Imperial College London
Department of Life Sciences

Exploring networks of the human microbiota in health and disease

Virginia Fairclough

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Life Sciences, Imperial College London, September 2016

The content of this thesis is entirely my own work, except where explicitly stated otherwise; all the assistance received in preparing this thesis and all sources have been acknowledged and referenced. This thesis has not been submitted for any degree or other purposes.

Abstract

The human microbiota changes throughout our lives and is crucial to our health. However, some perturbations to the microbiota can have life-changing implications for its host. We are still in the early stages of understanding the complexities of these crucial microbial communities; as research into the human microbiota progresses, more methods will be required to explore dependencies between constituents of the microbiota, with the aim of understanding how they interact with each other and their host.

In this thesis, I present new methods and tools for creating and analysing networks of the human microbiota. I present a new method for analysing and comparing microbial dependencies in disease cases when compared to controls. This method is applied to real data taken from patients, and I draw inferences from the results, which agree well with the existing scientific literature.

This analysis is followed by the generation of weighted metagenome-scale metabolic networks. This novel approach presents the metabolic capabilities of the microbiota in a new manner, whilst providing a basis for future comparisons between the metabolic profiles found within the microbiota of cases and controls.

As well as presenting the metabolic capabilities of the community as a whole, it would be beneficial to be able to model a particular metabolic symbiosis. However, the current bottleneck in this area of research is due to slow curation of genome-scale metabolic models. Therefore, this thesis ends with the presentation of a new tool to aid the curation of genome-scale metabolic models and its application, with the hope that it may be used to improve existing models prior to the creation of models of metabolic symbiosis.

Acknowledgements

This thesis would not have been possible without the help and support of many individuals.

To my supervisor, Dr John Pinney, I thank you for your time, support and encouragement. Thank you for helping me to build in confidence since I began working under your supervision, and for your patience on numerous occasions!

I would like to thank Professor Michael Stumpf for helping to secure funding for my PhD — without this, I would not have been able to do this research. Thank you, also, for your guidance and frank advice over the years.

I am grateful for the funding and support from the Department of Life Sciences and for the opportunities provided by the doctoral programme at Imperial College London.

I wish to thank Dr Will Bryant for the contribution of the SBML metagenome-scale metabolic model used in Chapter 5 of this thesis, and for his application of the eQuilibrator tool to my data, discussed in Chapter 8.

I would like to thank MSc students YunSoo Kim, Laura de Arroyo Garcia and Ana-Isabella Tanase who worked so hard to re-implement my differential networks method in order to make it publicly available via an online interface. You were fantastic to work with and I wish you every success with your future studies and careers.

I would like to thank Dr Suhail Islam for his technical support throughout my PhD.

Special thanks go to Dr Adam MacLean, Dr Ann Babtie and Dr Aaron Sim for voluntarily proof-reading parts of my thesis and providing invaluable feedback. Your encouragement throughout my PhD has been incredible.

To Dr Paul Kirk, I owe thanks for his clear explanations of mathematics and statistics in the early years, and for his friendship ever since.

To Rob Johnson, I must say, it looks like we have finally made it — I could not have asked for a better comrade.

Dr Aaron Sim, I must thank you again for your patience, discussions and support throughout my studies, but most of all, thank you for your friendship.

To Dr Derek Huntley, thank you for your friendship, advice and numerous cups of tea!

To all members of the Theoretical Systems Biology group, past and present, I owe you all so much for your support and good advice throughout my studies; my PhD years would not have been the same without you all.

Of course, I will be forever grateful for the continued support from my family. Sal and Dad, you have always taught me to *just keep going*, and here we are!

Finally, I must thank my husband, Toby, for his enduring support throughout my PhD. You have been my sanity throughout this work, for which I will be eternally grateful; I could not have done any of this without you.

Nomenclature

16S rRNA 16S ribosomal ribonucleic acid

EBI European Bioinformatics Institute

FBA Flux balance analysis

FTP File Transfer Protocol

GO Gene Ontology

IBD Inflammatory bowel disease

IBS Irritable bowel syndrome

iCD Ileal Crohn's disease

IGT Impaired glucose tolerance

IL Interleukin

KDE Kernel density estimation

KEGG Kyoto Encyclopedia of Genes and Genomes

LB Lysogeny broth

MNXR MetaNetX reaction

MRCA Most recent common ancestor

NCBI National Center for Biotechnology Information

NF-κB Nuclear factor kappa-light-chain-enhancer of activated B cells

NGT Normal glucose tolerance

non-iCD Non-ileal Crohn's disease

OTU Operational taxonomic unit

PBMC Peripheral blood mononuclear cell

PCA Principal Component Analysis

QIIME Quantitative Insights Into Microbial Ecology

SBML Systems Biology Markup Language

SCFA Short chain fatty acid

T2D Type II diabetes

TNBS 2,4,6-trinitrobenzenesulphonic acid

UC Ulcerative colitis

ZIP Zero inflated Poisson

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The human microbiota is the collection of all micro-organisms residing within the human host. It is a thriving ecosystem of microbes that differs across body sites and individuals, and it is becoming increasingly evident that the human microbiota is of supreme importance to its host. It changes throughout our lives, adapting and affecting us from birth through infant development, adolescence, adulthood and old age. Perturbations to the microbiota can have life-changing implications for its human host — at birth, the colonisation process is key to starting life on the right path and infant dysbiosis (an unhealthy microbiota) has been linked to health issues observed later in adulthood, and even plays a role in personality development. We are becoming increasingly aware of its broad impact upon the host, with dysbiosis at different body sites being associated with a multitude of disease phenotypes, some of which are completely life changing for the sufferers. However, we are still in the very early stages of exploration, understanding, and knowledge of the microbiota and, as more is known, we become increasingly aware that there is more and more to discover. In this thesis, I discuss the new methods and tools that I have developed in order to aid further understanding of the microbiota, and the insight these have offered.

1.1 Thesis outline

I begin in Chapter 2 by providing the background and theory that have formed the foundation and motivation for my research. This includes an introduction to the microbiota, introductory theory behind some relevant statistical and modelling methods (such as correlation analyses, differential networks and SBML models), as well as a brief summary of some of the existing bioinformatics tools and databases of relevance to my work.

In Chapter 3, I briefly discuss my use of correlation analyses and proceed to discuss the method that I subsequently developed, which makes use of proportionality measures to investigate potential symbioses within the microbiota. This work is based upon the use of 16S rRNA gene abundance data from case and control studies. An applied example of this method is presented in Chapter 4.

Chapter 5 presents an exploration study that complements the method discussed in the two preceding chapters; the study discussed here uses the same type of data alongside growing online bioinformatics databases in order to investigate the inferred metabolic profiles of the microbiota in disease cases compared to controls. This is done by creating weighted metagenome-scale metabolic models in the SBML format, which has never previously been done. Issues surrounding comparisons of these networks are discussed in the context of future work.

In Chapter 6, I present the development of a software application tool that I created in order to aid the curation of metabolic models in the SBML format, such as the metagenome-scale model that I used in my work in Chapter 5. It is a graphical user interface that makes it easy to transfer stoichiometric information between reactions in genome-scale metabolic models with different namespaces, without the need to access the SBML model files directly. The motivation for this project was the desire to make the process of model curation faster and easier in order to help reduce the bottleneck that slows down the production of genome-scale metabolic models. Until this process is more efficient, and until the number of fully curated models is increased, there is a limit to what we can do with these models. In order to demonstrate the utility of this

tool, I provide a straightforward applied example in Chapter 7 using *Escherichia coli* models. This is a straightforward example so, in Chapter 8, I provide a more complex example of how the tool was used to guide early-stage curation processes in a model of *Salmonella enterica* Typhi. Here, the new software application was used to guide and initiate a number of steps to improve the draft model undergoing curation.

Finally, in Chapter 9, I conclude with a summary and a discussion of the future prospects for this area of study.

Chapter 2

Background Theory

2.1 The microbiota

Interactions between the microbiota and its host can affect the health and phenotype of the host in a variety of ways. These interactions have been shown to be greatly affected by environmental factors such as diet, as well as the genetic profile of the host (Human Microbiome Project Consortium, 2012; Cho & Blaser, 2012; Faith *et al.*, 2011). Changes or imbalances in the host's microbiota can have a dramatic effect upon host development, behavioural patterns and physical health. In fact, the relationship between the host and its microbiota is so fundamental to human health that it is argued by some that the microbiota could be considered as an organ within the human body, where *eubiosis* (a healthy microbiota) allows effective development and helps to prevent diseases; conversely, *dysbiosis* (an imbalanced microbiota) may cause diseases or disorders to occur (Buccigrossi *et al.*, 2013; Cho & Blaser, 2012). Dysbiosis in infants can alter the way in which the immune system develops, and can affect the way in which it responds to microbes with huge ramifications throughout childhood and/or adulthood (Human Microbiome Project Consortium, 2012). For example, some alterations in the microbiota of an adult host are significantly associated with the manifestation of diseases and disorders such as obesity, diabetes, psoriasis, inflammatory arthritis and other autoimmune diseases, inflammatory bowel

disease (IBD) and asthma (Balter, 2012; Greenblum *et al.*, 2012; Collins & Bercik, 2009; Cho & Blaser, 2012). It has even been shown that alterations in the microbiota of infant mice can alter brain development and thereby affect their behaviour in adulthood, showing the importance of eubiosis in the microbiota from birth (Diaz Heijtz *et al.*, 2011).

The investigation of the effects of microbial communities upon the health of the host is now a popular area of research, and rightly so. The number and diversity of wet-lab experiments has soared and the published research covers a vast array of different foci. There have been breakthroughs with the use of fecal transplants to treat recurring *Clostridium difficile* infections caused by dysbiosis, and other studies have shown how traits such as obesity can be transferred by performing fecal transplants from human subjects to mice (Aroniadis *et al.*, 2016; Costello *et al.*, 2016; Ridaura *et al.*, 2013). It is now of interest to see whether fecal transplants from healthy donors to diseased individuals can help with cases of IBD, irritable bowel syndrome (IBS) and a myriad of other diseases. Alongside the excitement brought by these interesting studies, there have been numerous researchers who have used their concern for ethics and safety to channel their research into developing optimal microbial suspensions and growth media to allow for the development of safe substitutes for fecal transplants without the risk of parasite transfer from donor to recipient (Petrof *et al.*, 2013; Orenstein *et al.*, 2016). Many other studies have focused on identifying the differences in the composition of the microbiota in disease cases compared to controls, which often rely on the use of 16S rRNA gene sequencing or metagenomic shotgun sequencing (see Section 2.2 below for details) to identify constituents of the microbiome (Cho & Blaser, 2012; Shankar *et al.*, 2013; Hamady & Knight, 2009; Morgan *et al.*, 2012). Research into microbiota composition has led to investigations into how the accuracy and depth of bacterial identification methods alters our understanding of the ecological system (Li *et al.*, 2016; Lundin *et al.*, 2012). Yet others have begun to explore how the metabolism of the communities in dysbiosis differ from that of the system in eubiosis, with the hope of beginning to understand how these microbial ecosystems function. Some of these studies are based on the mass spectrometry of samples from patients while others rely on abundance data and databases of metabolic information (Preidis *et al.*, 2016; Gutiérrez-Díaz *et al.*, 2016; Morgan *et al.*, 2012).

Studies such as these demonstrate the importance of considering the ecology of the microbial ecosystems in order to identify how the profile of dysbiosis differs from that of eubiosis. More specifically, if we can begin to understand the metabolic interactions within (and metabolic capabilities of) these communities, it may shed some light on the pathological determinants of dysbiosis, and guide us in therapeutics for specific dysbiosis-induced conditions.

As a consequence of the increased interest and number of studies being carried out upon the topic of the microbiota, there are now whole databases of microbiota data available online, such as the Human Microbiome Project and the European Bioinformatics Institute (EBI) Metagenomics database (Turnbaugh *et al.*, 2007; Mitchell *et al.*, 2016). Existing resources such as these can be used in conjunction with modelling and statistics in order to further this system-focused ecological research. For example, a multitude of databases now exist that contain sequence data, genomes, gene expression data, gene annotations, metabolic pathways and more (Kanehisa *et al.*, 2016; Kumar *et al.*, 2012; Caspi *et al.*, 2016; Chelliah *et al.*, 2013). This wealth of information can be integrated and used to create or enhance models of the microbiota for healthy and disease states.

2.2 Data collection

There are two key methods for data collection when studying the microbiota: small-subunit ribosomal RNA (rRNA) gene sequences or metagenomic shotgun sequencing. The former method is used to characterise the constituents of the microbiome (i.e. produce presence/absence and abundance data) whereas the latter is typically used to identify the genetic functionality of the microbiome via the identification of genes present in the microbiota as a whole (Hamady & Knight, 2009). Further details are discussed below and a comparison of the two methods is shown in Table 2.1.

2.2.1 16S rRNA data

In studies of the microbiota using this method, the 16S rRNA gene sequence (i.e. the DNA code for the 16S rRNA gene) is used to identify the bacteria present within the samples taken from the microbiota. The bacteria are consequently represented by operational taxonomic units (OTUs) that serve as a way of classifying the bacteria. Typically, reads of 250 bases are used for the identification of a species; the level of accuracy achieved by using these reads is suitable for comparing communities but can only serve as a guide when trying to infer phylogenetic relationships between reads (Hamady & Knight, 2009). However, the taxonomic assignment of these 16S rRNA reads are not any more accurate or reliable when longer reads are used (Hamady & Knight, 2009). Furthermore, different primers can lead to different biological conclusions and primer bias can lead to misrepresentation of specific data – for example, some groups may be entirely excluded due to failure to be amplified (Hamady & Knight, 2009). Therefore, only datasets that have been produced using the same primers and amplification protocol can be compared, thereby making statistical comparisons of data from different studies difficult. Whilst this method of 16S rRNA gene sequencing is a much cheaper way of determining the bacterial and archaeal constituents of a microbial population when compared to full metagenomic shotgun sequencing, it must also be noted that eukaryotic and viral genomes will not be detected using this method, which provides limitations to the application of these data when analysing communities where viral or eukaryotic involvement is thought to play a pivotal role.

