approach (the worker framework) for a computer cluster.

2.4.2 Multiple Surrogacy

Recently, Van der Elst et al. (2018) proposed a multiple surrogacy approach in which a joint surrogate, for a true endpoint, can be identified and evaluated. In **Chapter 4**, we focus on slightly different approach in which the surrogacy value of a secondary biomarker is evaluated, given that a primary (or a known) biomarker for the primary endpoint is already identified. To fix notation, let Y be the primary endpoint, X_1 a primary biomarker and X_2, a secondary biomarker. We assume that the surrogacy value of X_1 was evaluated using the joint model formulated in equation (2.3) and our aim is to identify a second biomarker and to evaluate its added surrogacy value given X_1. We use a *partial surrogacy* measure to estimate the adjusted association between X_2 and Y, adjusting for X_1 and the fingerprint effect. We term this association, the *partial adjusted association*. A special case of partial surrogacy is *orthogonal surrogacy* in which X_1 and X_2 are independent but both can jointly be used as a biomarker for Y.

Chapter 3

High Dimensional Surrogacy: Computational Aspects of an Upscaled Analysis

3.1 Introduction

The analysis, presented in Perualila *et al.* (2016a, 2016b), was focused on one fingerprint feature and the joint model, formulated in equation (2.3), was fitted using a `for loop` over the gene expression data. For the analysis of the EGFR project with 35 compounds and a transcriptomics matrix of 3595 features, the modeling procedure requires 377 seconds for a particular fingerprint feature. The analysis implemented as a part of the QSTAR project, was done for all the transcriptomics data, but for only one fingerprint feature. However, in a typical drug discovery experiment, there are thousands of transcriptomic features as well as hundreds or thousands of fingerprint features that should be analyzed. As a result, the data analysis procedure proposed by Perualila *et al.* (2016a, 2016b), although can be implemented for any number of fingerprint feature, cannot be used in practice due to high computation time per analysis.

Our aim, in this chapter, is to upscale the data analysis procedure proposed by Perualila *et al.* (2016a, 2016b) to the setting where large data matrices (for transcriptomics and fingerprint features) are included in the analysis and to explore different computational configurations to reduce the computation time from few hours per study to few seconds. One solution, to reduce the computation time, is to parallelize the data analysis

procedure. In this case, the entire data analysis task is divided into smaller chunks of less computationally intensive tasks and those small jobs can be solved in parallel, using many nodes in a supercomputer (the Flemish Supercomputer Centre, VSC cluster, in our case).

The joint model proposed by Perualila *et al.* (2016a, 2016b) was implemented using the `R` software (R Core Team, 2016). Within the `R` environment several packages, e.g. `foreach`, `snow`, `parallel` etc., can be used for parallelization (Schmidberger et al., 2009, Vera et al., 2008). The analysis presented in this chapter is focused on the reduction of computation time. We investigate the influence of different computational configurations, i.e., the usage of the different settings for parallelization, on the computation time. Two `R` packages, `foreach` and `parallel`, are compared with the performance of the *worker* framework (`https://www.vscentrum.be/cluster-doc/running-jobs/worker-framework`) in a supercomputer environment. The EGFR project, presented in Chapter 2, is used as a case study to compare the computation time obtained for different computational configurations.

This chapter is organized as follows. Computational issues and different solutions to resolve the problem of long computation time using parallelization are presented in Section 3.2 while Section 3.3 presents the data analysis results. Finally, we discuss the results in Section 3.4.

3.2 Computational Aspects of the Joint Model

Within the QSTAR setting, a gene-specific joint model, for a given fingerprint feature, was fitted using only one `for loop` over the genes' dimension and it required 377 seconds (in a laptop with i5 processor) to complete the entire analysis for all the genes. However, for an analysis consisting of both fingerprint feature and gene expression dimensions, two `for loops`, one loop over the gene expression dimension and another over the fingerprint feature dimension should be implemented. The double `for loop` provides a simple solution, but comes with a price of very long computation time. In a `for loop`, the computations are executed sequentially, therefore the total computation time is linearly proportional to the total number of computations, i.e, the computation time for one fingerprint feature (one `for loop`), multiplied by number of fingerprint features. For the EGFR project an implementation of the joint model using a double `for loop` requires 14.5 hours which makes the analysis over the dimension of the fingerprint feature not practical.

The joint model specified in equation (2.1) is fitted using the `gls()` function in `R`. However, few other functions are used for other purposes e.g., preparing the data frame suitable for the `gls()` function, gathering the results in a dataframe for easier visualization of the parameter estimates etc. We categorize all the functions used for the

analysis into three groups: (a) `gls()`, (b) `anova()` and `summary()` functions and (c) all other functions e.g., `data.frame()` and `cor()`. Note that the functions `summary()` and `anova()` are used to extract the relevant output, p-values, parameter estimates, etc., from the `gls` object. Figure 3.1 shows that more than 80% of the total computation time, of a single run, is consumed by the `gls()`.

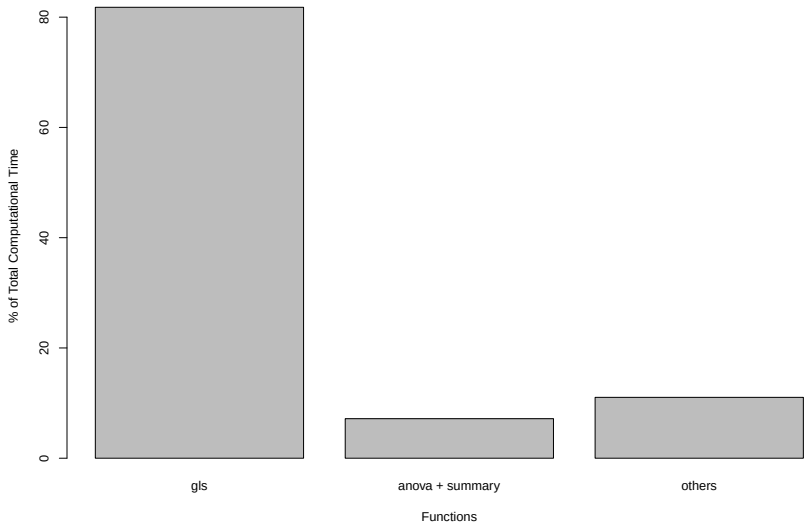


Figure 3.1: Distribution of computation time by function category for a single `for loop` over the gene expression dimension applied to the EGFR dataset.

In order to reduce the total computation time for all the genes, one can only parallelize the analysis for different genes since it is no longer possible to speed up the `gls()`, within `R`.

3.2.1 Reduction in Computation Time using Parallelization

In this section, we compare the two approaches for parallelization in order to reduce the computation time for the joint model. In Section 3.2.3 we use the most popular `R` packages for parallelization, `foreach` and `parallel`, while in Section 3.2.4 a user-specific parallelization framework is used. The different methods for parallelization are compared to a benchmark run of the joint model for a single fingerprint feature using a single `for loop` over the transcriptomics dimension when the model is fitted in a laptop or in a super computer.

Efficiency

In order to compare between different configurations in which the model is submitted, we use the computation times in different settings, all consisting of K fingerprint features and J genes, and in addition we define an efficiency measure (Azimfar, 2015) given by,

$$E_n = \frac{S_n}{n}, \tag{3.1}$$

where

$$S_n = \frac{T_1}{T_n}. \tag{3.2}$$

Here, T_1 and T_n are computation times when one and n cores are used, respectively. Hence, E_n is the average gain, per core, in computation time when n cores are used compared to the case when one core is used.

Throughout this chapter, the following definitions are used:

Master-Slave Framework

One of the most common paradigms in the context of parallel programming is the "master-slave" framework (Sahni and Vairaktarakis, 1996). The *master* is a core that divides the bigger and more complex main problem into smaller subproblems and supplies them to the other cores called *slaves*. The slaves, after completion of the smaller problems assigned to them, return the results to the master.

Load-Balancing

In the context of computing, *load-balancing* is the way to distribute the workload to all the available resources. The main objective of load-balancing is to optimize the usage of different resources.

Jobitem

Any computing job, submitted to a computer cluster, requires a certain number of nodes and a certain number of cores on each node. A *jobitem* refers to the part of the job which is scheduled to be executed in one single core.

3.2.2 Benchmark Computation Time for the EGFR Study

As mentioned above, for the EGFR study, the computation time of the joint model for a single fingerprint feature, using a single `for loop` over the transcriptomics dimension is

equal to 377.13 seconds (Table 3.1) using a laptop with i5 processor. The same data analysis procedure was implemented using a ivybridge processor in the Flemish super computer (VSC cluster) and the computation time was reduced to 259.35 seconds (Table 3.1).

3.2.3 Parallelization using R packages

Few R packages are available for parallelization: the `foreach` (Calaway et al., 2015, Steve, 2012) and the `parallel` are the most commonly used packages to achieve the goal. The main function in the `foreach` package, `foreach()`, has similar syntax as `for loop`. The `clusterApply()` of the `parallel` package is used frequently to reduce the computation time by distributing the task in a so called master-slave framework and the load-balancing is added, on top of it, by the function `clusterApplyLB()`. Table 3.1 displays the computation times of the same data analysis problem described in Section 3.2.2 when the existing R packages for parallelization are used in both a laptop (with 2 i5 cores) and in the VSC (each node with 20 ivybridge cores).

R function	No. of Physical Cores used	Computation Time (in Sec.)	Speed up	Efficiency (in %)
`for()`	1 (laptop i5)	377.13	1	100
`foreach()`	2 (laptop i5)	380.53	0.99	49.5
`clusterApply()`	2 (laptop i5)	247.59	1.52	76
`clusterApplyLB()`	2 (laptop i5)	233.77	1.61	80.5
`for()`	1 (VSC ivybridge)	259.35	1	100
`foreach()`	20 (VSC ivybridge)	272.02	0.95	4.75
`clusterApply()`	20 (VSC ivybridge)	84.73	3.06	15.3
`clusterApplyLB()`	20 (VSC ivybridge)	40.19	6.45	32.27

Table 3.1: Computational time (in seconds) for different R packages in different scenarios. Benchmark time for the laptop is 377.13 seconds and for the VSC cluster is 259.35 seconds.

Note that in contrast to the exceptions, for both laptop and VSC, the `foreach()` function increased the computation time when 2 and 20 cores were used on a laptop and VSC, respectively. Since the job had to be distributed among multiple cores (2 and 20 cores, for the laptop and the VSC, respectively) and the results had to be gathered from multiple cores as well, this adds an overhead, resulting in an increase in the total computation time. Both `clusterApply()` and `clusterApplyLB()` outperformed the

`for loop` and led to substantial reduction in computation time. The largest reduction in computation time was achieved for `clusterApplyLB()` on the VSC using 20 ivybridge cores for which

$$S_{20} = \frac{259.35}{40.19} = 6.45 \text{ and } E_{20} = \frac{6.45}{20} \times 100\% = 32.27\%.$$

This implies that the computation time of a 20 core configuration is 6.45 time faster than the benchmark computation time with a reduction of 32.27% per core. One of the main drawbacks of using the above `R` packages is that the parallelization is done implicitly, i.e., the user cannot specify the configuration to divide and distribute the entire job according to his/her plan. Due to the fact that the program accesses the different cores multiple times, an overhead time is added to the total computation time, as illustrated above. In addition, the `R` packages, for parallelization, perform well for a machine with a certain number of cores, but they fail to work well when a cluster, of several different machines, is considered. The number of cores one node/machine can have is limited. On the other hand, there is no limitation to the number nodes a computer cluster can have and therefore running the program in a computer cluster can lead to a reduction in computation time.

3.2.4 A User-Specific Parallelization Framework in a Cluster

In this section, we discuss a user-specific parallelization procedure, the worker framework (Bex, VSC website, 2017) in which the configuration of the parallelization procedure is not defined by an `R` package but rather designed specifically for a particular data analysis problem by the user. Table 3.2 lists the different set of files that are required for this framework compared to an implementation done in one computer.

The worker framework is similar to the common master-slave framework (Sahni and Vairaktarakis, 1996). Within the worker framework, when n cores are reserved for a job, one core works as master to distribute the total work among the remaining $n-1$ workers and the master gathers and combines the results from all the workers when all the jobitems are completed. In the context of our analysis, a jobitem is a `for loop` executed on a worker. In order to implement the parallelization procedure in the VSC cluster, we need to select the number of genes assigned to a `for loop` in one core.

Table 3.3 presents the computation time for different configurations of the parallelization procedure. For example, when one gene was used as a jobitem (i.e., one core is used to fit a joint model for only one gene), the procedure consists of 3595 jobitems, each of one gene, the total computation time is 339.93 seconds. The computation time is reduced to 19.47 seconds when the number of genes in a jobitem is equal to 190 and the parallelization procedure (Table 3.3) consists of 19 parallel jobitems. Note that the total computation time can be decomposed into actual computation time, which is the

Implementation in	Required set of files
Laptop/Desktop	- jm.R : Rscript, with loops over genes & features, to implement the joint model. - .rda data file : contains the EGFR dataset.
VSC cluster	- jm.R : Rscript to implement the joint model for one (or more) genes for a single feature i.e. no loop or single loop of smaller length than before. - .rda data file : contains the EGFR dataset. - params.csv: contains the parameter values, used to parallelize the code. - jm.pbs: a batch file with information on which module to load, which Rscript to use etc., in order to execute our code in the cluster.

Table 3.2: Set of files required for different types of implementation of the joint model.

dominant component in the total computation time and the time required to combine the results from different jobitems. For example, consider two configurations for the parallelization procedure: the first consists of 175 genes per jobitem and 21 jobitems while the second consists of 190 genes per jobitem and 19 jobitems (rows 10 and 12 in Table 3.3, respectively). The time needed to combine the results for the first configuration is equal to 0.41 seconds, which is slightly faster than the time needed to combine the results for the second configuration, 0.47. However, the latter configuration is preferred due to faster actual computation time, 19 seconds compared with 30 seconds for the first configuration. Thus, depending on the context, one can decide which part of the total computation time, actual time required for the analysis (third column in Table 3.3) or the total computation time (fifth column in Table 3.3), should get the priority in the configuration.

Similar to Section 3.2.3 load-balancing affects the results. The user needs to specify how the load-balancing will be incorporated while running the `RScript` in parallel. As shown in Table 3.3, different configurations for the size of the jobitems per worker were investigated and among them the worst load-balancing occurred for 20 jobitems. This is due to the fact that each jobitem takes similar time to be completed. At first 19 (out of 20 jobitems) are executed in 19 cores parallelly; but when the last one is being executed in 1 core, remaining 18 workers are in idle state for a certain time, resulting in an increase in the total computational time (see, for example, the last two rows in Table 3.3). The framework performs most efficiently when there are 19 jobitems as there are 19 workers

executing them in parallel and no resources are being wasted. Thus, one can neither have too many nor too few jobitems to have a better load-balancing.

In addition, for every jobitem, R needs to be loaded in one worker before the jobitem starts. The first initialization of R takes around 0.5 seconds, but the loading time reduces in later stages because then it is loaded from the cache memory instead of the disk memory. Thus, an educated guess needs to be made while designing different computational configurations. For example, if a jobitem takes 500 seconds to complete, then the 0.5 seconds loading time of R is neligible in that setting (less than 0.1% of the total computational time of the jobitem). However, if a jobitem takes 5 seconds then a loading time of 0.5 seconds takes 10% of the computation time.

# of genes/ jobitem	# of jobitems	Comp. Time for all jobitems	Time to combine jobitems	Total Comp. Time	Efficiency (in %)
1	3595	316	23.93	339.93	3.81
10	360	49	3.99	52.99	24.47
20	180	35	2.67	37.67	34.42
25	144	31	2.10	33.10	39.18
50	72	25	1.65	26.65	48.66
75	48	26	1.02	27.02	47.99
100	36	23	0.84	23.84	54.39
125	29	26	0.86	26.86	48.28
150	24	28	0.40	28.40	45.66
175	21	30	0.41	30.41	42.64
180	20	32	0.48	32.48	39.92
190	19	19	0.47	19.47	66.60

Table 3.3: Computation time in seconds and estimated efficiency in different configurations when 20 cores (i.e. 1 master and 19 workers) are used and the number of genes in each jobitem changes.

One advantage of worker framework is the ability to speed up of computation time by increasing the number of cores (until a certain level) for the analysis. Table 3.4 and Figure 3.2 present the computation results for different configurations when different number of cores, in the cluster, were. For each configuration, the analysis was repeated 5 times as the total computation time might have small variations, depending on the total amount of load on the VSC cluster (e.g., if other users are running other jobs, in the

# of Physical Cores used	# of genes per jobitem	Comp. Time for all jobitems(in Sec.)	Time to combine jobitems (in Sec.)	Total Comp. Time (in Sec.)	Efficiency (in %)
1	3595	259.35	0	259.35	100
3	1800	129	0.02	129.02	67.01
5	900	70	0.03	70.03	74.07
9	450	38	0.04	38.04	75.75
11	360	32	0.04	32.04	73.59
16	240	24	0.06	24.06	67.37
19	200	20	0.06	20.06	68.05
20	190	19	0.06	19.06	68.04
26	144	19	0.08	19.08	52.28
31	120	15	0.09	15.09	55.44
37	100	12	0.11	12.11	57.88
40	93	12	0.11	12.11	53.54

Table 3.4: Computational time for different configurations of number of cores and number of genes per core.

VSC cluster, which have high memory usage or other computational requirements). For the configuration with 9 cores and 450 genes per jobitem (9/450 from now onwards) the total computation was 1.84 times faster compared to the configuration 5/900. Although, by increasing the number of cores total computation time can be reduced, the efficiency of the job does not increase always. For example, the total computation time for the configuration 19/200 is 1.61 times higher than the computation time for configuration 37/100 but the first configuration is 1.21 times more efficient than the latter which uses almost double number of cores when compared to the first one. This illustrates the trade-off between efficiency and the total computation time. For example, in our case, the fastest computation time was obtained for a configuration of 37 cores (12.11 seconds), but maximum efficiency was obtained when 9 cores were used (with a total computation time of 38.04 seconds).

3.3 Application to the Data

In this section, we present the results obtained from the joint model applied to the EGFR study. As a benchmark analysis, we present the results for the joint model when it

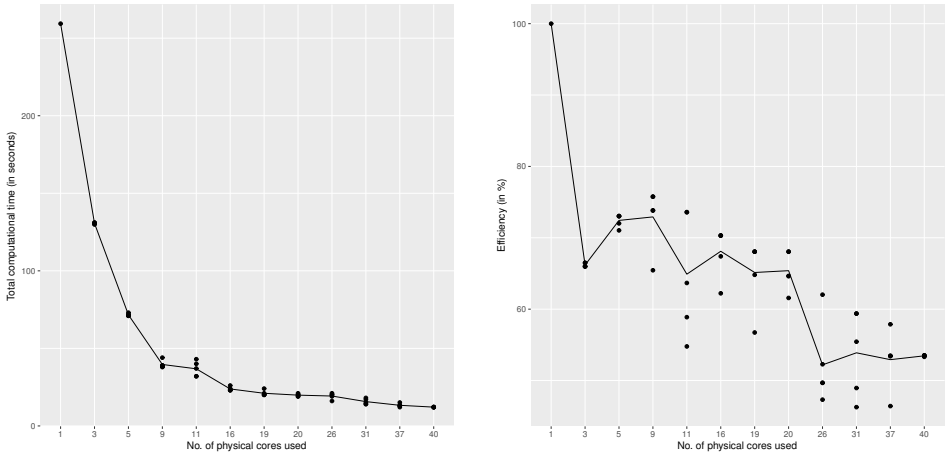


Figure 3.2: The worker framework for the EFGR study. Left panel: computation time for different combinations of number of cores/ genes per jobitem. Right panel: efficiency (in $\%$). For each configuration, the job was submitted 5 times. The solid line is the average of the 5 runs of the analysis using the same configuration.

is applied to one fingerprint feature (FF -442307337) while Section 3.3.2 presents the results when all 138 fingerprint features are included in the analysis.

3.3.1 Joint Modeling for One Fingerprint Feature

In this section, we briefly describe the analysis for a single fingerprint feature, FF -442307337, presented in Perualila *et al.* (2016a, 2016b). The aim of the analysis is to identify genes which are associated with pIC_{50} for the EGFR drug discovery project. As pointed out in Section 3.2.2, without implementation of parallelization the analysis was executed using a single `for loop` of 3595 genes and required 377.13 seconds to be completed. However, after parallelization the computation time reduced to 19.06 seconds when 20 cores were used (see Table 3.4), i.e.,19 jobitems each containing 190 genes. Table 3.5 displays a subset of genes which are both differentially expressed and significantly correlated with the pIC_{50}. For this subset of genes both null hypotheses, specified in equations (2.4) and (2.5) are rejected.

Figure 3.3 presents the 5 top genes mentioned in Table 3.5 and reveals that after correcting for a fingerprint feature effect the gene expression levels and pIC_{50} are still correlated. The adjusted association for all the genes for FF -442307337 ranges from -0.79 to 0.81.

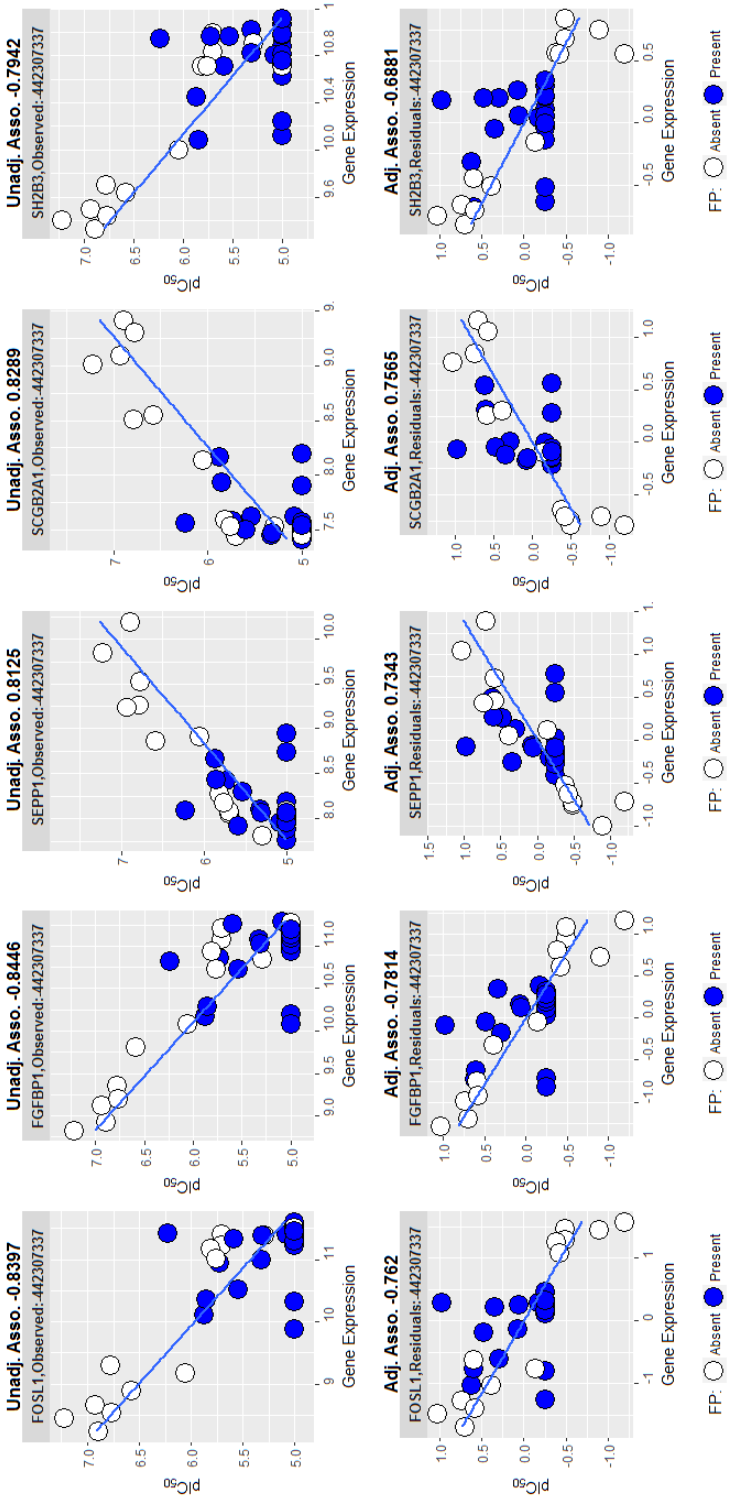


Figure 3.3: Top 5 differentially expressed genes with high adjusted correlation. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for the fingerprint feature effect.

Genes	$\hat{\alpha}_j$	adj. p-value	$\hat{r}$	$\hat{\rho}$	adj. p-value
FOSL1	1.19	0.01	-0.84	-0.76	0.00
FGFBP1	0.79	0.01	-0.84	-0.78	0.00
SEPP1	-0.64	0.01	0.81	0.73	0.00
SCGB2A1	-0.61	0.01	0.83	0.76	0.00
SH2B3	0.61	0.01	-0.79	-0.69	0.00
SLCO4A1	0.60	0.01	-0.79	-0.70	0.00
PHLDA1	0.58	0.01	-0.85	-0.77	0.00
RRM2	0.56	0.02	-0.77	-0.70	0.00
TXNIP	-0.53	0.00	0.75	0.58	0.00
CDC6	0.52	0.01	-0.80	-0.73	0.00

Table 3.5: List of top 10 differentially expressed genes with high adjusted association (adjusted p-value < 0.05) after adjusting for the effect of FF -442307337. $\hat{r}$ is the observed Pearson correlation between the expression level and pIC_{50} before adjusting for the fingerprint effect.

3.3.2 Analysis for All Fingerprint Features

The EGFR project consists of 138 fingerprint features and our aim is to implement the joint model for all 138×3595 combinations of genes and fingerprint features. Following the analysis presented in Section 3.2.4, we choose a configuration of jobitems consisting of a `for loop` over 190 genes. For a given fingerprint feature, there were 19 jobitems for configuration 20/190. Hence, for all the 138 fingerprint features, $138 \times 19 = 2622$ jobitems were distributed, taking into account the load-balancing, 44 nodes were used in the VSC cluster each with 20 cores. This configuration requires 67 seconds to complete the entire analysis and 30.64 seconds to combine the results from all the jobitems. Hence, the total computation time is 67 + 30.64 seconds (compared to 14.5 hours when a double `for loop` is used). For the configuration 880/190, we used $880/139 = 6.33$ times more cores than the configuration 139/3595 (with one master and 138 workers and each worker with a `for loop` over 3595 genes for one of the 138 fingerprint features) and the analysis was $259.35/97.64 = 2.66$ times faster.

The complete analysis described above allows us to evaluate the performance of one gene across all fingerprint features. For example, Figure 3.4 visualizes the association of gene FGFBP1 with pIC_{50} for different chemical structures. Note that, as expected, the

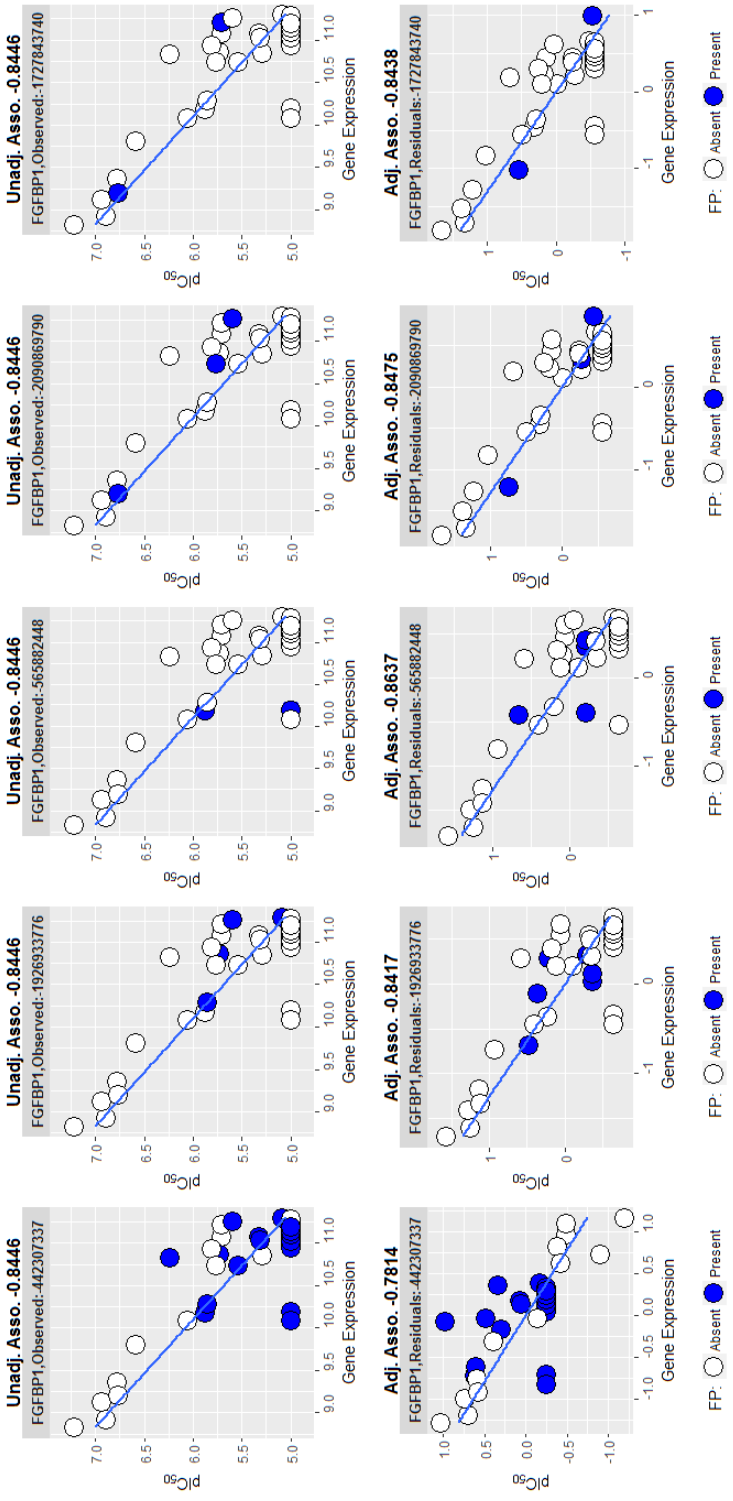


Figure 3.4: Gene FGFBP1 for different fingerprint features. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for the fingerprint feature effect.

unadjusted correlation is the same across all fingerprint feature but the adjusted correlation can vary due to different fingerprint feature effects. Furthermore, Figure 3.5 (right panel) shows the density estimate of the adjusted correlation across all fingerprint features and reveal that the lowest adjusted association is around -0.75.

Figure 3.6 presents a density estimate for the adjusted association for all genes for a specific fingerprint feature (FF -1926933776) and allows us to identify genes for which the expression levels are highly correlated with pIC_{50} after adjusting to the fingerprint feature effects. An example of a gene with positive (up regulated) and negative (down regulated) effect of the fingerprint feature, for two different fingerprint features, are shown in Figure 3.7.

3.4 Discussion

The difficulty to parallelize any job is related to the fact that there is no best method that can be used to achieve the "best" results and there are no specific guidelines since the configuration of the parallelization depends on the data analysis problem and the computing platform used for the analysis.

In this chapter, we focused on the comparison between a set of `R` functions developed for parallel programming to a worker framework configurations specified (and completely controlled) by the user. We have shown that the total computation time can be reduced from 14.5 hours, when a double `for loop` is implemented, to 97.64 seconds when the data analysis is executed in a computer cluster using parallel programing approach. Further, we have shown that, for a single `for loop` the computation time is reduced from 259.35 seconds to 19.47 seconds when it is executed as a single loop and in parallel master-slave configuration, respectively. Note that it is possible to modify the source code of different `R` functions or rewrite them in another low level language to reduce the computation time upto a certain extent, however that is not the optimized usage of the available resources in `R`. Our main aim was to keep the core of `R` same and then build on top of it. The analysis presented in this chapter was focused on an open source software and implementation using commercial software, such as SAS, was not investigated.

The analysis, presented in this chapter, reveals that currently `R` does not provide useful packages for parallelization. The existing packages can be used for parallelization but the master-slave framework that was developed for the particular joint modeling setting outperformed the `R` packages. The analysis conducted in this chapter requires an access to a computer cluster (VSC, in our case). Users without access to computer cluster can execute their analysis in publicly available cloud clusters such as *Amazon EC2 Cluster*, *Microsoft Azure Cluster*, etc. A possible implementation of the master-slave framework

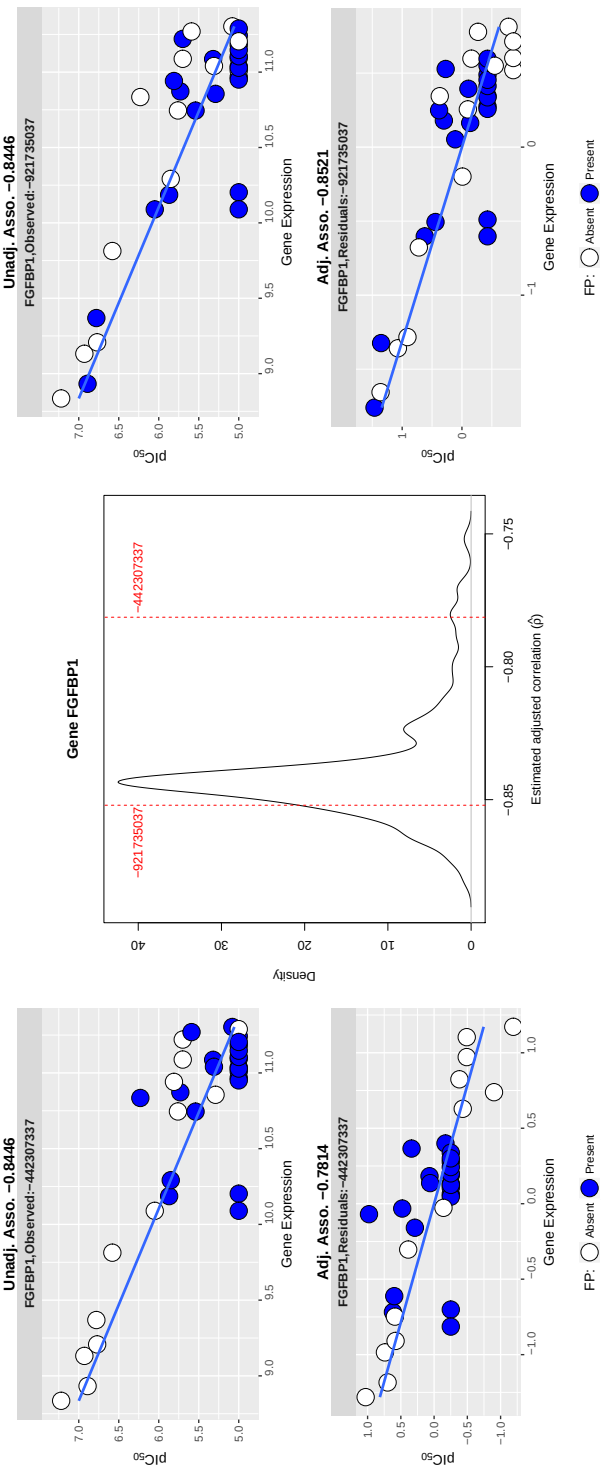


Figure 3.5: Adjusted association for gene FGFBP1. Left panel: gene expression vs pIC_{50} for gene FGFBP1 for FF -442307337. Middle panel: density of adjusted correlation for gene FGFBP1 for all fingerprint features. FF -442307337 and FF -921735037 are marked by the dashed lines at $\hat{\alpha} = -0.783$ and $\hat{\rho} = -0.852$, respectively. Right panel: gene expression vs pIC_{50} for gene FGFBP1 for FF -921735037.

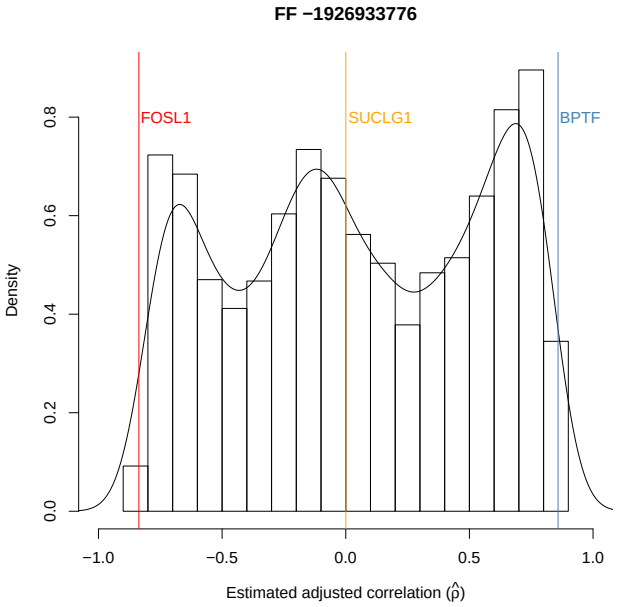


Figure 3.6: Adjusted correlation for all genes for a particular fingerprint feature (FF -1926933776). One gene with high positive adjusted correlation (BPTF), one gene with high negative adjusted correlation (FOSL1) and one gene with no significant adjusted correlation (SUCLG1) are marked with the coloured solid lines.

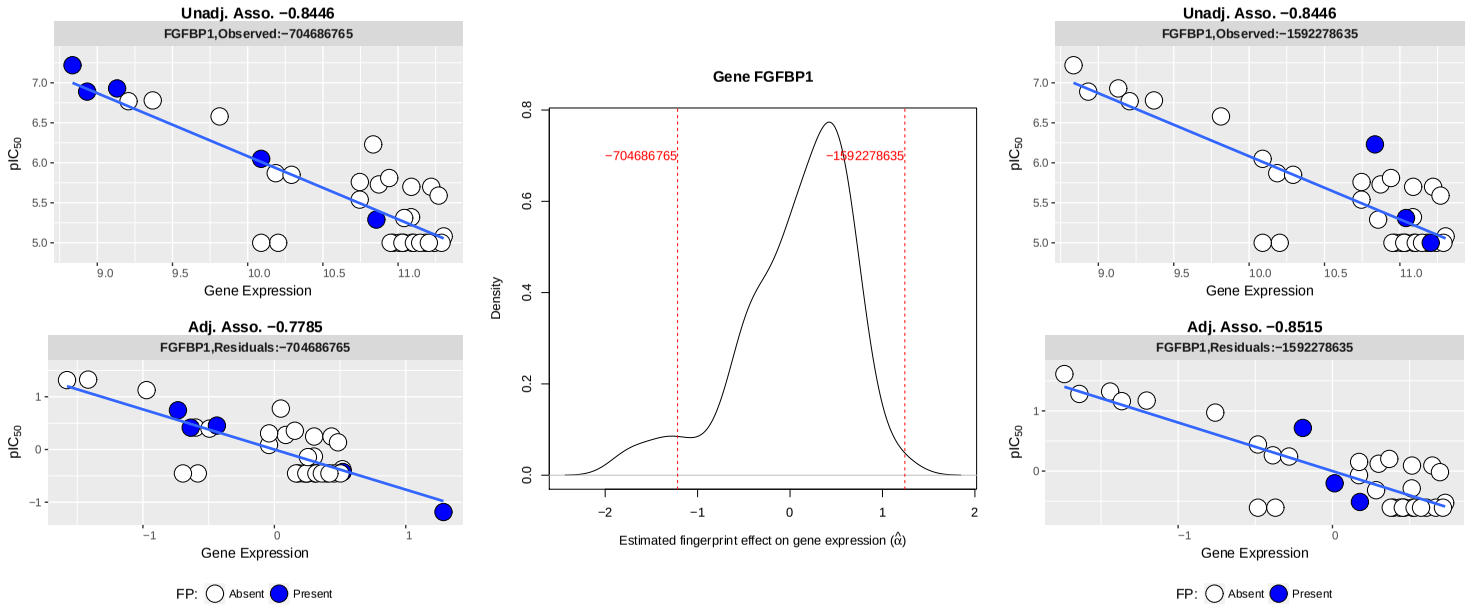


Figure 3.7: Fingerprint feature effect on gene FGFBP1. Left panel: gene expression vs pIC_{50} for gene FGFBP1 with negative effect for FF -704686765. Middle panel: density of estimated fingerprint feature effect on gene FGFBP1 for all fingerprint features. FF -704686765 and FF -1592278635 are marked by the dashed lines at $\hat{\alpha} = -1.21643$ and $\hat{\alpha} = 1.24239$, respectively. Right panel: gene expression vs pIC_{50} for gene FGFBP1 with positive effect for FF -1592278635.

using cloud computing platform is a subject of an ongoing research.
We have shown that upscaling an analysis from a setting of a single `for loop` to the setting, which become currently a standard in data analysis, of multiple `for loops` allows to analyze large scale datasets in short time, but requires a careful plan of the data analysis configuration. The analysis presented in Buyse and Molenberghs (1998) was focused on the evaluation of a single surrogate endpoint. In this chapter, we have shown that a similar analysis, upscaled for the evaluation of 496110 possible biomarkers, can be done in 97.64 seconds. The question about how to develop a surrogate, when more than one biomarker is available, will be addressed in the next chapter.