2.2.2 Metagenomic shotgun sequencing data

In metagenomic shotgun sequencing, samples are taken directly from the environment, all of the DNA within the sample of the community is sheared randomly and then these fragments are sequenced. The overlapping reads from these fragments are used to reconstruct the genomes from that environment, thereby making this method expensive and time consuming due to the large amount of sequencing required and the number of reads that need to be realigned and constructed into contiguous sequences. Therefore, the computing power required to do this is

immense as it involves such a large amount of data; for example, in a study by *Moore et al.*, 193 million reads were produced from just thirty one samples (Moore *et al.*, 2015). However, this method allows us to observe and predict the genes present within that community across all genomes present in a given sample, including eukaryotic and viral genomes.

Table 2.1: *Comparison of 16S rRNA gene sequencing data and metagenomic shotgun sequencing data.*

16S rRNA data	**Metagenomic shotgun sequencing data**
Classifies bacterial and archeael constituents	Characterises gene content of all sampled organisms
Constituents are typically represented by OTUs	Genes can be represented by orthologue groups
Amplify a region of one gene	Amplify all genetic material
Requires gene-specific primers and can exhibit primer bias	Primers are not gene-specific
Focus on taxonomic relationships within the community	Focus on genetic functionality of the community
No sequence reconstruction required	Contig construction required (using overlapping sequence fragments)
Reasonable computing power required	Vast amounts of computing power required for reconstructions
Relatively cheap	Relatively expensive

The vast majority of studies of the human microbiota are based upon 16S rRNA gene sequence data rather than metagenome shotgun sequencing due to its cheaper cost and the ability to still broadly identify the constituents of the microbiota. Therefore, only 16S rRNA sequence data has been used for the studies in this thesis.

2.3 Analysing data from the microbiota

Statistical methods can be used to analyse microbiota data and consequently identify any associations between community profiles and disease states, age of the host or other variables. Some of these statistical methods are exploratory methods for data visualisation (e.g. Principal Component Analysis (PCA) for dimension reduction) whilst others allow conclusions to be drawn and hypotheses tested. Ecological methods are often applied in order to asses the diversity of microorganisms identified within the microbiota (e.g. observations of statistically significant differences in OTU abundances in cases and controls using Wilcoxon tests with

multiple testing correction). Other common methods involve analysing the distance between case/control profiles, using methods such as the unique fraction metric (UniFrac), which measures the phylogenetic distance between sets of taxa in a tree (Clooney *et al.*, 2016; Caporaso *et al.*, 2010; Lozupone & Knight, 2005). This method can be used in conjunction with Monte Carlo simulations to determine whether the the fraction of the tree unique to one environment is greater than would be expected by chance when two communities are compared (Caporaso *et al.*, 2010; Lozupone & Knight, 2005).

Whilst most existing microbiome research has focused on identifying the differences in constituents of the microbiome in cases and controls, there are far fewer methods and analyses that allow us to assess how the members of a given community interact. Therefore, the implementation and application of a new method is required for identifying and comparing these interactions (which could manifest as associations or dependencies between variables) in case/control data. This requirement is addressed by the new method presented in Chapter 3 of this thesis. Some common statistical analyses that are used for analysing community data or identifying associations between variables are discussed below in order to give some background and context to the method developed in Chapter 3.

2.3.1 PCA and data transformation

In order to explore and visualise potential clusters in community data according to metadata classes, PCA can be applied. PCA is a dimension reduction method for high-dimentional data sets (i.e. data sets where multiple parameters are measured for a single sample) for which there may be many samples. For example, in microbiota data, there are many (hundreds or thousands) of different types of bacteria/OTUs for which we have an abundance value in each sample. In datasets such as this, PCA allows data visualisation by identifying the axes of greatest variance within the data, where the first principal component (PC1) represents the axis of greatest variance and all subsequent principal components represent a decreasing degree of variance in turn. Typically the first two principal components are chosen for visualisation and

can be represented by plotting their eigenvector weights (loadings). These weights represent the contribution of each variable (e.g. OTU) to these axes of greatest variance. After projection onto the first two principal components, it is sometimes possible to see separation of classes within the dataset (e.g. cases/controls, age groups).

An example of the application of PCA to existing microbiota data can be found within the research carried out by Shankar *et al.* (2013). They carried out PCA on scaled data taken from patients suffering from IBS and healthy controls, and they subsequently showed the two classes segregating into distinct clusters.

PCA uses a Euclidean distance measure between samples (Ramette, 2007; Legendre & Gallagher, 2001). However, ecological data (including microbiota samples) often have a high proportion of zero values or low abundances, which can distort the visualisation in PCA. This is because the differences between low-abundance species contribute more to the measured distance between points than differences between high-abundance species; i.e. low-abundance species have a disproportionately large impact upon PCA (Ramette, 2007; Legendre & Gallagher, 2001). Therefore, it is appropriate to transform the data; for example, the Hellinger and chord transformations give low weights to rare species (Ramette, 2007; Legendre & Gallagher, 2001).

2.3.2 Correlation

In order to measure or identify associations between pairs of variables, correlation might be used to quantify the dependence between pairs of variables within a dataset.

The Pearson correlation coefficient ($r_{X,Y}$) measures the linear relationship between a given pair of variables and is calculated as the covariance of two random variables ($cov(X, Y)$) divided by the product of their standard deviations (σ_X and σ_Y), as shown in Equation 2.1 (Quinn & Keough, 2002).

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

Refer to Appendix A.1 for mathematical definitions of standard deviation and covariance.

In the context of data from the microbiota, some studies have been conducted where correlation has been used to identify potential associations between OTUs within a given community. For example, in the previously mentioned study by Shankar *et al.* (2013), they showed correlations in abundance between different genera within the gut microbiota in adolescents, and analysed the differences in correlations observed in inflammatory bowel disease patients (IBS patients) and healthy individuals (controls). They consequently showed that more OTUs are observed to have correlated abundances within the healthy individuals and identified the hubs within the correlation network, such as *Ruminococcus*. Correlation between pairs of OTU abundances within samples may suggest that they interact in some way, or there may be a factor that affects both of them simultaneously. A review of correlation methods was recently published by Weiss *et al.* (2016), where they assessed the performance of such analyses in the context of the microbiota.

2.3.3 Partial correlation

It is likely that various OTUs within a microbial community interact with one another in a complex manner and affect each others' abundance, meaning that each pairwise correlation will be affected by other OTUs. In such cases as these, it could be argued that it is preferable to use partial correlation in place of standard correlation because partial correlation is the correlation that remains between two variables (e.g. OTUs) after the effects of other variables have been removed by regression (Opgen-Rhein & Strimmer, 2007). In the case of microbiota data, partial correlation may be used to investigate pairwise associations between OTUs with any confounding effects from other OTUs accounted for. See Appendix A.2 for a mathematical definition of regression and partial correlation.

2.3.4 Measures of goodness of fit to proportionality

Whilst correlation measures are a common method used for analysing associations between variables, it is not always the most appropriate measure for all data types. Lovell *et al.* (2015) demonstrated that correlation is not subcompositionally coherent; if we calculate the correlation between variables when only considering subsets of the full set of variables, we obtain different results than if we consider all of the variables. This is of note when using 16S rRNA gene sequence data, as we are aware that amplification methods are not perfect and can exhibit amplification bias in such a way that we may sometimes obtain only a subset of all constituents of the microbiome (Sim *et al.*, 2012).

Lovell *et al.* (2015) also go on to illustrate how correlated relative abundances tell us nothing about the nature of dependencies between the underlying absolute data unless the total level of all variables is equal across all experimental conditions. In the case of 16S rRNA gene sequence data, this would require all amplification runs to have the same total number of reads produced. This is an unrealistic expectation for such data as these.

Therefore, Lovell *et al.* (2015) propose the use of a measure of goodness of fit to proportionality (hereafter referred to as a *proportionality measure*) in place of correlation in order to investigate potential dependencies between relative variables, such as 16S rRNA abundance data. A measure of goodness of fit to proportionality assesses how closely the data points from two variables lie on a straight line represented by a linear equation of the form shown in Equation 2.2. If two variables are perfectly proportional, the quantity of one (e.g. A) can be predicted exactly by the quantity of the other (e.g. B) by multiplying its value by a given constant (i.e. $A = mB$). Therefore, if one is zero, the other must also be zero, and the proportional relationship between two variables, x and y, can be represented by a linear equation through the origin:

$$y = mx \tag{2.2}$$

The measure of goodness of fit to proportionality proposed by Lovell *et al.* (2015) gives rise

to two proportionality values (ϕ) for each given pair of variables (A and B) by the following equations:

$$\phi_{(A,B)} = \frac{\mathrm{Var}(\log(A/B))}{\mathrm{Var}(\log A)} \tag{2.3}$$

$$\phi_{(B,A)} = \frac{\mathrm{Var}(\log(B/A))}{\mathrm{Var}(\log B)} \tag{2.4}$$

This provides an alternative method to correlation analyses for the identification of potential dependancies between two variables when only relative data is available.

Unlike correlation measures, where $r_{(A,B)} = r_{(B,A)}$, the two ϕ values are not symmetrical unless A and B are perfectly proportional, in which case $\phi_{(A,B)} = \phi_{(B,A)} = 0$. The more proportional the two variables, A and B, the lower the ϕ values. Conversely, the higher the ϕ values, the more asymmetrical they may become and the less defined the AB-relationship becomes.

This measure of goodness of fit to proportionality may be used in conjunction with differential networks in order to analyse changes in associations between OTUs in cases when compared to controls, as demonstrated in the new method presented in Chapter 3 of this thesis.

2.3.5 Differential networks

Static networks are useful for visualising biological systems (e.g. protein-protein binding; molecular interactions; bacterial associations) and are ideal for analysing large sets of data, such as association measures within 16S rRNA gene abundance data. When two static networks need to be compared to each other (e.g. a static network for case data compared to a static network for control data), it is possible to create a differential network that shows the differences between the two static networks. An interesting study by Valcárcel *et al.* (2011) used partial correlation networks to analyse changes in metabolite levels in pre-diabetic (case) and healthy (control)

patients, and showed that these networks were more sparse than the networks created using correlation. Furthermore, Valcárcel *et al.* (2011) went on to show that differential networks (which show the changes in partial correlation between cases and controls) could highlight key changes in metabolite associations across the two classes. Similarly, differential networks can be produced to analyse the differences in associations between species under different conditions or in different classes from within the microbiota data, and can therefore be used to identify changes in OTU dependencies that are indicative of certain disease states. Potential measures of association or dependencies include correlation, partial correlation and measures of goodness of fit to proportionality. All three of these measures of association have their own benefits and limitations.

2.4 Modelling metabolism

The methods discussed above can be used to identify dependencies between bacteria and explore how these dependencies change in disease cases when compared to controls. However, it would be useful if we could determine the nature of these dependencies. For example, it is already known that constituents of the human gut microbiota cross-feed via metabolic symbiosis (Koropatkin *et al.*, 2012). Therefore, in order to characterise these dependencies, it is worth combining the microbiota data with existing information about the metabolic pathways and capabilities of the microbiota constituents. Metabolic models have been used to look at the metabolic capabilities of single cells, including quantitative studies such as flux balance analysis (FBA), where metabolic flux is modelled within a given cell under steady state assumptions (Orth *et al.*, 2010). Few models of metabolic symbiosis have been made to date, but there is scope for these types of models when assessing dependencies within the microbiota (Heinken *et al.*, 2013; Bordbar *et al.*, 2010). Therefore, these genome-scale metabolic models, FBA and models of symbiosis are discussed below.

2.4.1 Genome-scale metabolic models

With the advancements of modern-day sequencing and easy access to genomic information, it is possible to assess the genome and metabolic networks belonging to a given bacterium and subsequently use this knowledge to model its genetic and metabolic processes (Terzer *et al.*, 2009; Thiele & Palsson, 2010). Once communities of OTUs associated with disease symptoms have been identified, they can be investigated further by using metabolic networks, which are constructed in the Systems Biology Markup Language (SBML) format (Hucka *et al.*, 2003). A model for a given bacterium contains information about reactants, products and the stoichiometric values for all known reactions within the modelled cell. Each reaction is modelled with its set of reactants and products (see Figure 2.1 a)), and its own constraint parameters that allow for quantitative analyses based upon real-life conditions: the upper bound, lower bound and flux bound. These parameters can be used to constrain the flux through reactions and can be used as a proxy for enzyme activity or rate of a reaction (Brandes *et al.*, 2012). All of the reactions within the cell are modelled together in a large bipartite network, as shown in Figure 2.1 b).The stoichiometric information for every reaction is used in constraint based analyses to gain quantitative insight into cellular metabolic profiles. For example, methods such as flux balance analysis (FBA) can be used to model a cellular system in different circumstances (Schellenberger *et al.*, 2011; Orth *et al.*, 2010), giving rise to flux profiles that describe the cell's metabolism under particular conditions.

2.4.2 Flux balance analysis (FBA)

Flux balance analysis is a quantitative method that allows us to determine the flow of metabolites through genome-scale metabolic models. This is done via the use of a stoichiometry matrix containing the stoichiometries of the metabolites in all of the modelled reactions, which act as one set of constraints on the model. They ensure that the total amount of each metabolite being produced throughout all reactions of the model balances with its use as a reactant throughout the model, thereby ensuring a steady state. Therefore, these models do not rely on kinetic

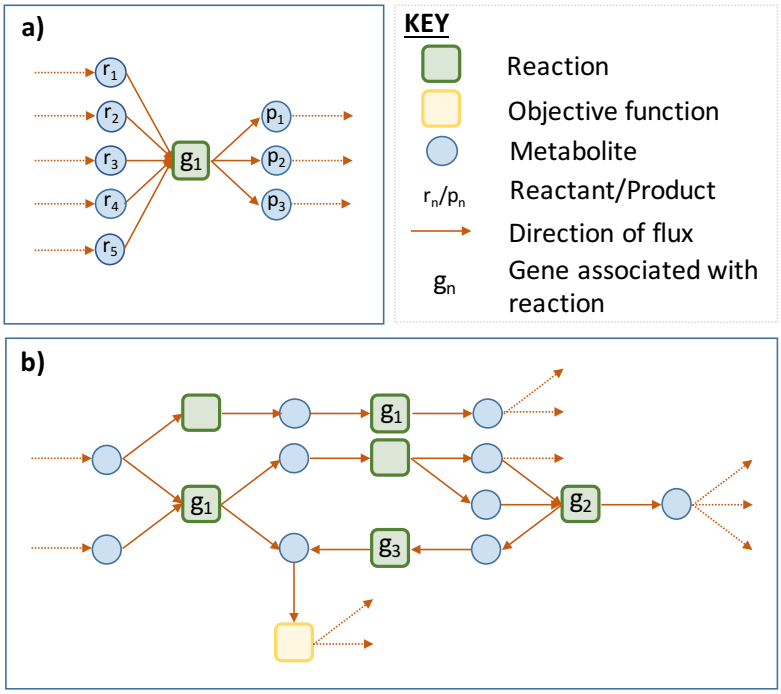


Figure 2.1: *a) A single reaction is shown as a green box with arrows showing the direction of flux (i.e. the reaction direction). Therefore, the reactants ($r_1 - r_5$) are upstream of the reaction and the products are downstream. The direction of flux is determined by the upper and lower flux constraints for each reaction. Dotted arrows represent an extension to the rest of the metabolic network. b) Numerous reactions and metabolites are joined together to show how multiple reactions come together to form a bipartite network where reactions are shown as green boxes and are separated by metabolites, which represent the reactants and products of the reactions that they are linked to. The direction of the arrows show the direction of flux through the reactions. Dotted arrows represent an extension to the rest of the metabolic network. The yellow box represents the reaction that is set as the objective function during FBA runs.*

parameters, which are hard to measure (Orth *et al.*, 2010).

The stoichiometric matrix, $\mathbf{S}$, contains n reactions, each of which is represented by a different column of the matrix. Similarly, each row of the matrix represents one of the m metabolites. Each entry in the matrix represents the stoichiometry of a given metabolite in a given reaction, where negative values represent reactants that are consumed, positive values represent products and zeros imply that a given metabolite is not present within a given reaction, which gives rise to a sparse matrix (Orth *et al.*, 2010).

$$\mathbf{S} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Another set of reaction-specific constraints lies in the form of the upper and lower flux bounds defining the maximum and minimum fluxes through each reaction. By setting both of these parameters to zero for a given reaction, this provides us with a representation of a reaction knock out by blocking flux through that reaction (Orth *et al.*, 2010).

Within the model, we also define the biochemical composition of the cell via a biomass reaction. This is represented as an extra column of stoichiometric coefficients within the stoichiometry matrix and it is scaled such that the flux through this reaction represents the exponential growth rate of the modelled organism (Orth *et al.*, 2010).

Together, these sets of stoichiometric and bound constraints define the solution space. This solution space represents all of the allowable flux values through the reactions of the model. These flux values represent the rates of reactions within the model and therefore also represent the rate of production and use of metabolites in the cell (Orth *et al.*, 2010). These internal fluxes can be represented by a flux vector, $\mathbf{v}$.

We can use the reaction constraints (i.e. mass balance) in order to define a system of linear equations that represent the system. By setting the biomass reaction as the objective function for which we must maximise the flux, we can use linear programming in order to determine the maximum possible growth rate and the associated internal flux values of the model (Orth *et al.*, 2010). Given the steady state assumption of FBA, we can solve the following equation:

$$\mathbf{Sv} = 0 \tag{2.5}$$

By setting upper and lower bounds of single reactions to zero (which mimics a reaction knock-

out), we can also investigate essential gene profiles using FBA methods. Essential gene predictions can be compared to experimental (wet-lab) data to validate genome-scale metabolic models. This method of model validation is used in Chapters 7 and 8 to explore and validate genome-scale metabolic models that were undergoing curation.

2.4.3 Modelling symbioses

Genome-scale metabolic models take a long time to create and curate using the literature and experimental methods. Automated methods have been developed to create very basic draft models (Swainston *et al.*, 2011), and now there are certain online resources that are dedicated to generating and providing access to metabolic models, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), BioModels and MetaCyc databases (Kanehisa *et al.*, 2016; Chelliah *et al.*, 2013; Caspi *et al.*, 2016). The bottleneck that is prohibiting wider use of these models is currently the time required to curate these basic draft models. More tools are required to aid and speed up the model curation process, because if this process was more efficient, quantitative methods would be more accurate and it would be easier to expand these models to incorporate more cells and therefore model symbioses. Some studies have shown that models of metabolic symbiosis between two cells can be generated by merging individual models of two single cells (Heinken *et al.*, 2013; Bordbar *et al.*, 2010), but the methods used to create these models are time-consuming due to extensive manual curation required at each stage of the merging process. Ideally, there needs to be a platform for automatically generating models of symbiosis as well as tools to aid faster model curation for the existing basic draft models. Ultimately, such models could be produced for pairs of bacteria that are observed to co-exist in disease-associated microbiota (determined via analysis of the OTUs discussed above) in order to characterise or investigate potential metabolic interactions between pairs of cells (e.g. mutualism through metabolic exchange, parasitism/syntrophic relationships). By modelling the metabolism of (parts of) bacterial communities associated with diseases, our understanding of such diseases may increase and lead to the identification and development of potential therapeutic methods

and effective treatments.

Although the existing concept of merging single-cell models to produce models of metabolic symbiosis is interesting and has great potential, the time investment currently required to produce them means that they are most likely to be produced and used for cases where we are already certain of existing metabolic reactions between species that we wish to model for exploratory reasons. It is less practical to invest such time in producing these models for investigating whether metabolic interactions are the cause of observed associations between species within microbial communities. It would also currently be too complex and time-consuming to produce a community-scale metabolic model in this manner. Therefore, the application of these merged methods is somewhat limited whilst single model curation remains a slow process. Any progress in curation methods will be valuable to the community in getting closer to the possibility of quantitative community-scale models of metabolic interactions. This was the motivation behind the software application presented in this thesis (see Chapter 6 for documentation and implementation details). This software application was created to aid model curation by making it easy to transfer information from curated models to draft models undergoing development. The application of this tool is presented via two examples: a simple example is presented in Chapter 7 and a complex example is presented in Chapter 8.

However, methods investigating population dependencies within communities and metabolic differences between communities are currently likely to provide more insight than merged metabolic models, given the time it takes to create each curated, accurate genome-scale metabolic model.

2.5 Inferring metabolic capabilities

Given the current limitations of the merged metabolic model, it is worth exploring metagenome-scale metabolic models as a way of assessing the metabolic capabilities of a whole community. These large-scale models could be created by combining sequencing, genomic, and metabolic

information in order to construct an inferred metabolic network representing the metabolic capabilities of the microbiota (Morgan *et al.*, 2012), which is discussed below.

2.5.1 Integrating sequencing data with genomic information

In order to analyse community-scale metabolism, it is possible to integrate 16S rRNA gene sequence data with existing genetic and metabolic information in order to gain an overview of the metabolic capabilities of a given community. For example, Morgan *et al.* (2012) used 16S rRNA gene sequence data to compare the gut microbiota of healthy individuals to that of patients suffering from inflammatory bowel disease (IBD) . They identified the relative abundances of the various different bacteria present within the two populations and subsequently used the sequenced genomes of these bacteria (or their close relatives) in order to build what they refer to as an 'inferred metagenome'. They then linked the relative abundances of genes within each inferred metagenome (one for cases and one for controls) to information contained within the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) databases in order to compare these abundances between the case/control communities. Their results showed differences in oligosaccharide transport, the sulfate transport system and the polysaccharide catabolic process when comparing the gut microbiota of ileal Crohn's disease patients with that of healthy controls.

The concept of the inferred metagenome could be valuable for extracting more information from simple 16S rRNA gene sequence data without the added cost of full metagenomic sequencing. Furthermore, linking metabolic information to a metagenomic profile allows us to consider community-level metabolism in a different way. Instead of considering explicit relationships between the individual species within the community (as is done in the case of merged models), it allows us to view the metabolic capabilities of the communities as a whole, thereby potentially providing insight into the broad metabolic processes within the community.

The existing database resources (such as KEGG and GO) could be used at a finer level than those used by Morgan *et al.* in earlier work. Instead, it is possible to create metagenome-scale

metabolic networks for case/control groups that incorporate individual reaction abundances within cases or controls. These reaction networks could potentially allow for greater resolution than comparing broad categories such as KEGG modules, KEGG pathways and GO term abundances, as Morgan *et al.* have done. Work on these reaction-resolution metagenome-scale metabolic models is presented in Chapter 5.

2.5.2 Resources for metabolic modelling with inferred metagenomes

The sequence data and metabolic modelling resources that we can use to create these metagenome-scale models and analyses are introduced below.

Sequence data

There are a number of platforms for processing and analysing raw sequence data. Two commonly used platforms are Quantitative Insights Into Microbial Ecology (QIIME) and Mothur (Caporaso *et al.*, 2010; Schloss *et al.*, 2009). They can be used to analyse datasets produced by Sanger, PacBio, IonTorrent, 454, and Illumina (MiSeq/HiSeq) sequencing methods in order to cluster the sequences by similarity into OTUs, and subsequently phylogenetically classify these clusters of OTUs using a representative sequence (Caporaso *et al.*, 2010; Schloss *et al.*, 2009). This ultimately gives rise to OTU abundance tables. Furthermore, these tools provide in-built downstream analysis tools, such as PCA for data exploration, generation of phylogenetic trees and ecological analyses such as measuring diversity metrics (Caporaso *et al.*, 2010).

As well as cluster-based methods, it is also possible to classify sequences using a sliding window of k-mers (with a default length of 31 bases) that are compared to a pre-built k-mer database, as is the case with Kraken (Wood & Salzberg, 2014). For each overlapping k-mer in a given nucleotide sequence, the underlying k-mer database is queried in order to identify the most recent common ancestor (MRCA) of organisms containing the given query k-mer. By analysing the list of MRCAs identified using the overlapping k-mers, it is possible to apply a weight to

each of these MRCAs and subsequently determine the most likely taxonomic classification of the sequence (Wood & Salzberg, 2014).

Metabolic modelling

Various resources exist to provide information about metabolic pathways and to aid the creation and curation of metabolic models. Five key databases are the MetaCyc, KEGG, MetRxn, MetaNetX and BioModels databases (Caspi *et al.*, 2016; Kanehisa *et al.*, 2016; Kumar *et al.*, 2012; Ganter *et al.*, 2013; Chelliah *et al.*, 2013). Each will be described briefly below.

MetaCyc

MetaCyc is primarily a qualitative database of non-redundant metabolic pathways and enzymes that have been experimentally validated and curated (Caspi *et al.*, 2016). However, it also contains information about individual reactions (including any associated enzyme EC numbers, where known), compounds, macromolecular complexes and genes, and limited quantitative data, such as some enzyme kinetics (Caspi *et al.*, 2016). This data is mostly from micro-organisms and plants, representing over 2400 organisms in total (Caspi *et al.*, 2016).

Kyoto Encyclopedia of Genes and Genomes (KEGG)

The KEGG database contains genomic, chemical and systemic information that allows us to link genetic information to cellular functions (Kanehisa *et al.*, 2016). For example, by analysing which genes are present within a genome, we can determine which enzymes can be produced by a cell with that genome, and therefore determine the metabolic capabilities of that cell.

The database is split into different sections that contain different types of information, including KEGG pathways, KEGG modules (e.g. groups of pathways forming functional units or groups of molecules forming complexes) and the KEGG Orthology (KO) system, which allows us to link genomes, metabolic pathways and metabolic networks via genetic homology (Kanehisa *et al.*, 2016). The database can be exploited to interpret sequencing data and model cellular metabolism via the automated generation of metabolic models or determine lists of reactions

associated with a given organism (Swainston *et al.*, 2011).

MetRxn

MetRxn combines information from multiple databases and genome-scale metabolic models into one database with standardised metabolite and reaction descriptions (Kumar *et al.*, 2012). This is incredibly valuable due to the fact that different databases use different namespaces, many of which are used across different models and they can be difficult to reconcile when comparing or merging models (Kumar *et al.*, 2012). The metabolite entries within the MetRxn database have matched synonyms, resolved protonation states and each one represents a unique chemical structure, making it possible to download standardised versions of existing models to enable faster, accurate comparisons of models from different databases (Kumar *et al.*, 2012). They also provide some exploratory tools for comparing models and subsequent visualisation of the results (Kumar *et al.*, 2012).

MetaNetX

MetaNetX provides an online platform for analysing and manipulating biochemical pathways and genome-scale metabolic models via the integration and standardisation of data from public resources via a unique and common namespace (Ganter *et al.*, 2013). Users can upload their own models to map them onto the common namespace or access one or more of the hundreds of models and pathways via the database built from the public resources. These publicly available models can be used for interactive comparison and analysis, including (but not limited to) flux balance analyses and detection of dead-end pathways (Ganter *et al.*, 2013).

BioModels

The BioModels database contains computational models of biological systems and processes, providing access to manually curated models taken from peer-reviewed scientific literature, some of which may be enriched with cross-data (Chelliah *et al.*, 2013). The database allows its users to store their models as well as retrieve others and provides extra services such as programmatic access for large-scale data retrieval. Their search criteria for retrieving models allow the user to search the curated models by taxonomy or by using the Gene Ontology tree

to select models of particular biological processes, cellular components or molecular functions (Chelliah *et al.*, 2013).

As well as manually curated genome-scale metabolic models, BioModels also contains automatically generated metabolic models that are created using pathway information from within other databases, such as the KEGG database. These automatically generated models are contained within their own part of the BioModels database: Path2Models (Chelliah *et al.*, 2013).

2.6 Conclusion

In summary, the microbiota is of supreme importance to its host and by studying dysbiosis, potential microbial symbioses and the metabolic capabilities of microbial communities, we may aid the development of treatments for dysbiosis-associated diseases. First we must begin by understanding the structure and ecology of microbial communities in dysbiosis, which is where we shall begin in Chapter 3, by using differential networks and the proportionality measure to explore the differences between eubiosis and dysbiosis.