Chapter 4

Single, Multiple, Partial and Orthogonal Surrogacy: A Joint Modeling Approach

4.1 Introduction

An analysis of modern drug discovery studies requires data integration from different sources. For example, in order to investigate potential toxicity effects in early drug development, researchers need to analyze chemical structure (fingerprint feature) of the compounds, phenotypic bioactivity (bioassay readouts) data for targets of interest and transcriptomic data. Within the QSTAR modeling framework (Perualila-Tan et al., 2016a, Perualila et al., 2016b), the main goal of the analysis is to identify and construct transcriptomic biomarker(s) for a phenotypic response, taking into account possible effect of the chemical structure on both endpoints. As described in Chapter 2 and 3, for this type of data analysis problems, Perualila et al. (2016b) proposed a joint model that allows detection of genes for whose the expression level is associated with the phenotypic response. In the previous chapter, the joint model was fitted as a gene-specific model, implying that a single biomarker can be identified for the phenotypic endpoint and the association between the genes was investigated in a secondary analysis using a pathway analysis in order to discover common biological pathways among the genes associated with the phenotypic response.

In contrast to the models discussed in Chapter 2 and 3 which were focused on the

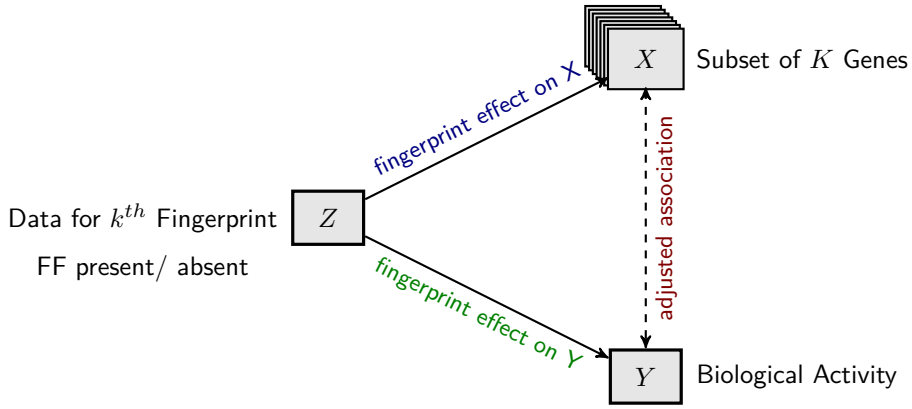


Figure 4.1: Data structure for the analysis of multiple biomarkers in drug discovery. The phenotypic variable $\mathbf{Y}$ and the fingerprint variable $\mathbf{Z}$ are $1 \times I$ vectors while the biomarker matrix is a $K \times I$ matrix $\mathbf{X}$ which contains information about the K genes.

detection of a single biomarker (per model), the analysis presented in this chapter is focused on a setting, illustrated in Figure 4.1, in which there are multiple candidates that can be used as biomarkers and our aim is to use a subset of variables as a joint biomarker or to use a new biomarker in addition to a known biomarker(s). In both cases, the attention is placed on surrogacy measures that can be used in order to quantify the surrogacy value of the subset of biomarkers or the additional biomarkers. We refer to the first case as *multiple surrogacy* while the latter is termed *partial surrogacy*. In both cases, the biomarkers under consideration can be correlated or not. The case in which a subset of K biomarkers are not correlated is termed *orthogonal surrogacy*. Let Y be the primary endpoint, $X_1, X_2, \cdots, X_K$ be a set of K candidates to be used as biomarkers for Y and Z be a condition. Throughout this chapter, we assume that the surrogacy is assessed using a joint model, similar to the model specified in equation (2.1), for the $K+1$ variables.

Alonso et al. (2016) and Perualila et al. (2016b), in the context of single high dimensional surrogacy, used the adjusted association to measure individual level surrogacy, i.e. to estimate the correlation between the biomarker and the primary endpoint, taking into account the condition effect. In this chapter, we use the *multiple adjusted association* (Van der Elst et al., 2018) to measure the joint surrogacy value of a subset of K biomarkers and the *partial adjusted association* (Sengupta et al., 2018) to quantify the surrogacy value of X_K and Y when the effect of both Z and $X_1, X_2, \cdots, X_{K-1}$ are taken into account.

This chapter is organized as follows. The modeling approach within the setting of high dimensional biomarkers is presented in Section 4.2 while the results of the analysis of

the EGFR study is presented in Section 4.3. A discussion of the method and the results are provided in Section 4.4.

4.2 High Dimensional Surrogacy

In this section, we present the modeling approach for single, multiple and partial surrogacy within the setting of high dimensional biomarker experiments. We briefly describe the joint modeling approach, proposed by Perualila-Tan et al. (2016a), for a single biomarker for high dimensional data in Section 4.2.1 while multiple, partial and orthogonal surrogacy are discussed in Section 4.2.2 and 4.2.3, respectively.

4.2.1 Modeling Approach for A Single Surrogate

The joint model for a single biomarker was described in Section 2.3. Briefly, we consider a joint model given by,

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{X_j} + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \mathbf{\Sigma}_j\right].$$

As pointed out in Chapter 2 and 3, biomarker detection is based on two sets of hypotheses,

$$H_{0_j} : \alpha_j = 0,$$
$$H_{1_j} : \alpha_j \neq 0,$$

and

$$H_{0_j} : \rho_j = 0,$$
$$H_{1_j} : \rho_j \neq 0.$$

Here, α_j and ρ_j are gene-specific fingerprint effect and adjusted association, respectively. We can classify genes into four different subclasses (Table SA1 in Appendix A), based on the inference results of the above hypotheses.

4.2.2 Multiple Surrogacy

The joint model, described in the previous section, allows us to identify a single biomarker for biological readout pIC_{50}. In this section, we discuss the case that a subset of K genes are used as a joint surrogate for the pIC_{50}. Note that, we assume that the subset is known. For example, a subset of K genes of the same biological pathway. Using the terminology introduced by Van der Elst et al. (2018) we term the subset of the K genes, a multiple surrogate for pIC_{50}. A joint model can be formulated in the following way:

$$\begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{Ki} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 + \alpha_1 Z_i \\ \mu_2 + \alpha_2 Z_i \\ \vdots \\ \mu_K + \alpha_K Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \mathbf{\Sigma} \right], \tag{4.1}$$

with a $(K+1)\times(K+1)$ covariance matrix, $\mathbf{\Sigma}$ and correlation matrix, $\mathbf{P}$ given by, respectively,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} & \sigma_{1y} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2k} & \sigma_{2y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} & \sigma_{ky} \\ \sigma_{y1} & \sigma_{y2} & \dots & \sigma_{yk} & \sigma_{yy} \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1k} & \rho_{1y} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2k} & \rho_{2y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & \rho_{kk} & \rho_{ky} \\ \rho_{y1} & \rho_{y2} & \dots & \rho_{yk} & \rho_{yy} \end{pmatrix}. \tag{4.2}$$

The adjusted correlation between two biomarkers and between the jth biomarker and the pIC_{50} are respectively given by,

$$\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} \quad \text{and} \quad \rho_{yj} = \sigma_{yj}/\sqrt{\sigma_{yy}\sigma_{jj}}. \tag{4.3}$$

Following Van der Elst et al. (2018), the covariance matrix specified in equation (4.2), $\mathbf{\Sigma}$, can be rewritten as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{X,X} & \mathbf{\Sigma}'_{X,Y} \\ \mathbf{\Sigma}_{X,Y} & \sigma_{Y,Y} \end{pmatrix}, \tag{4.4}$$

and the multivariate adjusted association, denoted by γ^2, is given by,

$$\gamma^2 = \rho^2_{Y,X_1,X_2,\dots,X_K} = \frac{\Sigma_{X,Y}\Sigma^{-1}_{X,X}\Sigma'_{X,Y}}{\sigma_{Y,Y}}. \tag{4.5}$$

Note that for a single biomarker $\sqrt{\gamma^2}$ is same as the adjusted association.

4.2.3 Partial Surrogacy

The joint model, formulated in equation (4.1), allows us to evaluate the surrogacy value of a single biomarker or a subset of K genes as a multiple biomarker for the biological readout, respectively. Often in drug discovery experiments, the primary interest is to estimate the surrogacy effect of the Kth biomarker, given the surrogacy effect of $K-1$ biomarkers. For simplicity, let us assume that $K=2$, i.e., we consider a case with two potential biomarkers and let us assume that the first biomarker is already identified. Our

main interest is to estimate the added surrogacy value of the second biomarker given the first biomarker. For $K = 2$, the joint model can be rewritten as

$$\begin{pmatrix} X_{1i} \\ X_{2i} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 + \alpha_1 Z_i \\ \mu_2 + \alpha_2 Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \mathbf{\Sigma_{12}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1y} \\ \sigma_{21} & \sigma_{22} & \sigma_{2y} \\ \sigma_{y1} & \sigma_{y2} & \sigma_{yy} \end{pmatrix} \right]. \tag{4.6}$$

We term the surrogacy effect of X_2 given X_1 and Z, the *partial surrogacy effect*. The partial surrogacy effect can be estimated using the partial correlation which is, in our setting, the *partial adjusted association*, $\rho_{Y,X_2|X_1,Z}$ that measures the adjusted association between the second biomarker and the biological readout, given the chemical structure and the first biomarker in the model. The partial adjusted association (Pindyck and Rubinfeld, 1976, Johnston, 1984), among the quadruplet $(Y_i, X_{1i}, X_{2i}, Z_i)$, can be estimated by,

$$\rho_{Y,X_2|X_1,Z} = \rho_{Y,X_2|X_1} = \frac{\rho_{y2} - \rho_{y1}\rho_{12}}{\sqrt{(1-\rho_{y1}^2)(1-\rho_{12}^2)}}, \tag{4.7}$$

where, $\rho_{y\ell}$ and $\rho_{\ell j}$ are given in equation (4.3). In order to determine the partial correlation between two variables, given a third variable, the effect of the third variable needs to be removed from the other two. Let us assume that both biological readout Y and X_2 are fingerprint adjusted and are regressed on (fingerprint adjusted) X_1, that is,

$$\begin{aligned} Y_i^* &= \alpha_0 + \alpha_1 X_{1i}^* + \epsilon_{1i}, \\ X_{2i}^* &= \beta_0 + \beta_1 X_{1i}^* + \epsilon_{2i}. \end{aligned} \tag{4.8}$$

The partial adjusted association is the estimated correlation between ϵ_{1i} and ϵ_{2i} (Pindyck and Rubinfeld, 1976). Figure 4.2 illustrates three scenarios of partial surrogacy. The left panels in Figure 4.2 show the pairwise scatterplots between the three variables, Y^*, X_1^*, X_2^*, i.e., the variables after adjusting for the grouping variable and right panels display the residuals obtained from the regression models specified in equation (4.8).

In the first scenario, presented in Figure 4.2a, all three variables are independent while in Figure 4.2b the variables are highly correlated with each other. Thus, for both cases the partial adjusted association between Y and X_2, given X_1 (and Z) is close to zero, indicating that in both scenarios X_2 is not contributing as a biomarker for Y given that X_1 is used as a biomarker. Note that in the first scenario, the adjusted association of X_1 and Y is expected to be close to zero as well which is not the case under the second scenario. Figure 4.2c illustrates a third scenario in which the partial adjusted association between Y and X_2, after removing the effect of X_1 and Z from both the variables, is relatively high indicating that X_2 can be used as a biomarker for Y in addition to X_1.

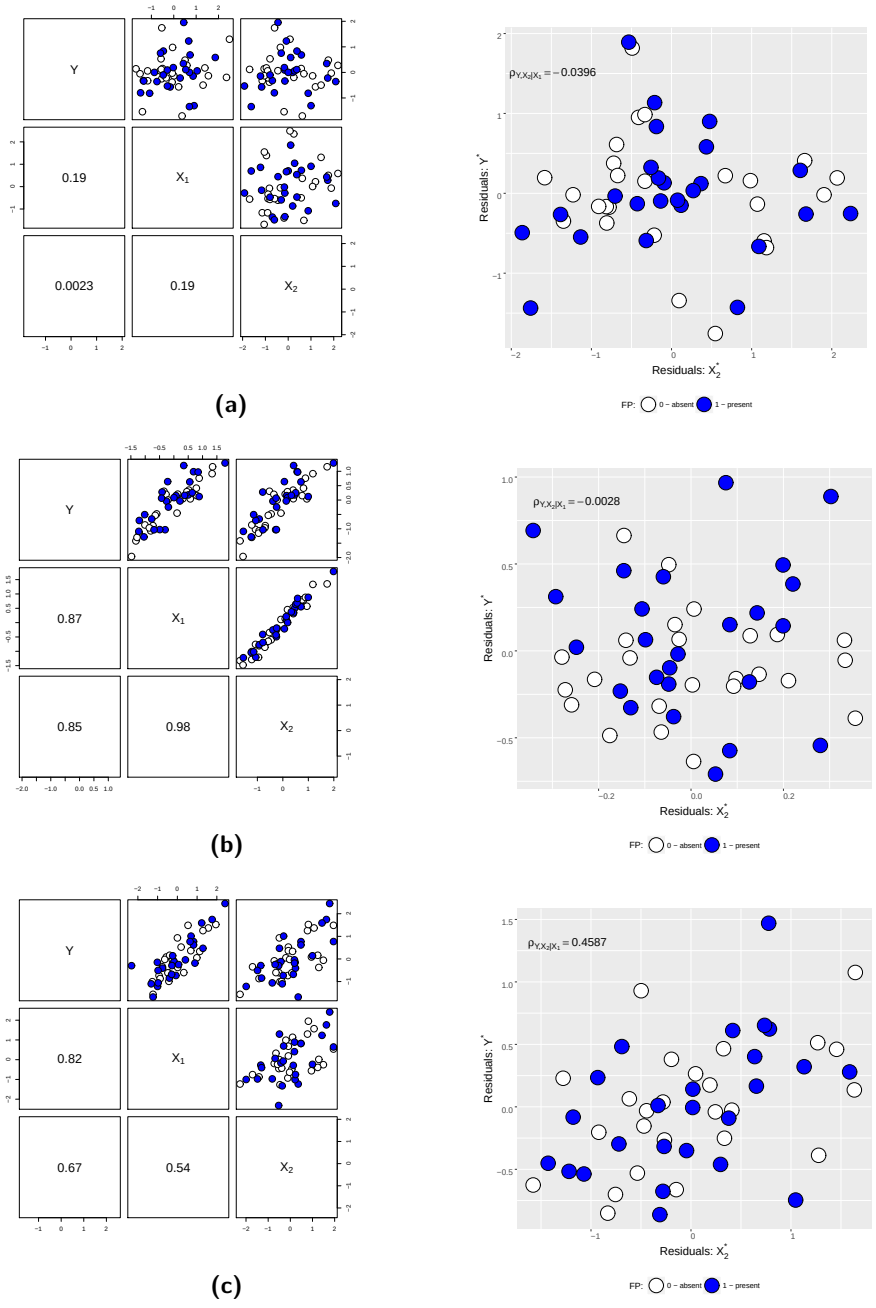


Figure 4.2: Graphical illustration of partial surrogacy for three hypothetical scenarios. Panel a: low correlation between the three variables and low partial adjusted correlation. Panel b: three correlated variables and low partial adjusted correlation between all the variables results in a low partial correlation. Panel c: three correlated variables and relatively high partial adjusted correlation.

4.2.4 Orthogonal Surrogacy

Orthogonal surrogacy is a special case of partial surrogacy in which X_1 and X_2 are uncorrelated but both are correlated with Y. In this case, the partial adjusted association between X_2 and Y is expected to be high since X_1 does not explain the variation of X_2. Note that in that case the covariance matrix in equation (4.6) is reduced to,

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & 0 & \sigma_{1y} \\ 0 & \sigma_{22} & \sigma_{2y} \\ \sigma_{y1} & \sigma_{y2} & \sigma_{yy} \end{pmatrix}. \tag{4.9}$$

This implies that the partial adjusted association in equation (4.7) can be rewritten as

$$\rho_{Y,X_2|X_1,Z} = \rho_{Y,X_2|X_1} = \frac{\rho_{y2}}{\sqrt{(1-\rho_{y1}^2)}}. \tag{4.10}$$

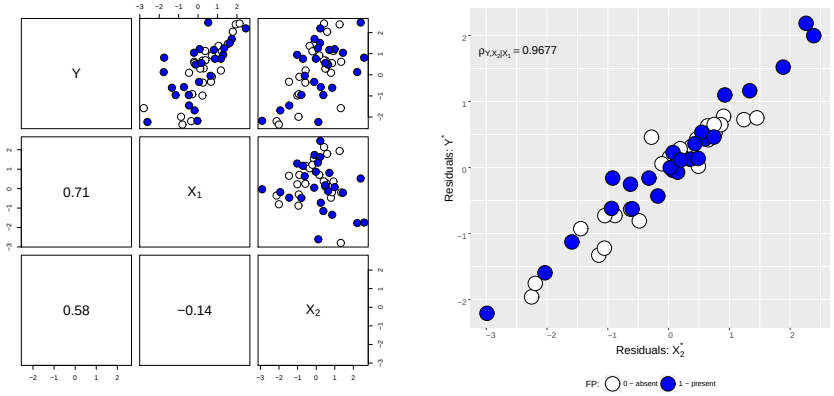


Figure 4.3: Graphical illustration of orthogonal surrogacy.

Figure 4.3 displays a hypothetical scenario for orthogonal surrogacy for the case with,

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 1.29 & -0.18 & 1.05 \\ -0.18 & 1.29 & 0.86 \\ 1.05 & 0.86 & 1.71 \end{pmatrix}, \quad \hat{\mathbf{P}} = \begin{pmatrix} 1 & -0.14 & 0.71 \\ -0.14 & 1 & 0.58 \\ 0.71 & 0.58 & 1 \end{pmatrix}.$$

For this example, the adjusted association is $\hat{\rho}_{Y,X_1} = 0.71$, the multiple adjusted association is equal to $\hat{\gamma}^2 = 0.95$ and the adjusted partial association is equal to $\hat{\rho}_{Y,X_2|X_1} = 0.97$. As shown in Figure 4.3b, the residuals obtained from the regression models, specified in equation (4.8), are highly correlated.

A test for orthogonal surrogacy can be based on a likelihood ratio test for the models, specified in equation (4.6) and equation (4.9).

$$H_0 : \sigma_{12} = 0,$$
$$H_1 : \sigma_{12} \neq 0.$$

An extension of the analysis for the case of $K > 2$ or the case in which a subset of K_2 biomarkers is considered in addition to a subset of K_1 biomarkers is straight forward.

4.3 Application to the Data

4.3.1 Single Surrogacy

The joint model formulated in equation (2.1) was applied to the EGFR data using the `R` package `IntegratedJM` (Perualila et al., 2016b, Sengupta and Perualila, 2017). In total 3595 gene-specific models were fitted and for 1192 genes (Table 4.1), the null hypotheses $H_0 : \alpha = 0$ and $H_0 : \rho = 0$, specified in equation (2.4) and (2.5), were rejected. Figure 4.4 and Table 4.2 display the top 5 genes which are differentially expressed as well

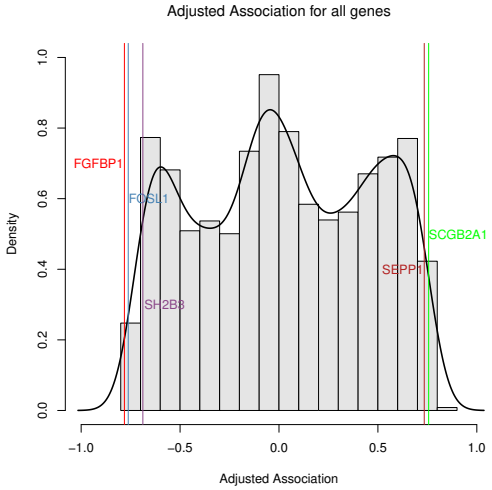
		$\rho_{Y,X}$	
		$\neq 0$	0
α	$\neq 0$	1192	109
	0	688	1606

Table 4.1: Results for FF -442307337 (EGFR) at 5%.

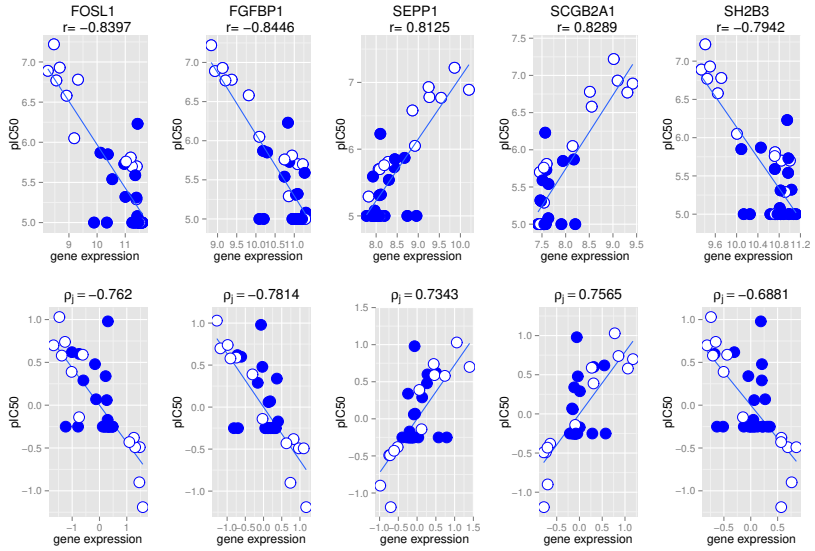
as significantly associated with pIC_{50}. Most of these genes are known to participate in biological processes involving cell proliferation (positive and negative), survival and differentiation. Note that both FOSL1 and FGFBP1 genes were identified as biomarkers for pIC_{50} by Verbist et al. (2015) and Perualila-Tan et al. (2016a). For illustration, in the next section the gene FOSL1 is used as a primary biomarker (X_1, using the notation in Section 4.2.3).

4.3.2 Multiple Surrogacy

The joint model specified in equation (4.6) was fitted to all 3494 subsets of two genes, i.e., the quadruplets $(FOSL1_i,\ X_{ji}, pIC_{50i}, Z_i), j = 1, \ldots, 3494$. Table 4.3 lists the top 5 genes with highest multiple adjusted association. The gain in surrogacy value, for using X_2 as a biomarker together with FOSL1, is equal to $\rho^2_{Y,FOSL1,X_2} - 0.5776$. Figure 4.5 displays the density plot for the multiple adjusted association for all the remaining 3594



(a)



(b)

Figure 4.4: Adjusted association. Panel a: density of adjusted association for all the genes with the top genes marked by their level of association. Panel b: the correlation between the gene expression and the inhibitory activity against EGFR, given by the pIC_{50}, of the compounds grouped by the substructure FF -442307337. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for the effect of FF -442307337.

Genes	FP-Effect	adj. p-val(α)	Unadj.Asso.	$\rho_{Y,X}$	adj. p-val($\rho_{Y,X}$)
FOSL1	1.19	0.01	-0.84	-0.76	0.00
FGFBP1	0.79	0.01	-0.84	-0.78	0.00
SEPP1	-0.64	0.01	0.81	0.73	0.00
SCGB2A1	-0.61	0.01	0.83	0.76	0.00
SH2B3	0.61	0.01	-0.79	-0.69	0.00

Table 4.2: Top 5 differentially expressed genes with high adjusted correlation for FF -442307337 (EGFR) at 5% FDR.

Genes	ρ_{Y,X_2}	ρ^2_{Y,X_2}	$\rho^2_{Y,FOSL1,X_2}$	Gain in Surrogacy Value
MPHOSPH9	-0.2586	0.0669	0.6881	0.1105
TOP2A	-0.3474	0.1207	0.6876	0.1100
MYO6	0.7319	0.5357	0.6833	0.1057
PNISR	0.7562	0.5718	0.677	0.0994
EREG	-0.5987	0.3518	0.6714	0.0938

Table 4.3: Top 5 genes, sorted according to their multiple adjusted association, when used together with FOSL1. $\rho^2_{Y,FOSL1} = 0.5776$.

genes, given FOSL1 and shows that γ^2 ranges from 0.5806 to 0.6881, indicating that the gain in surrogacy value ranges between 0.003 and 0.1105.

4.3.3 Partial Surrogacy

Similar to the previous section, we consider the FOSL1 gene as the primary biomarker and use the joint model specified in equation (4.6) in order to estimate the partial adjusted association. Figure 4.6a and Table 4.4 present the density estimate for the distribution of the partial adjusted association and the top 5 genes, with highest partial correlation, respectively.

Partial surrogacy effect of the gene MPHOSPH, given FOSL1 and fingerprint feature, is displayed in Figure 4.6b. Adjusted association between pIC_{50} and MPHOSPH9 is negative with a smaller magnitude when compared with the adjusted association between FOSL1 and pIC_{50}. However, the two genes are positively associated with each other, resulting in a positive partial adjusted association between MPHOSPH9 and the pIC_{50}, given the chemical structure and FOSL1.

An example of the TCIRG1 gene for which the estimated partial adjusted association is zero, is shown in Figure 4.7. Note that, even though the gene is originally negatively

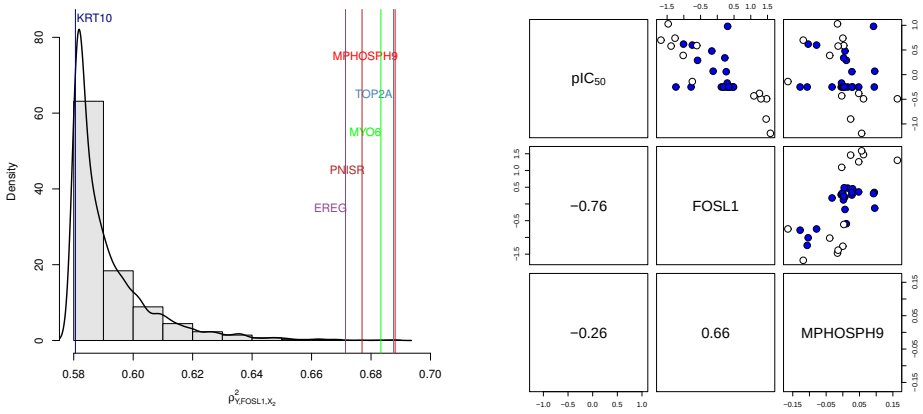


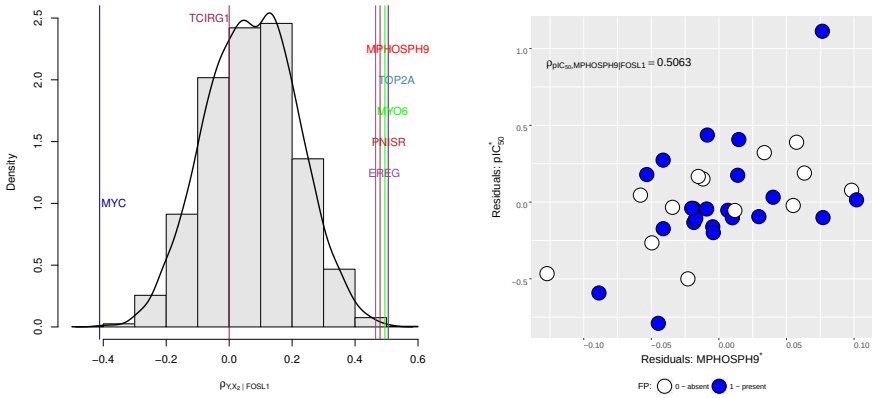
Figure 4.5: Multiple Surrogacy. Left panel: density of multiple adjusted association for all the 3494 pairs, when used as a biomarker in combination with FOSL1. Right panel: three-way association between fingerprint feature adjusted genes MPHOSPH9 and FOSL1 and fingerprint feature adjusted pIC_{50}.

Genes	ρ_{Y,X_2}	$\rho_{Y,X_2\|FOSL1}$
MPHOSPH9	-0.2586	0.5063
TOP2A	-0.3474	0.5051
MYO6	0.7319	0.4949
PNISR	0.7562	0.4794
EREG	-0.5987	0.4654

Table 4.4: Top 5 genes, sorted according to the partial correlation given FOSL1.

associated with pIC_{50}, after adjusting for gene FOSL1 and the fingerprint feature it becomes partially uncorrelated with pIC_{50}.

Verbist et al. (2015) identified both the genes FOSL1 and FGFBP1 as good biomarkers (also shown in Table 4.2). However, as displayed in Figure 4.8 (left panel), gene FGFBP1 is highly correlated ($\rho_{FOSL1,FGFBP1} = 0.95$) with gene FOSL1. Thus, there is no significant gain in surrogacy value by using them together as a joint biomarker even though both the genes can play an important role when the biological processes are considered.



(a) **(b)**

Figure 4.6: Partial surrogacy in the EGFR dataset. Panel a: density of partial adjusted association for all the genes, with selected genes marked by their level of association, given FOSL1. Panel b: partial correlation between MPHOSPH9 and pIC_{50}, given FOSL1.

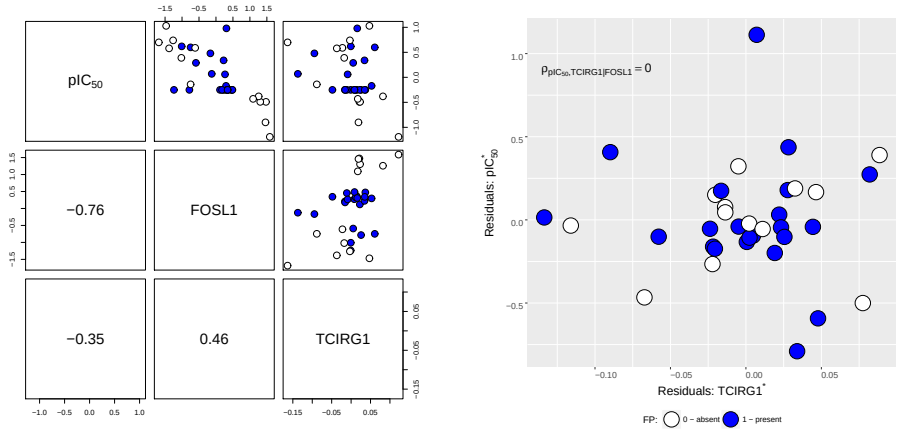


Figure 4.7: Gene TCIRG1. Left panel: three-way association between fingerprint feature adjusted genes TCIRG1 and FOSL1 and fingerprint feature adjusted pIC_{50}. Right panel: gene TCIRG1 has 0 partial surrogacy effect on pIC_{50}, given FOSL1.

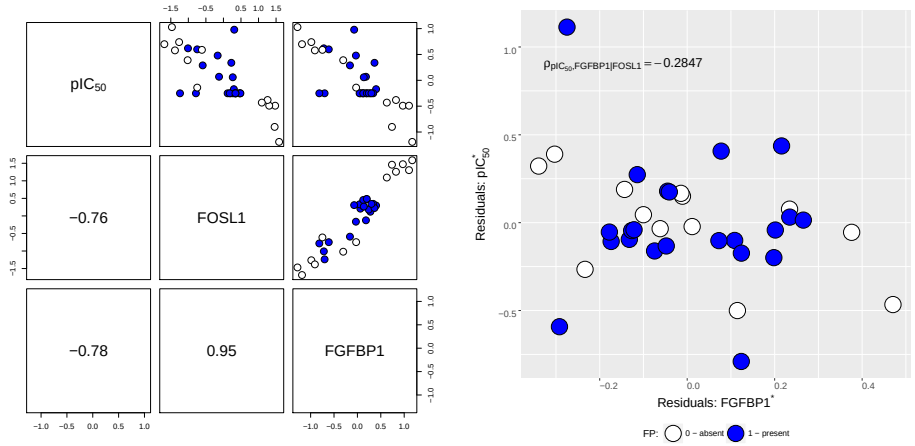


Figure 4.8: Gene FGFBP1. Left panel: three-way association between fingerprint feature adjusted genes FGFBP1 and FOSL1 and fingerprint feature adjusted pIC_{50}. Right panel: gene FGFBP1 has low partial surrogacy effect ($\rho_{Y,FGFBP1|FOSL1} = -0.2847$) on pIC_{50}, given FOSL1. $\gamma^2_{Y,FOSL1,FGFBP1} = 0.6146$. Gain in surrogacy value = 0.6146 - 0.5776 = 0.037.

4.4 Discussion

The single surrogacy approach within the high dimensional surrogacy was proposed, in the context of drug discovery experiments, by Perualila *et al.* (2016a,2016b). In this chapter, we extended the modeling approach to allow the evaluation of a subtest of K biomarkers. Van der Elst et al. (2018) focused on the joint surrogacy effect, i.e. the multiple surrogacy effect, in order to evaluate a subset of K biomarkers as a joint surrogate.

In drug discovery, often the primary interest is to estimate the added surrogacy effect of the Kth biomarker given that $K-1$ biomarkers were already identified. We have shown that the partial adjusted surrogacy approach can be used to estimate the association between the Kth biomarker and the response given the effect of $K-1$ biomarkers on both endpoints. The same methodology can be applied when the biomarkers are outcomes from different platforms, for example, transcriptomics data, metabolomics data etc.

Part II

Detection of Biomarkers in Microbiome Intervention Studies

Chapter 5

Microbiome Intervention Studies: An Introduction

5.1 Introduction

Human microbiome is defined as the collection of microorganisms or microbes which inhabit different sites of the human body (e.g., skin, gut etc.) in large number. Typically, microbiome cells in the human body outnumber human cells in a ratio of 10 to 1 (Sender et al., 2016). Researchers have reported that microbiome is related to different diseases (Table 5.1). Furthermore, Grice and Segre (2012) showed that microbiome may contribute to a link between genetic variation and diseases. Different sequencing methods, such as 16s RNA (Eckburg et al., 2005), shotgun (Ranjan et al., 2016), have been developed to measure microbiome data. Details about the 16s RNA sequencing method is given in Section 5.2.1.

According to The Human Microbiome Project Consortium (2012), different types of microorganisms which are part of the human microbiome form an ecosystem and, as shown in Figure 5.1, there is a hierarchy in the ecosystem which is known as the phylogenetic tree or often referred to as taxonomy (Matsen, 2015). The lowest level in the hierarchical ecosystem is the bacteria species and the highest level is the kingdom, representing the microbiome activity in the host. The microbiome measurements are collected at species level, using the methods mentioned above, and summarized for different levels in the hierarchy. We elaborate on the different microbiome measurements in Section 5.2.

The Human Microbiome Project (HMP) (Turnbaugh et al., 2007, The Human Microbiome Project Consortium, 2012) began with the focus on analyzing microbiome ecosys-

Publication	Associated Disease(s)
Kostic et al. (2014)	Inflammatory Bowel Disease
Llorente and Schnabl (2015)	Liver Disease
Parekh et al. (2015)	Obesity, Metabolic Syndrome and Gastrointestinal Disease
Sanz et al. (2015)	Metabolic Disease
John and Mullin (2016)	Obesity
Tedjo et al. (2016)	CrohnâĂŹs Disease
Johnson et al. (2017)	Metabolic Disease
Pascal et al. (2017)	Crohn's Disease

Table 5.1: Different diseases associated with human microbiome.

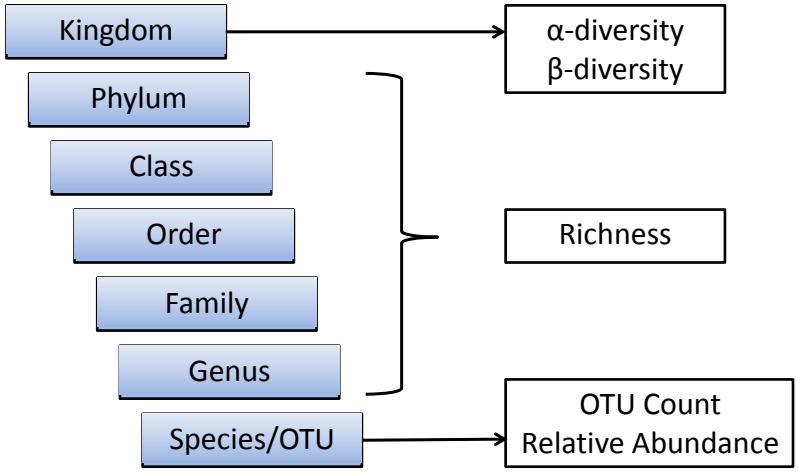


Figure 5.1: Hierarchical microbiome ecosystem with available measurements at different levels. OTU count is the count of a specific type of bacteria. Richness is the number of active OTUs at different levels of the microbiome ecosystem. α-diversity is the richness at the Kingdom level at a specific site (e.g., skin, gut etc.) and β-diversity is the difference in richness between different sites.

tem and linking these datasets to different diseases (Bäckhed et al., 2012). For example, Llorente and Schnabl (2015), Parekh et al. (2015) and John and Mullin (2016) showed a link between gut microbiome data and liver disease, gastrointestinal disease and obesity, respectively. Rooks and Garrett (2016) and Round and Mazmanian (2009) investigated the link between microbiome ecosystem and the immune system of the host. In the context of pharmaceutical industry, as discussed by Valencia et al. (2017), the number of microbiome-based therapeutic studies are increasing rapidly. Figure 5.2 provides an overview of the ongoing microbiome experiments in different phases of clinical and preclinical trials related to different therapeutic areas. In parallel, experiments to study the effect of early-life or short-term perturbation on microbiome composition are being conducted as well (Jakobsson et al., 2010, Gasparrini et al., 2016).

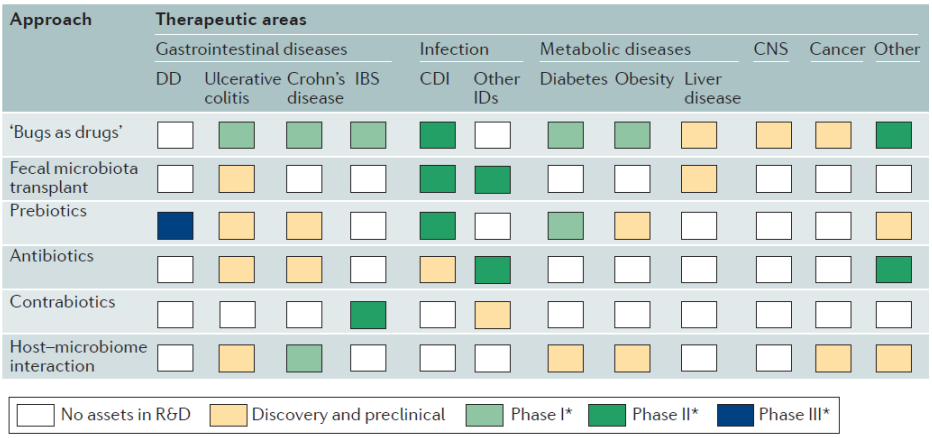


Figure 5.2: Overview of microbiome projects concerning different therapeutic areas and focusing on different disease (Valencia et al., 2017). CDI, Clostridium difficile infection; CNS, central nervous system; DD, digestive disorders; IBS, irritable bowel syndrome; ID, infectious diseases.