After identifying potential altered dependencies between bacteria in disease cases, it is possible to model the metabolic capabilities of these communities, with the potential to relate that information to the altered dependencies. This is done by using the 16S rRNA taxonomic and abundance data to add weights to the metagenome-scale metabolic network, which could consequently reveal areas of metabolism that are significantly increased or decreased in the microbiota of disease cases. An new method for generating these inferred metanetworks is presented in Chapter 5 and potential analyses of these networks is discussed.

Finally, it may be possible to combine the information from the inferred metanetworks and dependency analyses to hypothesise about potential metabolic symbioses between constituents of the microbiome. Therefore, it would be beneficial to be able to generate genome-scale metabolic models that incorporate two or more cells with the aim of modelling metabolic symbioses. Given the limitations currently involved with using the semi-automatically gener-

ated models, the first step towards making these symbiotic models is the improved curation of single-cell genome-scale metabolic models. This was the motivation for the production of the curation software (ROOMM — Reaction Operations of Metabolic Models), which is presented and validated in Chapters 6 to 8.

Chapter 3

Differential networks: Development

3.1 Introduction

As discussed in Chapter 1, the human microbiota is the collection of all micro-organisms residing within the human host. It is a thriving ecosystem of microbes that differs across body sites and individuals, and it is becoming increasingly evident that the human microbiota is of supreme importance to its host. When the microbiota is disrupted and relative abundances of its constituents are shifted, this is known as dysbiosis, whereas the microbiota in healthy individuals is referred to as eubiosis.

Dysbiosis is often associated with a number of diseases and disorders, such as obesity, diabetes, psoriasis, inflammatory arthritis (and other autoimmune diseases), inflammatory bowel disease (IBD) and asthma (Buccigrossi *et al.*, 2013; Cho & Blaser, 2012; Human Microbiome Project Consortium, 2012; Balter, 2012; Greenblum *et al.*, 2012; Collins & Bercik, 2009). This is not to say that dysbiosis is the cause of such diseases — it may be the case for some, but for others dysbiosis might only be a signature of the disease, or one of the many contributing factors causing the disease phenotype. By understanding these dysbiotic communities better, we may be able to deduce some of the ecological mechanisms at play that determine how dysbiosis and associated disease symptoms arise.

Many studies of the microbiota compare relative abundances of different bacteria in dysbiosis and eubiosis, in order to highlight the signature of dysbiosis. However, comparisons of relative bacterial abundances in dysbiosis and eubiosis do not provide an immediate understanding of the interactions between bacterial populations within the microbiota, or the biochemical implications of dysbiosis. The aim of this work is to try to identify potential associations or interactions between pairs of bacteria in the microbiota (classified and clustered at the genus level) and identify how these associations have changed in dysbiotic communities when compared to eubiotic controls. The existing studies identifying the changes in associations between bacteria in cases and controls focus on the use of correlation to detect the associations. However, in this chapter, I explain why correlation (the most common measure of association) is ill-suited to identifying these associations within the microbiota, and proceed to focus exclusively on the development and application of my methods that instead make use of proportionality measures. The use of proportionality to identify the bacterial associations is novel, and so the associated statistical framework presented in this chapter is also new. However, this framework can also be applied using a measure of correlation, thereby allowing direct comparisons of the proportionality and correlation approaches. This chapter therefore includes:

- **Section 3.2:** an explanation of the problems with correlation and justification for the use of a proportionality measure with relative data,
- **Section 3.3:** validations for the application of the proportionality measure to microbiota data (with further details provided within Appendix B),
- **Section 3.4:** the presentation of the final method that has been developed for producing proportionality differential networks,
- **Section 3.5:** a null test to validate the final method,
- **Section 3.6:** a discussion of the limitations of the final method.

3.2 Correlation vs. goodness of fit to proportionality

It is possible to create differential networks using correlation and two-way permutation tests in order to investigate how potential dependencies between bacteria change within the microbiota when comparing disease cases to controls. However, the resulting differential networks are incredibly dense, and are not accurate measures of changes within the system. This is due to very high numbers of false positive results. Weiss *et al.* (2016) observed this issue alongside varied sensitivity and precision when comparing different correlation methods for analysing the microbiota, and they have highlighted the need for better dependency measures for microbiota data.

Some of the observed false positive results arise due to confounding variables caused by inter-dependencies between bacteria in the microbiota, which could lead to the overestimation of correlation between pairs of bacteria. Therefore, it is possible to apply partial correlation instead, with the aim of regressing away the variance caused by confounding variables. Although the resulting networks using partial correlation measures are less dense than the original correlation networks, they are still too dense to be useful or meaningful.

High levels of false positive results can also be due to problems arising from relative data. Lovell *et al.* (2015) have highlighted that correlation is not subcompositionally coherent when assessing relative data; this is an important factor to consider when assessing 16S rRNA gene sequence data, due to the effects of potential primer bias.

Here, it worth sharing an adapted example from their publication. Let's suppose that we use two amplification methods to subsequently identify correlation between bacteria (A – D) in a few samples (1 – 3). We split the samples in order to run them through two amplification and sequencing protocols. In the first protocol, we might use a primer that is able to successfully amplify all bacteria found within the samples. After amplifying and sequencing, we determine that there are four different bacterial populations (A – D) in the samples and normalise our results in each sample in order to ensure that the results are comparable across samples, as

shown in Table 3.1.

	A	**B**	**C**	**D**
Sample 1	0.1	0.2	0.1	0.6
Sample 2	0.2	0.1	0.1	0.6
Sample 3	0.3	0.3	0.2	0.2

Table 3.1: Hypothetical compositional abundances of four bacterial species in three samples.

However, let's suppose that in the second protocol, we use a different primer that has a natural bias. Here, bacterium D is not detected, giving rise to the composition shown in Table 3.2 after normalising:

	A	**B**	**C**
Sample 1	0.25	0.5	0.25
Sample 2	0.5	0.25	0.25
Sample 3	0.375	0.375	0.25

Table 3.2: Hypothetical compositional abundances of three bacterial species in three samples. These data represent a subset of those in Table 3.1.

The results from the second protocol are a subcomposition of those from the first, and both datasets were derived from the same original samples. Therefore, we would expect any statements we deduce about the associations between bacteria A, B, C and D to agree across both protocols. However, if we calculate the correlation of A – D using the relative data from the two protocols, we observe different results. The correlation values returned from the first protocol are shown in Table 3.3 while those from the second protocol are shown in Table 3.4.

	A	**B**	**C**	**D**
A	1	0.5	0.866	-0.866
B	0.5	1	0.866	-0.866
C	0.866	0.866	1	-1
D	-0.866	-0.866	-1	1

Table 3.3: Hypothetical correlation of the compositional abundances from Table 3.1, relating to four bacterial species in three samples.

Lovell *et al.* (2015) also illustrated why the use of correlation to measure dependencies between pairs of variables is inappropriate for relative data (such as 16S rRNA gene sequence data): even

	A	**B**	**C**
A	1	-1	NA
B	-1	1	NA
C	NA	NA	1

Table 3.4: Hypothetical correlation of the compositional abundances from Table 3.2, relating to three bacterial species in three samples.

if we observe correlation in our relative data, this does not necessarily mean that we observe the same correlation in the underlying absolute abundances that give rise to the relative data, thereby resulting in large numbers of false positive results.

As an illustration of how the appearance of relative data differs from absolute data, we can look at an example. If we imagine that there are five species of bacteria (A – E) with absolute counts and relative counts, they might show values such as those shown in Table 3.5.

	Absolute	**Relative**
A	20	0.097
B	22	0.106
C	40	0.193
D	45	0.217
E	80	0.386
Total	207	1

Table 3.5: Hypothetical absolute and relative abundances of five bacterial species in a single sample.

Now let's imagine that we had another sample where E is undetected, either because it is not present or because of experimental error. The data might look like those shown in Table 3.6.

	Absolute	**Relative**
A	18	0.136
B	24	0.182
C	45	0.341
D	45	0.341
E	0	0
Total	132	1

Table 3.6: Hypothetical absolute and relative abundances of five bacterial species in a second single sample.

Whilst the absolute abundance of A is lower in Table 3.6 when compared to Table 3.5, the relative abundance is higher. Similarly, the absolute abundance of D is unchanged, while the relative abundance has increased. Therefore, the relative abundances do not reflect the same direction of difference as the underlying absolute data, demonstrating why relative data should be treated with caution. Here, if this data was extrapolated, the absolute abundances would suggest that A and B are negatively correlated, whilst the relative data would suggest that they are positively correlated.

Instead of using correlation, Lovell *et al.* (2015) proposed that a measure of goodness of fit to proportionality may be used to assess dependencies between pairs of variables when using relative data. This is because if two variables are strongly proportional, the ratio of the relative abundances for A and B will remain fairly constant across all samples, as will the ratio of the absolute abundances, A/B. The more variable the ratio of A/B across all samples (i.e. the more it deviates from a constant value), the less strongly proportional they are. The reader is referred back to Section 2.3.4 for the full introduction to the proportionality measure proposed by Lovell *et al.* (2015).

Here, I have used the goodness of fit to proportionality measure and bootstrap subsampling methods to generate differential networks. The resulting differential networks are weighted networks where the edges represent a significant change in proportionality between each pair of variables under two different circumstances or environments. In the case of the human microbiota, we can create differential networks to show how associations between pairs of bacterial genera change in disease cases (dysbiosis) compared to controls (eubiosis). These observed changes in associations may allow us to hypothesise about potential changes in symbiotic relationships within the microbiota during dysbiosis. Ultimately, if we can identify how these proportionality relationships change in dysbiosis, we may better understand the ecology of the microbial community and how this affects the phenotype of the human host. Validation studies investigating the use of the proportionality measure and bootstrap resampling to assess genus associations within the microbiota are summarised in Section 3.3 and further details are presented in Appendix B.

3.3 Validation studies

Although correlation measures are definitely inappropriate for use with 16S rRNA gene abundance data, it is also worth validating the use of proportionality measures for identifying associations between constituents of the microbiota.

3.3.1 Validation studies using simulated data

Unlike the correlation coefficient (r), the proportionality measure (ϕ) does not have a clear and intuitive interpretation beyond the fact that when $\phi = 0$, two variables are perfectly proportional. In order to calibrate ϕ against typical features of the microbiota data, it is possible to simulate data with known distributions and known relationships between variables.

For example, the presence of low-abundance bacteria gives rise to very zero-rich datasets, which requires important consideration because ϕ calculations cannot utilise zero values due to their use of logarithmic transformations. For a given pair of variables, any samples where one or both abundance values are zero (i.e. where $A = 0$ and/or $B = 0$) are referred to as 'zero pairs' and must be removed from the analysis because they cannot be used to calculate ϕ. Examples of these zero pairs are shown in Table 3.7.

Sample ID	Genus A	Genus B
Sample 1	0	19
Sample 2	74	0
Sample 3	185	21
Sample 4	0	0
Sample 5	115	10
⋮	⋮	⋮

Table 3.7: Example abundance data for two genera, referred to here as A and B. Samples 1, 2 and 4 (shown in grey) are referred to as 'zero pairs' for these two genera due to the fact that each of these samples contains at least one 0 value. These samples cannot be used to calculate ϕ for this given pair of genera because $\log 0$ is undefined. Conversely, samples 3 and 5 are defined as 'non-zero pairs' because both genera have non-zero values that can be used to calculate ϕ.

By creating datasets with varied distributions and transformations, we can reflect the sort of

data that is observed for high- and low-abundance bacteria within the microbiota, including zero-rich bacteria that give rise to many zero pairs when calculating ϕ values. Validation studies using simulated data reflecting potential relationships between these high- and low-abundance bacteria are described in detail in Appendix B.0.1, with a focus on the effects of omitting zero-rich data. The results of these validation studies showed that zero-rich data give rise to lower ϕ values; this observed result was then tested using real microbiota data, which is discussed next.

3.3.2 Observing distributions of ϕ values for varying levels of sparsity

Real microbiota datasets always contain some zero-rich variables due to the presence of rare bacteria that are present in very few individuals. However, the presence of zero-rich bacteria will influence the overall distribution of ϕ values observed within a real dataset, given that the simulated data validations showed that zero-rich data tend to give rise to lower ϕ values. In order to confirm this using real microbiota data, the distributions of ϕ values were observed for different levels of sparsity in 16S rRNA gene sequence data.

In order to observe the effect of high levels of zeros upon the distributions of ϕ measurements calculated from real microbiota data, we observed the distributions of ϕ values produced by real data with different proportions of zeros in it. To do this, a threshold was applied such that we began by including all pairs of genera when we calculate the pairwise ϕ values for a given dataset by calculating ϕ for every pair of genera. This was then repeated including only pairs of genera that have more than 10% of samples providing non-zero pairs. Again, this was repeated including only pairs of genera that have more than 20% of samples providing non-zero pairs, and so on for each 10% interval up to 100% non-zero pairs of data. Therefore, as we progressed, the data being observed was less and less sparse.

Here, this thresholding method was applied to 16S rRNA count data for two previously pub-

lished datasets from a study of the gut microbiota in inflammatory bowel disease (IBD) (Morgan *et al.*, 2012). It was applied to 27 samples taken from healthy controls and 43 samples taken from patients with ileal Crohn's disease (iCD).

Figure 3.1 shows the distributions of ϕ values between all pairs of genera identified within the 27 healthy individuals (top row) and the 43 patients with ileal Crohn's disease (iCD; bottom row) when we include all data (left column) and when we include only the data with at least 20% non-zero pairs (right column). When we include all pairwise ϕ values for each dataset, we observe more values at the lower end of the range. Therefore, the ϕ values show skewed distributions for both datasets (healthy and iCD) with an over-representation of highly proportional pairs of bacterial genera. However, there is also a progressive reduction in low ϕ values as sparsity decreases; as we eliminate pairs of genera that have zero pairs in their data, there are fewer values observed at the lower end of the range. These results agree with the observations in the validation studies using simulated data (discussed above and in Appendix B.0.1) where zero-inflated data showed lower ϕ values.

3.3.3 Observing the effects of sample size upon ϕ values

Due to the evidence that zero-rich variables skew the distributions of observed ϕ values in real data due to their lower ϕ values, it was imperative to determine whether this bias is due to characteristics of the zero-rich data or the reduction in sample size due to the removal of zero pairs from the analysis. For this reason, simulated zero-rich data was used to show how sample size affects observed ϕ values, thereby confirming that the skew observed due to zero-rich data is due to the reduced sample sizes. The issue of zero-rich data is therefore overcome by increasing sample size. The validation studies carried out to investigate this matter and draw these conclusions are presented in Appendix B.0.3.

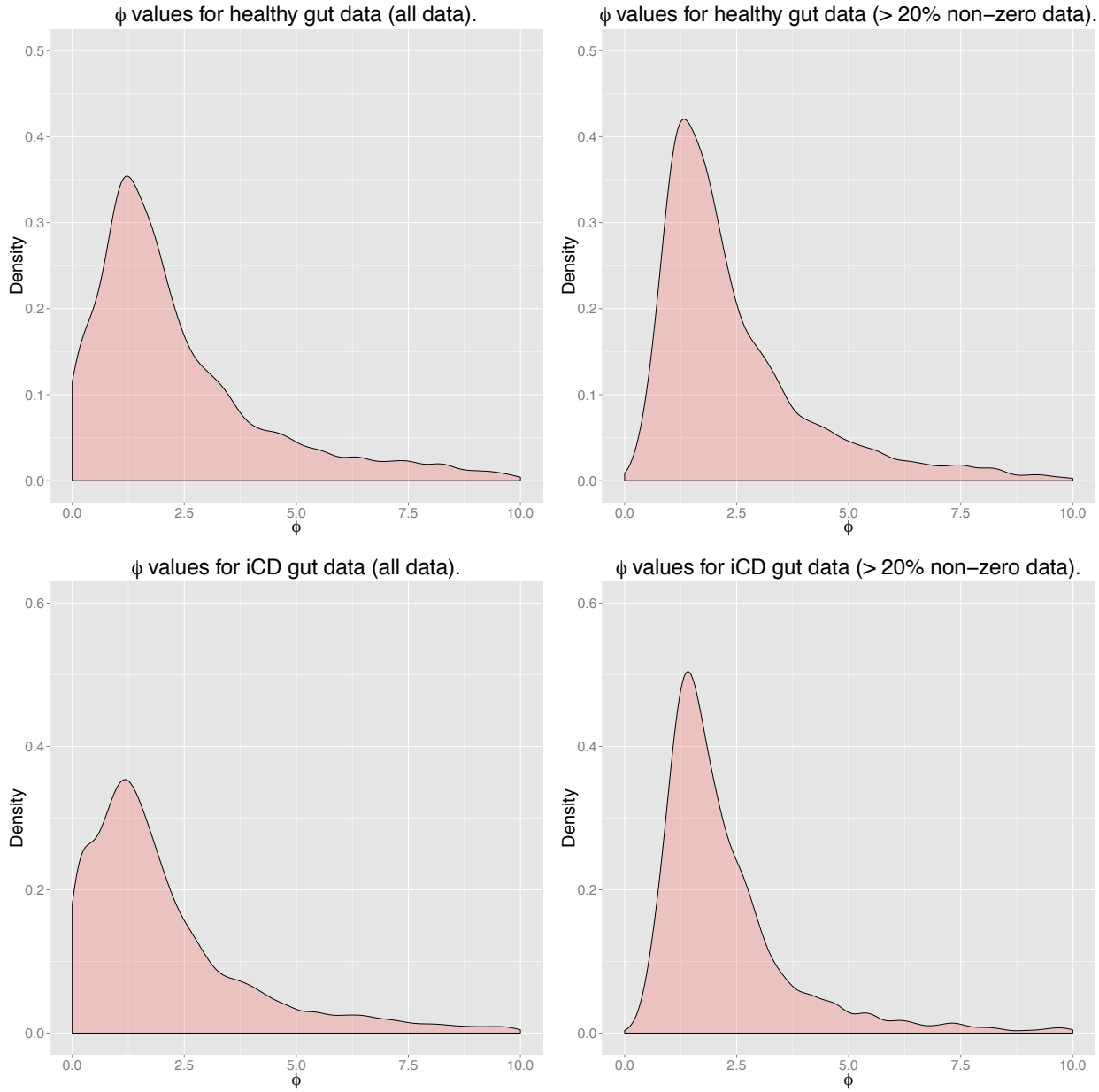


Figure 3.1: *Density plots showing the distributions of ϕ values for two datasets.* **Top left:** *ϕ values for all pairs of genera identified within samples taken from healthy individuals. All pairwise ϕ values were included, regardless of the proportion of non-zero data available for the calculation of ϕ.* **Top right:** *ϕ values for pairs of genera identified within samples taken from healthy individuals where $>$ 20% of the data pairs available for the calculation of ϕ were non-zero values.* **Bottom left:** *ϕ values for all pairs of genera identified within samples taken from individuals with ileal Crohn's disease (iCD). All pairwise ϕ values were included, regardless of the proportion of non-zero data available for the calculation of ϕ.* **Bottom right:** *ϕ values for pairs of genera identified within samples taken from individuals with iCD where $>$ 20% of the data pairs available for the calculation of ϕ were non-zero values.*

3.3.4 Validating the use of bootstrap empirical distributions

Finally, empirical bootstrap distributions of ϕ values can be used to perform a non-parametric test to compare ϕ values in case and control datasets. However, bootstrap resampling involves sampling with replacement, which means that the same sample can be selected numerous times within a single bootstrap run, thereby potentially increasing the apparent proportionality of the data. For this reason, a validation test using simulated data was carried out to ensure that there is no directional bias in the bootstrap estimator of ϕ. This test was carried out for varying levels of proportionality and different numbers of bootstrap repeats; no directional bias was observed in ϕ, so bootstrap resampling was not deemed inappropriate for generating empirical distributions of ϕ values. This validation study is presented in detail in Appendix B.0.4.

3.4 Proportionality Differential Networks: The method pipeline

Here, the overall structure of the proportionality differential network method is introduced. The main steps of the method pipeline are described in order to give the reader a high-level understanding of each step. This section lays the groundwork in order for the details of the method to be discussed in Section 3.4.2.

3.4.1 Method outline

The overall structure of the differential network method is shown in Figure 3.2. The basic logic is based upon the idea of comparing the ϕ values for a given pair of genera in the case dataset to the equivalent ϕ values in controls (where ϕ is calculated using the goodness of fit to proportionality, as discussed previously in Section 2.3.4). This comparison is done using the

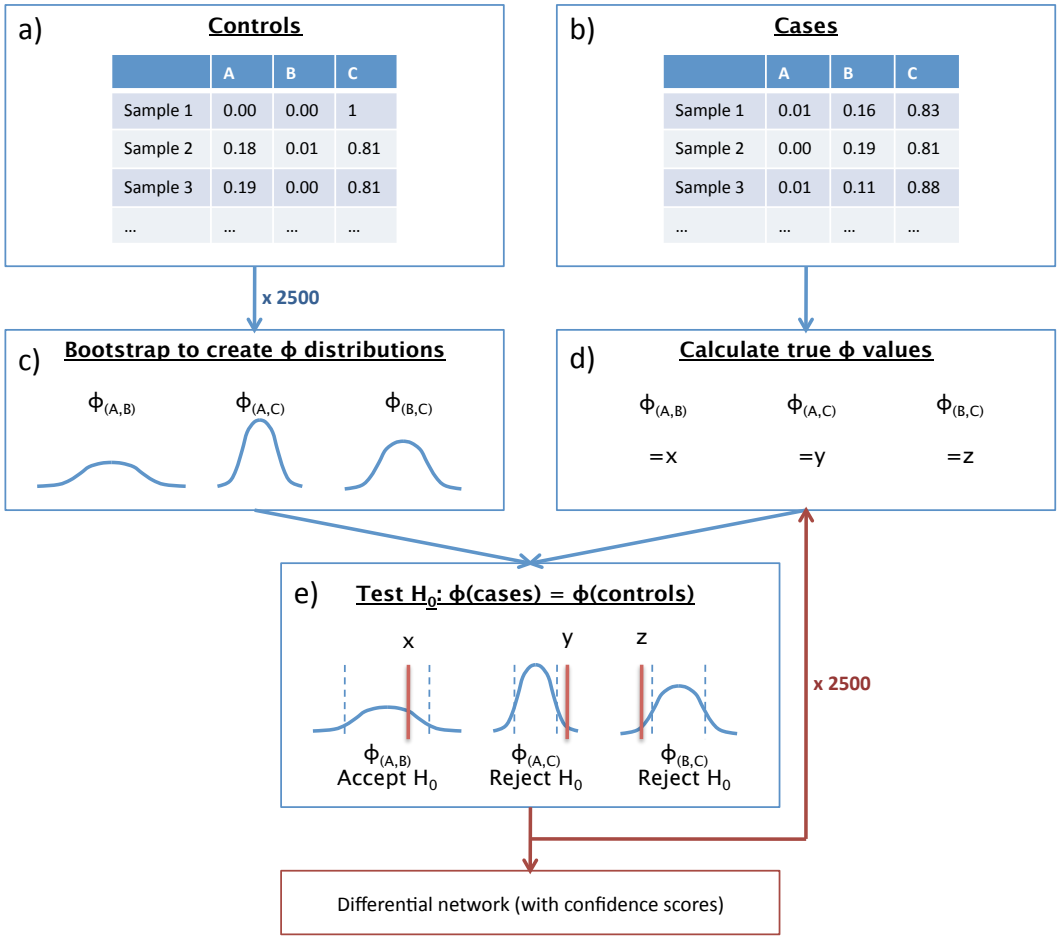


Figure 3.2: **a) & b)** *The input data for case and controls is in the form of two tables with bacterial genus assignments across the columns and each sample in a different row.* **c)** *For each pair of genera within the control dataset, bootstrap resampling is used to sample the data prior to the calculation of the associated ϕ values. This process of bootstrap resampling and ϕ calculation is repeated 2500 times in order to produce null distributions of ϕ values for each pair of genera, as shown by the distribution curves illustrated in blue.* **d)** *For each pair of genera within the case data, samples are subsampled without replacement prior to calculating the pair of 'true' ϕ values for the given genus pair. This enables matching of sample size in cases to that in controls in order to allow for a fair comparison of the resulting ϕ values.* **e)** *The 'true' ϕ values (x, y and z, shown as red lines) are compared to the corresponding bootstrap distributions of control ϕ values (shown as blue distribution curves) to determine whether they are statistically significantly different to the control values. This is determined by whether the case values fall within the tails of the control distributions (marked by blue dashed lines).* **Red box & arrows:** *The process of subsampling from cases followed by the comparison of case ϕ values to the control ϕ distributions is repeated 2500 times (shown by the red arrow). This is due to the fact that the necessary subsampling of case data (to maintain sampling distribution similarity with controls) means that the result may be subtly different each time this process occurs. By repeating the process, we can generate a measure of certainty (confidence scores) for these edges in the final differential network (marked by the red box).*

case ϕ value (Figure 3.2 d)) and an empirical bootstrap distribution that is constructed for the equivalent ϕ value in the control dataset (Figure 3.2 c)), as shown in Figure 3.2 e).

For the case data, ϕ values are measured for each pair of genera, which gives rise to two directed proportionality values for each pair of genera; for example, genera A and B would give rise to $\phi_{(A,B)}$ and $\phi_{(B,A)}$. The collection of all ϕ values for all pairs of genera can be represented as a directed network with two opposing, weighted edges (ϕ values) between each pair of genera (nodes).

For the control data, the samples for each pair of genera are subsampled using bootstrap subsampling and the resulting ϕ values are calculated. As with the case data, this gives rise to two directed proportionality values for each pair of genera. This process of bootstrap resampling and ϕ calculation is repeated to generate an empirical distribution of ϕ values for each directed edge within the control network.

The case and control networks can then be used to create a differential network. For each edge in the case and control networks, we begin with the null hypothesis that there is no significant difference between the ϕ value in cases when compared to the equivalent edge in controls. Therefore, if the case ϕ value lies within the tails of the control distribution (i.e. beyond a given threshold, α), we reject the null hypothesis, which states that there is no significant difference between the case and control ϕ values. This comparison step is illustrated in Figure 3.2 e).[1] However, in order to correct for multiple testing, a p-value must be calculated for each edge before declaring which edges are significantly different.

The p-value is determined by observing where a given case ϕ value (ϕ_{case}) falls within the corresponding ordered population of control bootstrap ϕ values (Figure 3.2 e)). If ϕ_{case} is

[1] Although we could generate a bootstrapped empirical distribution for each pair of genera in cases and compare it to the corresponding bootstrapped empirical distribution for controls using the Kolmogorov-Smirnov (KS) test, this method would be inappropriate here. This is because the KS test is dependent on sample size, so the larger the sample size (i.e. the more bootstrap repeats we run), the more significant the observed difference will be between the case and control distributions. This means we can make any result significant by increasing the number of bootstraps until the p-value falls below our chosen significance threshold, no matter how small the difference may be between the two distributions. The only case where there would be no difference between the two distributions is if we had exactly the same data for the two sets.

smaller than the median of the bootstrap control population, N is the number of values that lie to the left of ϕ_{case} in the bootstrap control population. Conversely, if ϕ_{case} is larger than the median of the control ϕ bootstrap population, N is the number of values that lie to the right of ϕ_{case}. The p-value is then determined as:

$$p = \frac{2N}{n} \tag{3.1}$$

where n is the number of bootstrap repeats (i.e. the number of values) forming the bootstrap control ϕ distribution. The set of p-values for all pairs of genera are then adjusted using the false discovery rate (FDR) to account for multiple testing (Storey, 2003).

The adjusted p-value for each edge allows us to test the following null hypothesis:

> H_0 : The proportionality measure in cases is not significantly different from the corresponding proportionality measure for controls.