The immune system of a host is assumed to be affected by the microbiome (Rooks and Garrett, 2016, Levy et al., 2017) and in addition the host's life habits such as food intake, sports and exercise, drug intake etc. can affect the development of the host's microbiome community. Figure 5.3 illustrates the experimental setting for microbiome intervention studies that will be analyzed in this part of the thesis. Our main goal is to investigate how the intervention factors influence the microbiome and other clinical endpoint(s), as well as the association between microbiome and the clinical endpoint(s), taking the intervention effect into account.

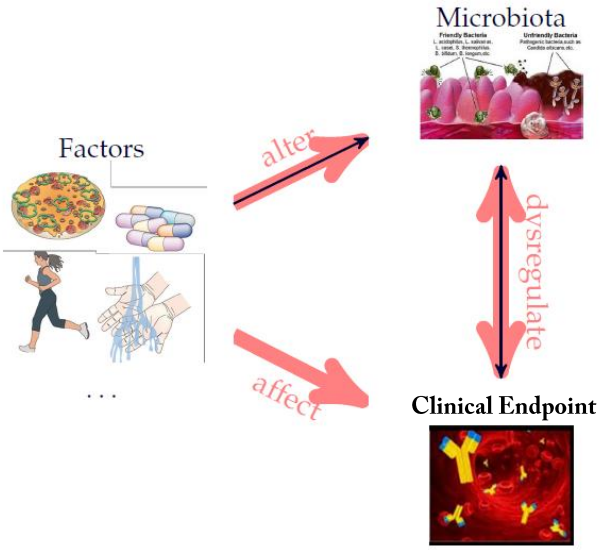


Figure 5.3: The experimental setting for the microbiome intervention experiments analyzed in the second part of the thesis. Our aim is to model the association between clinical variables such as immunological variables, time to develop a disease, etc. and the microbiome variables (at different levels of the hierarchy of the microbiome ecosystem), taking into account the intervention effect.

5.2 Microbiome Measurements at Different Levels of the Phylogenetic Tree

Throughout this part of the thesis, different microbiome measurements, corresponding to the different levels of the hierarchical microbiome ecosystem, are used.

5.2.1 Operational Taxonomic Unit (OTU)

As shown in Figure 5.1, the lowest level in the hierarchy of the microbiome ecosystem is the species level or the Operational Taxonomic Unit (OTU). A standard method to measure microbiome at OTU level is the 16s sequencing method (Eckburg et al., 2005). First, the DNA (collected from a subject under investigation) is sequenced. After the sequencing step, a clustering analysis is conducted where sequences with 97% similarity are clustered as one OTU. The basic microbiome measurement for an OTU is the count

of a specific type of bacteria, related to that specific OTU, in a particular subject. For an experiment with n subjects for which m OTUs are measured, the $m \times n$ microbiome data matrix is given by,

$$\mathbf{X} = \left.\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ . & . & . & . \\ x_{j1} & x_{j2} & \cdots & x_{jn} \\ . & . & . & . \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}\right\} m \text{ features} . \tag{5.1}$$

Here, x_{ji} is the count of the jth OTU (feature) for the ith subject (sample). Relative Abundance (RA) of any OTU, for a given sample, is calculated by dividing the raw count of the OTU by the *total library size*, i.e., the total count of all the OTUs, for that particular sample. Relative abundance of the jth OTU for ith sample is given by,

$$x'_{ji} = x_{ji} / \sum_{\ell=1}^{m} x_{\ell i}. \tag{5.2}$$

5.2.2 α-Diversity

Let x''_{ji} be an indicator variable defined by,

$$x''_{ji} = \begin{cases} 1, & \text{if the } x_{ji} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

The α-diversity (Morgan and Huttenhower, 2012) for the ith sample is defined as the number of active OTUs for that sample, given by,

$$d_i = \sum_{\ell=1}^{m} x''_{\ell i}. \tag{5.4}$$

Different measures of α-diversity such as Chao1 (Chao, 1984), Shannon Index (Shannon, 1948), Simpson Index (Simpson, 1949) etc. have been developed. The Chao1 measure for α-diversity corrects the observed diversity measure for singletons and doubletons in microbiome count data. The Shannon Index for the ith subject is defined by,

$$H'_i = \sum_{j} (x'_{ji} ln(x'_{ji})). \tag{5.5}$$

The Simpson Index for the ith subject is given by,

$$D_i = \sum_j x'^2_{ji}, \tag{5.6}$$

where x'_{ji} is the relative abundance of the jth OTU for the ith subject.

5.2.3 Family Level Richness

A microbiome family is defined as a collection of Genera which consist of different Species (Figure 5.1). The richness of a family is the α-diversity defined at family level and can be used as a measure of the microbiome activity for a family. For an experiment with p families, the richness matrix is given by,

$$\mathbf{F} = \left.\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ . & . & . & . \\ f_{j1} & f_{j2} & \cdots & f_{jn} \\ . & . & . & . \\ f_{p1} & f_{p2} & \cdots & f_{pn} \end{pmatrix}\right\} p \text{ families}, \tag{5.7}$$

where f_{ji} is the total number of active OTUs, belonging to the jth family, for the ith sample and is defined by,

$$f_{ji} = \sum_{\ell \in j} x''_{\ell i}. \tag{5.8}$$

5.3 Microbiome Intervention Studies

5.3.1 TransPAT Study

5.3.1.1 Study Design

The transPAT experiment (Figure 5.4) is an animal study that was conducted in order to investigate the influence of antibiotic treatment on both immune system and micrbiome of the subject. The transPAT dataset contains information about 15 germ-free mice that were recipients of cecal contents of donor mice. The cecal contents of mice, exposed to a single tylosin pulse for 5 days, were transferred to 7 germ-free mice, which comprised the PAT (Pulsed Antibiotic Treatment) group. The remaining 8 control mice received cecal contents of donor mice that were not exposed to antibiotics. The microbiome data, used for the analysis presented in this part of the thesis, were measured at day 1, 6, 12 and 20 during the experiment and Immunoglobulin A (IgA) level was measured at day 20. For an elaborate discussion about the transPAT study design, we refer to Ruiz et al. (2017).

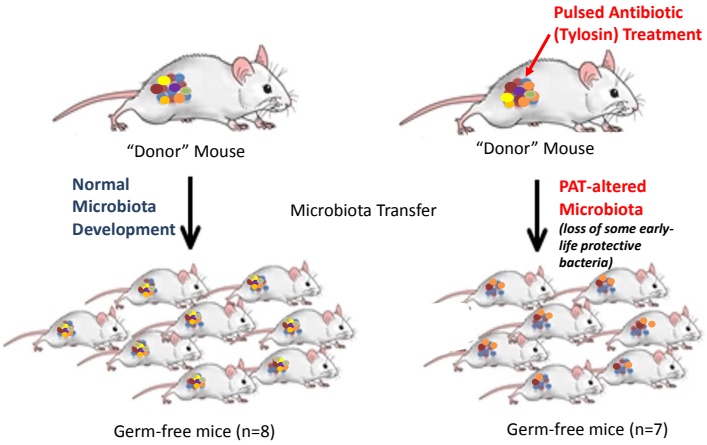


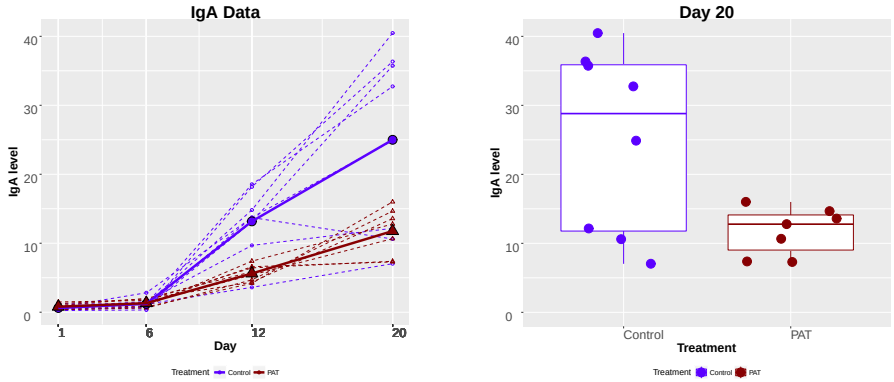
Figure 5.4: Illustration of the transPAT study (Ruiz et al., 2017).

Immunological Data

The level of secretory Immunoglobulin A (IgA), the most abundant of immunoglobulins in the gut lumen, can indicate the immunological condition of the host and is considered as an important line of defense protecting the host against pathogenic bacteria. As shown in Figure 5.5, all mice started with low IgA levels at the beginning of the experiment. Thereafter, IgA levels of the two groups evolved differently over time. As mentioned above, for the analysis presented in this part of the thesis, the interest lies only on the fecal IgA measurements at day 20 where the largest difference in IgA level was observed between treatment groups.

Microbiome Data

The microbiome sample of each mouse was sequenced at a given timepoint. The sequences are clustered and assigned to an operational taxonomic unit (OTU) with representation $> 0.01\%$ in relative abundance. The OTU counts were recorded for 355 OTUs. The library size, i.e., the depth of coverage of the sequencing varies across samples (McMurdie and Holmes, 2014). Hence, in order to make the microbial abundance comparable across samples, the relative abundance is used. OTU counts in the datasets have a high proportion of zeros. Since the analysis is feature-specific, features having zero counts in at least 70% of the samples within each treatment group (i.e., at least $n_{Control} = 6$ and $n_{PAT} = 5$ with zero counts) were excluded. In total, 30, 56, 67 and 87 OTUs analysed at day 1, 6, 12 and 20, respectively (Figure SB1 in Appendix B).



(a) **(b)**

Figure 5.5: Immunological data for the transPAT experiment. Panel a: the IgA level over time. Panel b: boxplot of the IgA level of 15 mice by treatment for day 20.

5.3.1.2 Data Structure

Per day, the transPAT data structure consists of,

$$\mathbf{X_t} = \left.\begin{pmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ . & . & . & . \\ x'_{j1} & x'_{j2} & \cdots & x'_{jn} \\ . & . & . & . \\ x'_{m1} & x'_{m2} & \cdots & x'_{mn} \end{pmatrix}\right\} m \text{ features}, \quad \mathbf{Y}' = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \quad \mathbf{Z}' = \begin{pmatrix} z_1 \\ z_2 \\ . \\ . \\ . \\ z_n \end{pmatrix}. \tag{5.9}$$

Here, $\mathbf{X_t}$, is the matrix with relative abundance of m OTUs (rows) and n subjects (columns), $\mathbf{Y}$ is the IgA vector and $\mathbf{Z}$ is an indicator vector for treatment such that,

$$Z_i = \begin{cases} 1, & \text{if the } i\text{th subject belongs to the PAT group ,} \\ 0, & \text{if the } i\text{th subject is from the Placebo group.} \end{cases}$$

Since the variation in the IgA levels at day 20 can be presumed to be influenced by or associated with the microbiota at an earlier timepoint, the association between the relative abundance from the previous timepoints ($t = 1, 6, 12, 20$ days) and the IgA level at day 20, accounting for the effect of the treatment, are of interest. Figure 5.6 displays an example of the relative abundance for OTU 997439 against the IgA level at different days.
The analysis presented in this part of the thesis is focused on the detection of microbiome

biomarkers (at different levels of the hierarchical ecosystem) for IgA. In **Chapter 6**, we analyze the transPAT study using a joint model, formulated in Chapter 2, for microbiome measurements and IgA. A non parametric approach is presented in **Chapter 7** while **Chapter 8** presents a hierarchical Bayesian model.

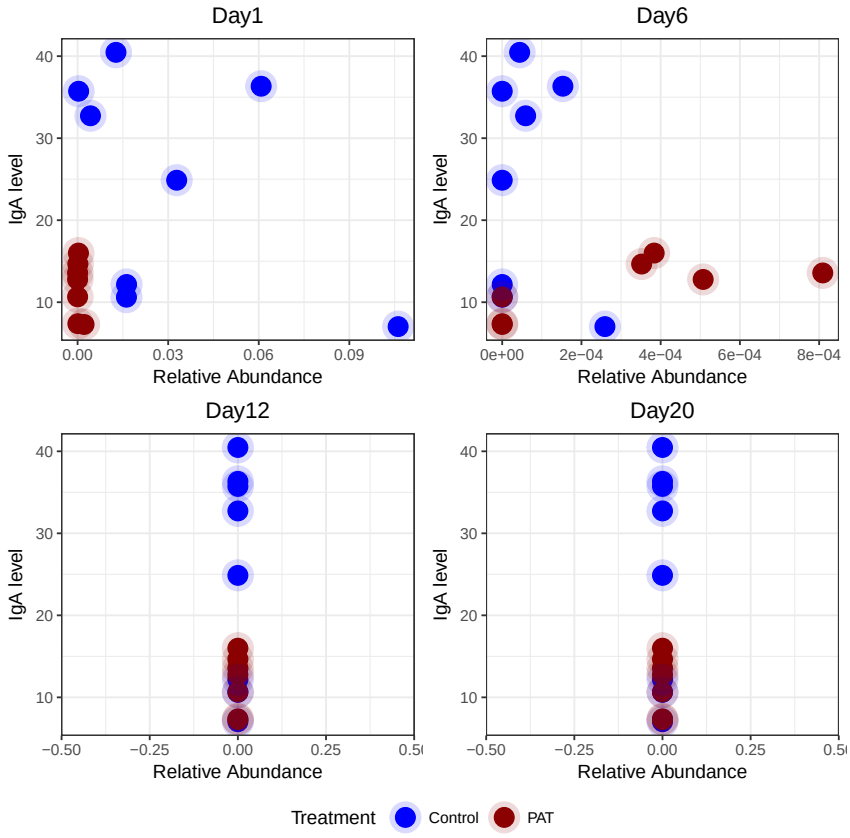


Figure 5.6: TransPAT study: OTU 997439. Relative abundance (X_j) versus IgA level (Y), per timepoint, colored by treatment (Z).

5.3.2 Type 1 Diabetes (T1D) Dataset

5.3.2.1 Study Design

The T1D study is a 30-weeks experiment conducted to investigate the function of microbiome in the context of the development of Type 1 Diabetes. The main objective of the study was to develop an early pulsed antibiotic treatment (PAT) model to further assess early-life effects on T1D onset in mice. In total, 79 mice were randomized into

two treatment groups - one group received the antibiotic (3PAT) and the other group (3PATCON) received placebo. OTU abundance data were collected for 348 OTUs and was measured in mice at 21, 35 and 49 days. As shown in Figure 5.7a, after 49 days, the mice were followed up for a period of 30 weeks during which their T1D status was monitored.

5.3.2.2 Time to Type 1 Diabetes

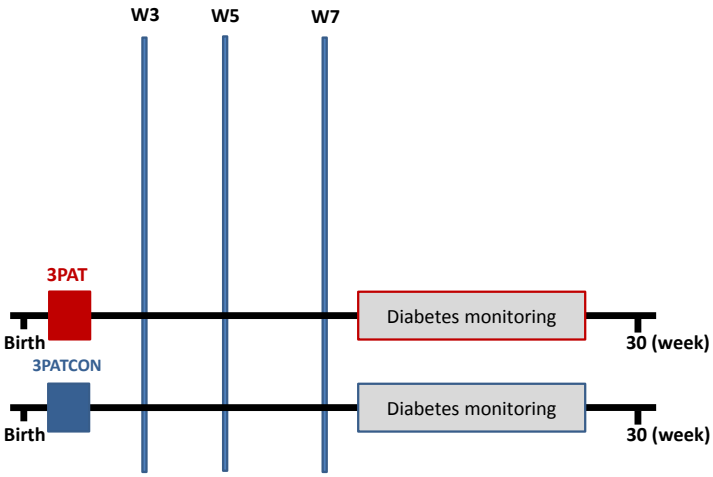
At week 7 (day 49), none of the animals had diabetes. Animals that did not develop diabetes within the follow up period were considered censored. Figure 5.7b displays the Kaplan-Meier (K-M) curves for the two treatment groups and shows that the K-M curve of the control group is found to have higher survival time. Log rank test (p-value: 0.006) indicates a significant difference in time to develop T1D between the two groups. The median time to T1D is equal to 19.5 weeks and 30 weeks for the 3PAT and 3PATCON, respectively.

5.3.2.3 Microbiome Data

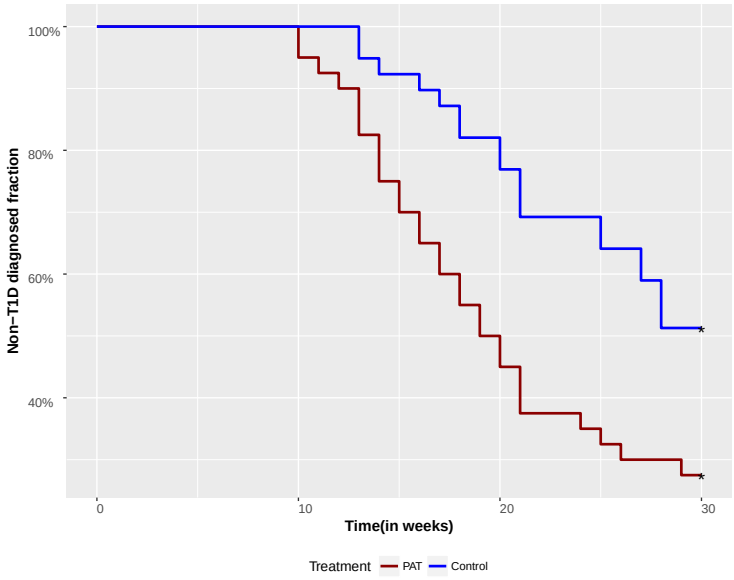
The microbiome measurement that we used in this study is the family richness which is a measurement of microbiome activity per family. Figure 5.8 displays the observed richness of all the families over time and reveals that microbiome activity varies between families. Per day, the T1D study's data structure consists of a $p \times n$ microbiome matrix given in equation (5.7). In addition, time to develop T1D, censoring and treatment vectors are given below,

$$\mathbf{S}'(\mathbf{t}) = \begin{pmatrix} S_1(t) \\ S_2(t) \\ . \\ . \\ . \\ S_n(t) \end{pmatrix}, \qquad \boldsymbol{\delta}' = \begin{pmatrix} \delta_1 \\ \delta_2 \\ . \\ . \\ . \\ \delta_n \end{pmatrix}, \qquad \mathbf{Z}' = \begin{pmatrix} z_1 \\ z_2 \\ . \\ . \\ . \\ z_n \end{pmatrix}. \tag{5.10}$$

The analysis of the T1D data is presented in **Chapter 8** in which a hierarchical Bayesian model for the time to develop T1D and microbiome data is formulated and used in order to detect microbiome biomarkers for the time to develop T1D.



(a)



(b)

Figure 5.7: T1D study. Panel a: study design. Panel b: time to develop T1D. Kaplan-Meier curve by treatment group. Log rank test p-value = 0.006.

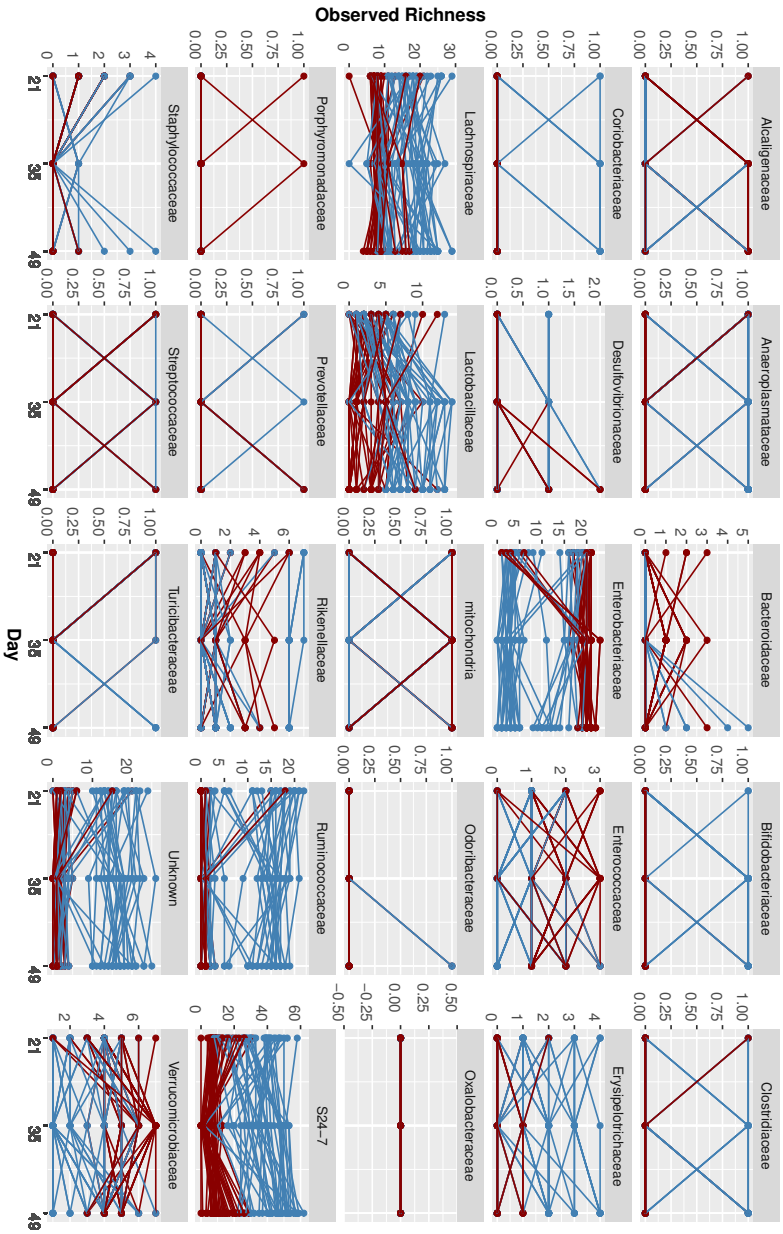


Figure 5.8: The TID study. Richness over time per family.

5.4 Modeling Approaches for the Analysis of Microbiome Data

Different modeling approaches, presented in Figure 5.9, are used in the second part of the thesis. In Chapter 6 we use the joint model, specified in equation (2.3), for the analysis of OTU level, Family level and Kingdom level data. In Chapter 7, we present non parametric approaches applied to OTU level data. In both Chapter 6 and 7, the parameters of primary interest are the adjusted association between the microbiome measurement and IgA and the treatment effect upon the microbiome variable. In Chapter 8 and 9, we present hierarchical path analysis models that can be used to estimate the direct effect of the microbiome variable on clinical variable of interest.

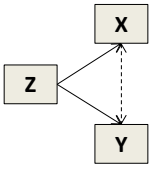
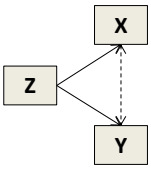
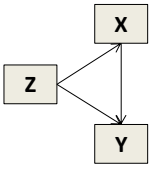
Chapters	**Chapter 6**	**Chapter 7**	**Chapter 8**	**Chapter 9**
Distributions	Normal / Normal	Non Parametric	Poisson / Normal	Poisson / Survival
Model Structure				
Estimated Parameters	• Effect of Z on X • Adjusted association between X and Y	• Effect of Z on X • Adjusted association between X and Y	• Direct and indirect effect of Z on Y • Direct effect of X on Y	
Level in the Microbiome Ecosystem	• Kingdom level • Family level • OTU level	• OTU level	• Kingdom level • Family level	

Figure 5.9: Biomarker detection in microbiome intervention studies. Approaches for the analysis of microbiome data at different levels of the microbiome ecosystem. $\mathbf{X}$ denotes the microbiome measurement at a certain level, $\mathbf{Y}$ denotes a clinical endpoint or a biological response and $\mathbf{Z}$ indicates the treatment variable.

Chapter 6

Development of Microbiome Biomarkers for IgA: A Joint Modeling Approach

6.1 Introduction

Over the past few years, there has been an increase in interest to study the associations between compositions of microbial communities and different diseases (Kostic et al., 2014, Parekh et al., 2015, John and Mullin, 2016, Young, 2017, Wang et al., 2017). Although methods to identify different compositions of microbial communities across diseases levels are well developed, the development of new methods to identify microbiome biomarkers, i.e., methods to model the association between the microbiome variables and a clinical variable is of primary interest.

In this chapter we present a joint model (Perualila-Tan et al., 2016a, Perualila et al., 2016b) that can be used to identify high dimensional microbiome biomarkers for the immune system which is measured using intestinal Immunoglobulin A (IgA) levels. For the analysis presented in this chapter we used the microbiome measurements in the first 4 timepoints (days 1, 6, 12 and 20) and the IgA level at day 20. Our goal is to link between the microbiome measurements and the IgA, taking into account that the treatment may influence both the microbiome and IgA variables. The time-specific joint model presented in this chapter allows to model two types of relationships between IgA level and the microbiome data: (1) an association which is driven by the treatment effect and (2) an

association reflecting the correlation between the microbiome and IgA. The analysis we present in this chapter is timepoint specific, i.e., at each day of the study, we model the association between the microbiome measurements and IgA at day 20.
The proposed joint model is flexible in the sense that it can accommodate microbiome measurements in different resolutions. Figure 6.1a and Figure 6.1b show the longitudinal measuerments of IgA and the boxplot of IgA at day 20 which is used as the clinical variable of primary interest for the analysis. Figure 6.1c shows an example of the change in relative abundance over time for OTU 264734. As pointed out in Chapter 5, as a measure of microbiome activity at a family level we use the family level richness, shown in Figure 6.1d for the *S24-7* family. The family level richness is the number of OTUs, belonging to a specific family, that have non zero counts. Figure 6.1e shows the richness profiles of the *Dehalobacteriaceae* family and reveals a non active family. The issue of active and non active families is discussed further in Section 2 of Appendix C. Finally, Figure 6.1f shows the α-diversity (Morgan and Huttenhower, 2012) profiles for the study which is the richness at a kingdom level.

This chapter is organised as follows. In Section 6.2, the joint model is formulated while in Section 6.3 the model is applied to different levels of the microbiome ecosystem, i.e., relative abundance at OTU level, richness at a family level and α-diversity at kingdom level. Finally, we discuss the results in Section 6.4.

6.2 Modeling Approach

6.2.1 A Joint Model for Microbiome Measurements and IgA

For the analysis presented in this chapter, we use the joint model formulated in Chapter 2 and used previously in Chapter 3 and 4. As mentioned in the previous section, in the current application the joint model is applied to three different levels of the hierarchical microbiome ecosystem: OTU, family and kingdom level. Let us focus on the OTU level in which relative abundance was used as the microbiome measurement. Let X_{ji}, $i = 1,2,\cdots,n$, $j = 1,2,\cdots,m$, be the relative abundance measures for the jth OTU and the ith subject, Y_i be the IgA measurement at day 20 and Z_i be the treatment group indicator of the ith subject.

For a given timepoint, the OTU-specific joint model also allows us to test which OTU is differentially abundant and which OTU is correlated with the IgA measurement, taking into account a possible effect of the treatment on both variables. Following Perualila *et*

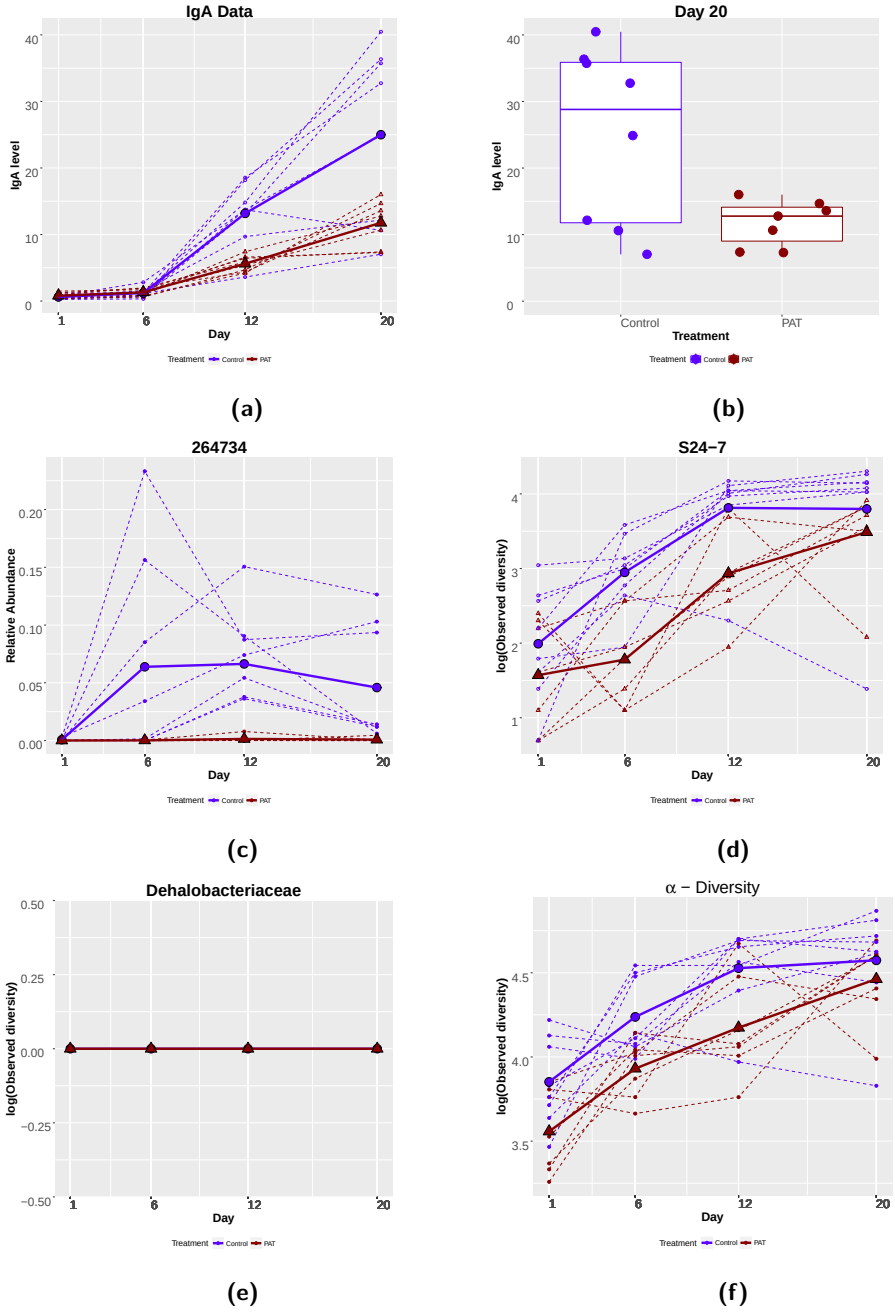


Figure 6.1: The transPAT data. Panel a: the IgA level over time. Panel b: boxplot of the IgA level of 15 mice by Treatment for Day 20. Panel c: example of an OTU over time. Panel d: active *S24-7* family over time. Panel e: non active *Dehalobacteriaceae* family over time. Panel f: α-diversity, in log scale, over time. For all the figures dashed lines and solid lines represent subject profiles and mean profiles, respectively.

al. (2016a, 2016b) the joint model is formulated as follows:

$$\begin{pmatrix} X_{ji} \\ Y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_j + \alpha_j Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \mathbf{\Sigma}_j\right], \tag{6.1}$$

where the OTU-specific covariance matrix, $\mathbf{\Sigma}_j$, given in equation (2.2). The parameters α_j and β represent the treatment effects for the jth OTU and IgA data, respectively and μ_j and μ_Y are the mean relative abundance for the jth OTU and the average of the IgA data, respectively, for mice in the PAT group. Estimation and inference for the treatment effects and the adjusted association are done in the same way as described in Section 2.3.

Note that for the analysis at family level, X_{ji} is the richness of the jth family for the ith subject while for the analysis at kingdom level, X_i is the α-diversity for the ith subject.

6.3 Application to the TransPAT Data

The joint model specified in equation (6.1) is fitted to the transPAT data using different resolutions. In Section 6.3.1, we present the results of the analysis when α-diversity is used as a biomarker while in Section 6.3.2, we present the results obtained for family level richness. The analysis of the OTU level data is presented in Section 6.3.3. For the kingdom level analysis, Shannon index (Shannon, 1948) is used as a measurement of α-diversity. For the family level analysis, log(richness) is used as a measure of microbiome activity while relative abundance is used for the analysis at OTU level.

6.3.1 Analysis of α-Diversity

In this section, the Shannon index is used as microbiome measurement for α-diversity. Note that for this resolution, for each timepoint, the microbiome matrix $\mathbf{X}$ is reduced to an $n \times 1$ vector.

Figure 6.2 and Table 6.1 show there is a significant difference between α diversity of the two treatment groups on all the timepoints and the adjusted association is significant on days 1, 12 and 20. Table SC1 and Table SC2, in Appendix C, present the results when observed and Chao1 measure of α-diversity are used, respectively. Multiplicity adjustment is done using the FDR method (Benjamini and Hochberg, 1995) with overall error rate of 10%.

Shannon Index for α-Diversity						
Day	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
1	-0.62	0.00	0.00	-0.63	0.01	0.02
6	-0.42	0.01	0.01	0.06	0.81	0.81
12	-0.63	0.01	0.01	0.46	0.06	0.08
20	-0.60	0.01	0.01	0.61	0.01	0.02

Table 6.1: Parameter estimates obtained for the Joint model using the Shannon index for α-diversity.

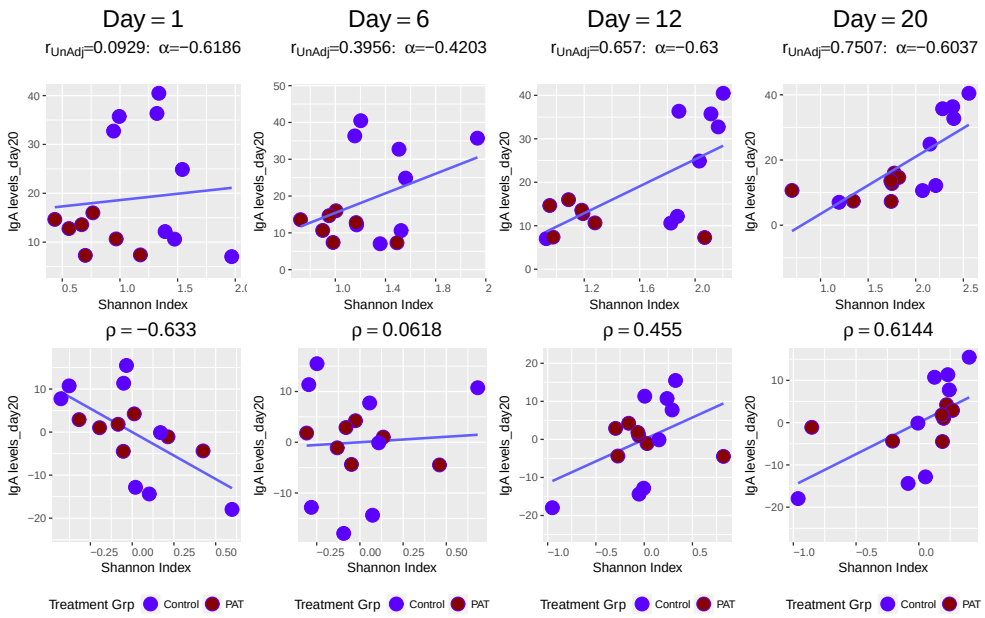


Figure 6.2: Shannon index for α-diversity versus IgA, per timepoint. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment effects.

6.3.2 Analysis at Family Level

Richness at a family level is a measurement for microbiome activity of that family. As shown in Figure SC1 (Appendix C), some families are not active and therefore are not included in the analysis. Note that we use the family level richness on a log scale. Further,

families with more than 70% zeros per treatment group per timepoint were excluded. After filtering 10, 10, 6 and 8 (out of 30 in total) families were analyzed at day 1, 6, 12 and 20, respectively and multiplicity adjustment is done using the FDR method (Benjamini and Hochberg, 1995) with overall error rate of 10%.

6.3.2.1 A Joint Model for the S24-7 Family and IgA

Table 6.2 displays the results for the *S24-7* family. Note that though this family is not significantly differentially abundant on day 1 and 20, it is found to be significant at day 6 and 12. Figure 6.3 shows the scatterplot of log(richness) and IgA at each timepoint. The adjusted association is found to be significant only at the last timepoint. This is probably due to the outlying observations in the treatment groups.

	S24-7 Family					
Day	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
1	-0.42	0.30	0.49	-0.41	0.10	0.49
6	-1.17	0.00	0.00	-0.03	0.92	0.99
12	-0.88	0.01	0.07	0.40	0.11	0.22
20	-0.31	0.48	0.77	0.53	0.03	0.05

Table 6.2: Parameter estimates obtained from the model applied to the *S24-7* family at different timepoints.

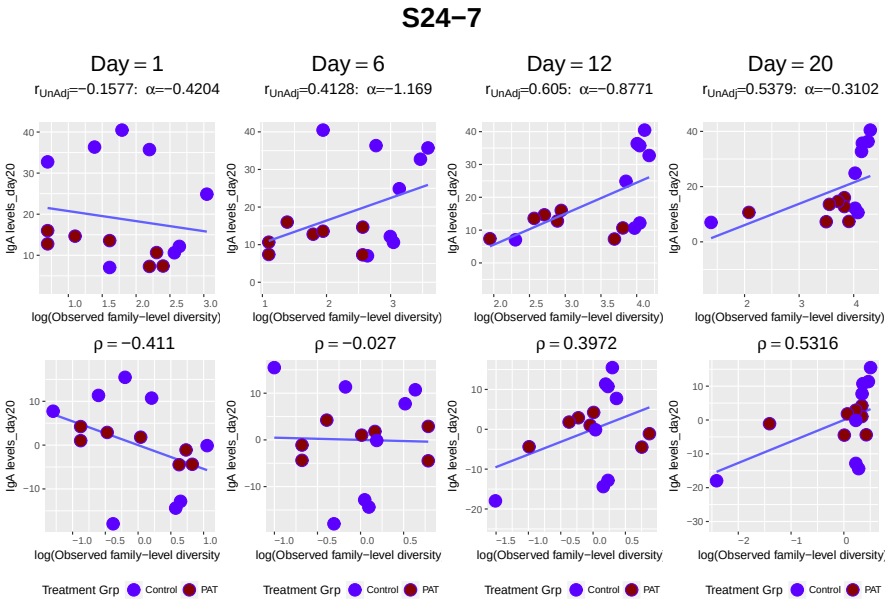


Figure 6.3: Observed richness (on log scale) of the *S24-7* family against IgA over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

6.3.2.2 Other Families

Table 6.3 displays the results for all the families. The treatment effect on two families, namely *Erysipelotrichaceae* and *Verrucomicrobiaceae*, is found to be significant at day 1 while at day 6 only the *S24-7* family is found to have a significant treatment effect. None are significant at day 12 and day 20. When the adjusted association is tested, none of the families are found to have a significant adjusted association at day 1 and day 12. At day 6 and 20 three families are found to have significant adjusted association. Among these families *Lachnospiraceae* (Figure SC2 in Appendix C) is found to have significant adjusted association at both day 6 and day 20 (Figure SC3 in Appendix C). A similar analysis was conducted for other richness measurements (Chao1 and Shannon index) (Chao, 1984, Shannon, 1948) and the results are shown in Table SC3 and Table SC4, respectively, in Section C.2.2 of Appendix C.