Therefore, the association between a given pair of genera was deemed to be significantly different when $p_{adjusted} < 0.05$ for $\phi_{(A,B)}$ and $\phi_{(B,A)}$. This gives rise to a single undirected edge between the given pair of genera and the resulting differential network is undirected; the edge weight is the change in ϕ, which is the difference between ϕ_{case} and the median of the control distribution of ϕ values (shown in Figure 3.3), denoted as δ.

3.4.2 Addressing the details

Whilst the comparison of ϕ values in cases to the equivalent bootstrap distributions in controls provides the basic premise of this method, there are a number of other factors that need to be taken into account when generating the differential networks. This includes:

- addressing the effects of zero values (due to the fact that $\log 0$ is undefined), such as handling data with insufficient non-zero pairs (which often gives rise to $\phi = \infty$ or causes

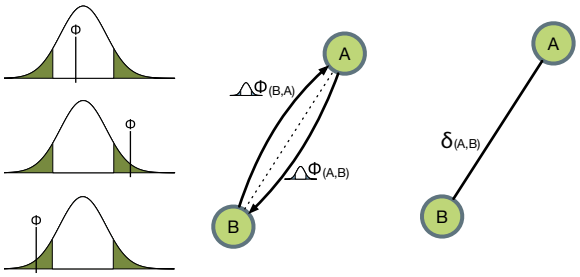


Figure 3.3: **Left:** *Each distribution represents the bootstrap distribution of ϕ values for a given directed edge in the control network. The vertical line represents the ϕ value for the corresponding edge for the case network. In green, the tails of the distributions beyond a given threshold, α, are marked to show the regions where a ϕ_{case} value would be considered to be statistically significantly different from control ϕ values. If the ϕ_{case} shown by the vertical line does not fall within the tails of the distribution, the case value is not deemed statistically significantly different to the controls.* **Centre:** *The pair of control edges are compared to their associated case values. The dotted line represents the resulting undirected edge within the differential network.* **Right:** *If both the directed edges are significantly different in cases when compared to controls, a single undirected edge is added to the differential network. The change in ϕ, denoted by δ, is the difference between ϕ_{case} and the median of the control distribution of ϕ values.*

ϕ to be undefined),

- ensuring the sampling distributions for cases and controls are as similar as possible by matching sample sizes via subsampling of cases, and subsequently making use of this process to provide a measure of certainty,

- making sense of the asymmetrical ϕ values for each pair of genera, and calculating the effect size, δ (the change in ϕ between cases and controls), in light of repeated subsampling of cases.

These factors and the resulting solutions are all discussed below.

Handling data with insufficient non-zero pairs

Due to the fact that $\log 0$ is undefined, 0 values cannot be included in the calculations of ϕ. Therefore, this is dealt with by only including samples where both variables have non-zero values, which is determined independently for each pair of genera. As discussed previously, data rich in 0 values tend to show a bias towards lower ϕ values due to reduced sample sizes.

In some instances, a genus that is very rich in zeros will not contain enough non-zero pairs of data in order to calculate ϕ when paired with another zero-rich genus. This gives rise to an undefined ϕ value (returned as 'NA'), which causes that pairwise comparison to be excluded from the analysis. Therefore, in this instance, the bias towards low ϕ values in zero rich data will not be observed, and therefore will not skew the results.

If two genera in the first dataset are zero-rich but there is sufficient data to calculate ϕ, the resulting ϕ value may be biased towards a low value. However, if the two genera are also zero-rich in the second dataset, it will have a similar bias, thereby making the ϕ values comparable across the case and control datasets. Once again, in this instance, the bias towards low ϕ values in zero rich data will not skew the results.

However, if the two genera in the second dataset are not also zero-rich, they will not be biased towards a low ϕ value. If they are high-abundance genera in one dataset but not in the other, they are also likely to display altered relationships, so we would expect to see an edge between these genera in the differential network, regardless of whether or not the reduction in sample size skews the ϕ values for the zero-rich data.

In other cases, the removal of zero pairs may leave only low abundance data and therefore one or both variables will only have abundance values of 1, which gives rise to a genus that has an abundance variance of 0 across its non-zero samples. Ultimately, this can cause ϕ to be undefined (in this case returned as 'NaN' values) or $\phi = \infty$. Instances where $\phi = \infty$ will be shown in the differential network where their δ value will have a magnitude of ∞, so they are easily identified. Any pairs of genera with undefined ϕ do not have edges within the differential network and a comparison cannot be made between cases and controls. In these instances, omission from the network is appropriate because we do not have enough data with which to come to an accurate conclusion about whether ϕ is different in cases compared to controls.

Sampling distributions and measuring stability

In order to ensure that the ϕ values in cases (ϕ_{case}) are comparable to the corresponding ϕ values within the control bootstrap distribution ($\phi_{control}$), we must keep the sampling distributions as similar as possible for each ϕ comparison. Therefore, we must use the same number of samples to calculate ϕ_{case} as is used to calculate each $\phi_{control}$ value. This is done by subsampling from cases without replacement and is explained fully below.

The number of samples available to calculate the ϕ values for a given pair of genera is dependent upon the number of non-zero pairs of data available for that pair of genera in cases and controls. If we define $U_{controls}$ as the number of usable samples for a given pair of genera in controls and U_{cases} as the number of usable samples for the same pair of genera in cases, the maximum possible number of samples that can be used for bootstrap sampling of controls and subsampling of cases, u, is defined as:

$$u = \min(U_{controls}, U_{cases}) \tag{3.2}$$

When $u = U_{controls}$, (i.e. the number of samples is limited by controls) we will obtain subtly different results each time the analysis is run due to the fact that we are subsampling from cases in each run. Whilst this might not seem ideal, we can capitalise on this fact because it allows us to incorporate a measure of certainty, S, if we run the comparison analysis (i.e. subsample the case dataset) numerous times (see red arrows and red box in Figure 3.2):

$$S = \frac{x}{N} \tag{3.3}$$

where x is the number of times a given edge is determined to be significantly different in cases when compared to the control distribution, and N is the number of subsampling repeats.

However, when $u = U_{cases}$, this would give rise to a certainty score of 1 because the subsampled set will be the same as the original set every time the analysis is run. Therefore, the number of samples for subsampling and bootstrapping each edge, y, is defined as $y = u - 1$. This allows

for the certainty to be measured for all edges, regardless of whether u is limited by cases or controls. By including this subsampling, we can make the sampling distributions for cases and controls more comparable whilst assessing the certainty of each edge in the differential network.

In an ideal world, we would not subsample at all and would just compare the ϕ values calculated from the raw data for one phenotype (e.g. cases) against the empirical bootstrap distributions for the other (e.g. controls) — we subsample because we have to match the usable sample size for cases and controls. Subsampling without replacement was originally chosen as a way of culling the dataset down to size to match the sample size of the comparison (bootstrapped) set. The decision to exploit the variability in taking the subset by repeated subsampling came from the fact that each run of the algorithm would bring a subtly different result due to subsampling. However, the fact that the observed variance of the statistic is dependent on the subsample size when sampling without replacement means that, in future developments of the method, the method would be improved by subsampling with replacement in order to capture the true variance of the data. In this case, we would also be able to maintain use a sample size of u instead of $u - 1$.

Addressing the asymmetry of ϕ and calculating effect size

Given that each pair of genera gives rise to two ϕ values (two directed edges), we must reconcile these two edges to produce a final undirected differential network. Therefore, if both of the ϕ values are identified as being statistically significantly different between cases and controls (i.e. $p_{adjusted}$ <0.05 for $\phi_{(A,B)}$ and $\phi_{(B,A)}$, as described in Section 3.4.1), that pair of genera is given an edge in the undirected differential network. This process is shown in Figure 3.3 above.

The effect size, δ, is measured for each directed edge for two genera (A and B) by calculating the difference between the case ϕ value (ϕ_{cases}) and the median of the empirical control distribution ($\tilde{\phi}_{controls}$) each time the cases are subsampled, as follows:

$$\delta_{(A,B)} = \tilde{\phi}_{(A,B)controls} - \phi_{(A,B)cases} \tag{3.4}$$

$$\delta_{(B,A)} = \tilde{\phi}_{(B,A)controls} - \phi_{(B,A)cases} \tag{3.5}$$

For the finished undirected differential network, the mean of all $\delta_{(A,B)}$ and $\delta_{(B,A)}$ values for each pair of genera is used to calculate the effect size for the undirected network ($\delta_{(B,A/B,A)}$):

$$\delta_{(B,A/B,A)} = \frac{\bar{\delta}_{(A,B)} + \bar{\delta}_{(B,A)}}{2} \tag{3.6}$$

3.5 Null test

In order to check that we do not get excessive false positive results when using the proportionality measure to create differential networks, the finished method described in Section 3.4 was tested using two null tests.

The first test was implemented using previously published 16S rRNA gene count data taken from patients suffering with ulcerative colitis (UC) (Morgan *et al.*, 2012). Of the 76 samples taken from patients, 15 were biopsy samples and 61 were stool samples. These biopsy and stool samples were each split into two random sets (8 and 7 biopsy samples; 30 and 31 stool samples) which were combined to make two mixed datasets containing 38 samples each, as shown in Table 3.8.

Null set	**Biopsy**	**Stool**
Set 1	7	31
Set 2	8	30

Table 3.8: 76 samples taken from patients with ulcerative colitis (UC) were split into two groups of samples containing 38 samples each, formed from a mixture of biopsy and stool samples. The biopsy/stool composition of these two groups of samples are shown here.

The first set was submitted as the 'control' set to undergo bootstrap resampling and the second set was submitted as the 'case' set. These sets were run with 2500 iterations (i.e. 2500 bootstrap repeats for *Set 1* and 2500 resampling repeats for *Set 2*).

The total number of possible edges in the network is dependent on the number of genera within

the dataset, n, and is given by Equation 3.7, where n is the number of genera within the UC dataset:

$$\frac{n(n-1)}{2} \tag{3.7}$$

For the UC dataset above, the total number of possible edges is 18721 ($n = 194$). The first test revealed 57 edges with a non-zero certainty score in the differential network. Therefore, the 57 observed edges with non-zero certainty scores in the network represent 0.3% of the possible edges.

Ideally, no edges would be returned for the null test due to the fact that the FDR multiple testing correction has been applied within the differential network method. However, even though the data was split evenly to contain similar proportions of stool/biopsy samples in the two sets, the datasets are still different to one another. Therefore, the fact that we still observe edges within the resulting network suggests that the method is highly sensitive to small differences in the two datasets.

In order to check that these results arise in the first null test due to high sensitivity, we also ran a null test where the case and control datasets were identical. Therefore, the comparison of a dataset containing 18 healthy samples was compared against itself by submitting the same data for cases and controls. As expected, this returned no edges within the resulting differential network.

3.6 Limitations

The proportionality measure is more appropriate than correlation when identifying dependencies in relative data but, as with all methods, the proportionality differential network method has its own set of limitations. These limitations include:

- comparison limitations — the method can only compare bacterial relationships that are present in cases and controls,

- the omission of low abundance bacteria,
- the identification of only positive proportional interactions,
- the asymmetry of ϕ values,
- the asymmetry of the method.

These limitations are discussed below.

3.6.1 Comparison of bacteria intersection

The comparison of the case and control networks of bacterial associations can only be conducted for pairs that are present both in cases and controls. Therefore, this differential network method causes us to lose information about bacteria that are present in one dataset but not in the other. It can be argued that some of these bacteria may be some of the most important, as they may be some of the key bacteria that differentiate the cases from controls. However, it is also known that many bacteria observed in individuals simply represent the known interpersonal variation of bacteria. When this is the case, these bacteria are less likely to be of interest with respect to disease mechanisms caused by dysbiosis. Also, these bacteria will tend to be very zero-rich due to the fact that they will only be observed in a few individuals, so even if they could be included in the analysis, they would most likely have undefined ϕ values.

3.6.2 Omission of low abundance bacteria

When genera are zero-rich (due to high interpersonal variation across hosts) but are observed within cases and controls, they have the potential to be included within the differential networks. However, if there are not enough positive (non-zero) samples to calculate ϕ for a given pair of genera, ϕ is undefined and that edge is omitted from the analysis. However, omission of such edges as these is unlikely to be detrimental, as these edges are likely to represent the

interpersonal variation between hosts rather than the broader differences between cases and controls.

3.6.3 Limitation of ϕ to identify only positive proportional interactions

It is only possible to compare positive linear associations between genera in cases to those observed in controls, due to the way ϕ is calculated. ϕ measures the goodness of fit of the genus data to a line following the general equation of $y = mx$, where y represents the count variables for one genus and x represents the count values for another genus. The fact that the count data must always be positive (i.e. lie within the positive quartile of a graph) means that any negative linear associations between pairs of genera will not fit a line of $y = mx$ but would instead fit a line of $y = mx + c$, where c must be a positive integer. The addition of a positive, non-zero integer immediately implies that y is not perfectly proportional to x and therefore gives rise to a greater (non-zero) ϕ value. This non-zero ϕ value cannot be differentiated from a higher ϕ value caused by a non-linear relationship or by increased levels of noise in the data. This consideration is important because some interactions between bacteria may result in non-linear abundance relationships between bacteria. For example, if one population of bacteria is dependent upon another for the acquisition of a particular metabolite, we may observe a saturation point where this metabolite is no longer the limiting factor for the growth of the dependent bacteria, thereby showing a non-linear dependency thanks to the observation of a saturation curve. Given the linear nature of the ϕ measurement, the nature of the change in dependancy cannot be determined by ϕ alone, and further analysis is required. This is one of the reasons why this differential network method is best suited as an exploratory and hypothesis-generating method prior to further analysis. However, given that correlation measures are not appropriate for use with relative data and are still limited to linear or monotonic relationships, the ϕ measure is still a more appropriate measure than correlation.