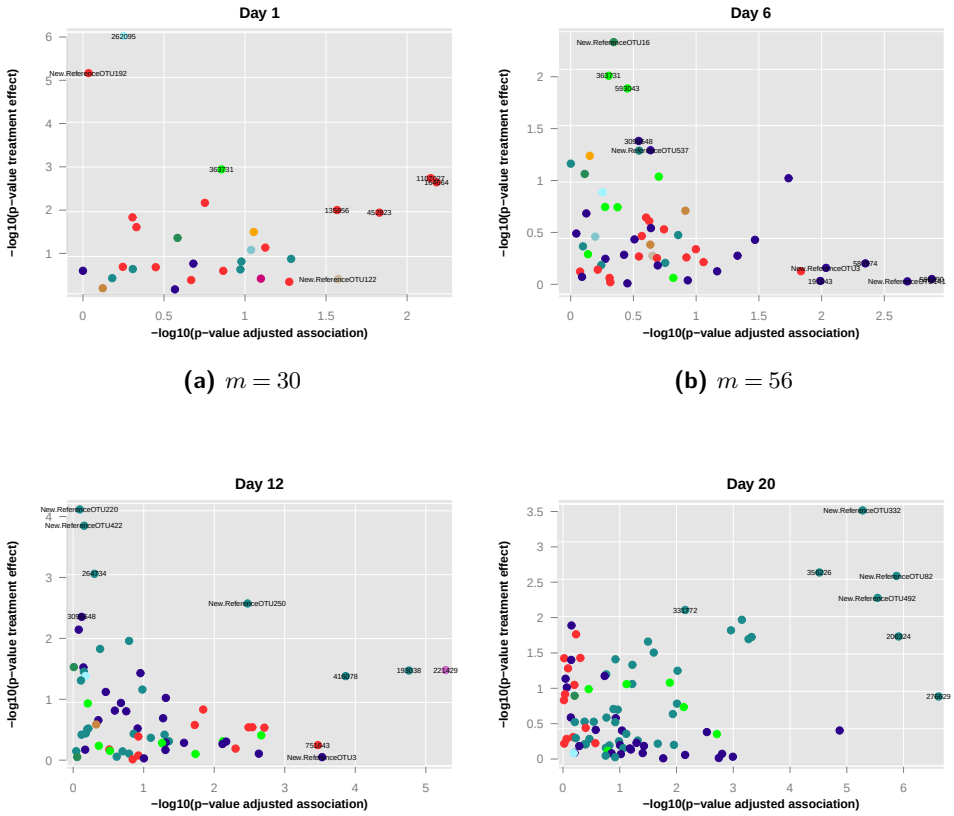
6.3.3 Analysis at OTU Level

Figure 6.4 presents the scatterplots for -log10(p-values) of the adjusted association versus -log10 (p-values) of the treatment effect per timepoint. OTUs in the upper left corner, e.g., 262095 at day 1 (Figure 6.5a), New.ReferenceOTU220 at day 12 (Figure 6.5b) are

			Day 1			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Erysipelotrichaceae	-1.09	0.00	0.00	-0.25	0.32	0.64
Verrucomicrobiaceae	-0.97	0.00	0.00	-0.32	0.21	0.52
Bifidobacteriaceae	-0.00	0.03	0.10	-0.04	0.89	0.89
Lactobacillaceae	-0.07	0.08	0.18	-0.41	0.10	0.49
Lachnospiraceae	-0.60	0.09	0.18	0.05	0.85	0.89
S24-7	-0.42	0.30	0.49	-0.41	0.10	0.49
Enterobacteriaceae	-0.15	0.43	0.60	0.36	0.15	0.49
Unknown	-0.20	0.48	0.60	-0.19	0.47	0.67
Ruminococcaceae	-0.06	0.74	0.82	-0.15	0.56	0.70
Turicibacteraceae	-0.00	0.93	0.93	-0.22	0.40	0.67
			Day 6			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-1.17	0.00	0.00	-0.03	0.92	0.99
Turicibacteraceae	0.00	0.17	0.60	0.61	0.01	0.04
Ruminococcaceae	-0.29	0.21	0.60	0.10	0.69	0.86
Unknown	-0.21	0.40	0.60	0.23	0.37	0.64
Erysipelotrichaceae	-0.18	0.42	0.60	-0.12	0.64	0.86
Lachnospiraceae	-0.09	0.42	0.60	0.66	0.00	0.03
Enterococcaceae	0.00	0.48	0.60	0.22	0.38	0.64
Verrucomicrobiaceae	0.03	0.48	0.60	0.00	0.99	0.99
Lactobacillaceae	0.12	0.65	0.72	0.57	0.02	0.05
Bifidobacteriaceae	0.00	0.80	0.80	0.38	0.13	0.31
			Day 12			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-0.88	0.01	0.07	0.40	0.11	0.22
Erysipelotrichaceae	0.20	0.11	0.32	-0.00	0.99	0.99
Verrucomicrobiaceae	0.03	0.48	0.95	-0.16	0.54	0.81
Unknown	0.12	0.65	0.97	0.60	0.01	0.06
Lachnospiraceae	0.02	0.84	0.97	-0.07	0.77	0.93
Lactobacillaceae	0.01	0.97	0.97	0.42	0.08	0.22
			Day 20			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Lachnospiraceae	0.15	0.04	0.35	0.67	0.00	0.02
Erysipelotrichaceae	0.20	0.11	0.35	0.11	0.67	0.68
Lactobacillaceae	0.28	0.13	0.35	0.33	0.19	0.30
Ruminococcaceae	-0.23	0.40	0.77	0.62	0.01	0.02
S24-7	-0.31	0.48	0.77	0.53	0.03	0.05
Verrucomicrobiaceae	-0.01	0.73	0.84	-0.11	0.68	0.68
Unknown	-0.11	0.74	0.84	0.65	0.00	0.02
Rikenellaceae	-0.02	0.94	0.94	0.17	0.52	0.68

Table 6.3: Parameter estimates for all the families at different timepoints. Results are sorted according to the adjusted p-values for the treatment effect α.

differentially abundant but are conditionally independent of the IgA level. The OTUs at the bottom right, e.g., 221429 at day 12 (Figure 6.6a), 276629 at day 20 (Figure 6.6b) have significant adjusted association but are not differentially abundant.



(a) $m = 30$ **(b)** $m = 56$

(c) $m = 67$ **(d)** $m = 87$

Figure 6.4: Scatterplot (per day) of -log(p-values) for adjusted association versus -log(p-values) for treatment effect. For each OTU, raw p-values are presented here. OTUs in the upper left corner and bottom right corner are significant with respect to treatment effect and adjusted association, respectively. OTUs in the upper right corner are significant for both treatment effect and adjusted association. Panel a: day 1. Panel b: day 6. Panel c: day 12. Panel d: day 20.

Per day, as shown in Table 6.4, we can classify the OTUs into four subclasses based on the inference results for the null hypotheses, specified in equation (2.4) and (2.5). For day 1, 9 OTUs are found to be differentially abundant. None of the OTUs are found to be both differentially abundant and significantly correlated for day 1 and 6 while for

OTUs, New.ReferenceOTU250 at day 12 and New.ReferenceOTU332 at day 20, both the null hypotheses in equation (2.4) and (2.5) were rejected. Figure 6.7 shows an example of the OTU New.ReferenceOTU513, from the *Lachnospiraceae* family, which is highly associated with the IgA level from day 12 onwards (lower panels in Figure 6.7) but not differentially abundant between the two groups (upper panels in Figure 6.7).

		$H_0 : \rho = 0$							
		Day 1		Day 6		Day 12		Day 20	
		R	NR	R	NR	R	NR	R	NR
$H_0 : \alpha = 0$	R	0	11	0	0	1	5	3	0
	NR	0	19	3	53	18	43	25	59

Table 6.4: Inference results per timepoint. R: rejected. NR: not rejected.

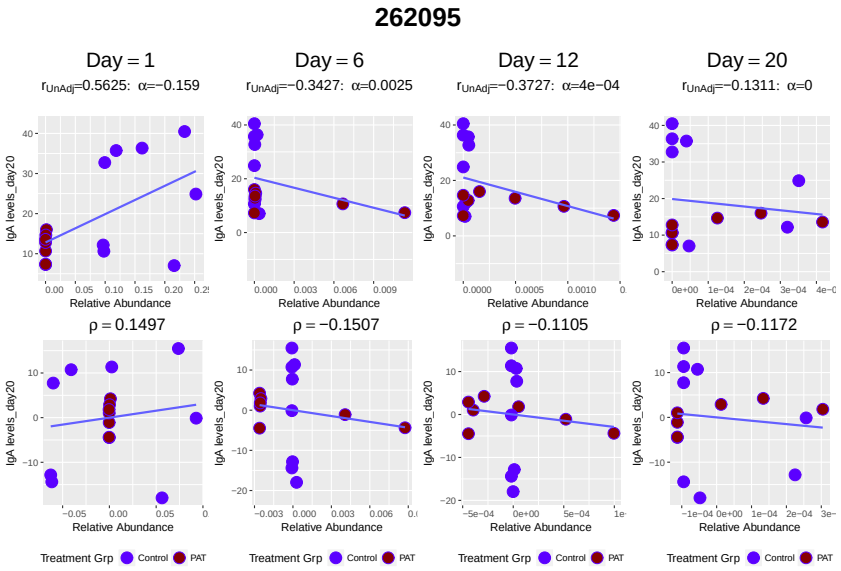
We notice that from day 12 onwards, all the differentially abundant OTUs belong to the *S24-7* family. Moreover, OTUs from this family have higher relative abundance in the control group (negative treatment effect, Table 6.3). This implies that OTUs from S24-7 family, might not be thriving in the host system with PAT-altered microbiota. Table SC5, in Appendix C, shows the results for top 10 OTUs, per timepoint.

Table SC6, in Appendix C, provides the estimates of the adjusted association by day, for the top 10 OTUs. For the earlier timepoints, no OTUs are found to be significantly associated at day 1 or at day 6 while few OTUs belonging to different families are found to be significant at day 12 for the IgA level. The OTUs from the *S24-7* family, dominate at day 20.

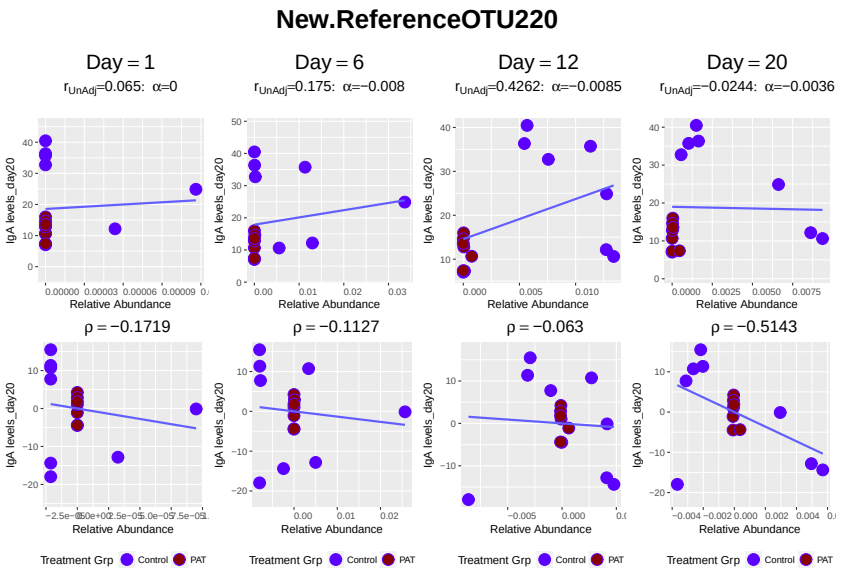
6.4 Discussion

Similar to previous chapters, a joint model for the high dimensional microbiome data and IgA was formulated and used for the detection of microbiome biomarkers. We have shown that, depending on the research question of interest, the joint model can be used to identify features which are differentially abundant and significantly associated with IgA. The joint model was applied to three different levels of the phylogenetic tree: (1) OTU-level, (2) Family-level and (3) Kingdom-level.

The main criticism to use the joint model, specified in equation (6.1), is the fact that it assumes a bivariate normal distribution for the microbiome feature and IgA which might not be the case for the microbiome data, on all levels of the hierarchical ecosystem, since it is based on count data. In the next two chapters we address this point. In Chapter 7,

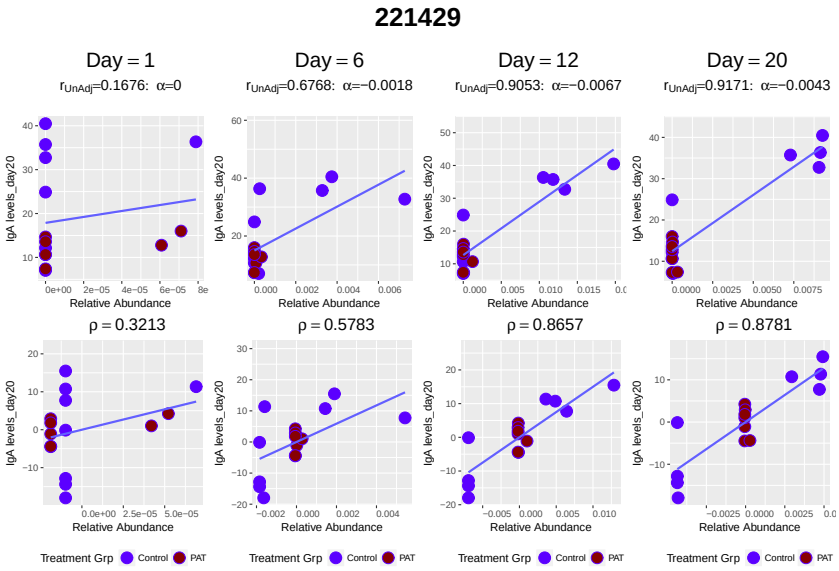


(a)

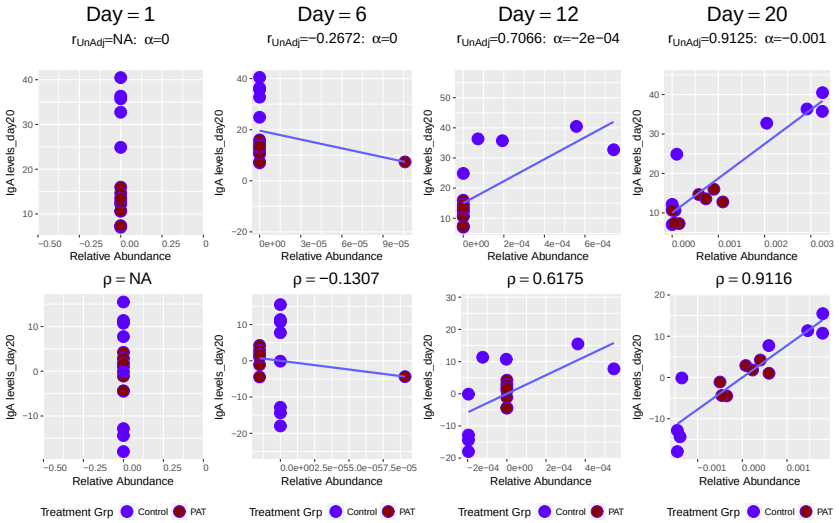


(b)

Figure 6.5: Two differentially abundant (FDR = 0.10) OTUs over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment effects. Panel a: OTU 262095 - differentially abundant on day 1. Panel b: OTU New.ReferenceOTU220 - differentially abundant on day 12.



(a)



(b)

Figure 6.6: Two significantly (FDR = 0.10) correlated OTUs over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment effects. Panel a: OTU 221429 - significantly correlated on day 12. Panel b: OTU 276629 - significantly correlated on day 20.

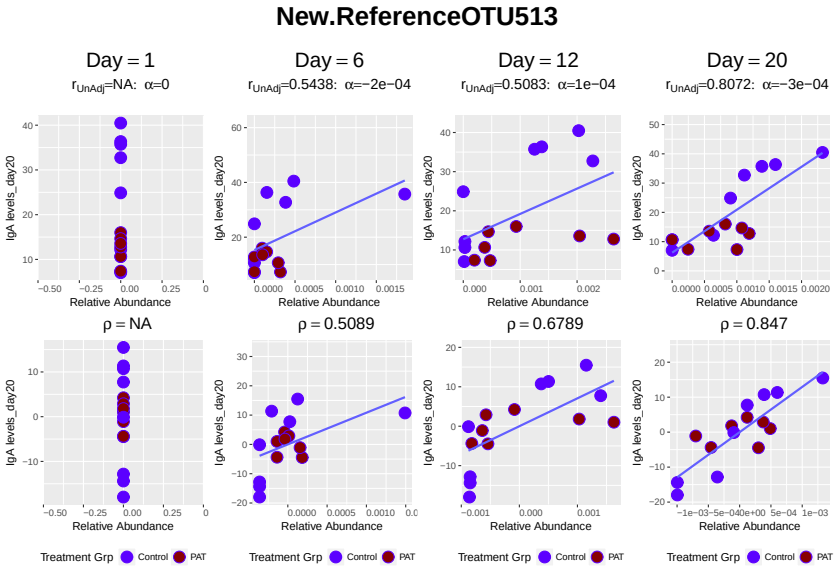


Figure 6.7: Relative abundance of New.ReferenceOTU513 against IgA over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment effects.

we present a non parametric approach in which inference for both α_j and ρ_j is conducted using non parametric and resampling based tests while in Chapter 8 we present a joint Poisson / Normal model for the microbiome measurements and IgA.

Chapter 7

Development of High Dimensional Microbiome Biomarkers: A Non Parametric Approach

7.1 Introduction

In the context of microbiome studies, the main disadvantage of the joint model formulated in Chapter 6 is the fact that it assumes an underlying bivariate normal distribution for the IgA and the microbiome measurements (for the analysis at all levels of the hierarchical microbiome ecosystem). Since the microbiome measurements are based on counts, with possibly excess zero counts at an OTU level, the joint model in equation (6.1) might not be appropriate. In this chapter, we focus on a non parametric approach that allows us to depart from the distributional assumptions. Our main interest in this chapter is still to estimate the parameters α_j and ρ_j and to test the hypotheses $H_0 : \alpha_j = 0$ and $H_0 : \rho_j = 0$ while taking into account possible high proportion of zeros for the OTU count data. Non parametric tests, such as the Wilcoxon rank-sum test, two-part Wilcoxon test (Lachenbruch, 2001, Wagner et al., 2011), truncated Wilcoxon rank-sum test (Hallstrom, 2010) will be used to test for differential abundance. The latter two take into account the proportion of zeros in the sample. For the adjusted association, Spearman rank correlation is used for estimation. For both the parameters, α_j and ρ_j, a permutation test is used

for inference.

The transPAT data, introduced in Chapter 5, is used as a case study in this chapter. Similar to Chapter 6, the analysis is conducted at each day separately with IgA at day 20 as the immunological variable of interest. Figure 7.1 displays an example of the relative abundance of OTU 997439 against the IgA level at day 20. We notice that the PAT group contains high proportion of zeros in day 1. Even though PAT group becomes relatively abundant at day 6, both groups have zero counts at day 12 and 20.

In Section 7.2 we present the non parametric tests used for the data analysis followed by the results presented in Section 7.3. Finally, discussion and concluding remarks are given in Section 7.4.

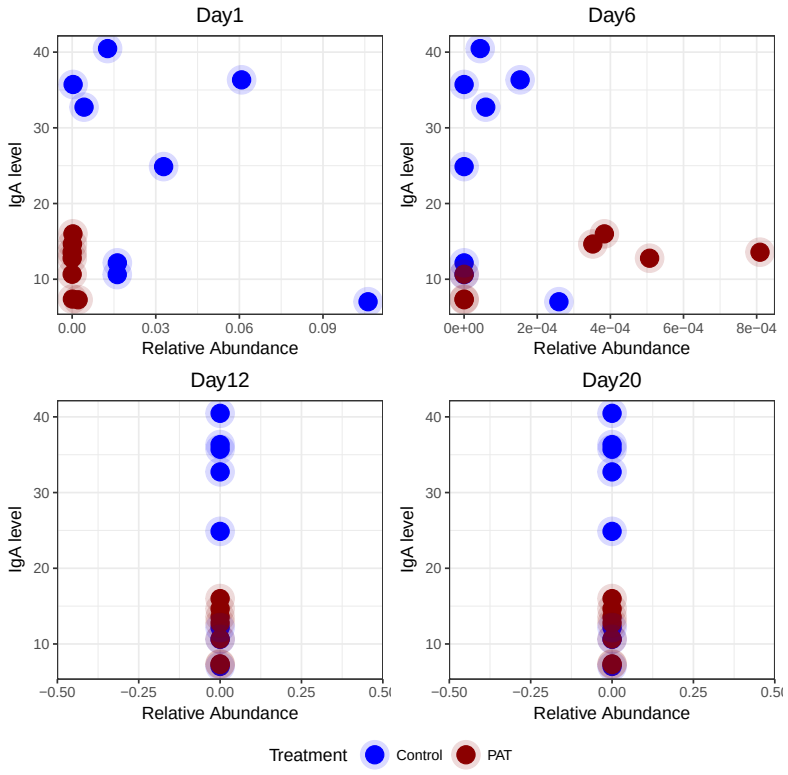


Figure 7.1: Relative abundance for OTU 997439 versus IgA level, per timepoint, colored by treatment.

7.2 A Non Parametric Approach for the Detection of Microbiome Biomarkers

7.2.1 Estimation

The effect of the treatment on X and Y, denoted by $\tilde{\alpha_j}$ and $\tilde{\beta}$, respectively, can be estimated by the difference between the median value of the two groups,

$$\begin{aligned} \tilde{\alpha_j} &= \tilde{X}_{j2} - \tilde{X}_{j1}, \\ \tilde{\beta} &= \tilde{Y}_2 - \tilde{Y}_1, \end{aligned} \tag{7.1}$$

where $\tilde{Y}_k$ and $\tilde{X}_{jk}$, $k = 1, 2$, are the median relative abundance and IgA measurements of the kth group, respectively. The adjusted association, denoted by $\tilde{\rho_j}$, is the association between the two endpoints, relative abundance and IgA, after correcting for the treatment effects. The Spearman rank correlation of the adjusted data of the two endpoints provides the estimates of $\tilde{\rho_j}$, that is:

$$\tilde{\rho}_{sj} = \text{cor}[(X_{jik} - \tilde{X}_{jk}, Y_{ik} - \tilde{Y}_k] = 1 - \frac{6\Sigma_{i=1}^{N} {D_i}^2}{N(N^2 - 1)}, \tag{7.2}$$

where D_i is the difference between the ranks of $(X_{jik} - \tilde{X}_{jk})$ and $(Y_{ik} - \tilde{Y}_k)$ in their separate rankings.

7.2.2 Inference

7.2.2.1 Treatment effects on X and Y

Wilcoxon-Rank Sum Test

The Wilcoxon rank sum test (Wilcoxon, 1945, Mann and Whitney, 1947) can be used to test whether the distribution of the measurements differs between two groups. It is valid for data from any distribution and is much less sensitive to outliers than the two-sample t-test. Under the assumption that the shape of the distribution is same for both groups then an OTU specific hypothesis, to test for location shift (Wild and Seber, 2000), can be formulated as follows:

$$H_{0_j} : \tilde{\mu}_{j2} = \tilde{\mu}_{j1},$$

$$H_{1_j} : \tilde{\mu}_{j2} \neq \tilde{\mu}_{j1}.$$

Here, $\tilde{\mu}_{j2}$ and $\tilde{\mu}_{j1}$ are the location parameters of the PAT and control group, respectively. In more general case, the null and the alternative hypotheses can be formulated in terms of the distribution in each group, that is,

$$\begin{aligned} H_{0_j} &: h_2(X_j) = h_1(X_j), \\ H_{1_j} &: h_2(X_j) \neq h_1(X_j), \end{aligned} \tag{7.3}$$

where $h_k(X_j)$ is the distribution function of measurements in the kth treatment group, $k = 1, 2$. The Wilcoxon rank sum statistic is given by,

$$S = R - N(2N+1)/2, \tag{7.4}$$

where R is the sum of the ranks of the control group and N is the total sample size (Hallstrom, 2010).

Two-Part Wilcoxon

For the OTU level analysis, the count data consists of a substantial proportion of zero counts. A large number of zeros, treated as ties, can lead to the reduction of power of Wilcoxon test (Hallstrom, 2010). To overcome this problem, Lachenbruch (1976) proposed the two-part Wilcoxon test that compares the distributions with respect to their proportion of zeros using a Z-test as well as the distribution of non-zero measurements, using a Wilcoxon test, simultaneously. Lachenbruch (2001) defined the probability distribution of the kth group as,

$$f_k(X_j, d_j) = [p_k^{1-d_j}\{(1-p_k)h_k(X_j)\}^{d_j}], \quad k = 1, 2,$$

where p_k is the binomial probability of an indicator variable d_j,

$$d_j = \begin{cases} 1 & X_j \text{ is non-zero observed count,} \\ 0 & X_j \text{ is zero or missing or below the limit of detection.} \end{cases}$$

The hypotheses for the two-part test are specified as follows:

$$\begin{aligned} H_{0_j} &: p_2 = p_1 \cap h_2(X_j) = h_1(X_j), \\ H_{1_j} &: p_2 \neq p_1 \cup h_2(X_j) \neq h_1(X_j). \end{aligned} \tag{7.5}$$

Lachenbruch (1976, 2001) proposed a two-part statistic, given by,

$$V^2 = B^2 + U^2, \tag{7.6}$$

where B is a binomial test statistic for testing the difference in proportion of zero and U is the Wilcoxon rank-sum statistic based on the non-zero observations. Note that, in some

cases, B is not defined. For example, if one or both of the groups have no observations equal to zero, B^2 is not defined. In these cases, $B^2 = 0$ and the test statistic U^2 is identical to the Wilcoxon test statistic as long as there are continuous observations. The two-part test statistic is asymptotically distributed as χ^2_2 (Lachenbruch, 1976, 2001, Wagner et al., 2011).

Truncated Wilcoxon-Rank Sum Test

Let n_{01j} and n_{02j} be the number of zeros in each group, for the jth OTU and let $n_{0j} = \min(n_{01j}, n_{02j})$. Hallstrom (2010) proposed the truncated Wilcoxon rank-sum test which removes an equal number of zeros, n_{0j} from each group. The Wilcoxon test is then performed on the truncated dataset, $n_j = N - n_{0j}$ observations. An OTU-specific test statistic is given by,

$$W_{tj} = \frac{s_j}{\sqrt{Var(s_j)}}, \qquad s_j = r - n_j(2n_j + 1)/2, \tag{7.7}$$

where, r is the sum of the ranks of the observations in the first group, after truncation and details about $Var(s_j)$ is available in Hallstrom (2010). Moreover, Hallstrom (2010) showed that this test recovers much of the power loss from the standard application of the Wilcoxon test using simulation studies and pointed out that in contrast with the two-part test that assumes independent errors for both parts of the test, the truncated Wilcoxon test is relatively unaffected when the non-zero scores are dependent on the proportion of zeros.

7.2.2.2 Adjusted Association between the OTU and the IgA level

As mentioned above, the adjusted association $\tilde{\rho}_j$ measures the association between IgA level and relative abundance adjusting for the treatment effect. In this chapter, given that the adjusted data may not meet the assumption of normality and/or linear association, the adjusted Spearman's correlation ($\tilde{\rho}_{Sj}$) was used to measure the association between the two endpoints. The following hypotheses are formulated to test if the relative abundance of the OTUs is correlated with the IgA level after adjusting for the treatment effects,

$$\begin{aligned} H_{0_j} &: \tilde{\rho}_{Sj} = 0, \\ H_{1_j} &: \tilde{\rho}_{Sj} \neq 0. \end{aligned} \tag{7.8}$$

For the truncated microbiome dataset discussed in the previous section, the corresponding IgA level of the retained observations was used to calculate the Spearman's correlation. We refer to this statistic as the truncated Spearman's rank correlation ($\tilde{\rho}_{TSj}$). As in the previous chapters, the inference, for both relative abundance and correlation,

was adjusted for multiple testing, due to the large number of tests per day, using the Benjamini-Hochberg false discovery rate method (FDR, Benjamini and Hochberg, 1995).

7.2.2.3 Permutation Based Inference

Neuhäuser et al. (2005) described a permutation test, for the two-part test statistic defined by Lachenbruch (1976, 2001), where the permutation is done based on the group labels for the whole sample, i.e., all observations including zero and non-zero values. A similar permutation method is conducted for the analysis in this chapter. In the transPAT study for the jth OTU, we consider the triplet ($\mathbf{X_j}$, $\mathbf{Y}$, $\mathbf{Z}$), given by,

$$\mathbf{X_j'} = \begin{pmatrix} x_{j1} \\ x_{j2} \\ . \\ . \\ . \\ x_{j15} \end{pmatrix}, \qquad \mathbf{Y'} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_{15} \end{pmatrix}, \qquad \mathbf{Z'} = \begin{pmatrix} z_1 \\ z_2 \\ . \\ . \\ . \\ z_{15} \end{pmatrix}. \tag{7.9}$$

Permutation tests were conducted for all test statistics described in Section 7.2.2.1 and Section 7.2.2.2. For the remainder of this section we focus on the permutation test for SpearmanâĂŹs correlation.

Let t_{sobs} be the observed test statistic, calculated from the observed data ($\mathbf{X_j}$, $\mathbf{Y}$, $\mathbf{Z}$), for the hypothesis defined by equation (7.8) for SpearmanâĂŹs correlation test. For L permutations we permute the elements of $\mathbf{X_j}$, in equation (7.9). Let $\mathbf{X_{j\ell}}$, be the permuted vector for the ℓth permutation given by,

$$\mathbf{X_{j\ell}} = (x_{j\ell_1}, x_{j\ell_2}, \cdots, x_{j\ell_{15}}),$$

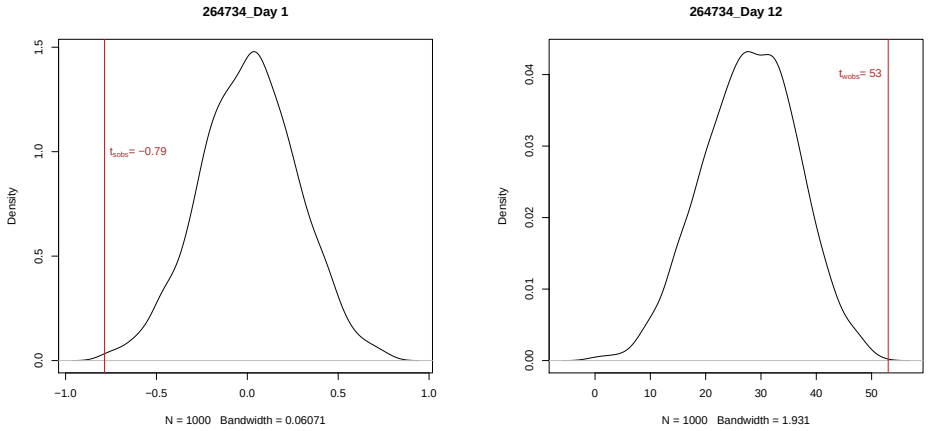
where $x_{j\ell_i}$ are the permuted values of $\mathbf{X_j}$. The test statistic is calculated for the triplet ($\mathbf{X_{j\ell}}$, $\mathbf{Y}$, $\mathbf{Z}$). Let $t_{s_1}, t_{s_2}, \cdots, t_{s_L}$ be the test statistics obtained for the L permuted datasets. The permutation based p-value for a two-sided alternative against the null hypothesis, in equation (7.8), is given by,

$$p_{perm} = \frac{p'_{perm}}{L+1}, \tag{7.10}$$

where p'_{perm} is given by,

$$p'_{perm} = 2 * min\left(\sum_{i=1}^{L} t_{s_i} \geqslant t_{sobs}, \sum_{i=1}^{L} t_{s_i} \leqslant t_{sobs}\right). \tag{7.11}$$

Figure 7.2a and 7.2b show a density plot, based on 1000 permutations, for the Spearman's correlation test statistic (for day 1) and Wilcoxon test statistic (for day 12) for OTU 264734, respectively. The vertical line in each plot represents the observed test statistic. Similar procedure was applied for the two parts Wilcoxon test, the truncated Wilcoxon test and the truncated Spearman's correlation test.



(a) **(b)**

Figure 7.2: Approximation of the distribution of the test statistic under the null hypothesis based on 1000 permutations for OTU 264734. Vertical lines mark the observed test statistic. Left panel: Spearman's correlation test at day 1. Right panel: Wilcoxon test at day 12.

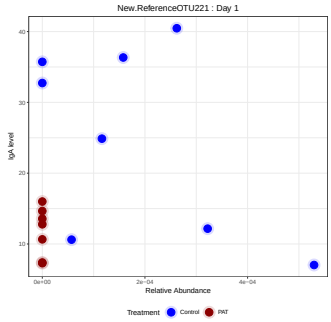
7.3 Application to the Data

7.3.1 Differentially Abundant OTUs

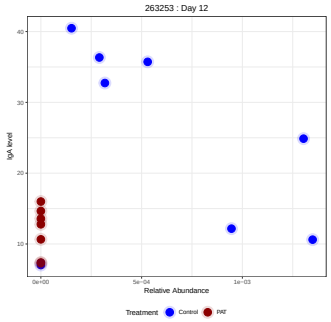
The microbiome composition based on relative abundance at day 6, 12 and 20 are more comparable than that of day 1, conditioning on the treatment group as shown in Figure SD1 in Appendix D. The cumulative relative abundance of OTUs belonging to family *Lactobacillaceae*, although dominated at day 1, reduced drastically at day 6 while *Verrucomicrobiaceae* and *S24-7* become more abundant from day 6 onwards.

In total, out of the 355 OTUs, 30, 56, 67 and 87 OTUs were analysed at day 1, 6, 12 and 20, respectively (Figure SD2 in Appendix D). Figure 7.3 shows examples of four OTUs as measured on day 1, 12 and 20 that were not included in the analysis due to

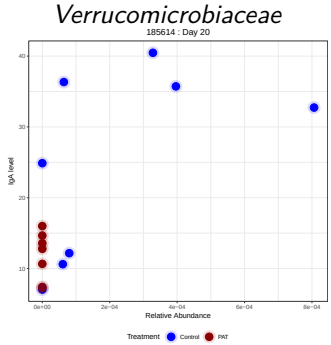
the fact that they have all zero measurements in the PAT group and a low proportion ($< 25\%$) of zero measurements in the control group. 6 out of 7 OTUs, belonging to the *Verrucomicrobiaceae* family, were filtered out on Day 1 with the NewReferenceOTU221 having all observations in PAT group equal to zero. They become abundant, having measurements for both groups, from day 6 onwards.



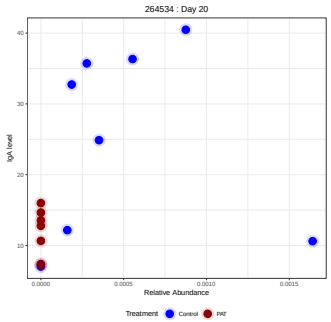
Day 1: NewReferenceOTU221- family *Verrucomicrobiaceae*



Day 12: 263253 - family *S24-7*



Day 20: 185614 - family *S24-7*



Day 20: 264534 - family *S24-7*

Figure 7.3: Examples of OTUs with 100% proportion of zeros in the PAT group and low proportion of zeros in the Control group. These OTUs were filtered out and were not included in the analysis.

Table 7.1 presents the OTUs that were found to be significantly differentially abundant by day. Figures 7.4 provides a graphical summary of the the results for all OTUs (FDR adjusted p-values) across all timepoints for the three tests discussed in Section 7.2.2.1. Differentially abundant OTUs are present only on days 1 and 12. Ten of the 30 OTUs analysed at day 1 were found to be differentially abundant according to the Wilcoxon and truncated Wilcoxon method, with seven of them belonging to family *Lactobacillaceae* (Table 7.1) and the remaining three OTUs are part of three different families. However, only 5 of these 10 OTUs remain significant when two-part Wilcoxon method is used.

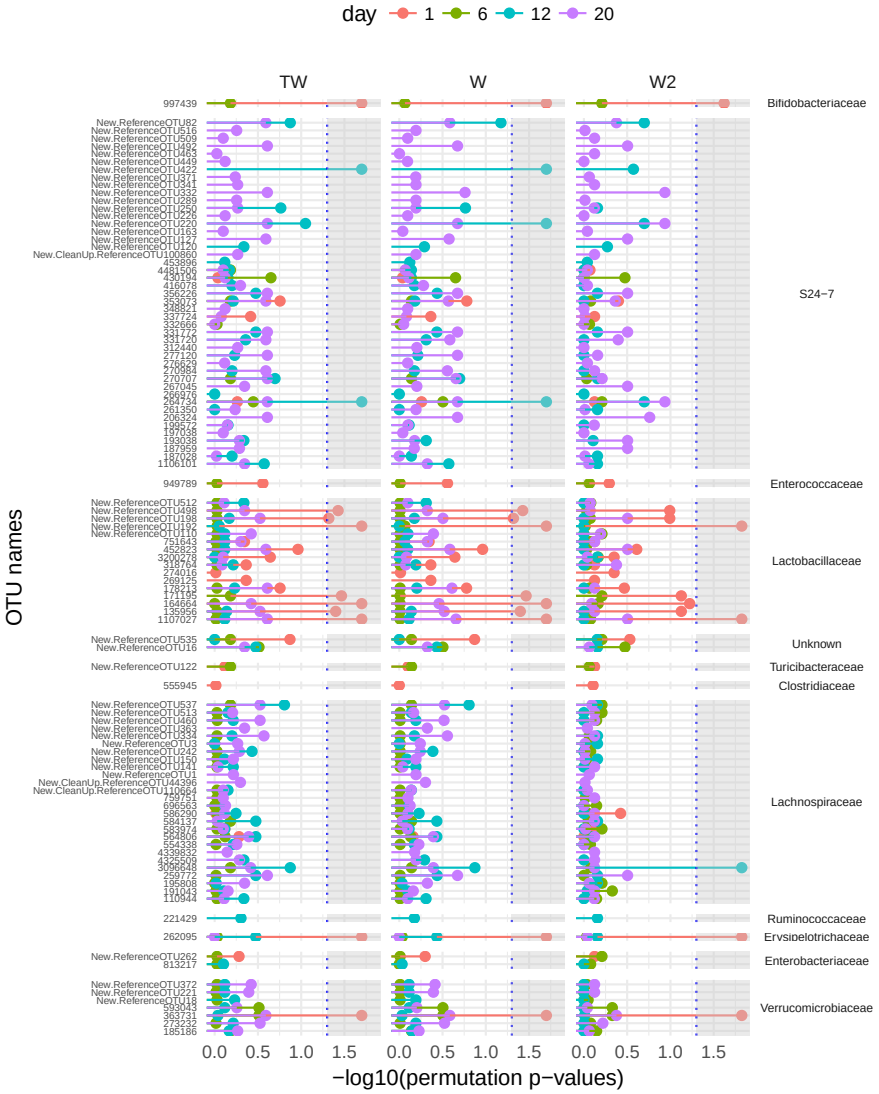


Figure 7.4: P-value by OTU (on -log(p-value) scale) for Truncated Wilcoxon(TW), Wilcoxon(W) and Two-Part Wilcoxon(W2) tests. The p-values used are adjusted for multiplicity within a method. Whenever a -log(p-value) appears in the gray area, there is a statistically significant difference between the two treatment groups.

	Estimates				adjusted p-values			
OTU names	$\tilde{\alpha}^M$	$\tilde{\alpha}^C$	p_{02}	p_{01}	p(W)	p(W2)	p(TW)	Family
Day 1								
1107027	0.24	0.24	0.00	0.00	0.00	0.00	0.00	Lactobacillaceae
262095	-0.14	-0.14	0.14	0.00	0.00	0.00	0.00	Erysipelotrichaceae
997439	-0.02	-0.02	0.43	0.00	0.00	0.02	0.00	Bifidobacteriaceae
363731	-0.01	-0.01	0.14	0.00	0.00	0.01	0.00	Verrucomicrobiaceae
New.RefOTU192	0.01	0.01	0.00	0.00	0.00	0.00	0.00	Lactobacillaceae
164664	0.00	0.00	0.00	0.00	0.02	0.06	0.02	Lactobacillaceae
171195	0.00	0.00	0.00	0.00	0.03	0.07	0.03	Lactobacillaceae
New.RefOTU498	0.00	0.00	0.00	0.00	0.04	0.10	0.04	Lactobacillaceae
135956	0.00	0.00	0.00	0.00	0.04	0.07	0.04	Lactobacillaceae
New.RefOTU198	0.00	0.00	0.00	0.00	0.05	0.10	0.05	Lactobacillaceae
Day 6								
None								
Day 12								
264734	-0.06	-0.06	0.00	0.00	0.00	0.20	0.00	S24-7
New.RefOTU422	-0.00	-0.00	0.57	0.00	0.00	0.27	0.00	S24-7
New.RefOTU220	-0.01	-0.01	0.43	0.12	0.00	0.20	0.09	S24-7
3096648	0.00	0.00	0.14	0.38	0.13	0.00	0.13	Lachnospiraceae
Day 20								
None								

Table 7.1: Significantly differentially abundant OTU's by day (FDR = 0.05). $\tilde{\alpha}^M$ and $\tilde{\alpha}^C$ are the estimates for the difference in medians of two groups, as given in (7.1), when the medians are calculated, respectively, with all observations and for only non-zero observations. p_{01} and p_{02} are the proportion of zeros in the control and PAT group, respectively. p(W), p(W2) and p(TW) indicate the adjusted p-values for Wilcoxon, two-part Wilcoxon and truncated Wilcoxon tests, respectively.

In contrast to the results in day 1, there is consensus among the three methods on day 6 and none of the OTUs were found to be significant by any of the three methods. On day 12, three of the 67 OTUs, all belonging to *S24-7* family, were identified as significantly differentially abundant by the Wilcoxon method and two of them were identified by the truncated Wilcoxon method. However, none of these three OTUs were found to be significant with two-part Wilcoxon method which identified another OTU (3096648) from *Lachnospiraceae* family to be significantly differentially abundant on day 12. None of the OTUs was found to be significant on day 20.

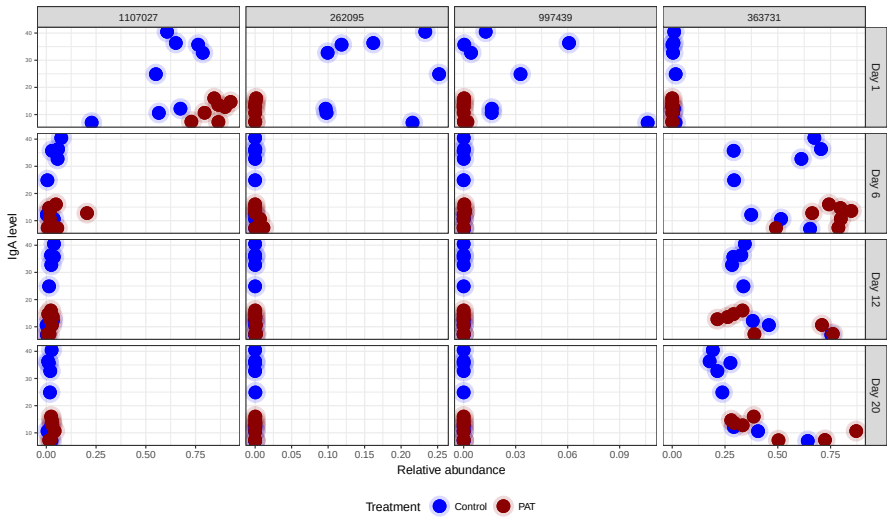
Figure 7.5 shows the examples of differentially abundant OTUs. For both day 1 and day 12, the top OTUs (OTU 1107207 and 264734) are relatively more abundant than the rest of the differentially abundant OTUs. Note that both OTUs recorded no zero count. Moreover, OTU 363731 became more abundant from day 6, but differential abundance was detected to be significant on Day 1.

7.3.2 OTUs Associated with IgA Level at Day 20

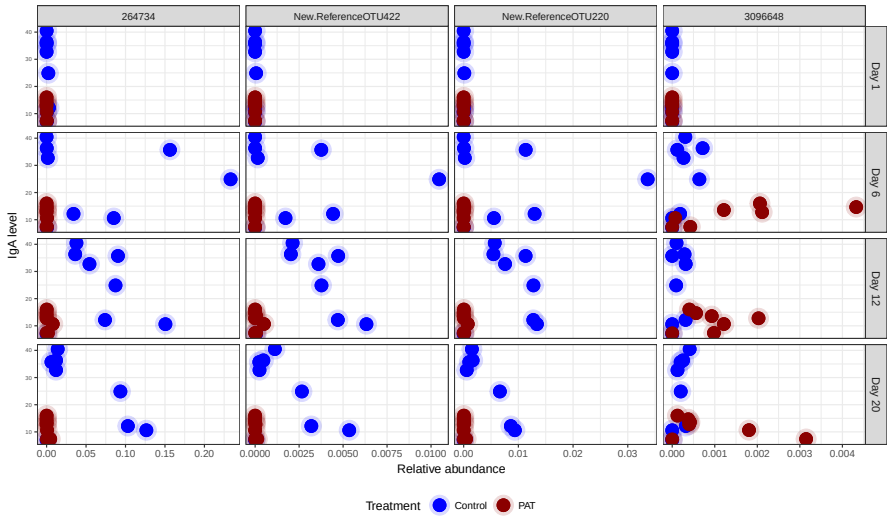
Figure 7.6 and Table 7.2 present the results for the adjusted association using Spearman and truncated Spearman test statistics. OTU 264734 and 586290, shown in Figure 7.7, were found to be significant on days 1 and 6, respectively. Note that the first was found to be significant only by the Spearman test.

As shown in Table 7.2, the number of significantly correlated OTUs increased over time. Interestingly, Table 7.2 and Figure 7.6 show that the S24-7 family becomes active and correlated with the IgA level over time. Figure 7.8 shows two OTUs, 276629 and New.ReferenceOTU82, from S24-7 family, both with significant correlation with the IgA level at day 20, as identified by both Spearman and truncated Spearman tests ($\tilde{\rho}_S = 0.93$ and $\tilde{\rho}_{TS} = 0.92$ for OTU 276629 and $\tilde{\rho}_{Sj} = 0.90$ and $\tilde{\rho}_{TSj} = 0.87$ for OTU New.ReferenceOTU82, respectively). Figure 7.9 and Figure 7.10 display these two OTUs over time. Note that on day 1 both OTUs are not abundant for both the treatment groups. From day 12 onwards, the abundance is increased among the PAT-altered mice as compared to that for the mice in the control group. Figure 7.11 shows OTU 185186 that was the only feature found to be associated with the IgA level for 2 timepoints, day 12 and day 20.

In general, Spearman test has a tendency to reject more null hypotheses than the truncated Spearman test, 7 compared to 3 on day 12 and 9 compared to 4 on day 20.



(a)



(b)

Figure 7.5: Examples of differentially abundant OTUs. Panel a: relative abundance vs. IgA level across 4 timepoints of top differentially abundant OTUs at Day 1. Panel b: relative abundance vs. IgA level across 4 timepoints of top differentially abundant OTUs at Day 12.

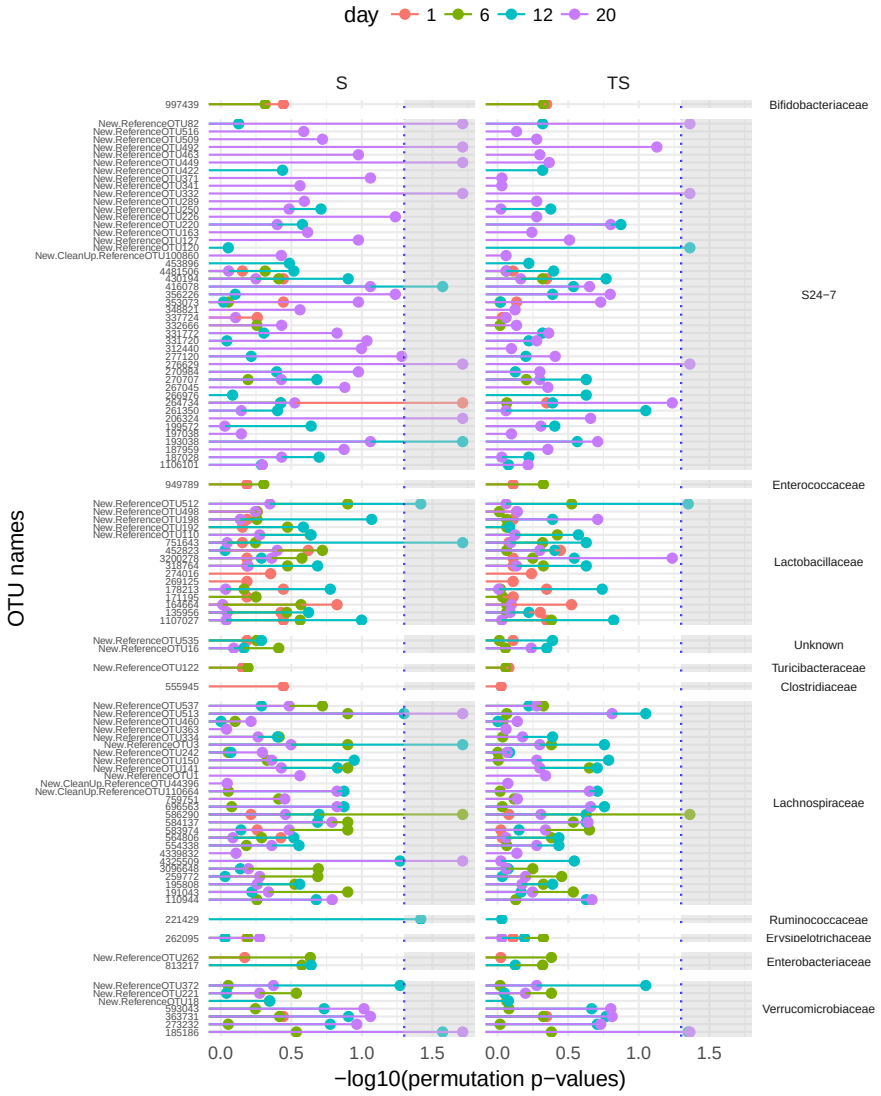


Figure 7.6: Adjusted p-values (FDR) for the correlation test on -log(p-value) scale for Spearman (S) and truncated Spearman(TS) tests. The p-values used are adjusted for multiplicity within a method.

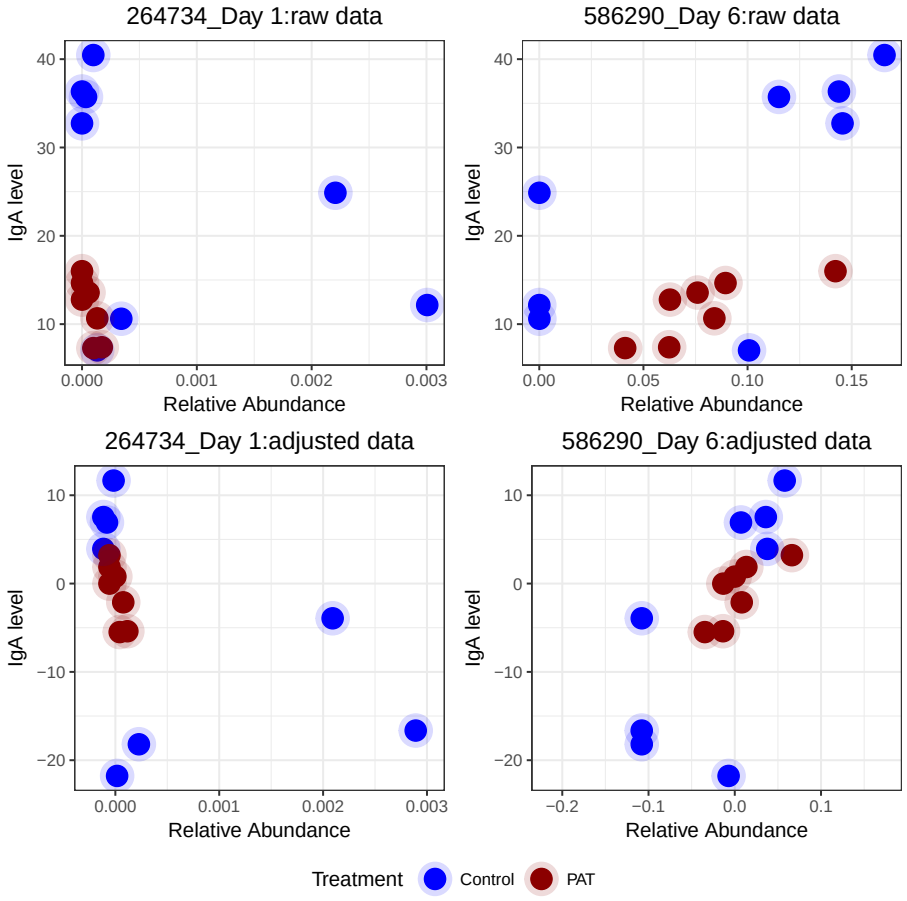


Figure 7.7: Relative abundance of OTU 264734 and OTU 586290 against IgA at day 1 and day 6, respectively. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

	Estimates				adjusted p-values		
OTU names	ρ_S	ρ_{TS}	$\tilde{\rho}_S$	$\tilde{\rho}_{TS}$	p($\tilde{\rho}_S$)	p($\tilde{\rho}_{TS}$)	Family
Day 1							
264734	-0.49	-0.33	-0.79	-0.43	0.00	0.45	S24-7
Day 6							
586290	0.63	0.63	0.82	0.82	0.00	0.00	Lachnospiraceae
Day 12							
751643	0.72	0.85	0.78	0.68	0.00	0.24	Lactobacillaceae
193038	0.89	0.91	0.93	0.62	0.00	0.27	S24-7
New.ReferenceOTU3	0.76	0.63	0.81	0.63	0.00	0.18	Lachnospiraceae
185186	-0.84	-0.84	-0.79	-0.79	0.03	0.04	Verrucomicrobiaceae
416078	0.54	0.84	0.82	0.60	0.03	0.29	S24-7
New.ReferenceOTU512	0.24	0.39	0.69	0.75	0.04	0.04	Lactobacillaceae
221429	0.60	0.79	0.81	0.14	0.04	0.93	Ruminococcaceae
New.ReferenceOTU120	0.41	0.32	0.07	-0.97	0.88	0.00	S24-7
Day 20							
185186	-0.81	-0.81	-0.78	-0.78	0.00	0.00	Verrucomicrobiaceae
276629	0.82	0.85	0.93	0.92	0.00	0.00	S24-7
New.ReferenceOTU82	0.85	0.82	0.90	0.87	0.00	0.04	S24-7
New.ReferenceOTU492	0.81	0.77	0.86	0.75	0.00	0.07	S24-7
New.ReferenceOTU513	0.79	0.78	0.82	0.76	0.00	0.15	Lachnospiraceae
New.ReferenceOTU332	0.79	0.76	0.82	0.84	0.02	0.04	S24-7
206324	0.63	0.66	0.80	0.66	0.02	0.22	S24-7
4325509	0.55	-0.18	0.77	0.20	0.02	0.95	Lachnospiraceae
New.ReferenceOTU449	0.65	0.18	0.67	0.57	0.02	0.43	S24-7

Table 7.2: Top significantly correlated OTUs by Day (FDR = 0.05). ρ_S and ρ_{TS} denote unadjusted Spearman correlation, $\tilde{\rho}_S$ and $\tilde{\rho}_{TS}$ denote the Spearman correlation estimates from Spearman and Truncated Spearman tests, respectively whereas p($\tilde{\rho}_s$) and p($\tilde{\rho}_{sT}$) are the corresponding p-values, respectively.

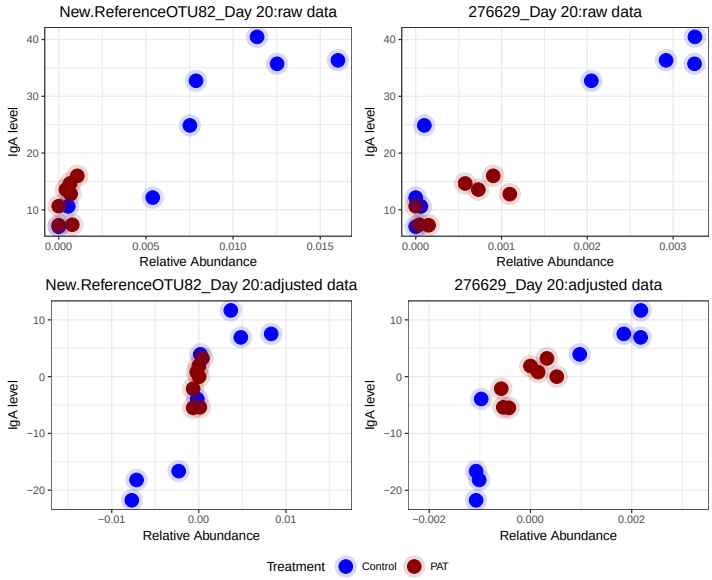


Figure 7.8: Example of two OTUs, New.RederenceOTU82 and OTU 276629, found to be significantly associated with IgA at Day 20 by both Spearman and truncated Spearman tests. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

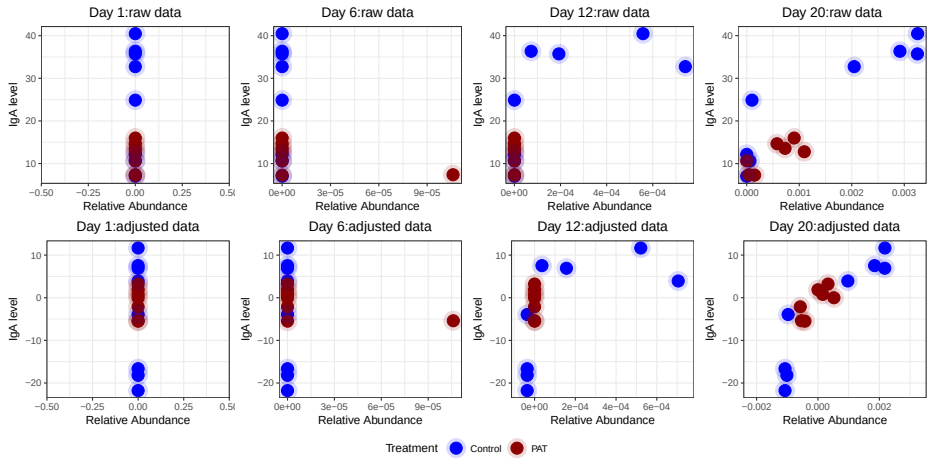


Figure 7.9: Relative abundance of OTU 276629 against IgA over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

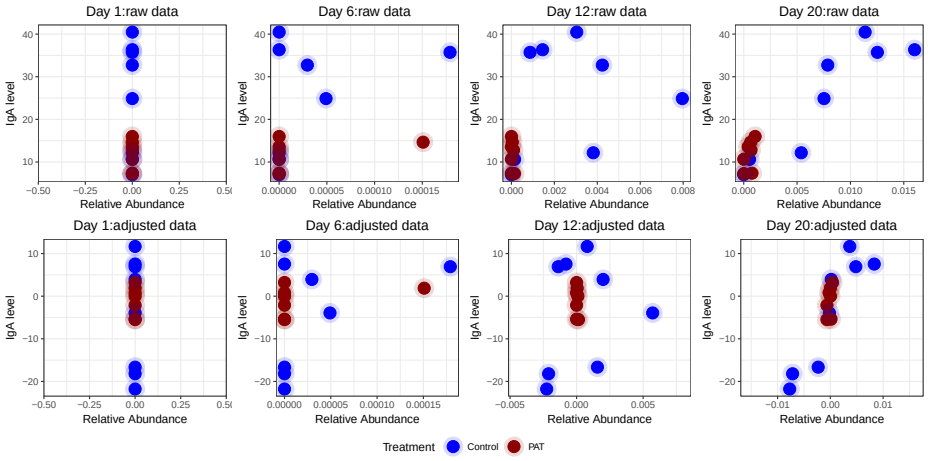


Figure 7.10: Relative abundance of New.ReferenceOTU82 against IgA over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

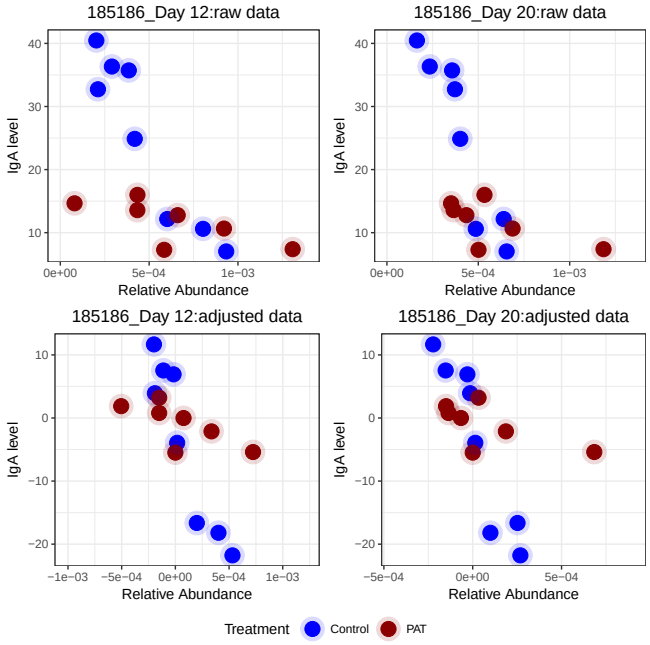


Figure 7.11: OTU 185186: significant adjusted association was detected from day 12 onwards. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

7.4 Discussion

For microbiome count data, with a lump at zero, it is difficult to model it with standard parametric models because all the distributional assumptions are no longer valid. Thus, we moved to non parametric models and defined three tests to test for differentially abundant OTUs and two tests to identify significantly associated OTUs. It was observed that the most of the significantly correlated OTUs on day 20 were from the *S24-7* family. Moreover, we presented the analysis at the OTU level in this chapter, but similar analysis can be done at any level of the hierarchical microbiome ecosystem e.g., family level, genus level, etc.

Chapter 8

Development of High Dimensional Microbiome Biomarkers for an Immune Response: Hierarchical Bayesian Approach

8.1 Introduction

In the previous chapters, the attention was placed on the triplet (X_{ji}, Y_i, Z_i), where $\mathbf{X}$ is an $m \times n$ biomarker matrix, $\mathbf{Y}$ is the clinical outcome and $\mathbf{Z}$ is the treatment (intervention) variable. The treatment effects on the biomarker and clinical outcome can be expressed, respectively, as,

$$\begin{aligned} g[E(X_{ji}|Z_i = 1)] - g[E(X_{ji}|Z_i = 0)] &= \alpha_j, \\ g[E(Y_i|Z_i = 1)] - g[E(Y_i|Z_i = 0)] &= \beta, \end{aligned} \tag{8.1}$$

where $g()$ denotes an appropriate link function, the index i indicates the ith subject, $i = 1, 2, \cdots, n$ and j denotes the jth microbiome feature, $j = 1, 2, \cdots, m$. As in the previous chapters, α_j and β are the treatment effects upon the microbiome and the clinical variable, respectively.

Up to this point in the thesis, all parametric models assumed a feature specific bivariate

Normal distribution for Y_i and X_{ji} with means given in equation (8.1) and the identity link function. In this chapter, we formulate a slightly different mean structure for the clinical outcome Y,

$$g[E(Y_i|X_{ji}, Z_i)] = \mu_Y + \beta Z_i + \gamma_j X_{ji}. \tag{8.2}$$

The mean structure for the microbiome biomarker remains the same as specified in equation (8.1) but the distribution and link function are changed. In this chapter, we use a Poisson distribution for the counts with log link function.
When both Y and X_j are assumed to follow a Normal distribution, the adjusted association ρ_j, given in equation (2.3), was used as a measure for individual level surrogacy based on the covariance matrix of the joint distribution $[Y_i, X_{ji}|Z_i]$. In this chapter, we proposed hierarchical Bayesian path analysis models (Congdon, 2014) in which the feature specific effect of the microbiome variable upon the clinical outcome is estimated using the parameter γ_j in equation (8.2) which is the direct effect of the biomarker on Y. We elaborate on this point in Section 8.2.1. Furthermore, this hierarchical path analysis model, formulated in Section 8.2, allows to evaluate a second level of association between Y and X_j, an association that is derived from the joint distribution of the treatment effects $[\alpha_j, \beta]$. This level of association corresponds to the trial level surrogacy (Alonso et al., 2016) in the clinical trial setting.
Similar to Chapter 6 and 7, the path analysis models, developed in this chapter, can be implemented at any level of the hierarchical microbiome ecosystem. For the analysis, presented in this chapter, the model is applied to family and kingdom level data. The transPAT data, presented in Chapter 2, is used for illustration. Figure 8.1 shows the family level richness (for the *S24-7* family) and α-diversity versus log(IgA).

This chapter is organized as follows. Section 8.2 describes the modeling approach while the results are discussed in Section 8.3. Final discussions about the method and the results are available in Section 8.4.

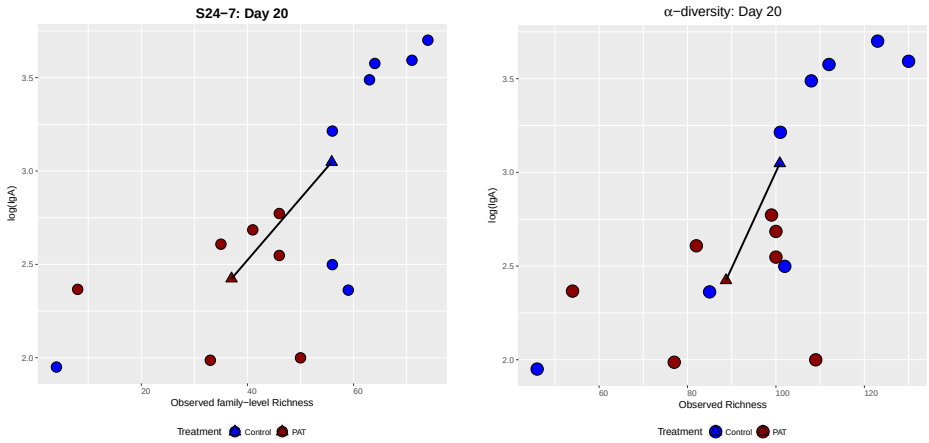
8.2 Hierarchical Model for Microbiome and IgA: A Path Analysis Approach

8.2.1 Model Formulation

In this section, we discuss the hierarchical Bayesian path analysis models for microbiome measurement and the clinical outcome log(IgA). We formulate a feature specific model

for a Poisson / Normal setting, given by,

$$\begin{aligned} X_{ji} &\sim \text{Poisson}(\lambda_i), \\ \log(\lambda_i) &= \mu_{X_j} + \alpha_j Z_i, \\ Y_i &\sim N(\mu_i, \tau), \\ \mu_i &= \mu_Y + \beta Z_i + \gamma_j X_{ji}. \end{aligned} \tag{8.3}$$



(a) **(b)**

Figure 8.1: The transPAT data. Panel a: observed family level richness for *S24-7* family against log(IgA). Panel b: observed α-diversity against log(IgA). The black line connects the means of the two treatment groups.

The path analysis model, specified in equation (8.3), is a partial mediation model. The indirect effect of Z on Y, which is mediated via X_j, is equal to $\alpha_j \times \gamma_j$ and the direct effect of Z on Y is equal to β. The parameter of primary interest, γ_j represents the direct effect of X_j on Y. Figure 8.2 illustrates the main difference between the joint model, used in Chapter 6, and the path analysis model formulated in equation (8.3). For the joint model, the association between Y and X_j is captured via ρ_j while in the path analysis model, the association is captured using the effect of X_j upon Y, i.e., γ_j. Note that a similar approach was used in Tilahun et al. (2010) who proposed an information theory approach (Alonso and Molenberghs, 2007) in the context of genomic biomarkers for depression. Tilahun et al. (2010) proposed to measure the degree of association between X_j and Y by comparing a model with mean structure given in equation (8.3) and a reduced model for which the mean structure is given by,

$$E(Y_i|X_{ji}, Z_i) = \mu_y + \beta Z_i. \tag{8.4}$$

Table 8.1 presents a set of 5 possible models. The mean structure for Y, specified for models 1 and 4, is similar to the mean structure specified in equation (8.3) while the mean structure for Y, specified for models 2 and 5, is similar to the mean structure of the reduced model in equation (8.4). The feature specific detection procedure proposed in Tilahun et al. (2010) was based on the so called R_{hj} (Alonso and Molenberghs, 2007). Feature with relatively high value of R^2_{hj} were considered as potential biomarkers for the clinical outcome. For the analysis, presented in this chapter, we use a different approach. All models specified in Table 8.1 and Table 8.2 are fitted and a model selection procedure, based on the deviance information criterion (DIC, Gelman et al., 2003, Spiegelhalter et al., 2014) is conducted to select the model with the best goodness of fit. Note that models 1, 3 and 4, which include a direct path from X_j to Y, indicate that the microbiome variable can be used as a biomarker to the clinical outcome.

(a) **(b)**

Figure 8.2: Modeling the relationship between $\mathbf{X_j}$ and $\mathbf{Y}$. For the family level analysis $\mathbf{X_j}$ denotes the richness of the jth family and for the kingdom level analysis $\mathbf{X}$ is a vector, containing α-diversity. Panel a: joint model formulated in equation (6.1). Panel b: path analysis model formulated in equation (8.3). The association between $\mathbf{X_j}$ and $\mathbf{Y}$ is measured using the adjusted association ρ_j in (a) and the direct effect γ_j in (b).

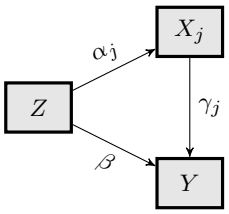
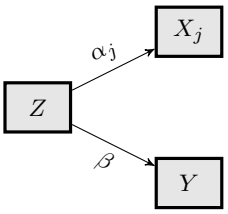
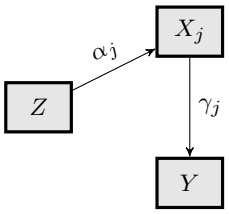
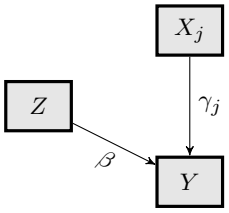
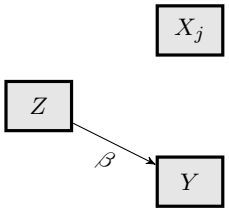
Models	Mean Structure	Correlation between Treatment Effects	Graphical Illustration
Model 1	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i + \gamma_j X_{ji}$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & 0 \\ 0 & d_\beta \end{pmatrix}$.	
Model 2	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & 0 \\ 0 & d_\beta \end{pmatrix}$.	
Model 3	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \gamma_j X_{ji}$.	NA	
Model 4	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j}$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i + \gamma_j X_{ji}$.	NA	
Model 5	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j}$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i$.	NA	

Table 8.1: Mean structure and covariance between the treatment effects for 5 possible models. The priors for the parameters α_j, β and γ_j are specified in Section 8.2.2.

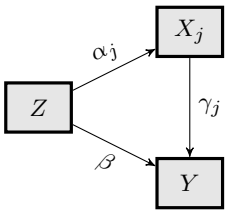
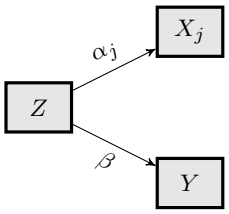
Models	Mean Structure	Correlation between Treatment Effects	Graphical Illustration
Model 6	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i + \gamma_j X_{ji}$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & d_{\alpha_j,\beta} \\ d_{\beta,\alpha_j} & d_{\beta} \end{pmatrix}$.	
Model 7	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim \text{N}(\mu_i, \tau)$, $\mu_i = \mu_Y + \beta Z_i$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & d_{\alpha_j,\beta} \\ d_{\beta,\alpha_j} & d_{\beta} \end{pmatrix}$.	

Table 8.2: Mean structure and covariance between the treatment effects for 2 possible models with correlation between the treatment effects α_j and β. The priors for the parameters α_j, β and γ_j are specified in Section 8.2.2.

8.2.2 Specification of the Prior Distributions

The experimental setting, discussed in this chapter, is similar to the single trial setting of the surrogacy framework in the sense that a single study was conducted to investigate the relationship between X_j and Y. However, the joint model, specified in equation (8.3), allows to estimate a second level of association related to the treatment effects. In order to complete the specification of the model, formulated in equation (8.3), we specify a bivariate Normal prior distribution for the treatment effects, that is,

$$\begin{pmatrix} \alpha_j \\ \beta \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D}_j \right), \tag{8.5}$$

with the covariance matrix defined by,

$$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & d_{\alpha_j,\beta} \\ d_{\beta,\alpha_j} & d_{\beta} \end{pmatrix}. \tag{8.6}$$

Note that similar prior model was proposed by Shkedy and Barbosa (2005) in the context of the meta-analytic approach for surrogacy. For the covariance matrix $\mathbf{D}_j$, we assume Wishart hyperprior distribution, given by,

$$\mathbf{D}_j^{-1} \sim \text{Wishart}(\mathbf{R}_{D_j}), \tag{8.7}$$

where $\mathbf{R}_{D_j}$ is given by,

$$\mathbf{R}_{D_j} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

In the context of the clinical trials setting with ℓ trials, Shkedy and Barbosa (2005) showed that the joint distribution of trial specific treatment effects, α_ℓ and β_ℓ can be used to estimate the trial level surrogacy using the correlation between the treatment effects. In our application, it can be derived from the covariance matrix $\mathbf{D}_j$ and is defined by,

$$\rho(\alpha_j, \beta) = \frac{d_{\alpha_j,\beta}}{\sqrt{d_{\alpha_j} d_\beta}}. \tag{8.8}$$

In Shkedy and Barbosa (2005), a measure for trial level surrogacy is equal to $\rho^2(\alpha_\ell, \beta_\ell)$[1]. For the current application, $\rho(\alpha_j, \beta)$ can be interpreted as the second level of association between X_j and Y that is derived by the treatment effect. Examples of two models are presented in Table 8.2. For the case in which the treatment effects are not correlated (i.e., Model 1 - Model 5 in Table 8.1), the covariance matrix is reduced to

$$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & 0 \\ 0 & d_\beta \end{pmatrix}. \tag{8.9}$$

Next, we specify a non informative prior for γ_j,

$$\gamma_j \sim \text{N}(0, 1000000). \tag{8.10}$$

Further, independent non informative flat priors were specified for the intercepts,

$$\begin{aligned} \mu_{X_j} &\sim \text{N}(0, 1000000), \\ \mu_Y &\sim \text{N}(0, 1000000). \end{aligned} \tag{8.11}$$

For the precision parameter, τ, in equation (8.3), flat hyperprior is specified using Gamma distributions,

$$\tau \sim \text{Gamma}(0.001, 0.001). \tag{8.12}$$

Note that for the kingdom level analysis, $\mathbf{X}$ is a vector, containing the observed α-diversity for the subjects and hence the index j should be dropped.

[1] Note that in Shkedy and Barbosa (2005) treatment effects were trial specific and therefore correlation was estimated from a joint distribution of ℓ random effects (representing ℓ trials).

8.3 Application to the Data

We present the results for the *S24-7* family in Section 8.3.1 while the results for the α-diversity are presented in Section 8.3.2. The 7 models, formulated in Table 8.1 and Table 8.2, were implemented using `WinBUGS` and `R` using the `R2WinBUGS` package. Three MCMC chains were run and each chain had 50000 iterations from which the first 20000 were used as the burn-in period. Example code to fit the models is given in Section E.2 in Appendix E.

8.3.1 Results for S24-7 Family

Table 8.3 presents the DIC values for all fitted models for the *S24-7* family at different timepoints. For the first two days Model 2 has the smallest DIC among the fitted models. This model implies a conditional independence between richness and log(IgA), i.e., given the treatment group, richness and log(IgA) are independent. Hence, no direct effect of X_j upon Y is included. At the last two days of the study, day12 and 20, Model 3, the complete mediation model, has the smallest DIC (217.366 and 244.923, respectively). This indicates that the effect of richness on log(IgA) is developed over time and it is present from day 12 onwards. It is important to mention that the complete mediation model implies that the treatment has only an indirect effect on log(IgA). Table 8.4 displays the posterior means and credible intervals for the selected models. Figure 8.3 shows the density estimate for γ_j obtained from Model 1 for all timepoints and reveals a shift to the right in the posterior distribution as time passes, which, as mentioned above, indicates the development of the effect of the richness on log(IgA) over time.

As mentioned in Section 8.2.2, the path analysis model specified in equation (8.3), allows to estimate the correlation between α_j and β. On days 1 and 6, Model 7, the conditional independence model with correlated treatment effect, has slightly lower DIC than Model 2 (133.589 and 133.866 for model 7 and 2 on day 1, respectively and 142.696 and 142.541 for Model 7 and 2 on day 6, respectively). Posterior mean for the correlation are equal to 0.139 and 0.287 on day 1 and 6, respectively. However, we notice that the credible intervals cover almost all the range of $[-1, 1]$, (-0.927, 0.968) and (-0.889, 0.979) on day 1 and 6, respectively.

Models	Day 1	Day 6	Day 12	Day 20
Model 1	134.222	144.621	219.381	245.929
Model 2	**133.866**	**142.696**	220.279	251.572
Model 3	138.442	143.659	**217.366**	**244.923**
Model 4	137.459	197.497	300.113	272.697
Model 5	137.018	195.496	300.935	278.257
Model 6	134.140	144.307	218.962	245.655
Model 7	**133.589**	**142.541**	220.115	251.414

Table 8.3: *S24-7* family. Deviance information criterion for the fitted models.

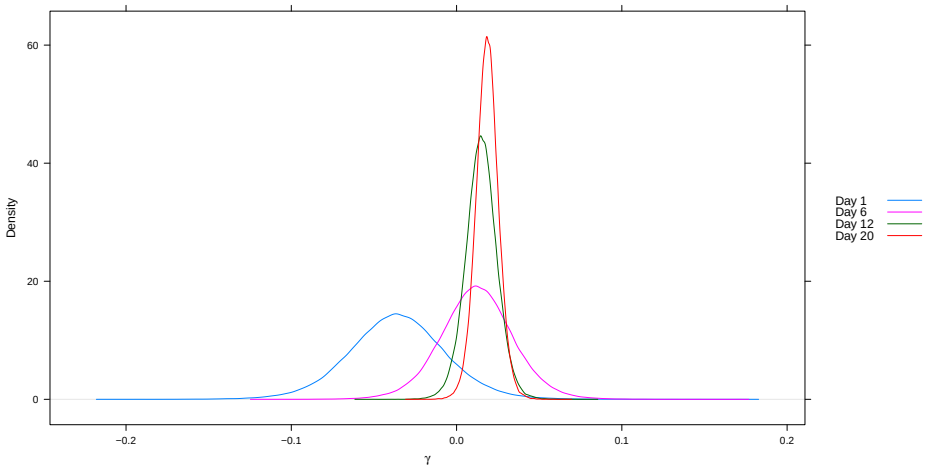


Figure 8.3: Density estimate of the direct effect γ_j, obtained from Model 1, over time.