3.6.4 Asymmetry of ϕ values

As discussed in Section 2.3.4, two ϕ values are calculated for each pair of genera, which are not symmetrical unless the two genera are perfectly proportional. Each single edge occurs in the undirected differential network when both of the directed ϕ values are significantly different in cases when compared to controls. There is no biological meaning behind the asymmetry of the pair of directed ϕ values — their asymmetry is simply a characteristic of the ϕ calculation. Therefore, the lack of biological meaning behind the asymmetry of ϕ is the reason why each pair of ϕ values is reconciled to give rise to a single undirected edge in the final differential network.

3.6.5 Asymmetry of the method

It is possible for the resulting differential network to differ when the comparison is run in reverse (i.e. when case data is bootstrapped and control ϕ values are compared to the bootstrapped case values). The reasons for this are most easily explained with the aid of a diagram, as shown in Figure 3.4, alongside a reminder of the method.

Each case ϕ value produced for a given edge (via subsampling) is compared to the corresponding bootstrap distribution for controls. The confidence value is determined by the proportion of subsampled ϕ values in cases that fall within the tails of the bootstrap ϕ distribution for controls. In Figure 3.4 (Top left), the distribution of bootstrap ϕ values for controls for a given edge is shown in blue. The tails of the distribution are shown with blue shading beyond the given threshold values, z_1 and z_2. The distribution of subsample ϕ values for the equivalent edges in cases is shown in green. Here, when the case values are compared to the control bootstrap distribution, the certainty value, s, will be equal to 1. This is because all $\phi_{case} > z_2$, so all ϕ_{case} values calculated after subsampling would be considered to be statistically significantly different to $\phi_{controls}$, as shown by the green hashed area.

If we now consider the reverse comparison shown in Figure 3.4 (Bottom left), we observe a similar result. Here, the control ϕ distribution is again shown in blue, but this time it will have been generated using subsampling without replacement. Each value within that control

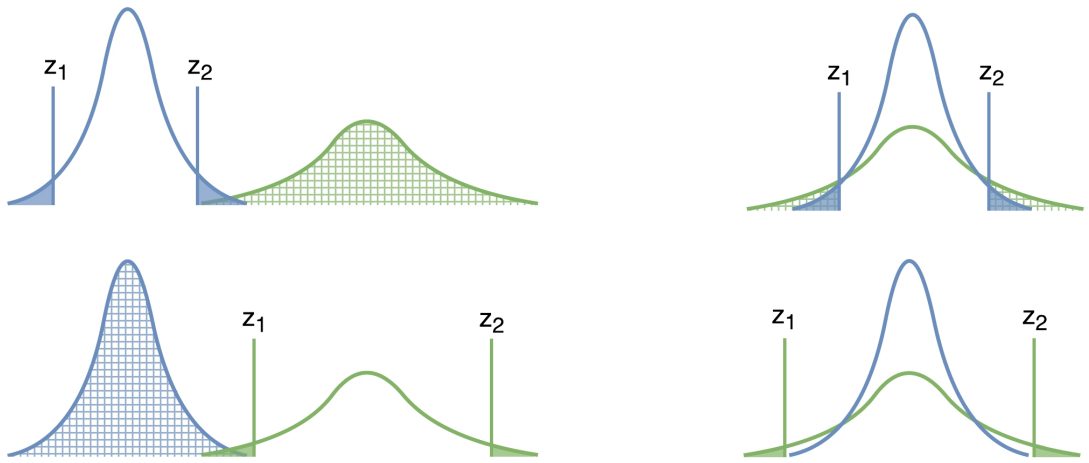


Figure 3.4: *Schematic diagram showing the comparison of ϕ_{case} values (shown as the blue distributions) to $\phi_{control}$ values (shown as the green distributions). The distributions showing coloured tails represent the bootstrap distribution where z_1 and z_2 represent threshold values determined using a given α probability cutoff value. These are the distributions against which the values from the other distribution will be compared. The alternative distributions (without marked z values) represent the distribution of values produced via subsampling without replacement to generate values to compare to the corresponding bootstrap distribution. The hashed areas represent the parts of these alternative distributions that lie to the left of z_1 and the right of z_2 (i.e. where these alternative distributions lie across the tails of the bootstrap distributions). The area of these hashed sections represent the certainty score, s, that is calculated for a given edge in the differential network method.* **Top left:** *Here, $\phi_{control}$ values are generated via bootstrap resampling to generate a distribution and identify cutoff z values. The ϕ_{case} values are generated via subsampling with replacement where each ϕ_{case} value is compared to the $\phi_{control}$ bootstrap distribution to determine whether the ϕ_{case} value is significantly different to those for the controls.* **Bottom left:** *Here, ϕ_{case} values are generated via bootstrap resampling to generate a distribution and identify cutoff z values. The $\phi_{control}$ values are generated via subsampling with replacement where each $\phi_{control}$ value is compared to the ϕ_{case} bootstrap distribution to determine whether the $\phi_{control}$ value is significantly different to those for the cases.* **Top right:** *Here, $\phi_{control}$ values are generated via bootstrap resampling to generate a distribution and identify cutoff z values. The ϕ_{case} values are generated via subsampling with replacement where each ϕ_{case} value is compared to the $\phi_{control}$ bootstrap distribution to determine whether the ϕ_{case} value is significantly different to those for the controls.* **Bottom right:** *Here, ϕ_{case} values are generated via bootstrap resampling to generate a distribution and identify cutoff z values. The $\phi_{control}$ values are generated via subsampling with replacement where each $\phi_{control}$ value is compared to the ϕ_{case} bootstrap distribution to determine whether the $\phi_{control}$ value is significantly different to those for the cases.*

distribution will be compared to the case ϕ distribution (shown in green), which will have been generated using bootstrap resampling. The tails of the case distribution are marked in green shading beyond the given threshold values, z_1 and z_2. Here, when the control values are compared to the case bootstrap distribution, the certainty value, s, will be equal to 1. This is because all $\phi_{control} < z_1$, so all $\phi_{control}$ values would be considered to be statistically significantly different to ϕ_{cases}, as shown by the blue hashed area.

For these first two examples, the comparisons of ϕ_{cases} to $\phi_{controls}$ are evenly matched such that the resulting edge would either be present in both differential networks or neither. However, there are situations where the comparison is not evenly matched, which is illustrated in Figure 3.4 (Top right and Bottom right).

Let us consider Figure 3.4 (Top right). Here, as in Figure 3.4 (Top left), the distribution of bootstrap ϕ values for a given edge in controls is shown in blue. The tails of the distribution are marked with blue shading beyond the given threshold values, z_1 and z_2. The distribution of ϕ values for the equivalent edge in cases is shown in green. The bounds of the case distribution encompass those of the control distribution. In this example, the corresponding edge in a differential network would have a non-zero confidence score (represented by the area under the curve marked with green hash). However, the effect size (i.e. the difference in median values) would be very small in this situation.

Conversely, if we now consider the reverse comparison shown in Figure 3.4 (Bottom right), we observe a different result. Here, the control ϕ distribution is again shown in blue, but this time it will have been generated using subsampling without replacement and each value within that distribution will be compared to the bootstrapped case ϕ distribution shown in green. The tails of the case distribution are marked in green shading beyond the given threshold values, z_1 and z_2. Here, when the control values (shown in blue) are compared to the case bootstrap distribution, the confidence value, s, is zero. This is because all $\phi_{control}$ values fall between z_1 and z_2 rather than in the tails. Therefore, no $\phi_{control}$ values would be considered to be significantly different to ϕ_{cases}.

As a result of the asymmetry of this method, one measure that could be considered in order to give symmetrical results would be to use the area under the ROC curve. The area under the ROC curve represents the probability that a random sample from one distribution is larger than a random sample from the other. Therefore, if the two distributions are identical, the area under the ROC curve would be 0.5. For a given pair of distributions, the area under the ROC curve would give x and $x - 1$, depending on which distribution is used as the reference distributions, thereby giving symmetrical results. Therefore, for example, if the two distributions did not overlap at all, the area under the ROC curve would be either 0 or 1, depending on which distribution is your reference, although by convention, the value above 0.5 would be used. This measure could potentially be applied in future developments of the differential network method.

However, due to the current asymmetry of this differential network method, it is important that the resulting networks are interpreted fully. This means that we must carefully consider the effect size and confidence scores in conjunction together in order to gain a fuller understanding of the changes in association. For example, if a given edge has a non-zero confidence score but small effect size, this may indicate that the ϕ distribution for the subsampled set is broad and therefore displays greater variability than the equivalent distribution for the bootstrapped population. Furthermore, this asymmetry may also be exploited to understand our data more fully. The analysis may be run twice such that one input set is subsampled the first time and bootstrapped the second time. The two resulting networks may be interpreted in conjunction to give us a better understanding of the comparison of ϕ values in cases to those in controls.

3.7 Conclusions and future work

In this chapter, I have discussed why correlation measures are inappropriate for measuring dependencies in relative data. I have subsequently explored the use of a proportionality measure, ϕ, in place of correlation and calibrated the measure against typical features and relationships observed in microbiota data. This process highlighted the need for substantially large datasets to overcome the reduction in sample size and its associated effects, which is caused by the loss

of zero pairs of data when calculating proportionality between each pair of variables.

An integral part of the method for creating proportionality differential networks involves the use of bootstrap resampling to build empirical distributions. This enables a p-value to be calculated, which acts as a guide for determining the significance in difference between ϕ in cases when compared to controls. Whilst no directional bias was observed in ϕ values generated via bootstrap resampling, the current method effectively discretises the ϕ value by representing the empirical distribution as a set of ϕ value estimates. Therefore, it would be interesting to observe how these p-values are affected by using kernel density estimation (KDE) to smooth the bootstrapped distributions (Parzen, 1962; Rosenblatt, 1956).

Furthermore, in order to make the case and control ϕ values comparable, it was imperative to match the case and control sampling distributions as closely as possible. Therefore, the same number of samples must be used to calculate ϕ in controls and cases. To address this problem, subsampling of cases was introduced, which subsequently allows for the calculation of certainty scores upon repeated subsampling. These certainty scores represent the proportion of times that a statistically significant difference is observed in cases and controls upon repeated subsampling of cases, as discussed in Section 3.4.2.

The certainty score is particularly useful when interpreted alongside the effect size, which represents the magnitude of change in ϕ in cases when compared to controls. The consideration of these measures can be further complemented by running the differential network analysis in reverse; we can run the differential network method where the cases are bootstrapped and controls are subsampled (*reverse*), rather than bootstrapping the controls (*forward*). This allows us to observe any differences in the forward and reverse differential networks and subsequently determine how the case and control ϕ distributions overlap, as discussed previously in Section 3.6.5. In the future, it could be interesting to measure the magnitude of the overlap in the case and control ϕ distributions, which would provide a measure that would be identical when the analysis is run in forward and reverse. This would result in forward and reverse differential networks that would essentially be identical, except for the effects of sampling variation.

Null tests were run to validate the final nuanced method (including the considerations of sparsity, sampling distributions and certainty), showing that the method is sensitive to small changes but gives the expected empty network when run with identical sets of data for the cases and controls.

Finally, it is important that we observe the results of this method when applied to real microbiota data. The human gut microbiota is one of the most extensively studied microbiota communities within the scientific literature. Therefore, in the next chapter, this differential network method is applied to microbiota data taken from inflammatory bowel disease patients and healthy controls, with the aim of identifying potential shifts in bacterial dependencies in cases when compared to controls.

Chapter 4

Differential networks: Application

4.1 Introduction

The proportionality differential network method described in Chapter 3 was developed and described in order to explore the ecology behind dysbiosis via the identification of changes in associations between pairs of bacterial genera. This method can be applied to any case/control datasets with enough samples in order to identify how some ecological relationships between constituents of the microbiota may have changed in disease cases when compared to healthy controls. It is important to identify these changes in associations/ecological relationships because they may reveal mechanisms of the disease caused by dysbiosis. For example, we often observe a different balance of bacteria in disease cases, such as IBD patients. It is thought that the inflammatory profile within the gut of the patient modulates (or is modulated by) the microbial constituents of the gut, and the altered mucus lining of the gut in IBD patients creates a different balance of substrates for gut bacteria when compared to healthy controls. This has a direct knock-on effect for metabolic 'food chains' across bacteria in the gut. These changes in gut environment give rise to massively altered bacterial constituents and metabolic profiles within IBD patients when compared to controls.

In this chapter, I demonstrate how this new differential network method may be used to explore

altered dependencies between bacteria in the microbiota in IBD cases. I have applied this proportionality differential network method to existing data that was published by Morgan *et al.* (2012). Their study involved the use of stool and biopsy samples taken from patients suffering with forms of inflammatory bowel disease (IBD) which were compared to samples taken from healthy controls. They had inferred the metabolic capabilities of the microbiota in cases by identifying whole KEGG pathways associated with the identified bacteria and then compared these pathways to those identified in the healthy controls.

Here, I have used the data published by Morgan *et al.* (2012) to generate proportionality differential networks; I subsequently discuss potential altered relationships between constituents of the gut microbiota in IBD patients when compared to healthy controls. Ultimately, changes in associations between bacteria may be related to changes in metabolic capabilities within dysbiosis, or linked to known altered abundances identified within the current scientific literature.

It is known that IBD is a complex disease with the involvement of genetic and environmental factors, as well as shifts in immunological profiles within the gut (Furusawa *et al.*, 2013; Sokol *et al.*, 2008). The gut microbiota and its role in IBD manifestations has been heavily explored using standard methods such as abundance comparisons. Whilst it is already known that the relative abundances of different genera are known to be different in IBD cases, the exact cause of dysbiosis is unknown, as are the ecological mechanisms that support the disease-associated dysbiosis (Morgan *et al.*, 2012). However, if we can begin to understand how the constituents of the microbiota interact with each other and their host in IBD cases, we may be able to shed more light on how the dysbiosis is related to the genetic, environmental and immunological factors at play.

4.2 The data

We obtained the raw 16S rRNA gene sequence data (V3 – V5 regions) for the 220 samples that were previously published. These samples were collected as a mixture of stool samples

and biopsies from individuals that were each classified into one of four categories: healthy individuals, ileal Crohn's disease patients (iCD), non-ileal Crohn's disease patients (non_iCD) or ulcerative colitis patients (UC) (Morgan *et al.*, 2012). The numbers of samples in each of these categories are shown in Table 4.1.