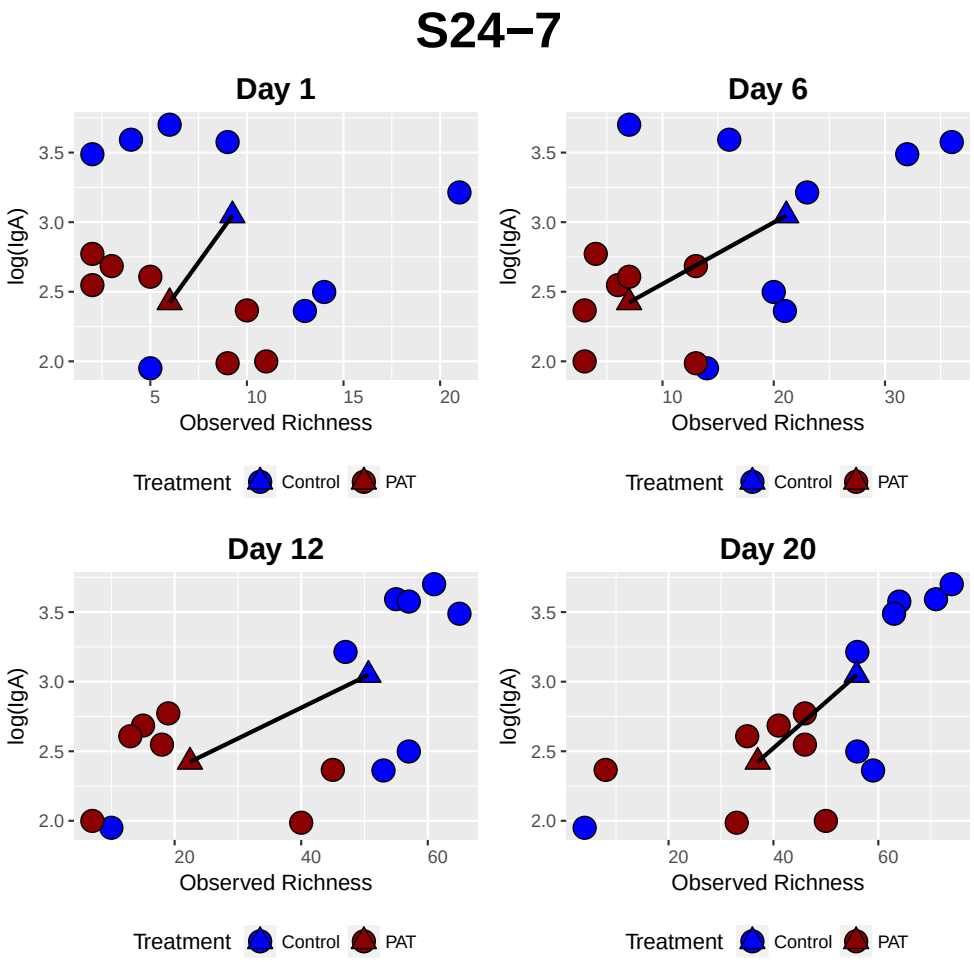


Figure 8.4: Observed family level richness of the *S24-7* family against log(IgA) across different timepoints.

8.3.2 Results for α-Diversity

Table 8.5 presents the deviance information criteria for all fitted models when α-diversity is used as a biomarker and shows that for days 1 and 20 Model 1 (partial mediation model) is the model with the best goodness of fit while for days 6 and 12 Model 3 (complete mediation model) is to be preferred. This implies that the direct effect of the microbiome variable is present from the beginning of the study at day 1 until day 20. Moreover, for days 6 and 12, the entire effect of the treatment on log(IgA) is mediated via the microbiome variable. Table 8.6 displays the posterior means and credible intervals for the

Timepoint	**Day 1**		**Day 6**		**Day 12**	**Day 20**
Best Model	Model 2	Model 7	Model 2	Model 7	Model 3	Model 3
Estimate(α)	-0.438	-0.408	-1.112	-1.078	-0.816	-0.413
95% C.I.(α)	(-0.824,-0.059)	(-0.786,-0.043)	(-1.438,-0.799)	(-1.401,-0.770)	(-1.001,-0.633)	(-0.567,-0.260)
Estimate(β)	-0.624	-0.536	-0.624	-0.564	NA	NA
95% C.I.(β)	(-1.228,-0.018)	(-1.090,0.033)	(-1.228,-0.018)	(-1.133,0.020)	NA	NA
Estimate(γ)	NA	NA	NA	NA	0.018	0.022
95% C.I.(γ)	NA	NA	NA	NA	(0.005,0.032)	(0.010,0.034)
Estimate(ρ)	NA	0.139	NA	0.287	NA	NA
95% C.I.(ρ)	NA	(-0.927,0.968)	NA	(-0.889,0.979)	NA	NA

Table 8.4: *S24-7* family. Posterior means for the selected model by day.

parameters of the selected models at each timepoint.
When the correlation between the treatment effects, is added to the models the conditional independence model with correlated treatment effects, α_j and β, has the smallest DIC on day 1 (151.824 and 151.721 for Model 1 and Model 7, respectively) and on day 12 (197.993 and 197.912 for Model 3 and Model 7, respectively). In other words, the direct effect of α-diversity which is captured via γ_j in Model 1 (day 1) and Model 3 (day 12), can now be explained via the correlation between the treatment effects. Note that, at day 20 Model 6 has slightly lower DIC than Model 1 (200.173 and 200.052 for Model 1 and Model 6, respectively). Both the direct effect of α-diversity on log(IgA) and the correlation between the treatment effects are present on day 20. However, as shown in Table 8.6, the posterior mean of the correlation is estimated with poor precision, i.e, credible intervals are (-0.934, 0.965), (-0.929, 0.966),(-0.945, 0.956) on day 1, 12 and 20, respectively.

Models	**Day 1**	**Day 6**	**Day 12**	**Day 20**
Model 1	**151.824**	153.612	198.753	**200.173**
Model 2	151.862	156.931	198.032	208.180
Model 3	156.728	**151.697**	**197.993**	201.990
Model 4	163.691	174.406	230.437	204.057
Model 5	163.650	177.636	229.622	211.972
Model 6	151.792	153.197	198.325	**200.052**
Model 7	**151.721**	156.832	**197.912**	208.059

Table 8.5: Deviance information criterion for α-diversity.

8.4 Discussion

In this chapter, we shifted from the Normal / Normal setting of the joint model, discussed in Chapter 6 to a Poisson / Normal setting in order to model directly the effect of the microbiome biomarker on log(IgA). Model selection was done based on DIC. We have shown that for the *S24-7* family the direct effect of the observed family level richness on log(IgA) develops over time. It might be because it takes some time for the microbiome to mature and once, the microbiome has stabilized, the richness of the *S24-7* family affects the immunity and alters the IgA level. Note that the direct effect of α-diversity

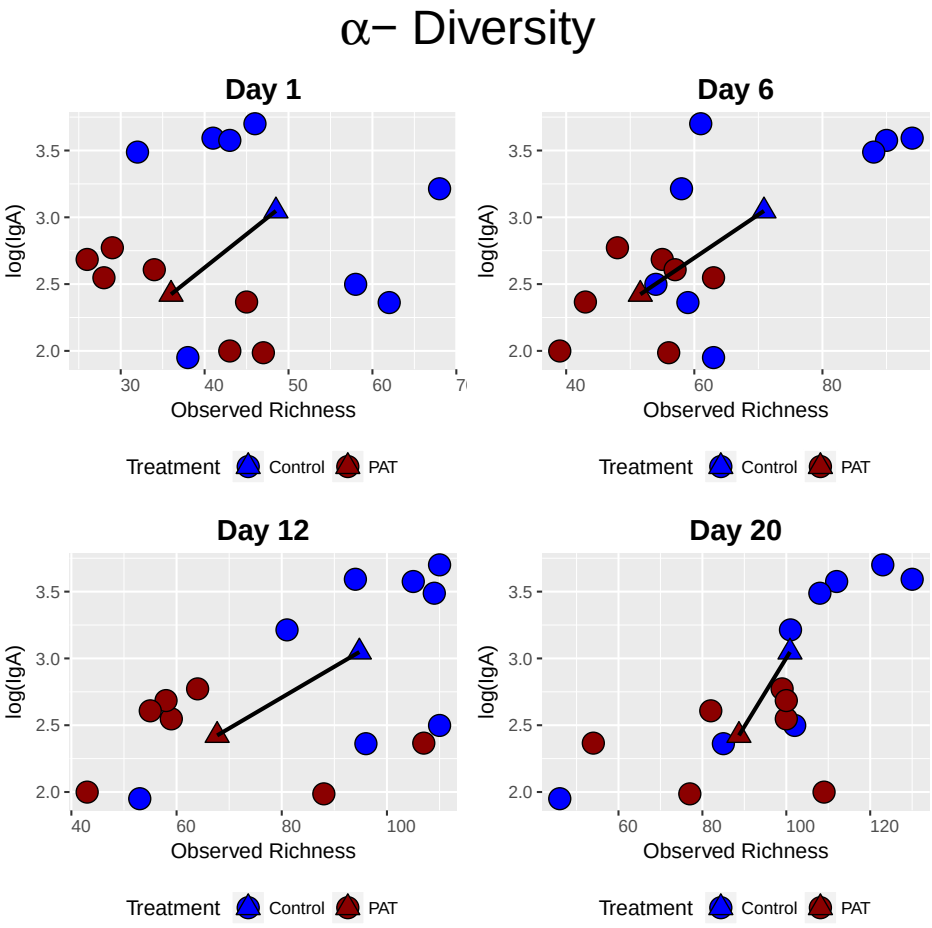


Figure 8.5: Observed α-diversity against log(IgA) across different timepoints.

Timepoint	**Day 1**		**Day 6**	**Day 12**		**Day 20**	
Best Model	Model 1	Model 7	Model 3	Model 3	Model 7	Model 1	Model 6
Estimate(α)	-0.299	-0.295	-0.319	-0.336	-0.334	-0.129	-0.127
95% C.I.(α)	(-0.458,-0.141)	(-0.454,-0.139)	(-0.450,-0.187)	(-0.451,-0.222)	(-0.450,-0.219)	(-0.233,-0.024)	(-0.232,-0.024)
Estimate(β)	-0.863	-0.531	NA	NA	-0.534	-0.423	-0.374
95% C.I.(β)	(-1.541,-0.185)	(-1.095,0.040)	NA	NA	(-1.093,0.036)	(-0.885,0.041)	(-0.812,0.072)
Estimate(γ)	-0.019	NA	0.026	0.013	NA	0.017	0.017
95% C.I.(γ)	(-0.047,0.009)	NA	(0.010,0.042)	(0.001,0.026)	NA	(0.006,0.027)	(0.007,0.027)
Estimate(ρ)	NA	0.107	NA	NA	0.118	NA	0.039
95% C.I.(ρ)	NA	(-0.934,0.965)	NA	NA	(-0.929,0.966)	NA	(-0.945,0.956)

Table 8.6: Posterior means obtained from the best model for observed α-diversity across all timepoints.

on log(IgA) is always present across different timepoints. The weakness of the modeling approach presented in this chapter is the fact that it is based on timepoint-specific models and not on a model which include time effect and fitted for all days together. This is a topic of an ongoing research.

The proposed Poisson / Normal setting allows to model a second level of association between the microbiome biomarker and log(IgA) via the correlation between the treatment effects, α_j and β. However, the precision with which $\rho(\alpha_j, \beta)$, the correlation between the treatment effects, can be estimated is poor as the credible interval for $\rho(\alpha_j, \beta)$ contains the entire range from -1 to 1. This could be a result of the small sample size (7 and 8 subjects, in the PAT and control groups, respectively) and should be investigated further via a simulation study.

Chapter 9

Development of Microbiome Biomarkers for Type 1 Diabetes

9.1 Introduction

The study we analyze in this chapter is a microbiome intervention study based on an animal model, developed in order to explore the association between antibiotic intake and type 1 diabetes. In total, 39 and 40 type 1 diabetes (T1D) free mice were randomized into two treatment groups (control and PAT) that received placebo and antibiotic, respectively. The treatment was administered to the subjects on three occasions before Week 3 (study design is shown in Figure 5.7a) and the subjects were followed up until week 30. The disease status was monitored for a period of 30 weeks and microbiome data were collected day 21, 35 and 49. The first T1D event occurred at week 10. An elaborate description of the data is given in Section 5.3.2.

Figure 9.1 reveals a clear treatment effect on the time to develop T1D between the two treatment groups. Median time to develop T1D is equal to 19.5 and 30 weeks in the antibiotic and placebo group, respectively, indicating that antibiotic intake increases the risk to develop the disease (log rank test p-value $= 0.006$). In total, 48.7% and 72.5% of the mice in the placebo and PAT group developed T1D, respectively. Figure 9.2 shows the time to develop T1D versus α-diversity (Panel a) and the richness of the *S24-7* family (Panel b). The main objective of the analysis presented in this chapter is to explore if microbiome variable(s), at different levels of the hierarchical microbiome ecosystem, can be used as a predictive biomarker for the time to develop T1D. Similar to Chapter 8, we use a path analysis model to jointly model the microbiome and time to develop T1D

data. Taking into account that the clinical variable of primary interest is time to event, we present a joint Poisson / Survival model for the microbiome and time to develop T1D variables. For the time to develop T1D a Weibull model is used. The modeling approach is discussed in Section 9.2. Similar to the previous chapter, the analysis is conducted for both α-diversity (Section 9.3.1) and family level richness (Section 9.3.2). We discuss the results in Section 9.4.

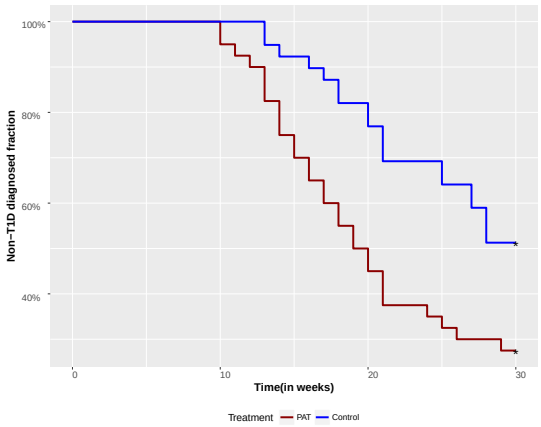


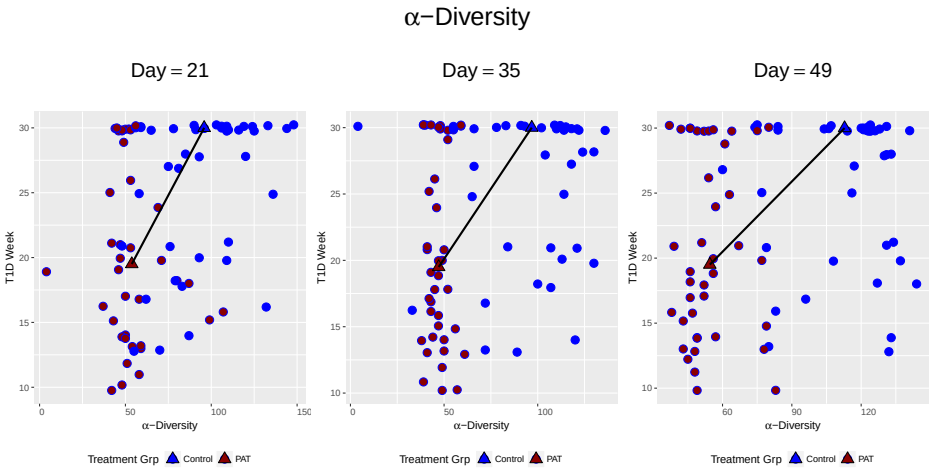
Figure 9.1: Time to develop T1D: Kaplan-Meier curve by treatment group. Log rank test p-value $= 0.006$.

9.2 A Poisson / Survival Path Analysis Model

Burzykowski et al. (2005) and Alonso et al. (2016) discussed a model for continuous endpoint with Normal distribution and a failure-time endpoint. For a single trial setting and for the ith subject, they proposed to model the time to failure, Y_i using proportional hazard model given by,

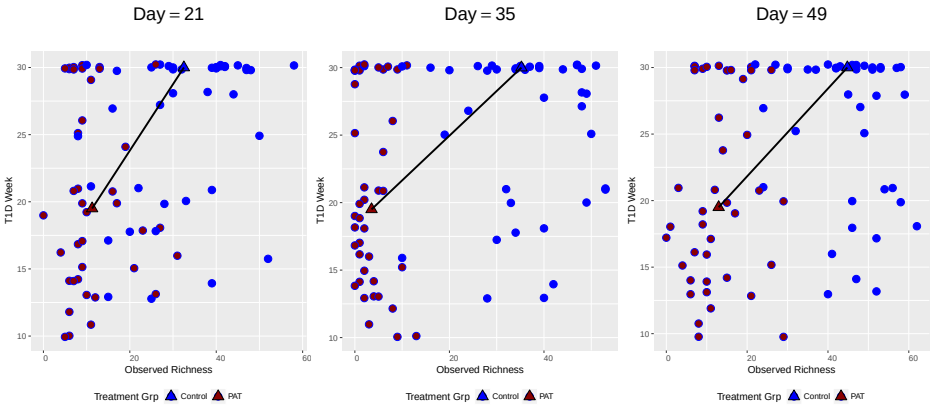
$$h_i(y|Z_i) = h_0(y)exp(\beta Z_i). \tag{9.1}$$

Here, β is the treatment effect upon the true endpoint and $h_0(y)$ is a baseline hazard function. For a mult trial setting, Burzykowski et al. (2005) formulated Weibull model to analyze time to event data using the treatment as a covariate and individual level surrogacy was estimated using copulas. Burzykowski et al. (2005) used proportional hazard model, specified in equation (9.1), with Weibull trial specific baseline hazard functions.



(a)

S24−7



(b)

Figure 9.2: T1D study. Panel a: scatterplot of observed α-diversity over time against the time to develop T1D. Panel b: scatterplot of observed family level richness of the *S24-7* family over time against the time to develop T1D.

In this section, we discuss several path analysis models for the analysis of the time to develop T1D and the microbiome variables. Similar to Chapter 8, family richness and α-diversity are used for illustration. For the microbiome variable X_{ji}, we define a feature specific Poisson model with linear predictor given by,

$$\begin{aligned} X_{ji} &\sim \text{Poisson}(\lambda_i), \\ \log(\lambda_i) &= \mu_{X_j} + \alpha_j Z_i. \end{aligned} \tag{9.2}$$

As in Chapter 8, X_{ji} is the richness of the jth family for the ith subject. For the case when α-diversity is used, $\mathbf{X} = (X_1, X_2, \cdots, X_n)$. Non informative independent priors were specified for the model parameters,

$$\begin{aligned} \mu_{X_j} &\sim \text{N}(0, 1000000), \\ \alpha_j &\sim \text{N}(0, 1000000). \end{aligned} \tag{9.3}$$

9.2.1 A Weibull Model for Time to Develop T1D

Klein and Moeschberger (2003) discussed Weibull regression models for the analysis of time to event data for which the hazard rate and survival function at time *y*, *h(y)* and *S(y)* respectively, are defined as,

$$\begin{aligned} h(y) &= \lambda\alpha y^{\alpha-1}, \\ S(y) &= exp(-\lambda y^{\alpha}), \end{aligned} \tag{9.4}$$

where $\lambda > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter. $\alpha = 1$ indicates constant hazard equivalent to a exponential model.
In the context of proportional hazard models λ can be parameterized in order to include covariates in the model (Dellaportas and Smith, 1993). For time to develop T1D, such a model can be expressed as,

$$\begin{aligned} Y_i &\sim Weibull(\mu_i, \theta), \\ \mu_i &= exp(\mu_Y + \beta Z_i + \gamma_j X_{ji}). \end{aligned} \tag{9.5}$$

Here, θ is the shape parameter and μ_i captures the effect of the covariates on time to develop T1D. The model formulated in equation (9.5) implies that the baseline hazard is given by,

$$h_0(y) = \theta y^{\theta-1},$$

and the survival function is calculated by,

$$S_i(y) = exp(-\mu_i y^{\theta}).$$

Following Dellaportas and Smith (1993), we assume a Gamma prior for the shape parameter, that is,

$$\theta \sim \text{Gamma}(1, 0.00001).$$

The prior distributions for the parameters are given by,

$$\mu_Y \sim \mathsf{N}(0, 1000000),$$
$$\beta \sim \mathsf{N}(0, 1000000),$$
$$\gamma_j \sim \mathsf{N}(0, 1000000).$$

Similar to the previous chapter, several path analysis models, presented in Table 9.1 and Table 9.2 can be fitted to the data. Model 1 and 2 assume that the treatment effects are independent while model 6 and 7 assume correlated treatment effects.

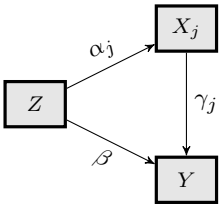
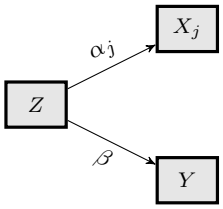
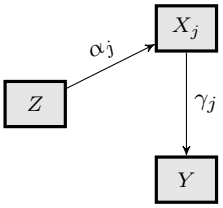
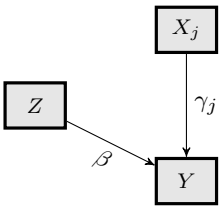
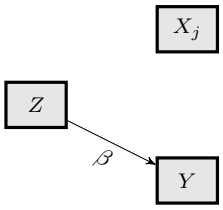
Models	Mean Structure	Correlation between Treatment Effects	Graphical Illustration
Model 1	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i + \gamma_j X_{ji})$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & 0 \\ 0 & d_\beta \end{pmatrix}$.	
Model 2	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i)$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & 0 \\ 0 & d_\beta \end{pmatrix}$.	
Model 3	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \gamma_j X_{ji})$.	NA	
Model 4	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j}$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i + \gamma_j X_{ji})$.	NA	
Model 5	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j}$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i)$.	NA	

Table 9.1: Mean structure and covariance between the treatment effects for Poisson / Weibull model. Hyperprior distribution for $\mathbf{D}_j$ is same as mentioned in Section 8.2.2.

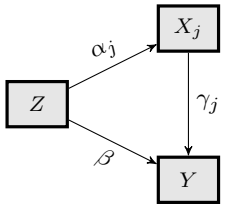
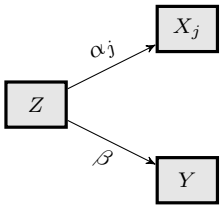
Models	Mean Structure	Correlation between Treatment Effects	Graphical Illustration
Model 6	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i + \gamma_j X_{ji})$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & d_{\alpha_j,\beta} \\ d_{\beta,\alpha_j} & d_\beta \end{pmatrix}$.	
Model 7	$X_{ji} \sim \text{Poisson}(\lambda_i)$, $\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i$, $Y_i \sim Weibull(\theta, \mu_i)$, $\mu_i = exp(\mu_Y + \beta Z_i)$.	$\mathbf{D}_j = \begin{pmatrix} d_{\alpha_j} & d_{\alpha_j,\beta} \\ d_{\beta,\alpha_j} & d_\beta \end{pmatrix}$.	

Table 9.2: Mean structure and covariance between the treatment effects for all Poisson / Weibull models with correlation between the treatment effects α_j and β. Hyperprior distribution for $\mathbf{D}_j$ is same as mentioned in Section 8.2.2.

9.3 Application to the Data

The models discussed above were implemented using `WinBUGS` and `R2WinBUGS`. Three MCMC chains were used to fit each model. A chain consists of 100000 iterations from which the first 50000 are used as a burn-in period. The DIC is used for model selection. The code to fit Model 1 and model diagnostics are presented in Section F.2 in Appendix F.

9.3.1 Results for α-Diversity

Table 9.3 displays DIC values for all fitted models and Table 9.4 presents the posterior means and credible intervals obtained from the selected models across different timepoints. Model 3 (the complete mediation model) has the best goodness of fit at day 21 while for the last two timepoints Model 2 (the conditional independence model) has the lowest DIC. This implies that at day 21 the entire treatment effect on the time to develop T1D is mediated through the microbiome variable. However, for the last two timepoints, these two variables become independent, given the treatment group.

When the models with correlation between the treatment effects were fitted, we noticed that the conditional independence model with correlated treatment effects has the minimum DIC at day 49. However, the correlation is estimated with a poor precision, i.e.,

with credible interval equal to (-0.980,0.880).

Model	Day 21	Day 35	Day 49
Model 1	1395.090	1328.190	1181.830
Model 2	1394.550	**1326.160**	**1179.680**
Model 3	**1394.270**	1329.720	1181.470
Model 4	1869.750	2018.950	2000.650
Model 5	1869.170	2017.060	1998.580
Model 6	1395.010	1328.250	1181.150
Model 7	1394.430	1326.270	**1179.430**

Table 9.3: α-diversity. Deviance information criterion obtained from the fitted models.

Timepoint	Day 21	Day 35	Day 49	
Best Model	Model 3	Model 2	Model 2	Model 7
Estimate(α)	-0.580	-0.580	-0.720	-0.727
95% C.I.(α)	(-0.633,-0.527)	(-0.633,-0.526)	(-0.775,-0.665)	(-0.778,-0.676)
Estimate(β)	NA	0.840	0.836	0.754
95% C.I.(β)	NA	(0.253,1.445)	(0.263,1.434)	(0.201,1.327)
Estimate(γ)	-0.014	NA	NA	NA
95% C.I.(γ)	(-0.025,-0.004)	NA	NA	NA
Estimate(ρ)	NA	NA	NA	-0.293
95% C.I.(ρ)	NA	NA	NA	(-0.980,0.880)

Table 9.4: Posterior means obtained from the model with the smallest DIC per timepoint.

9.3.2 Results for S24-7 Family

Table 9.5 and Table 9.6 summarize the results for the *S24-7* family for models in which the family level richness was used as the microbiome variable. Similar to the results obtained for the α-diversity, at day 21, Model 3 has the lowest DIC, indicating that the entire treatment effect on time to develop T1D is mediated via the microbiome. However, the microbiome variable and time to develop T1D become conditionally independent, given

the treatment, for the last two timepoints. When the models with correlation were fitted to the data, Model 7 achieves the lowest DIC at day 35. The results for other 2 active families, namely *Lachnospiraceae* and *Lactobacillaceae* are presented in Section F.1 of Appendix F.

Models	**Day 21**	**Day 35**	**Day 49**
Model 1	1142.790	1072.270	1073.370
Model 2	1142.630	**1071.450**	**1071.380**
Model 3	**1141.570**	1075.630	1074.190
Model 4	1566.220	2263.890	1807.000
Model 5	1566.100	2263.000	1805.060
Model 6	1142.770	1072.200	1073.140
Model 7	1142.350	**1071.240**	1071.420

Table 9.5: Richness of the *S24-7* family. DIC values for all fitted models.

Timepoint	**Day 21**	**Day 35**		**Day 49**
Best Model	Model 3	Model 2	Model 7	Model 2
Estimate(α)	-1.060	-2.305	-2.297	-1.242
95% C.I.(α)	(-1.168,-0.953)	(-2.481,-2.133)	(-2.471,-2.127)	(-1.241,-1.145)
Estimate(β)	NA	0.839	0.808	0.834
95% C.I.(β)	NA	(0.257,1.440)	(0.236,1.394)	(0.257,1.425)
Estimate(γ)	-0.031	NA	NA	NA
95% C.I.(γ)	(-0.053,-0.011)	NA	NA	NA
Estimate(ρ)	NA	NA	-0.487	NA
95% C.I.(ρ)	NA	NA	(-0.990,0.786)	NA

Table 9.6: The *S24-7* family. Posterior means and credible intervals obtained from the models with the smallest DIC across different timepoints.

9.4 Discussion

In this chapter, we presented a Poisson / Survival model for the analysis of α-diversity and family level richness and the time to develop T1D. Similar analysis can be conducted at an OTU level or at any other level of the phylogenetic tree. The disadvantage of the analysis presented in this chapter is that the models are fitted at each timepoint separately and therefore a possible correlation among the microbiome observation for the same subject over time is ignored. As in Chapter 8, the models discussed in this chapter can be used to estimate the correlation between the treatment effect upon the microbiome and time to develop T1D variables. However, posterior means for the correlation was estimated with poor precision and a simulation study should be conducted in order to investigate the source(s) of uncertainty related to this parameter.

Chapter 10

Discussion and Further Research

The research, presented in this thesis, is focused on biomarker detection and high dimensional surrogacy in the context of transcriptomic and microbiome data. Different settings and modeling approaches were presented across different chapters. In this chapter, we discuss additional research lines that can be further developed based on the research presented in this thesis.

10.1 High Dimensional Surrogacy

10.1.1 Computational Issues

Powerful hardware is expensive to buy and maintain and institutional cluster computing resources, such as the VSC cluster, are not always accessible for all users. Cloud computing, an internet based computing platform, which enables any user to use resources that are not physically available to the user, can be a solution whenever a large scale analysis should be conducted. In this section, we discuss possible research lines and current problems within the cloud computing framework.

10.1.1.1 Amazon and Microsoft Cloud

Amazon Web Services or *Microsoft Azure* provides a fast, cheap and easy to use computational resources in the cloud.
A *virtual appliance* is a pre-configured virtual machine image e.g., *Amazon Machine Image* (AMI). An AMI includes one of the common operating systems such as Windows, Linux, etc. and any additional software as per requirement. Once an AMI is created, the user

can configure and launch his/her instance (refers to a machine / node in the cloud).
Our aim is to set up a cluster in the Amazon cloud in order to conduct the analysis presented in Chapter 2. This can be done following the cluster computing toolkit, developed by the Software Tools for Academics and Researchers Group (STAR) group at the Massachusetts Institute of Technology (MIT). After setting up the cluster and installing the necessary softwares (e.g., `R`, RStudio, JAGS etc.) a MapReduce framework (Dean and Ghemawat, 2008) can be implemented using the `R` package `rmr2` (Revolution Analytics, 2015).
There are few advantages of having a cluster in the cloud:

1. there is a cluster specific to the user and the user is billed based on usage. Hence, there is no extra cost of maintaining the cluster.

2. there is no additional queueing time for the user as he/she can start and stop the cluster as per requirement.

3. it is no longer necessary to have `R` or other softwares installed in the user's laptop as it is possible to access the cluster in cloud from any device (for example, using smartphone, tab, laptop, desktop) with an active internet connection.

Even though there are technologies to create a cluster in cloud is available, the usage of Amazon cloud computing with `R`, in the context of the analysis presented in Chapter 2, is not recommended since none of the available technologies use an efficient approach for parallelization, e.g., the worker framework discussed in Chapter 2.

Microsoft Azure is a cloud platform that allows to use Azure's computing resources from the userâĂŹs local `R` session (Figure 10.1) using the `R` package `doAzureParallel` and provides the infrastructure to run massively parallel simulations on Azure directly from the local `R` session. Therefore it allows to upscale an analysis from a local `R` session as it builds on the similar framework as the `foreach` package and the user needs to make only small modifications to the code to adapt to this framework.

10.1.2 Modeling Issues

In the first part of the thesis, several models, in the context of drug discovery experiments, were discussed. In Chapter 2, the joint model given by,

$$\begin{aligned} X_{ji} &= \mu_{X_j} + \alpha_j Z_i + \epsilon_{X_j}, \\ Y_i &= \mu_Y + \beta Z_i + \epsilon_Y, \end{aligned} \tag{10.1}$$

was formulated and used to estimate the feature specific treatment effect upon the biomarker and the adjusted association.

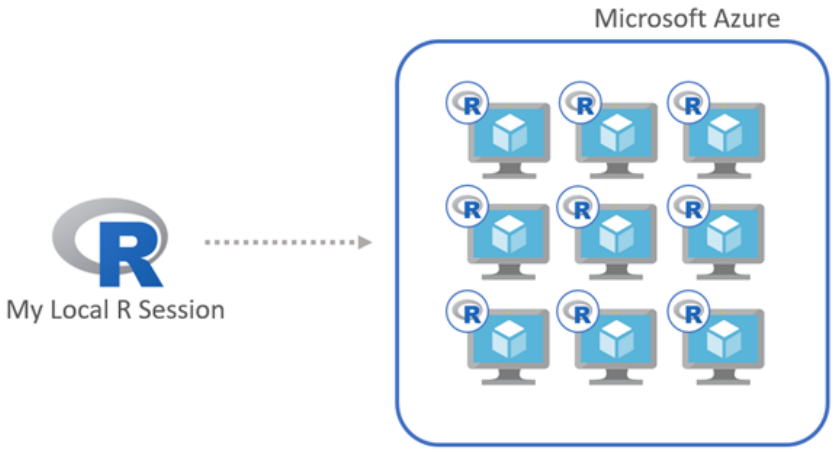


Figure 10.1: R in Microsoft Azure cloud. More details are available at `https://blogs.technet.microsoft.com/machinelearning/2017/03/16/take-advantage-of-scalable-cloud-compute-directly-from-your-r-session-with-doazureparallel/`.

The mean structures for the path analysis models, presented Chapter 8, are given by,

$$\begin{aligned} E(X_{ji}|Z_i) &= \mu_{X_j} + \alpha_j Z_i, \\ E(Y_i|Z_i, X_{ji}) &= \mu_Y + \beta Z_i + \gamma_j X_{ji}. \end{aligned} \tag{10.2}$$

Alonso and Molenberghs (2008) suggested an information theoretic approach to calculate individual level surrogacy using the squared informational coefficient of correlation (SICC),

$$\hat{R}^2_{indiv} = 1 - e^{-\frac{1}{n}G^2}.$$

Here, n is the number of subjects and G^2 is the log(likelihood ratio test), comparing the models formulated for the clinical outcome in equation (10.1) and (10.2),

$$\begin{aligned} M_0 &: E(Y_i|Z_i) = \mu_Y + \beta Z_i, \\ M_1 &: E(Y_i|Z_i, X_{ji}) = \mu_Y + \beta Z_i + \gamma_j X_{ji}. \end{aligned} \tag{10.3}$$

In the context of the multiple surrogacy setting, discussed in Chapter 3, when more than one biomarker is available, the mean structure for the clinical outcome in equation (10.2) can be expressed as,

$$E(Y_i|Z_i, X_{1i}, X_{2i}, \cdots, X_{Ji}) = \mu_Y + \beta Z_i + \sum_{j=1}^{J} \gamma_j X_{ji}. \tag{10.4}$$

Here i indicates the common dimension (e.g., subjects, compounds etc.). Penalized regression techniques such as LASSO (Tibshirani, 1996), can be used to fit models, specified in equation (10.4) when the number of predictors, J, is greater than the number of observations. The model specified in equation (10.4), can be further used to evaluate different measures of surrogacy. Following the information theoretic approach, the value of G^2 can be updated by log(likelihood ratio test), comparing the models specified in equation (10.3) and (10.4).

10.2 Analyzing Microbiome Data

10.2.1 Identifying Multiple Microbiome Biomarkers

The OTU level analysis, presented in Chapter 6 is based on an OTU specific joint model. Similar to Chapter 4, multiple microbiome biomarkers can be identified and evaluated using multiple and partial surrogacy. The same holds for a richness analysis at family level when few families are evaluated as a multiple biomarkers for IgA. The model formulated in equation (6.1) can be extended to include more than one family, that is,

$$\begin{pmatrix} X_{1i} \\ X_{2i} \\ Y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 + \alpha_1 Z_i \\ \mu_2 + \alpha_2 Z_i \\ \mu_Y + \beta Z_i \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1y} \\ \sigma_{21} & \sigma_{22} & \sigma_{2y} \\ \sigma_{y1} & \sigma_{y2} & \sigma_{yy} \end{pmatrix} \right].$$

Here, Y indicates the response variable (IgA or log(IgA)), X_1 and X_2 denote the relative abundance (or a normalized measure of richness) of two OTUs (or families).

10.2.2 Interaction between OTUs

The analysis presented in the second part of the thesis does not take into account a possible interaction between OTUs. Since, both transPAT and T1D datasets are longitudinal, one can use the methodology proposed by Shi et al. (2016) to estimate a sample based interaction effects. Let X_{jit} be the observed count for the jth OTU of the ith subject at time t. Shi et al. (2016) proposed the following model for simulated data,

$$\begin{aligned} &\Delta X_{jit} = log(X_{jit}) - log(X_{jit-1}), \\ &\tilde{X}_{jit} = \Delta X_{jit} / \Delta\, t, \\ &\tilde{X}_{jit} = \alpha_0 + \sum_{k=1}^{m} \alpha_{jkt} X_{kit} + \epsilon_{jit}, \end{aligned} \tag{10.5}$$

The model for $\tilde{X}_{jit}$ can be estimated using LASSO (Tibshirani, 1996) or Elastic Net models (Zou and Hastie, 2005). The parameter α_{jkt} captures the interaction between the jth and the kth OTUs at time t. Note that, although the model (Shi et al., 2016) was proposed and applied for simulated data (which typically has high number of timepoints in which data are generated) it was never fitted for observational longitudinal data which consists only small number of timepoints in which data are observed.

10.2.3 Longitudinal Analysis of Microbiome Data

The models presented in Chapter 6 - Chapter 8 were fitted at each timepoint separately and therefore did not take into account the longitudinal sequence of the data. Let us consider an intervention experiment in which the microbiome data and the primary outcome were measured over time. The Poisson / Normal path analysis model specified in equation (8.3) can be re written as,

$$\begin{aligned} &X_{jit} \sim \text{Poisson}(\lambda_{it}), \\ &\log(\lambda_{it}) = \mu_{X_{jt}} + \alpha_{jt} Z_i + a_i, \\ &Y_{it} \sim N(\mu_{it}, \tau_t), \\ &\mu_{it} = \mu_{Y_t} + \beta_t Z_i + \gamma_{jt} X_{jit} + b_i. \end{aligned} \tag{10.6}$$

X_{jit} and Y_{it} are the microbiome variable and log(IgA) of the ith subject at time t. Note that all parameters in equation (10.6) are timepoint specific. The same priors specified in Section (8.2.2) can be used for $\mu_{X_{jt}}$, α_{jt}, μ_{Y_t}, β_t and γ_{jt}. The parameters a_i and b_i are subject specific random effects for the microbiome feature and log(IgA), respectively. The two random effects can capture a possible correlation among the longitudinal measurements for both microbiome variable and log(IgA). A bivariate Normal prior can be formulated,

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D} \right), \tag{10.7}$$

with the covariance matrix defined by,

$$\mathbf{D} = \begin{pmatrix} d_{a,a} & d_{a,b} \\ d_{b,a} & d_{b,b} \end{pmatrix}. \tag{10.8}$$

For the precession matrix $\mathbf{D^{-1}}$, we can assume a Wishart hyperprior.

Bibliography

Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N., Shkedy, Z., and Van der Elst, W. (Eds.) (2016). *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. CRC Press.

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, 63, 180–186.

Alonso, A. and Molenberghs, G. (2008). Evaluating time to cancer recurrence as a surrogate marker for survival from an information theory perspective. *Statistical Methods in Medical Research*, 17, 497–504.

Amaratunga, D., Cabrera, J., and Shkedy, Z. (2014). *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data*. Wiley Series in Probability and Statistics.

Azimfar, S. J. (2015). Efficiency of parallel processing in multi-core processors. *Cumhuriyet Science Journal*, 36, 2535–2543. Special Issue: Technological Advances of Engineering Sciences.

Bäckhed, F., Fraser, C., Ringel, Y., Sanders, M., Sartor, R., Sherman, P., Versalovic, J., Young, V., and Finlay, B. (2012). Defining a healthy human gut microbiome: Current concepts, future directions, and clinical applications. *Cell Host & Microbe*, 12, 611–622.

Bai, J. P. F., Alekseyenko, A. V., Statnikov, A., Wang, I., and Wong, P. H. (2013). Strategic applications of gene expression: from drug discovery/development to bedside. *The AAPS Journal*, 15, 427–437.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Bex, G. J. Worker framework. Manual. `https://www.vscentrum.be/cluster-doc/running-jobs/worker-framework`.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69, 89–95.

Bruce, E., Autenrieth, R., Burghardt, R., Donnelly, K., and McDonald, T. (2008). Using quantitative structure-activity relationships (qsar) to predict toxic endpoints for polycyclic aromatic hydrocarbons (pah). *Journal of Toxicology and Environmental Health. Part A.*, 71, 1073–1084.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004). The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167, 103–124.

Burzykowski, T., Molenberghs, G., and Buyse, M. (Eds.) (2005). *The Evaluation of Surrogate Endpoints*. Springer-Verlag New York.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics*, 54, 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1, 49–67.

Calaway, R., Microsoft, and Weston, S. (2015). *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.3. `https://cran.r-project.org/web/packages/foreach/foreach.pdf`.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265–270.

Congdon, P. (2014). *Applied Bayesian Modelling*. John Wiley & Sons, Ltd.

Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51, 107–113.

Dellaportas, P. and Smith, A. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 42, 443–459.

Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S., Nelson, K., and Relman, D. (2005). Diversity of the human intestinal microbial flora. *Science*, 308, 1635–1638.

Gasparrini, A., Crofts, T., Gibson, M., Tarr, P., Warner, B., and Dantas, G. (2016). Antibiotic perturbation of the preterm infant gut microbiome and resistome. *Gut Microbes*, 7, 443–449.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman & Hall/ CRC, second edition.

Göhlmann, H. and Talloen, W. (2009). *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & Hall/CRC Mathematical & Computational Biology.

Grice, E. and Segre, J. (2012). The human microbiome: Our second genome. *Annual Review of Genomics and Human Genetics*, 13, 151–170.

Hallstrom, A. P. (2010). A modified wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine*, 29, 391–400.

Jakobsson, H., Jernberg, C., Andersson, A., Sjolund-Karlsson, M., Jansson, J., and Engstrand, L. (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE*, 5, 1–12.

John, G. and Mullin, G. (2016). The gut microbiome and obesity. *Current Oncology Reports*, 18, 1–7.

Johnson, E. L., Heaver, S. L., Walters, W. A., and Ley, R. E. (2017). Microbiome and metabolic disease: revisiting the bacterial phylum bacteroidetes. *Journal of Molecular Medicine*, 95, 1–8.

Johnston, J. (1984). *Econometric Methods*. McGraw-Hill Companies.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer.

Kostic, A., Xavier, R., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology*, 146, 1489–1499.

Lachenbruch, P. (2001). Comparisons of two-part models with competitors. *Statistics in Medicine*, 20, 1215–1234.

Lachenbruch, P. A. (1976). Analysis of data with clumping at zero. *Biometrische Zeitschrift*.

Levy, M., Blacher, E., and Elinav, E. (2017). Microbiome, metabolites and host immunity. *Current Opinion in Microbiology*, 35, 8–15.

Lin, D., Shkedy, Z., Molenberghs, G., Talloen, W., Gohlmann, H., and Bijnens, L. (2010). Selection and evaluation of gene-specific biomarkers in pre-clinical and clinical microarray experiments. *Online Jounal of Bioinformatics*, 11, 106–127.

Llorente, C. and Schnabl, B. (2015). The gut microbiota and liver disease. *Cellular and Molecular Gastroenterology and Hepatology*, 1, 275–284.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18, 50–60.

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, 45, 4350–4358.

Matsen, F. (2015). Phylogenetics and the human microbiome. *Systematic Biology*, 64, e26–e41.

McMurdie, P. and Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, 10, 1–12.

Morgan, X. and Huttenhower, C. (2012). Chapter 12: Human microbiome analysis. *PLOS Computational Biology*, 8, 1–14.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., and Prachayasittikul, V. (2009). A practical overview of quantitative structure-activity relationship. *EXCLI J.*, 8, 74–78.

Neuhäuser, M., Boes, T., and Jöckel, K. (2005). Two-part permutation tests for dna methylation and microarray data. *BMC Bioinformatics*, 6, 35.

Parekh, P., Balart, L., and Johnson, D. (2015). The influence of the gut microbiome on obesity, metabolic syndrome and gastrointestinal disease. *Clinical and Translational Gastroenterology*, 6, e91–e102.

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., Martinez, X., Varela, E., Sarrabayrouse, G., Machiels, K., Vermeire, S., Sokol, H., Guarner, F., and Manichanh, C. (2017). A microbial signature for crohn's disease. *Gut*, 66, 813–822.

Perualila, N. J., Shkedy, Z., Sengupta, R., Bigirumurame, T., Bijnens, L., Talloen, W., Verbist, B., Göhlmann, H. W., Kasim, A., and QSTAR Consortium (2016b). High-dimensional biomarkers in drug discovery: the qstar framework. In Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila,

N. J., Shkedy, Z., and Van der Elst, W. (Eds.), *Applied Surrogate Endpoint Evaluation Methods with SAS and R*, pages 275–309. CRC Press.

Perualila-Tan, N. J., Kasim, A., Talloen, W., Verbist, B., Göhlmann, H. W., QSTAR Consortium, and Shkedy, Z. (2016a). A joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. *Statistical applications in genetics and molecular biology*, 15, 291–304.

Pindyck, R. and Rubinfeld, D. (1976). *Econometric Models and Economic Forescasts*. Mcgraw-Hill Book Company, second edition.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8, 431–440.

R Core Team (2016). R: A language and environment for statistical computing. `https://www.r-project.org/`.

Ranjan, R., Rani, A., Metwally, A., McGee, H., and Perkins, D. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469, 967–977.

Ravindranath, A. C., Perualila-Tan, N., Kasim, A., Drakakis, G., Liggi, S., Brewerton, S. C., Mason, D., Bodkin, M. J., Evans, D. A., Bhagwat, A., Talloen, W., Gohlmann, H. W. H., Consortium, Q., Shkedy, Z., and Bender, A. (2015). Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Mol. BioSyst.*, 11, 86–96.

Revolution Analytics (2015). rmr2 package. A package that allows R developer to use Hadoop MapReduce. `https://github.com/RevolutionAnalytics/rmr2`.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50, 742–754.

Rooks, M. and Garrett, W. (2016). Gut microbiota, metabolites and host immunity. *Nature Reviews Immunology*, 16, 341–352.

Round, J. and Mazmanian, S. (2009). The gut microbiome shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9, 313–323.

Ruiz, E., Battaglia, T., Kurtz, D., Bijnens, L., Ou, A., Engstrand, I., Zheng, X., Iizumi, T., Mullins, J., L. Christian, M., Cadwell, K., Bonneau, R., Perez-Perez, I. Guillermo, and

Blaser, J. (2017). A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity. *Nature Communications*, 8, 518–531.

Sahni, S. and Vairaktarakis, G. (1996). The master-slave paradigm in parallel computer and industrial settings. *Journal of Global Optimization*, 9, 357–377.

Sanz, Y., Olivares, M., Angela, M.-P., and Agostoni, C. (2015). Understanding the role of gut microbiome in metabolic disease risk. *Pediatric Research*, 77, 236–244.

Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., and Mansmann, U. (2009). State of the art in parallel computing with r. *Journal of Statistical Software*, 31, 1–27.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, 14, 1–14.

Sengupta, R., Molenberghs, G., Alonso, A., Van der Elst, W., and Shkedy, Z. (2018). Single, multiple and partial surrogacy - a joint modeling approach. (In Preparation).

Sengupta, R. and Perualila, N. J. (2017). *IntegratedJM: Joint Modeling of the Gene-Expression and Bioassay Data, Taking Care of the Effect Due to a Fingerprint Feature.* R package version 1.6. `https://cran.r-project.org/web/packages/IntegratedJM/IntegratedJM.pdf`.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10, 1019–1040.

Shkedy, Z. and Barbosa, F. T. (2005). Bayesian evaluation of surrogate endpoints. In Burzykowski, T., Molenberghs, G., and Buyse, M. (Eds.), *The Evaluation of Surrogate Endpoints*, pages 253–270. Springer New York.

Simpson, E. (1949). Measurement of diversity. *Nature*, 163, 688.

Software Tools for Academics and Researchers Group. Starcluster. `http://star.mit.edu/cluster/docs/latest/manual/index.html`.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, 76, 485–493.

Steve, W. (2012). *Using the foreach package*. `https://cran.r-project.org/web/packages/foreach/vignettes/foreach.pdf`.

Tedjo, D., Smolinska, A., Savelkoul, P., Masclee, A., van Schooten, F., Pierik, M., and Penders, J. (2016). The fecal microbiota as a biomarker for disease activity in crohnŠs disease. *Scientific Reports*, 6, 35216–35225.

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486, 207–214.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Tilahun, A., Lin, D., Shkedy, Z., Geys, H., Alonso, A., Peeters, P., Talloen, W., Drinkenburg, W., Gohlmann, H., Gorden, E., Bijnens, L., and Molenberghs, G. (2010). Genomic biomarkers for depression: Feature-specific and joint biomarkers. *Statistics in Biopharmaceutical Research*, 2, 419–434.

Todeschini, R. and Consonni, V. (2009). In Mannhold, R., Kubinyi, H., and Folkers, G. (Eds.), *Molecular Descriptors for Chemoinformatics*. Wiley.

Turnbaugh, P., Ley, R., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449, 804–810.

Valencia, P., Richard, M., Brock, J., and Boglioli, E. (2017). The human microbiome: opportunity or hype? *Nature Reviews Drug Discovery*, 16, 823–824.

Van der Elst, W., Alonso, A., Geys, H., Meyvisch, P., Bijnens, L., , and Molenberghs, G. (2018). Univariate versus multivariate surrogate endpoints. (In Preparation).

Van Sanden, S., Shkedy, Z., Burzykowski, T., Göhlmann, H. W., Talloen, W., and Bijnens, L. (2012). Genomic biomarkers for a binary clinical outcome in early drug development microarray experiments. *Journal of Biopharmaceutical Statistics*, 22, 72–92.

Vera, G., Jansen, R., and Suppi, R. (2008). R/parallel Ű speeding up bioinformatics analysis with r. *BMC Bioinformatics*, 9, 390.

Verbist, B., Klambauer, G., Vervoort, L., Talloen, W., Consortium, Q., Shkedy, Z., Thas, O., Bender, A., Göhlmann, H., and Hochreiter, S. (2015). Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the qstar project. *Drug Discovery Today*, 20, 505–513.

Wagner, B., Robertson, C., and Harris, J. (2011). Application of two-part statistics for comparison of sequence variant counts. *PLOS ONE*, 6, 1–8.

Wang, B., Yao, M., Lv, L., Ling, Z., and Li, L. (2017). The human microbiota in health and disease. *Engineering*, 3, 71–82.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 80–83.

Wild, C. J. and Seber, G. A. F. (2000). *Chance Encounters: A First Course in Data Analysis and Inference*. John Wiley & Sons, New York.

Young, V. (2017). The role of the microbiome in human health and disease: an introduction for clinicians. *BMJ*, 356.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Appendix A

Single, Multiple, Partial and Orthogonal Surrogacy: A Joint Modeling Approach

This supplementary appendix contains additional materials related to Chapter 4.

A.1 Subclasses of Genes

Based on the results of hypotheses, specified in equation (2.4) and (2.5), genes can be classified into four different categories. Table SA1 presents these four categories with hypothetical examples. Under scenario (a) and (b) both α_j and β, are significantly different from zero. In scenario (a) the correlation between the gene expression and pIC_{50} is present but under scenario (b) the correlation between the expression levels and the bioactivity variable is due to the effect of the fingerprint feature, hence its adjusted association is zero, $\rho_j = 0$. From the point of view of the structural optimization in the early drug development, the association pattern observed in (b) is desirable while the one observed in (a) is an ideal predictive biomarker for pIC_{50}.

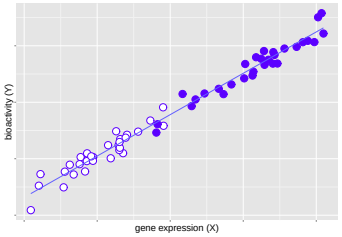
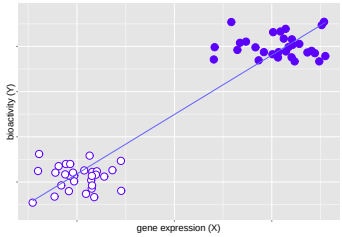
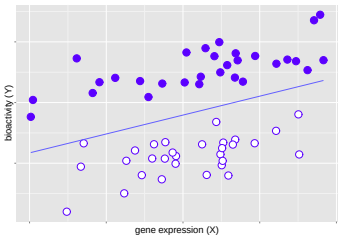
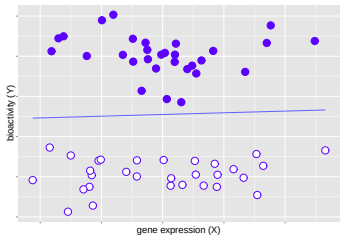
	$\rho_j \neq 0$	$\rho_j = 0$
$\beta \neq 0$ & $\alpha_j \neq 0$	 X and Y are correlated, the gene is differentially expressed.	 X and Y are correlated but are conditionally independent.
$\beta \neq 0$ & $\alpha_j = 0$	 X and Y are correlated, the gene is not differentially expressed.	 X and Y are uncorrelated.

Table SA1: Association patterns between gene expression and bioactivity: a hypothetical example.

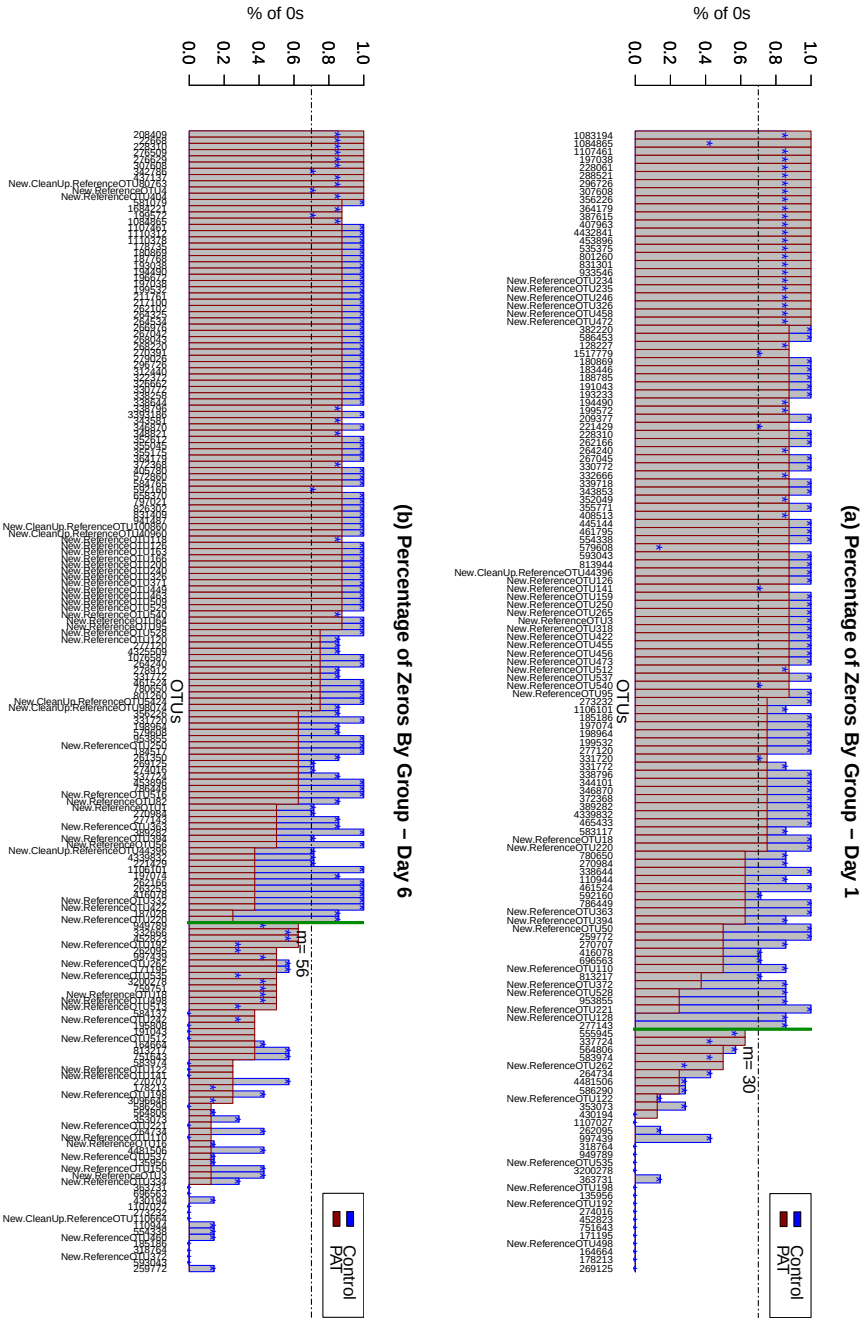
Appendix B

Microbiome Intervention Studies: An Introduction

This supplementary appendix contains additional materials related to Chapter 5.

B.1 OTU Filtering

As mentioned in Chapter 5, initial filtering using a 70% thresholding was conducted per timepoint. This implies that OTUs with more than 70% zero counts either in one or both treatment groups were not included in the analysis. Figure SB1 presents the proportion of zero counts, per treatment group, for all the OTUs, per timepoint.


% of 0s
0.0
0.2
0.4
0.6
0.8
1.0
OTUs
m= 56
Control
PAT
(b) Percentage of Zeros By Group – Day 6
% of 0s
0.0
0.2
0.4
0.6
0.8
1.0
OTUs
m= 30
Control
PAT
(a) Percentage of Zeros By Group – Day 1

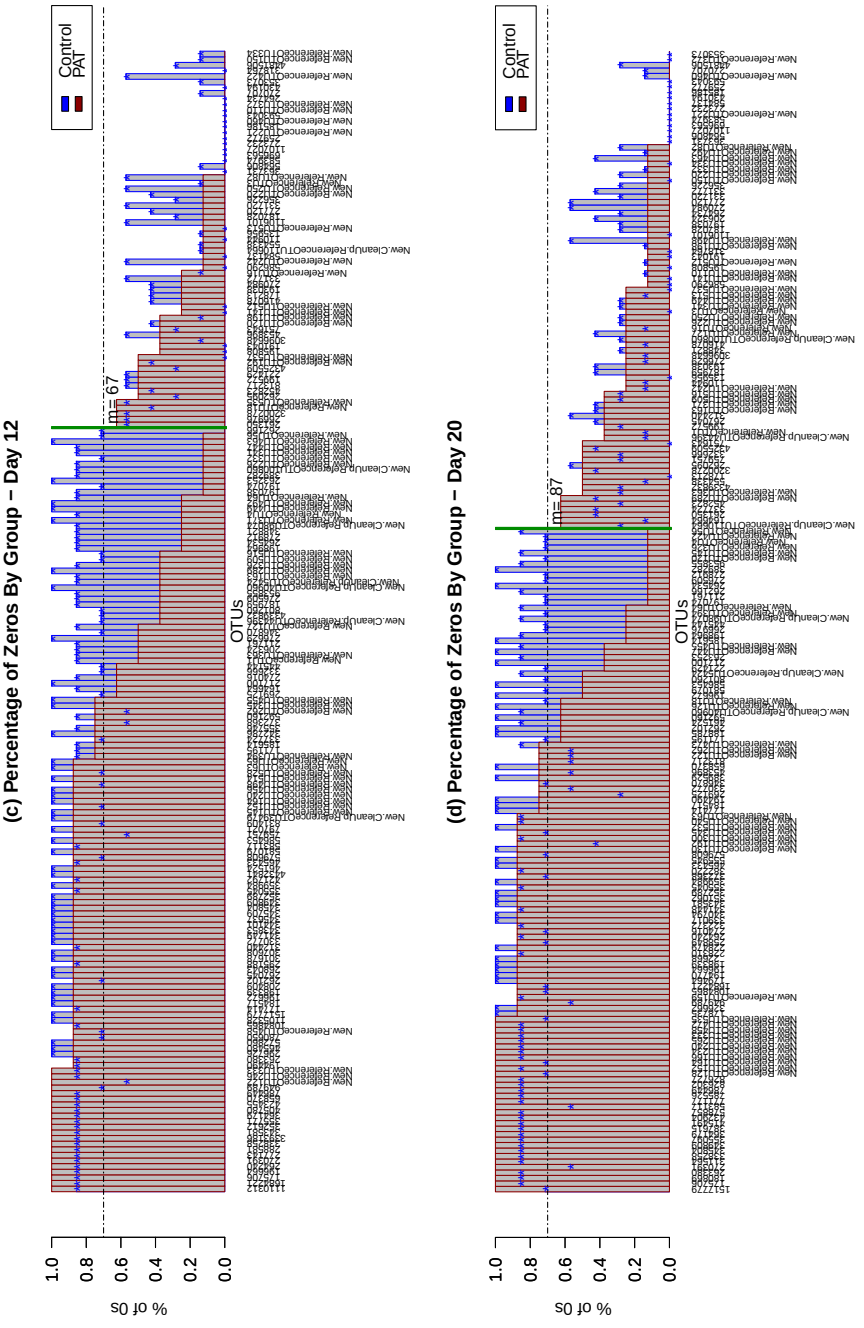


Figure SB1: Proportion of zero counts present in each group using the 70% threshold (horizontal line). All OTUs to the left of the green vertical line were not included in the joint model analysis. Panel a: day 1. Panel b: day 6. Panel c: day 12. Panel d: day 20.

Appendix C

Development of Microbiome Biomarkers for IgA: A Joint Modeling Approach

This supplementary appendix contains additional materials related to Chapter 6.

C.1 Family Level Richness

Figure SC1 presents the richness of all the families over time. As discussed in Chapter 6, many OTUs from the *Lachnospiraceae* family were found to be differentially abundant at different timepoints. Figure SC2 and Figure SC3 show that it is indeed one of the active families over time.

C.2 Application to the TransPAT Data

In this section, we present the results when different levels of the hierarchical microbiome ecosystem is analyzed.

C.2.1 Analysis of α-Diversity

Results for observed and the Chao1 measure of α-diversity is presented in Tables SC1 and SC2, respectively.

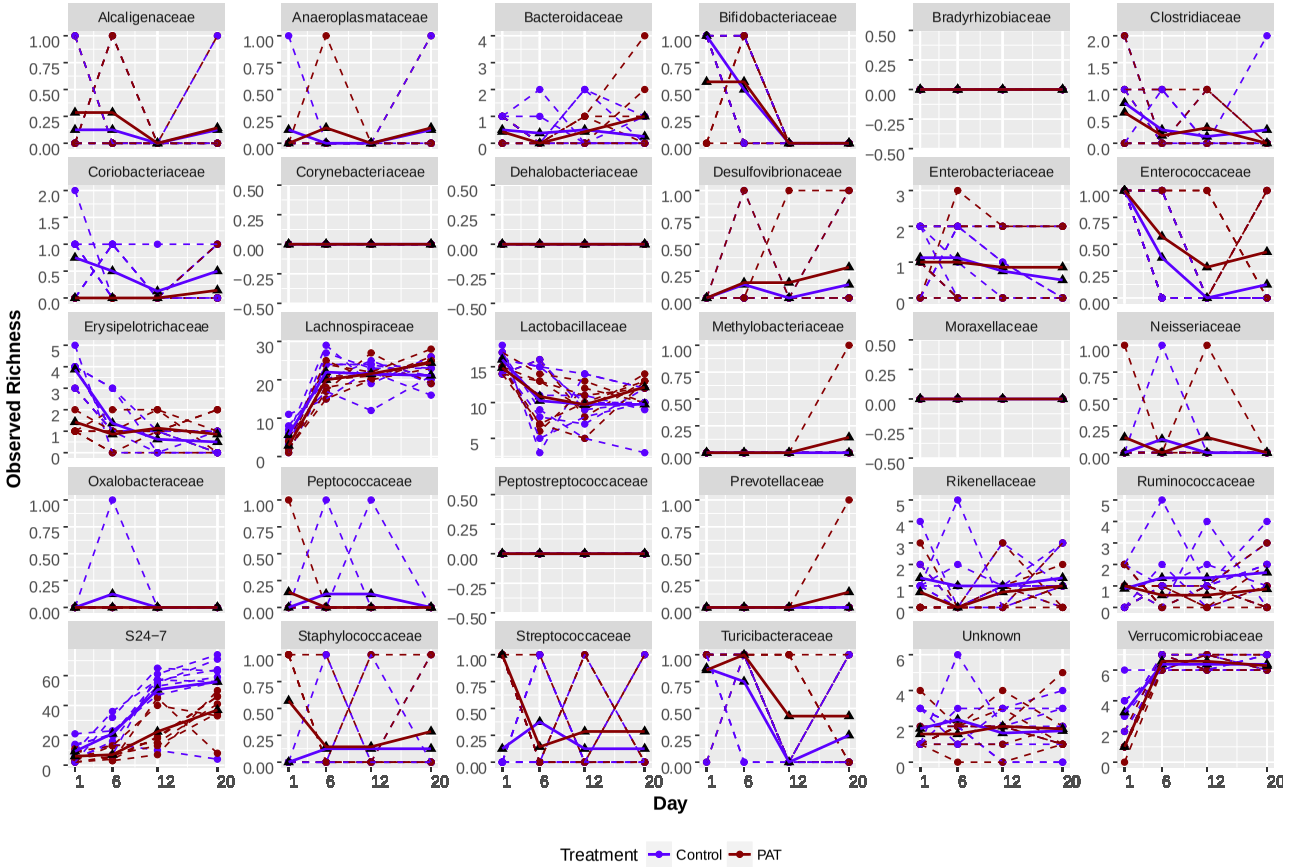


Figure SC1: Observed richness for all the families over time with dashed and solid lines representing subject mean profiles, respectively.

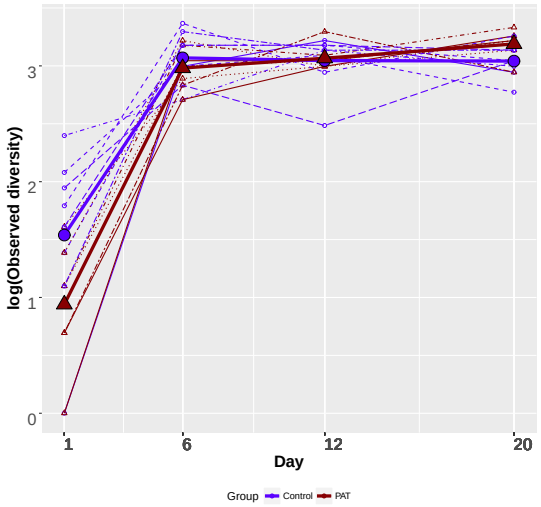


Figure SC2: Individual and mean family level richness profiles over time for the *Lachnospiraceae* family with dashed and solid lines representing subject profiles and mean profiles, respectively.

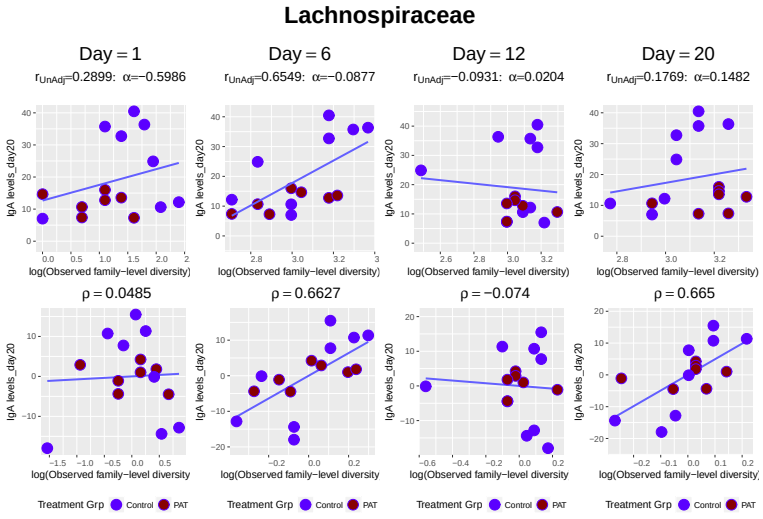


Figure SC3: Observed richness (on log scale) of *Lachnospiraceae* family against IgA level over time. Upper panels: scatterplot of the raw data. Lower panels: scatterplot of the residuals, after adjusting for treatment.

Observed α-Diversity						
Day	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
1	-0.29	0.03	0.05	-0.40	0.11	0.15
6	-0.31	0.01	0.03	0.55	0.02	0.04
12	-0.35	0.02	0.04	0.33	0.19	0.19
20	-0.11	0.47	0.47	0.66	0.00	0.01

Table SC1: Joint model results for observed α-diversity and IgA (FDR $= 0.10$).

Chao1 Measure for α-Diversity						
Day	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
1	-0.42	0.09	0.18	-0.34	0.18	0.24
6	-0.23	0.15	0.20	0.51	0.03	0.07
12	-0.31	0.05	0.18	-0.01	0.98	0.98
20	0.03	0.82	0.82	0.68	0.00	0.01

Table SC2: Joint Model Results for log(Chao1) measure of α-diversity and IgA (FDR $=$ 0.10).

C.2.2 Analysis at Family Level

This section presents the results when Chao1 measure (Table SC3) and Shannon Index (Table SC4) for family level richness was used as the microbiome measure for the joint model, specified in equation (6.1).

C.2.3 Analysis at OTU Level

Table SC5 and Table SC6 present detailed results for the analysis at OTU level discussed in Section 6.3.3.

			Day 1			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Verrucomicrobiaceae	-0.97	0.00	0.00	-0.32	0.21	0.67
Erysipelotrichaceae	-0.99	0.00	0.00	-0.22	0.38	0.67
Bifidobacteriaceae	-0.00	0.03	0.10	-0.04	0.89	0.89
Lachnospiraceae	-0.82	0.08	0.21	0.07	0.79	0.87
Lactobacillaceae	-0.11	0.13	0.26	-0.26	0.31	0.67
S24-7	-0.62	0.19	0.32	-0.28	0.28	0.67
Enterobacteriaceae	-0.25	0.28	0.40	0.43	0.08	0.67
Unknown	-0.19	0.59	0.74	-0.18	0.49	0.70
Ruminococcaceae	-0.10	0.74	0.82	-0.15	0.56	0.70
Turicibacteraceae	-0.00	0.93	0.93	-0.22	0.40	0.67
			Day 6			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-0.95	0.01	0.15	0.10	0.70	0.78
Turicibacteraceae	0.00	0.17	0.60	0.61	0.01	0.04
Ruminococcaceae	-0.35	0.23	0.60	0.16	0.54	0.77
Lachnospiraceae	-0.18	0.27	0.60	0.46	0.06	0.19
Unknown	-0.27	0.34	0.60	0.27	0.30	0.59
Lactobacillaceae	0.23	0.42	0.60	0.62	0.01	0.04
Enterococcaceae	0.00	0.48	0.60	0.22	0.38	0.64
Verrucomicrobiaceae	0.03	0.48	0.60	0.00	0.99	0.99
Erysipelotrichaceae	-0.12	0.63	0.70	-0.10	0.70	0.78
Bifidobacteriaceae	0.00	0.80	0.80	0.38	0.13	0.31
			Day 12			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-0.50	0.08	0.36	0.40	0.10	0.27
Erysipelotrichaceae	0.26	0.12	0.36	0.01	0.97	0.97
Verrucomicrobiaceae	0.03	0.48	0.62	-0.16	0.54	0.65
Unknown	0.18	0.53	0.62	0.56	0.02	0.10
Lachnospiraceae	-0.08	0.56	0.62	-0.34	0.18	0.27
Lactobacillaceae	0.09	0.62	0.62	0.34	0.17	0.27
			Day 20			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Lachnospiraceae	0.21	0.01	0.07	0.56	0.02	0.06
Erysipelotrichaceae	0.20	0.11	0.43	0.11	0.67	0.68
Lactobacillaceae	0.20	0.32	0.85	0.31	0.21	0.34
Ruminococcaceae	-0.13	0.69	0.92	0.54	0.02	0.06
Verrucomicrobiaceae	-0.01	0.73	0.92	-0.11	0.68	0.68
S24-7	-0.09	0.84	0.92	0.52	0.03	0.06
Unknown	-0.07	0.84	0.92	0.61	0.01	0.06
Rikenellaceae	-0.03	0.92	0.92	0.20	0.44	0.58

Table SC3: Parameter estimates for all the families at different timepoints with respect to log(Chao1) measure for observed diversity. Results are sorted according to the adjusted p-values for the treatment effect α.

			Day 1			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Lactobacillaceae	-0.04	0.56	0.67	-0.59	0.01	0.03
Lachnospiraceae	-0.14	0.60	0.67	0.44	0.07	0.10
S24-7	-0.14	0.67	0.67	-0.20	0.44	0.44
			Day 6			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
Unknown	-0.19	0.16	0.56	-0.16	0.52	0.53
S24-7	-0.35	0.19	0.56	0.42	0.09	0.53
Lactobacillaceae	0.05	0.29	0.59	0.32	0.20	0.53
Verrucomicrobiaceae	0.01	0.71	0.94	-0.18	0.49	0.53
Enterobacteriaceae	-0.04	0.78	0.94	-0.21	0.42	0.53
Lachnospiraceae	0.01	0.96	0.96	0.16	0.53	0.53
			Day 12			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-1.11	0.00	0.00	-0.44	0.07	0.18
Lactobacillaceae	0.04	0.20	0.50	0.09	0.74	0.88
Unknown	-0.16	0.30	0.51	0.57	0.01	0.07
Verrucomicrobiaceae	0.00	0.53	0.66	0.37	0.14	0.23
Lachnospiraceae	-0.01	0.94	0.94	-0.04	0.88	0.88
			Day 20			
Families	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)
S24-7	-0.64	0.01	0.02	0.29	0.26	0.54
Lachnospiraceae	0.27	0.04	0.08	0.02	0.94	0.94
Lactobacillaceae	0.02	0.69	0.87	0.21	0.40	0.54
Verrucomicrobiaceae	-0.00	0.87	0.87	0.22	0.39	0.54

Table SC4: Parameter estimates for all the families at different timepoints with respect to the Shannon index for observed diversity. Results are sorted according to the adjusted p-values for the treatment effect α.

			Day 1				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
262095	-0.16	0.00	0.00	0.15	0.56	0.66	Erysipelotrichaceae
New.ReferenceOTU192	0.01	0.00	0.00	0.03	0.92	0.95	Lactobacillaceae
363731	-0.01	0.00	0.01	-0.37	0.14	0.28	Verrucomicrobiaceae
1107027	0.24	0.00	0.01	0.62	0.01	0.11	Lactobacillaceae
164664	0.00	0.00	0.01	0.62	0.01	0.11	Lactobacillaceae
New.ReferenceOTU498	0.00	0.01	0.03	0.34	0.18	0.33	Lactobacillaceae
135956	0.00	0.01	0.04	0.53	0.03	0.16	Lactobacillaceae
452823	0.00	0.01	0.04	0.57	0.01	0.15	Lactobacillaceae
171195	0.00	0.01	0.05	0.17	0.49	0.62	Lactobacillaceae
New.ReferenceOTU198	0.00	0.02	0.07	0.19	0.47	0.62	Lactobacillaceae
			Day 6				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
New.ReferenceOTU16	0.00	0.00	0.24	0.19	0.45	0.67	
363731	0.22	0.01	0.24	0.17	0.50	0.68	Verrucomicrobiaceae
593043	0.00	0.01	0.24	0.24	0.35	0.57	Verrucomicrobiaceae
3096648	0.00	0.04	0.46	0.27	0.29	0.50	Lachnospiraceae
New.ReferenceOTU537	0.00	0.05	0.46	0.30	0.23	0.49	Lachnospiraceae
430194	-0.15	0.05	0.46	-0.27	0.28	0.50	S24-7
997439	0.00	0.06	0.46	0.10	0.70	0.82	Bifidobacteriaceae
264734	-0.06	0.07	0.48	-0.00	1.00	1.00	S24-7
New.ReferenceOTU535	0.00	0.09	0.48	0.07	0.77	0.85	
185186	0.00	0.09	0.48	0.32	0.20	0.49	Verrucomicrobiaceae
			Day 12				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
New.ReferenceOTU220	-0.01	0.00	0.00	-0.06	0.81	0.86	S24-7
New.ReferenceOTU422	-0.00	0.00	0.00	-0.10	0.70	0.81	S24-7
264734	-0.06	0.00	0.02	-0.17	0.50	0.67	S24-7
New.ReferenceOTU250	-0.01	0.00	0.04	0.66	0.00	0.02	S24-7
3096648	0.00	0.00	0.06	0.08	0.76	0.84	Lachnospiraceae
New.ReferenceOTU537	0.00	0.01	0.08	0.05	0.83	0.87	Lachnospiraceae
1106101	-0.01	0.01	0.11	0.35	0.16	0.29	S24-7
New.ReferenceOTU82	-0.00	0.01	0.13	0.21	0.42	0.60	S24-7
New.ReferenceOTU16	0.00	0.03	0.17	0.01	0.98	0.98	
259772	0.00	0.03	0.17	0.09	0.72	0.81	Lachnospiraceae
			Day 20				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
New.ReferenceOTU332	-0.00	0.00	0.03	0.87	0.00	0.00	S24-7
356226	-0.01	0.00	0.07	0.83	0.00	0.00	S24-7
New.ReferenceOTU82	-0.01	0.00	0.07	0.89	0.00	0.00	S24-7
New.ReferenceOTU492	-0.00	0.01	0.11	0.88	0.00	0.00	S24-7
331772	-0.01	0.01	0.14	0.62	0.01	0.03	S24-7
353073	-0.13	0.01	0.15	0.73	0.00	0.01	S24-7
259772	0.00	0.01	0.15	-0.10	0.71	0.80	Lachnospiraceae
331720	-0.01	0.02	0.15	0.71	0.00	0.01	S24-7
178213	0.00	0.02	0.15	-0.14	0.59	0.74	Lactobacillaceae
206324	-0.00	0.02	0.15	0.89	0.00	0.00	S24-7

Table SC5: Parameter estimates for top 10 differentially abundant OTUs (FDR = 0.10) at different timepoints. The results are sorted according to the adjusted p-values for the treatment effect α.

			Day 1				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
1107027	0.24	0.00	0.01	0.62	0.01	0.11	Lactobacillaceae
164664	0.00	0.00	0.01	0.62	0.01	0.11	Lactobacillaceae
452823	0.00	0.01	0.04	0.57	0.01	0.15	Lactobacillaceae
135956	0.00	0.01	0.04	0.53	0.03	0.16	Lactobacillaceae
New.ReferenceOTU122	-0.03	0.39	0.45	-0.53	0.03	0.16	Turicibacteraceae
353073	-0.00	0.13	0.27	-0.47	0.05	0.23	S24-7
274016	-0.00	0.45	0.49	-0.47	0.05	0.23	Lactobacillaceae
997439	-0.03	0.03	0.09	-0.42	0.09	0.25	Bifidobacteriaceae
178213	0.00	0.07	0.17	0.44	0.07	0.25	Lactobacillaceae
949789	0.01	0.08	0.18	-0.42	0.09	0.25	Enterococcaceae
			Day 6				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
586290	-0.00	0.89	0.97	0.71	0.00	0.06	Lachnospiraceae
New.ReferenceOTU141	-0.00	0.94	0.97	0.68	0.00	0.06	Lachnospiraceae
583974	-0.00	0.63	0.88	0.65	0.00	0.08	Lachnospiraceae
New.ReferenceOTU3	0.00	0.69	0.90	0.60	0.01	0.11	Lachnospiraceae
191043	-0.00	0.93	0.97	0.60	0.01	0.11	Lachnospiraceae
New.ReferenceOTU512	0.00	0.75	0.90	0.57	0.01	0.14	Lactobacillaceae
584137	0.00	0.10	0.48	0.56	0.02	0.15	Lachnospiraceae
New.ReferenceOTU513	-0.00	0.37	0.80	0.51	0.03	0.24	Lachnospiraceae
564806	-0.02	0.53	0.86	-0.48	0.05	0.29	Lachnospiraceae
195808	-0.00	0.75	0.90	0.45	0.07	0.38	Lachnospiraceae
			Day 12				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
221429	-0.01	0.03	0.17	0.87	0.00	0.00	Ruminococcaceae
193038	-0.00	0.03	0.17	0.84	0.00	0.00	S24-7
416078	-0.00	0.04	0.17	0.79	0.00	0.00	S24-7
751643	-0.00	0.57	0.76	0.76	0.00	0.00	Lactobacillaceae
New.ReferenceOTU3	-0.00	0.89	0.92	0.76	0.00	0.00	Lachnospiraceae
318764	-0.00	0.29	0.59	0.69	0.00	0.02	Lactobacillaceae
185186	0.00	0.39	0.66	-0.68	0.00	0.02	Verrucomicrobiaceae
New.ReferenceOTU513	0.00	0.78	0.87	0.68	0.00	0.02	Lachnospiraceae
New.ReferenceOTU250	-0.01	0.00	0.04	0.66	0.00	0.02	S24-7
New.ReferenceOTU198	0.00	0.29	0.59	0.67	0.00	0.02	Lactobacillaceae
			Day 20				
OTUs	α	$p(\alpha)$	adj-p(α)	ρ	$p(\rho)$	adj-p(ρ)	Family
276629	-0.00	0.13	0.34	0.91	0.00	0.00	S24-7
New.ReferenceOTU82	-0.01	0.00	0.07	0.89	0.00	0.00	S24-7
206324	-0.00	0.02	0.15	0.89	0.00	0.00	S24-7
New.ReferenceOTU492	-0.00	0.01	0.11	0.88	0.00	0.00	S24-7
New.ReferenceOTU332	-0.00	0.00	0.03	0.87	0.00	0.00	S24-7
New.ReferenceOTU513	-0.00	0.39	0.69	0.85	0.00	0.00	Lachnospiraceae
356226	-0.01	0.00	0.07	0.83	0.00	0.00	S24-7
New.ReferenceOTU127	-0.00	0.02	0.15	0.75	0.00	0.01	S24-7
277120	-0.00	0.02	0.15	0.74	0.00	0.01	S24-7
353073	-0.13	0.01	0.15	0.73	0.00	0.01	S24-7

Table SC6: Parameter estimates for top 10 associated OTUs (FDR = 0.10) with IgA at different timepoints. The results are sorted according to the adjusted p-values for the adjusted association ρ.

Appendix D

Development of High Dimensional Microbiome Biomarkers: A Non Parametric Approach

This supplementary appendix contains additional materials related to Chapter 7.

D.1 Microbiome Composition

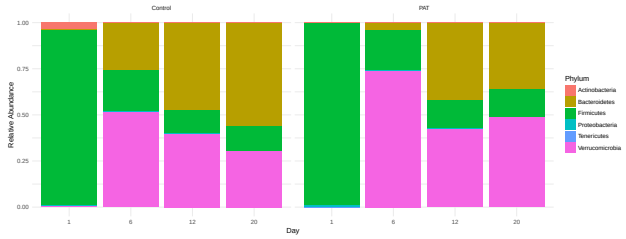
Figure SD1 displays the treatment groupwise composition of the micromiome data at different levels of the data across different timepoints and it is evident that at OTU level the two treatment groups have different composition. At a family level, the *S24-7* family is found to be less active at the beginning and became more abundant at later timepoints while an opposite pattern is observed for the *Lactobacillaceae* family.

D.2 Filtering OTUs

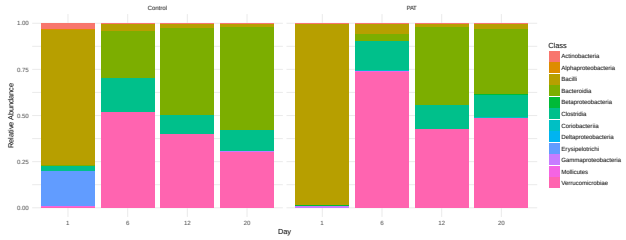
As mentioned before, the analysis is done after filtering out OTUs with high proportion of zeros. As mentioned in Appendix B, OTUs with zeros in at least 70% of the samples per treatment group were filtered out per timepoint. Figure SD2 displays the non-zero proportion of samples for all the OTUs, together with their corresponding families, across

different timepoints. The *S24-7* family and the *Lachnospiraceae* become more dominant at later timepoints. Hence, OTUs corresponding to these families were not included in the analysis for the earlier timepoints, but they were included at later stages. After filtering, 30, 56, 67 and 87 OTUs were included (Figure SD2) in the analysis at day 1, 6, 12 and 20, respectively. The families of these OTUs are displayed in Figure SD3 for each timepoint. Figure SD4 displays the treatment groupwise proportion of zeros for the filtered out OTUs across all the timepoints and it is worth noting that all of them have high proportion of zeros in the PAT group.

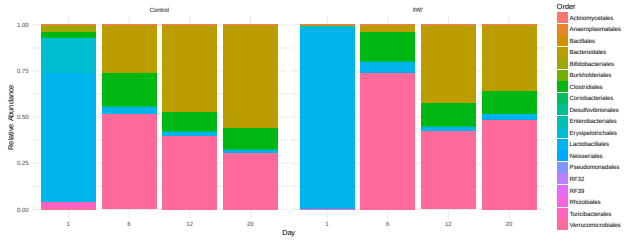
(a) Phylum



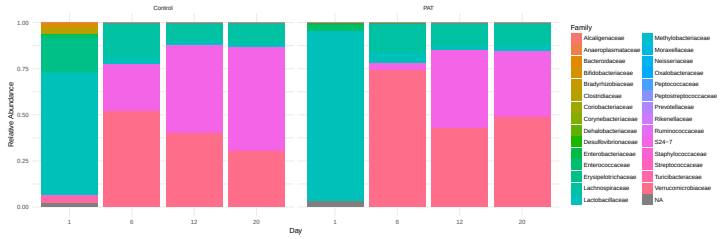
(b) Class



(c) Order



(d) Family



(e) OTU

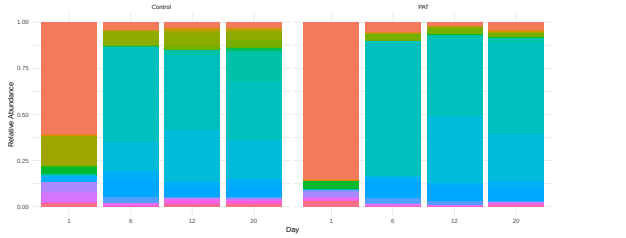


Figure SD1: The microbiome composition based on the relative abundance by phylogentic level of each treatment group for the first 4 days.

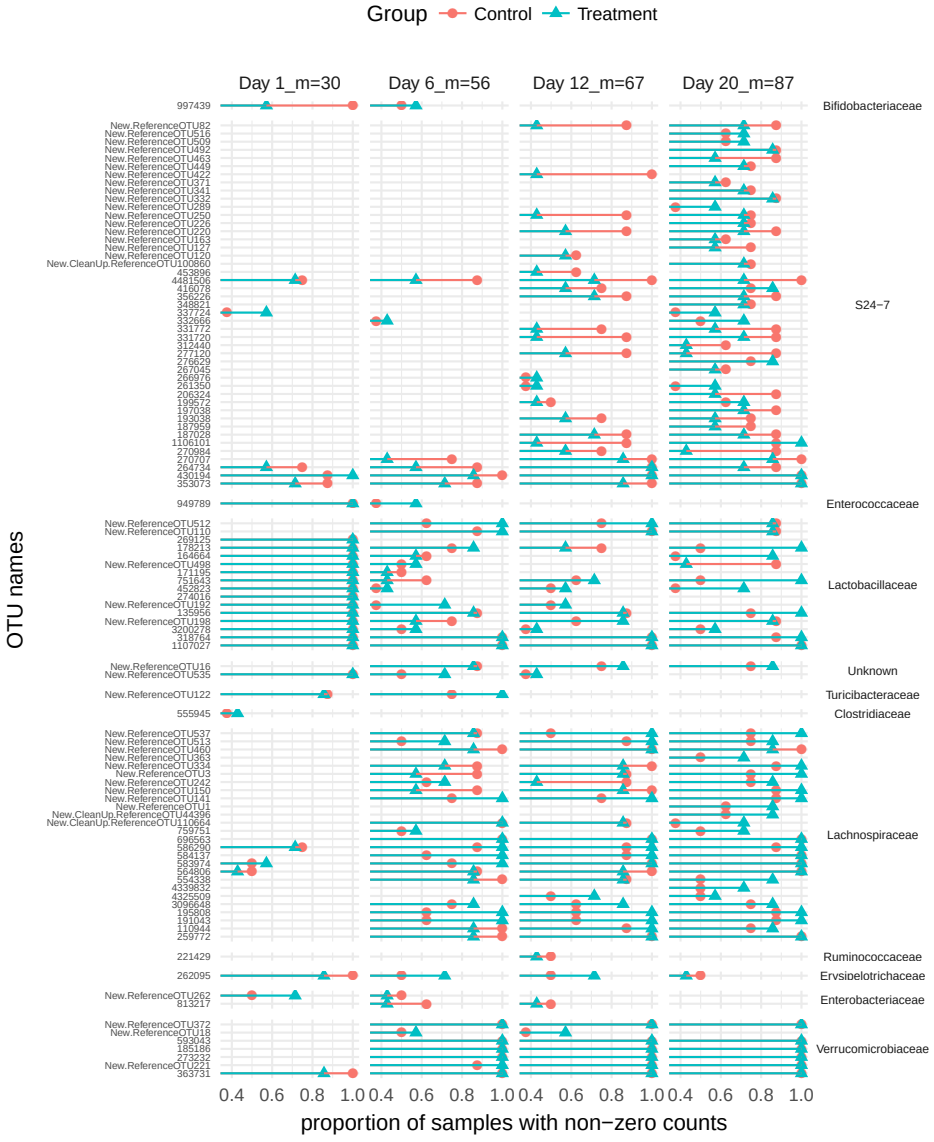


Figure SD2: OTUs included in the analysis. Both treatment groups have proportion of zeros of at most 70% per OTU.

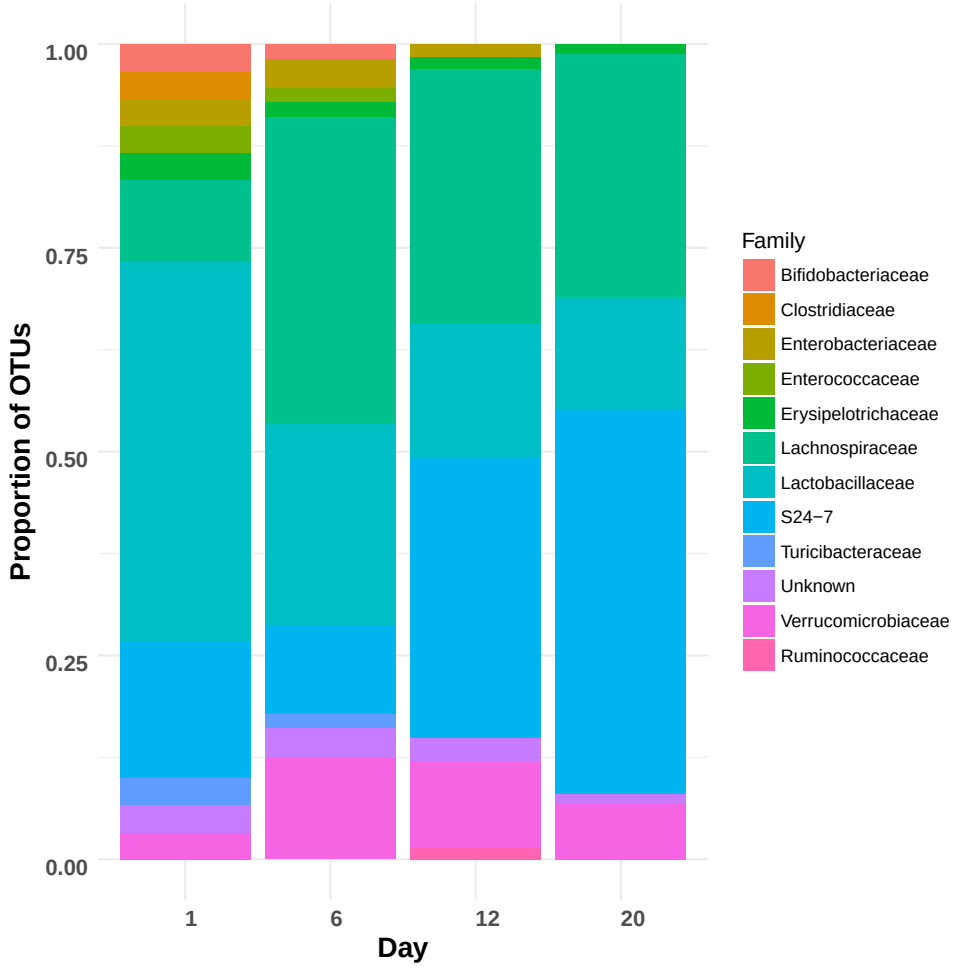


Figure SD3: Proportion of OTUs, belonging to each family, included in the analysis per day. All OTUs had at most 70% zeros per treatment group.

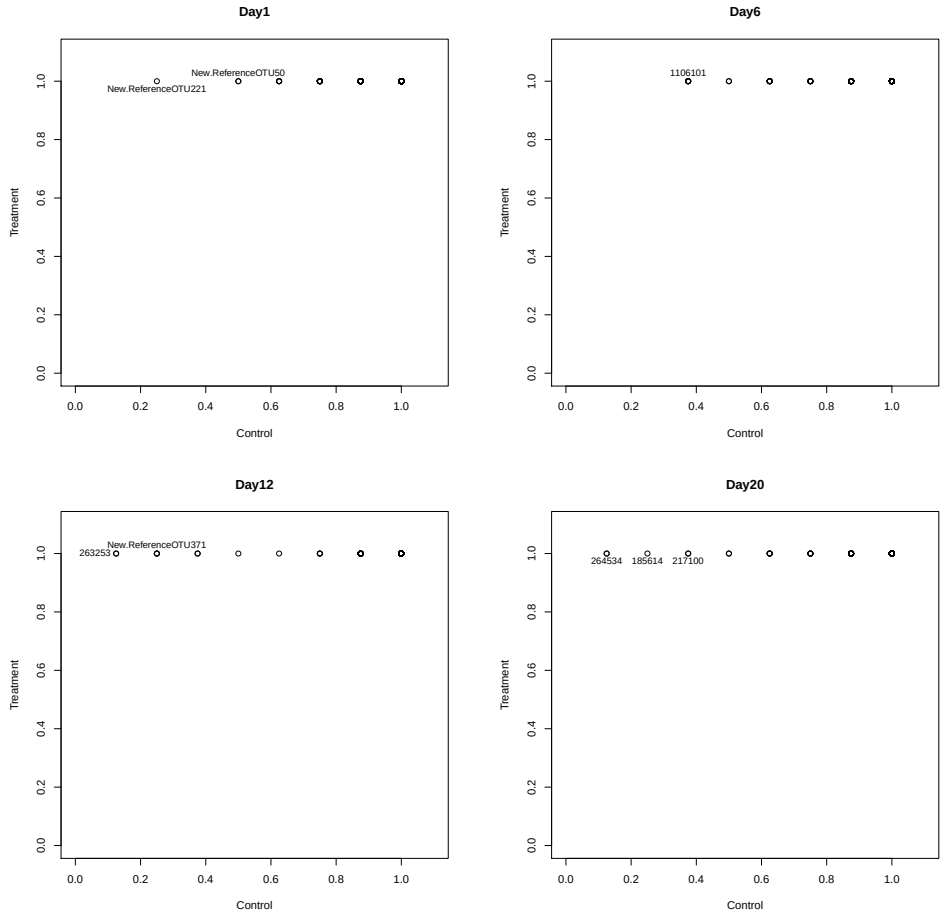


Figure SD4: Proportion of zeros of control versus PAT group for filtered out OTUs.

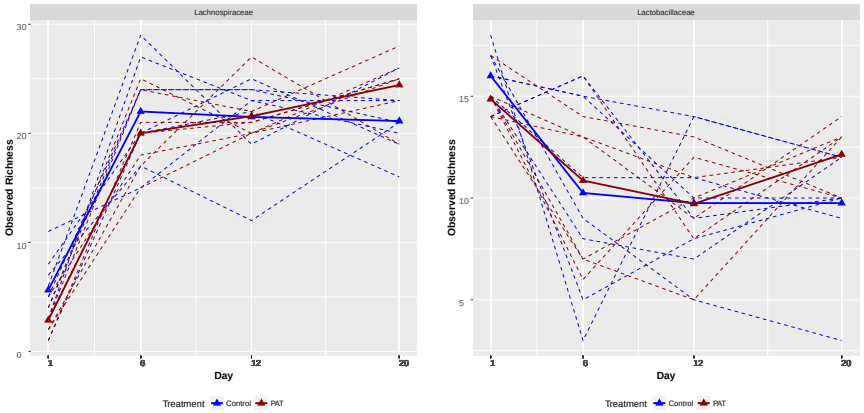
Appendix E

Development of High Dimensional Microbiome Biomarkers for an Immune Response: Hierarchical Bayesian Approach

This supplementary appendix contains additional materials related to Chapter 8.

E.1 Results for Other Active Families

Besides S24-7, *Lachnospiraceae* and *Lactobacillaceae* were two most active families. The results for these families are summarized in this section. For these two families no treatment effect on family level richness was evident on day 12 (Figure SE1). Table SE1 and Table SE3 validates this observation as Model 5 is found to be the model with the smallest DIC on day 12 for both these families.



(a) **(b)**

Figure SE1: Example of active families. Individual and mean longitudinal profiles of observed family level richness. Panel a: *Lachnospiraceae*. Panel b: *Lactobacillaceae*.

Models	Day 1	Day 6	Day 12	Day 20
Model 1	98.976	111.107	116.345	107.468
Model 2	**96.899**	117.189	114.311	111.896
Model 3	101.235	113.138	118.876	116.330
Model 4	103.894	**109.850**	114.386	**107.307**
Model 5	101.6	115.862	**112.274**	111.649
Model 6	98.774	110.822	116.140	107.557
Model 7	**96.695**	117.002	114.144	111.752

Table SE1: Deviance information criterion for the *Lachnospiraceae* family per timepoint.

Timepoint	**Day 1**		**Day 6**	**Day 12**	**Day 20**
Best Model	Model2	Model 7	Model 4	Model 5	Model 4
Estimate(α)	-0.692	-0.617	NA	NA	NA
95% C.I.(α)	(-1.238,-0.166)	(-1.146,-0.126)	NA	NA	NA
Estimate(β)	-0.624	-0.544	-0.460	-0.625	-0.996
95% C.I.(β)	(-1.224,-0.020)	(-1.108,0.035)	(-0.957,0.032)	(-1.229,-0.028)	(-1.591,-0.405)
Estimate(γ)	NA	NA	0.082	NA	0.112
95% C.I.(γ)	NA	NA	(0.023,0.142)	NA	(0.019,0.206)
Estimate(ρ)	NA	0.192	NA	NA	NA
95% C.I.(ρ)	NA	(-0.916,0.974)	NA	NA	NA

Table SE2: Parameter estimates from the best model for *Lachnospiraceae* family across all timepoints.

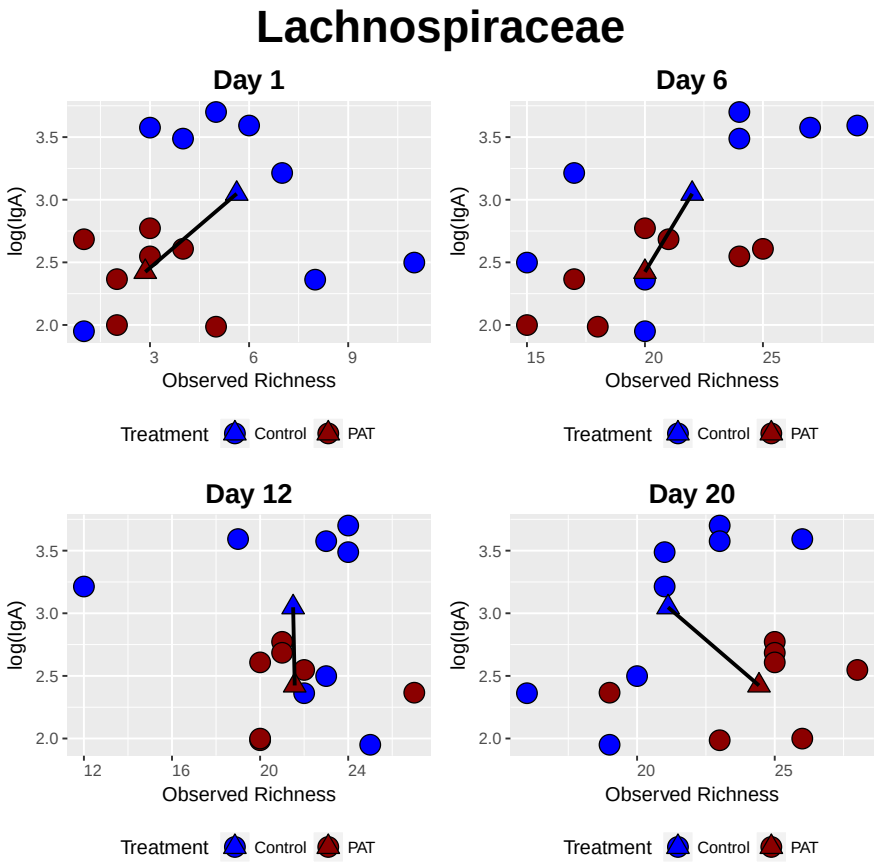


Figure SE2: Richness vs. log(IgA) for the *Lachnospiraceae* family across different time-points.

Models	Day 1	Day 6	Day 12	Day 20
Model 1	102.036	119.231	106.685	106.226
Model 2	102.464	121.467	106.157	104.634
Model 3	107.263	123.702	109.639	109.418
Model 4	**100.387**	**117.415**	104.738	106.243
Model 5	100.766	119.554	**104.114**	**104.565**
Model 6	101.953	119.023	106.471	106.064
Model 7	102.275	121.220	105.919	**104.438**

Table SE3: Deviance information criterion for the *Lactobacillaceae* family per timepoint.

Timepoint	**Day 1**	**Day 6**	**Day 12**	**Day 20**	
Best Model	Model 4	Model 4	Model 5	Model 5	Model 7
Estimate(α)	NA	NA	NA	NA	0.209
95% C.I.(α)	NA	NA	NA	NA	(-0.092,0.513)
Estimate(β)	-0.839	-0.662	-0.625	-0.625	-0.532
95% C.I.(β)	(-1.491,-0.191)	(-1.215,-0.117)	(-1.229,-0.027)	(-1.229,-0.027)	(-1.098,0.040)
Estimate(γ)	-0.188	0.062	NA	NA	NA
95% C.I.(γ)	(-0.446,0.069)	(-0.003,0.127)	NA	NA	NA
Estimate(ρ)	NA	NA	NA	NA	-0.074
95% C.I.(ρ)	NA	NA	NA	NA	(-0.960,0.940)

Table SE4: Parameter estimates from the best model for *Lactobacillaceae* family across all timepoints.

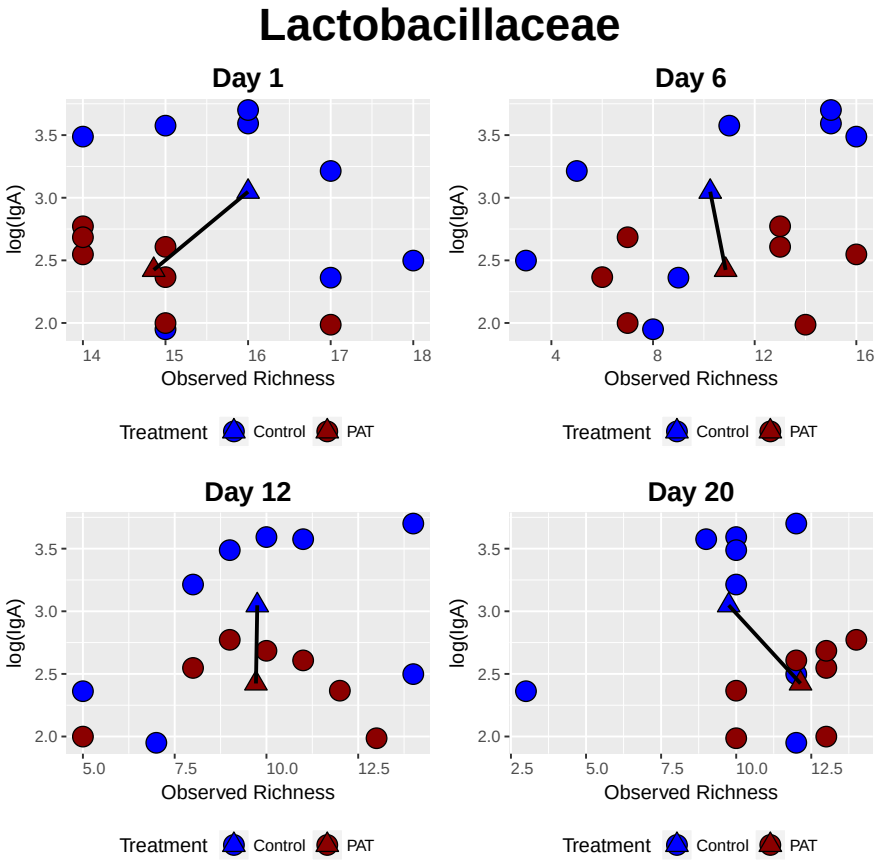


Figure SE3: Richness vs. log(IgA) for the *Lactobacillaceae* family across different time-points.

E.2 Implementation in WinBUGS

The path analysis model formulated in Section 8.2 was implemented in WinBUGS. The Poisson model for the richness, specified in equation (8.3),

$$X_{ji} \sim \text{Poisson}(\lambda_i),$$
$$\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i,$$

can be implemented in WinBUGS using the following code.

```
X[i] ~ dpois(lambda[i])
log(lambda[i]) <- muX + alpha*Z[i]
```

The Normal Model for log(IgA), specified in equation (8.3),

$$Y_i \sim N(\mu_i, \tau),$$
$$\mu_i = \mu_Y + \beta Z_i + \gamma_j X_{ji},$$

can be implemented using the following specification.

```
Y[i] ~ dnorm(mu[i],tau)
mu[i] <- muY + beta*Z[i] + gamma*Richness[i]
```

Prior distributions, specified in Section 8.2.2, can be implemented using the following code.

```
muY ~ dnorm(0.0,0.000001)
muX ~ dnorm(0.0,0.000001)
alpha ~ dnorm(0.0,0.000001)
beta ~ dnorm(0.0,0.000001)
gamma ~ dnorm(0.0,0.000001)
tau ~ dgamma(0.001,0.001)
sigma <- 1 / sqrt(tau)
```

Appendix F

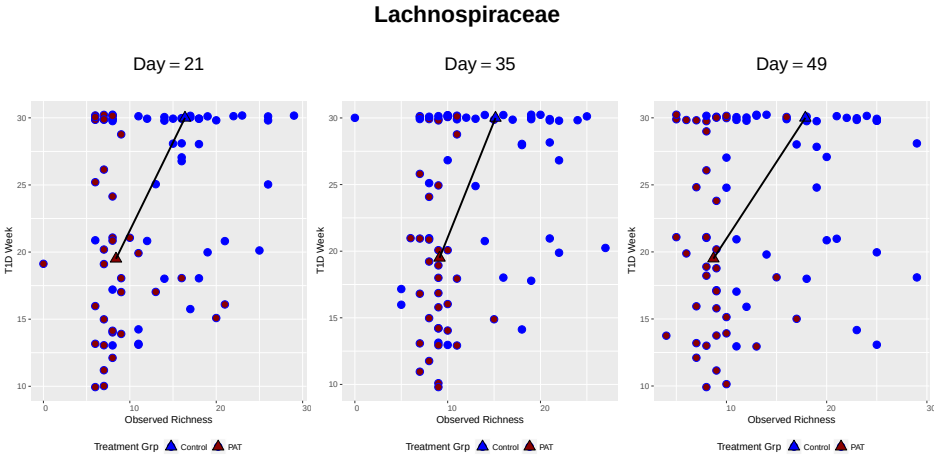
Development of Microbiome Biomarkers for Type I Diabetes

This supplementary appendix contains additional materials related to Chapter 9.

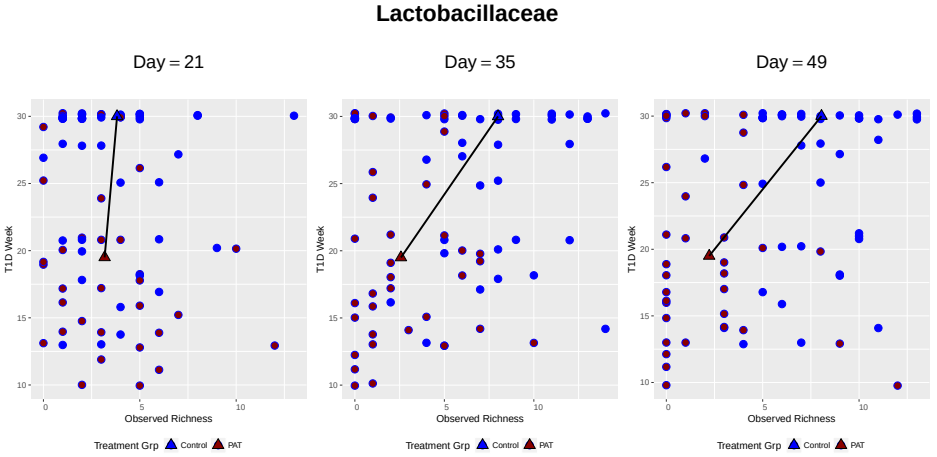
F.1 Results for Other Active Families

Besides S24-7, *Lachnospiraceae* and *Lactobacillaceae* were two most active families (Figure SF1). In this section we discuss the results for the Weibull regression model for these two families. Table SF1 and Table SF3 present the DIC values when all the models are fitted.

For the *Lachnospiraceae* family, Model 2 has the lowest DIC for all the timepoints. In other words, for all the timepoints the microbiome variable and the time to develop T1D are independent of each other, given the treatment. When the models with correlation are considered, the best model at day 21 includes the correlation between the treatment effects in the model. Tabel SF2 displays the parameter estimates from the selected model at each timepoint.



(a)



(b)

Figure SF1: Scatterplot of observed family level richness of other active families over time against the time to develop T1D. Panel a: *Lachnospiraceae* family. Panel b: *Lactobacillaceae* family.

Models	**Day 21**	**Day 35**	**Day 49**
Model 1	878.980	880.189	858.441
Model 2	**877.067**	**878.764**	**856.693**
Model 3	880.683	882.422	861.589
Model 4	980.55	937.931	984.308
Model 5	978.602	936.721	982.467
Model 6	878.944	880.080	858.436
Model 7	**876.930**	878.971	856.788

Table SF1: Comparing the models for the *Lachnospiraceae* family based on DIC values.

Timepoint	**Day 21**		**Day 35**	**Day 49**
Best Model	Model 2	Model 7	Model 2	Model 2
Estimate(α)	-0.669	-0.665	-0.508	-0.721
95% C.I.(α)	(-0.802,-0.538)	(-0.797,-0.533)	(-0.638,-0.378)	(-0.850,-0.592)
Estimate(β)	0.846	0.746	0.847	0.845
95% C.I.(β)	(0.261,1.443)	(0.192,1.326)	(0.269,1.447)	(0.268,1.439)
Estimate(γ)	NA	NA	NA	NA
95% C.I.(γ)	NA	NA	NA	NA
Estimate(ρ)	NA	-0.2705	NA	NA
95% C.I.(ρ)	NA	(-0.978,0.891)	NA	NA

Table SF2: Parameter estimates from the best model for the *Lachnospiraceae* family across all timepoints.

Similar to the *Lachnospiraceae* family, Model 2 is found to be the best model across all timepoints for the *Lactobacillaceae* family. However, when the models with correlation are fitted Model 7 has lower DIC value than Model 2 at day 21 and day 49. The parameter estimates from the selected models are available in Table SF4.

Models	Day 21	Day 35	Day 49
Model 1	786.549	829.823	802.999
Model 2	**785.956**	**827.797**	**800.989**
Model 3	793.536	831.871	803.935
Model 4	786.970	943.304	939.818
Model 5	786.166	941.299	937.793
Model 6	786.679	829.539	802.773
Model 7	**785.765**	827.868	**800.947**

Table SF3: Comparing the models for the *Lactobacillaceae* family based on DIC values.

F.2 Implemention in `WinBUGS`

Model 1, formulated in equation (9.2),

$$
\begin{aligned}
&X_{ji} \sim \text{Poisson}(\lambda_i),\\
&\log(\lambda_i) = \mu_{X_j} + \alpha_j Z_i,\\
&Y_i \sim Weibull(\theta, \mu_i),\\
&\mu_i = exp(\mu_Y + \beta Z_i + \gamma_j X_{ji}).
\end{aligned}
$$

can be implemented in `WinBUGS` in the following way,

```
# Model for Richness
X[i] ~ dpois(lambda[i])
log(lambda[i]) <- muX + alpha*Z[i]

# Model for time to develop T1D
t[i] ~ dweib(theta,mu[i])I(t.cen[i],)
mu[i] <- exp(muY + beta*group[i]+ gamma*Richness[i])
```

Timepoint	**Day 21**		**Day 35**	**Day 49**	
Best Model	Model 2	Model 7	Model 2	Model 2	Model 7
Estimate(α)	-0.186	-0.180	-1.137	-1.290	-1.273
95% C.I.(α)	(-0.423,0.051)	(-0.415,0.053)	(-1.363,-0.916)	(-1.530,-1.057)	(-1.513,-1.042)
Estimate(β)	0.848	0.726	0.844	0.839	0.782
95% C.I.(β)	(0.262,1.454)	(0.175,1.304)	(0.264,1.442)	(0.262,1.434)	(0.218,1.365)
Estimate(γ)	NA	NA	NA	NA	NA
95% C.I.(γ)	NA	NA	NA	NA	NA
Estimate(ρ)	NA	-0.084	NA	NA	-0.403
95% C.I.(ρ)	NA	(-0.960,0.936)	NA	NA	(-0.986,0.834)

Table SF4: Parameter estimates from the best model for the *Lactobacillaceae* family across all timepoints.

The priors for the parameters in the model can be specified by,

```
muX ~ dnorm(0.0,0.000001)
alpha ~ dnorm(0.0,0.000001)
muY ~ dnorm(0.0,0.000001)
beta ~ dnorm(0.0, 0.000001)
gamma ~ dnorm(0.0,0.000001)
theta ~ dexp(0.001)
```

Samenvatting

Één van de grootste uitdagingen tijdens geneesmiddel onderzoek is een zeer traag, kostelijk en inefficient ontwikkelingsproces. De keuze van de eindpunten om de doeltreffendheid van het geneesmiddel vast te stellen speelt een belangrijke rol en het beïnvloed de duur van het ontwikkelingsproces. Echter kan het meten van de eindpunten moeilijk, tijdrovend en duur worden. Een "surrogaat" eindpunt dient als een vervanging voor het "ware" eindpunt omdat het meestal goedkoper en gemakkelijker gemeten kan worden. Vooraleer een surrogaat als vervanging voor een waar eindpunt kan worded gebruikt moet het eerst worden gevalideerd. Statistische methoden voor de identificatie en evaluatie van surrogate eindpunten in gerandomiseerde klinische studies werden reeds de laatste drie decennia ontwikkeld (Prentice, 1989, Buyse and Molenberghs, 1998, Burzykowski et al., 2005, Alonso et al., 2016).
Een biomarker is een surrogaat voor bepaalde biologische processen in een therapeutisch interventie experiment. Het is gedefinieerd als een attribuut dat objectief wordt gemeten en wordt geëvalueerd als een indicator voor biologische of pathogene processen of pharmacologische responsen op verschillende types van interventies (Biomarkers Definitions Working Group, 2001). Dit impliceert dat alle surrogate markers biomarkers zijn maar niet alle biomarkers zijn surrogate markers.

Deze dissertatie behandeldt twee belangrijke onderwerpen - het bestuderen van gen expressie data en analyse van microbioom data. Het adresseert belangrijke problemen van de evaluatie van surrogate markers in hoogdimensionale transcriptie en microbioom data, maar is hiertoe niet gelimiteerd. Het werk bouwt verder op vroeger werk van verscheidene auteurs en gaat hier ook ruimschoots over.

Hoofdstuk 2 omvat het kader van gezamenlijk modelleren (joint modeling framework) gepresenteerd in Perualila *et al.* (2016a, 2016b), voor transcriptie data en het EGFR ontwikkelingsproject (Verbist et al., 2015) dat wordt gebruikt ter illustratie in het eerste

deel van de thesis. **Hoofdstuk 3** is gewijd aan computationele problemen. Een nieuw parallellisatie kader wordt voorgesteld dat toestaat dat analyse in Perualila *et al.* (2016a, 2016b) op te drijven. Het nieuwe parallellisatie kader wordt vergeleken met een aantal R software pakketen voor parallellisatie en verschillende configuraties voor parallellisatie worden onderzocht om, in het teken van de rekentijd, de beste oplossing te vinden voor het data analyse probleem.

In **Hoofdstuk 4** bespreken we verschillende types van surrogaten - enkelvoudige, veelvoudige, gedeeltelijke en orthogonale surrogatie. In dit hoofdstuk, breiden we het scenario van enkelvoudige surrogatie uit met een deelverzameling van ℓ mogelijke biomarkers en bespreken we het scenario van veelvoudige (Van der Elst et al., 2018), gedeeltelijke en orthogonale surrogatie.

Het tweede deel van de thesis is gefocused op microbiome interventie studies. De twee studies die worden geanalyseerd in dit deel van de thesis zijn dierlijke experimenten die werden uitgevoerd om het verband tussen het microbioome en klinische variabelen in kaart te brengen gebaseerd op een dierlijke model.

Hoofdstuk 5 introduceert de microbiooom data voor de twee interventie experimenten. Het eerste experiment werd uitgevoerd om het verband tussen het immuunsysteem van het subject (de host) en het microbioom te onderzoeken en het tweede experiment bestudeert de associatie tussen de tijd om type 1 diabetes (T1D) te ontwikkelen en het microbioom. Een parametrisch joint model en niet-parametrische benaderingen worden respectevelijk besproken in **Hoofdstuk 6** and **Hoofdstuk 7**. Bayesiaanse path analyse modellen voor Poisson en Normal verdeelde variabelen en Poisson en tijd tot event variabelen worden respectievelijk besproken in **Hoofdstuk 8** and **Hoofdstuk 9**. Een algemene discussie en een overzicht van toekomstige werkpunten van het onderzoek gepresenteerd in deze thesis worden gegeven in **Hoofdstuk 10**.