The reads from each sample were classified individually using the Kraken classification tool (Wood & Salzberg, 2014) and clustered according to their genus classification. For each sample, the number of sequences in each genus cluster represented the relative abundance of that genus within the given sample. Any reads that could not be classified as far as genus were discarded. The resulting abundance data was in the form of relative abundances of each genus within each of the 220 samples collected. Each comparative analysis was limited by the dataset with the smallest number of usable samples, which is influenced by the proportion of zero abundance counts, as discussed in Section 3.6 above.

Table 4.1: *Classes of samples split by sample type.*

Class label	**Number of samples**	**Total samples**	**Disease/control classification**
Healthy_stool	18	27	Healthy individuals (stool samples)
Healthy_biopsy	9		Healthy individuals (biopsy samples)
iCD_stool	0	43	Ileal Crohn's disease patients (stool samples)
iCD_biopsy	43		Ileal Crohn's disease patients (biopsy samples)
non_iCD_stool	61	76	Non-ileal Crohn's disease patients (stool samples)
non_iCD_biopsy	15		Non-ileal Crohn's disease patients (biopsy samples)
UC_stool	47	74	Ulcerative colitis disease patients (stool samples)
UC_biopsy	27		Ulcerative colitis disease patients (biopsy samples)

4.3 Applying the method

The proportionality differential network method described in Chapter 3 was reimplemented with an online front-end (with thanks to Laura de Arroyo Garcia, Yunsoo Kim and Ana-Isabella Tanase for their contributions under the supervision of the Pinney research group, Imperial College London), and was used to generate the differential networks for these data sets.

Each case dataset was compared to the corresponding control (details shown in Table 4.2) by applying the proportionality differential network method as described in Section 3.4. The method was also applied in both directions (i.e. bootstrapping controls to test cases against them and vice versa) to produce two differential networks for each comparison of cases and controls. This was done due to the asymmetry of the method, as discussed above in Section 3.6.5.

Morgan *et al.* (2012) have demonstrated previously that the observed microbiota is different for stool samples compared to biopsy. This is due to the fact that biopsies typically contain large parts of the mucosal epithelium, whilst stool samples contain far fewer bacteria from this mucosal layer. However, there are not enough samples in each group when samples are split into stool and biopsy groups. The success and failure of the associated dataset comparisons using the differential network method are shown in Table 4.2, which demonstrates the limitations of small datasets. The same parameters were used to generate each network:

$$n = 2500, \alpha = 0.05$$

where n is the number of iterations (i.e. the number of bootstrap and subsampling iterations) and α is the p-value threshold of statistical significance.

4.4 Results

Due to the limited number of samples available for the healthy datasets (stool and biopsy), no edges were returned within the differential networks for the split data — see Table 4.2. However, when the stool and biopsy data were assessed together in the mixed runs, there is enough data to produce three pairs of differential networks, as shown in Figures 4.1 to 4.6. These networks show shifts in genus abundance and genus associations between the datasets being compared.

Table 4.2: *Comparisons carried out using the proportionality differential network method. The comparisons fall into one of three groups: mixed, stool or biopsy (see third column). Mixed comparisons were carried out using stool and biopsy data combined. Stool comparisons were carried out using only stool samples. Biopsy comparisons were carried out using only biopsy samples. In the fourth column, a tick indicates that a proportionality differential network was returned and a cross indicates that there was an insufficient number of samples to produce any significant results in a differential network.*

Control group	**Case group**	**Sample type**	**Network returned**
Healthy	iCD	Mixed	✓
Healthy	non_iCD	Mixed	✓
Healthy	UC	Mixed	✓
Healthy_stool	non_iCD_stool	Stool	✗
Healthy_stool	UC_stool	Stool	✗
Healthy_biopsy	iCD_biopsy	Biopsy	✗
Healthy_biopsy	non_iCD_biopsy	Biopsy	✗
Healthy_biopsy	UC_biopsy	Biopsy	✗

4.4.1 Ileal Crohn's disease (iCD)

The resulting networks for the comparison between ileal Crohn's disease patients and healthy controls show all edges with a certainty score above 0.5 — these networks are shown in Figures 4.1 and 4.2. The corresponding full networks showing edges for all non-zero certainty scores are shown in Figures C.1 and C.2 within Appendix C.

4.4.2 Non-ileal Crohn's disease (non-iCD)

The resulting networks for the comparison between non-ileal Crohn's disease patients and healthy controls show all edges with a certainty score above 0.5 — these networks are shown in Figures 4.3 and 4.4. The corresponding full networks showing edges for all non-zero certainty scores are shown in Figures C.3 and C.4 within Appendix C.

4.4.3 Ulcerative colitis (UC)

The resulting networks for the comparison between ulcerative colitis patients and healthy controls show all edges with a certainty score above 0.5 — these networks are shown in Figures 4.5 and 4.6. The corresponding full networks showing edges for all non-zero certainty scores are shown in Figures C.5 and C.6 within Appendix C.

4.5 Discussion

IBD is known to be associated with local inflammation, disrupted mucus layers protecting the gut epithelium, and altered metabolite concentrations in the gut lumen, all of which are associated with alterations in the gut microbiota. Below, we will discuss the significance of the changes observed in the proportionality differential networks in the context of the mucus barrier, the immune system and the metabolites known to be associated with IBD.

4.5.1 Mucosal bacteria and IBD

There are many studies that have shown the importance of appropriate mucus levels throughout the gut (Jakobsson *et al.*, 2015; Derrien *et al.*, 2010; Tailford *et al.*, 2015; Wenzel *et al.*, 2015). Part of the function of the mucus layers is to provide a barrier between the bacteria within the gut and the immune system at the epithelial surface of the gut wall. Disruptions in this mucus barrier play a key role in shaping the immune response and inflammation observed in IBD patients (Furusawa *et al.*, 2013; Malago & Sangu, 2015).

The mucus layers present on the epithelial surface vary longitudinally along the gastrointestinal tract. In the colon of healthy individuals, there are two layers of mucus — the outer layer provides a nutrient-rich habitat for bacterial colonies to grow and acts as a lubricant for easier expulsion of faecal matter, whilst the inner layer is firmly attached to the epithelium and is virtually free of bacteria (Jakobsson *et al.*, 2015; Derrien *et al.*, 2010; Tailford *et al.*, 2015).

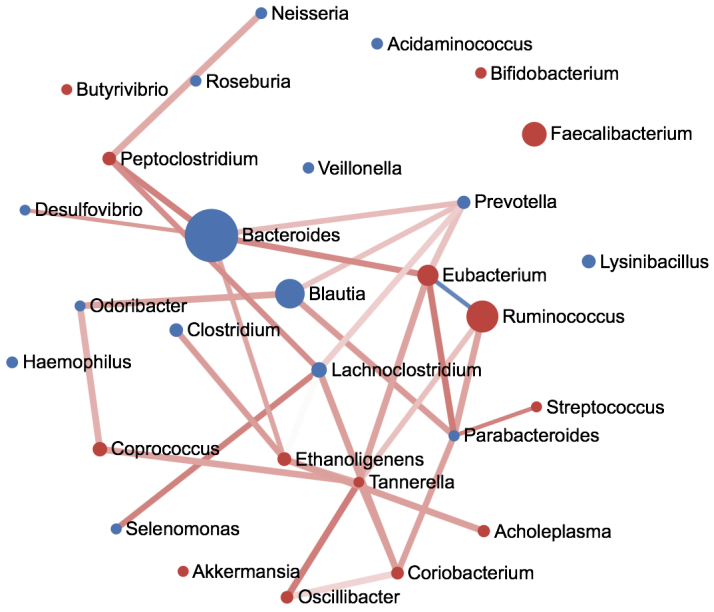


Figure 4.1: **Healthy/iCD differential network:** *Bootstrap = Healthy; Subsampling = iCD; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in iCD patients; Blue nodes = Increase in abundance in iCD patients. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in iCD patients; Blue edges = Increase in ϕ (decrease in association) in iCD patients; The darker the edge, the greater the change in ϕ.*

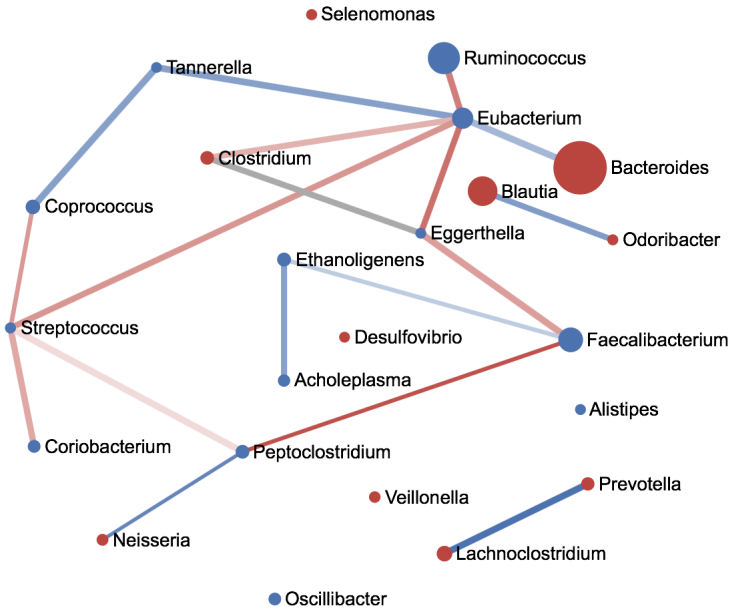


Figure 4.2: **iCD/Healthy differential network:** *Bootstrap = iCD; Subsampling = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in healthy individuals; Blue nodes = Increase in abundance in healthy individuals. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in healthy individuals; Blue edges = Increase in ϕ (decrease in association) in healthy individuals; The darker the edge, the greater the change in ϕ.*

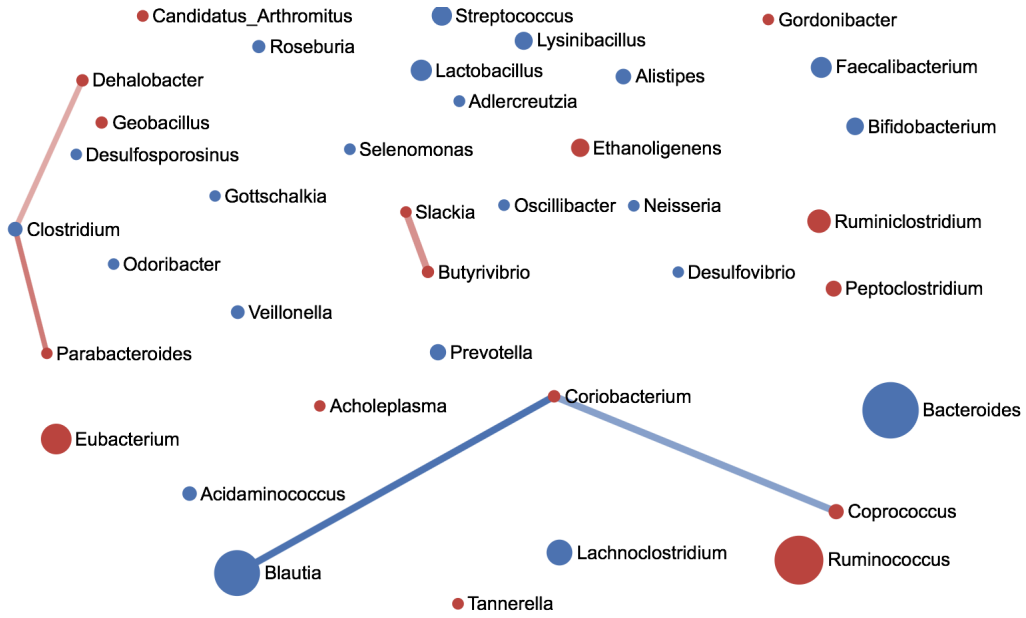


Figure 4.3: **Healthy/Non-iCD differential network:** *Bootstrap = Healthy; Subsampling = Non-iCD; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in non-iCD patients; Blue nodes = Increase in abundance in non-iCD patients. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in non-iCD patients; Blue edges = Increase in ϕ (decrease in association) in non-iCD patients; The darker the edge, the greater the change in ϕ.*

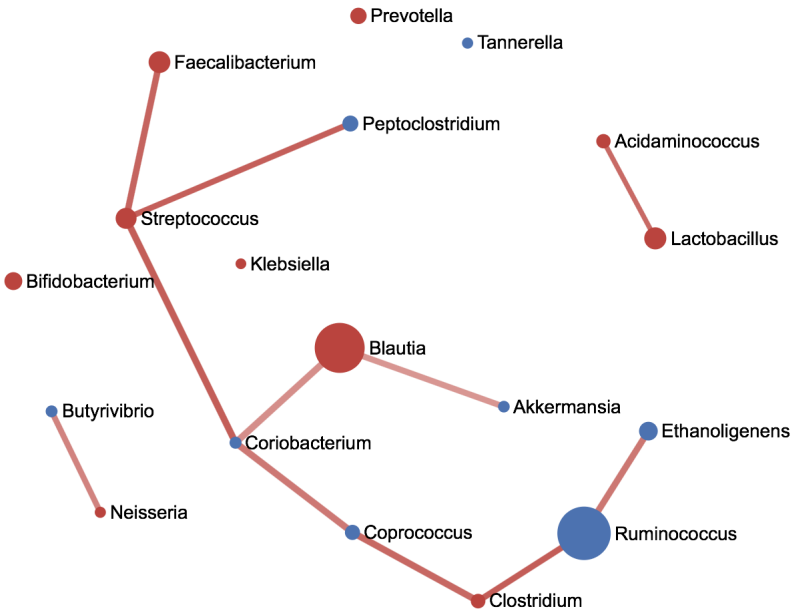


Figure 4.4: **Non-iCD/Healthy differential network:** *Bootstrap = Non-iCD; Subsampling = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in healthy individuals; Blue nodes = Increase in abundance in healthy individuals. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in healthy individuals; Blue edges = Increase in ϕ (decrease in association) in healthy individuals; The darker the edge, the greater the change in ϕ.*

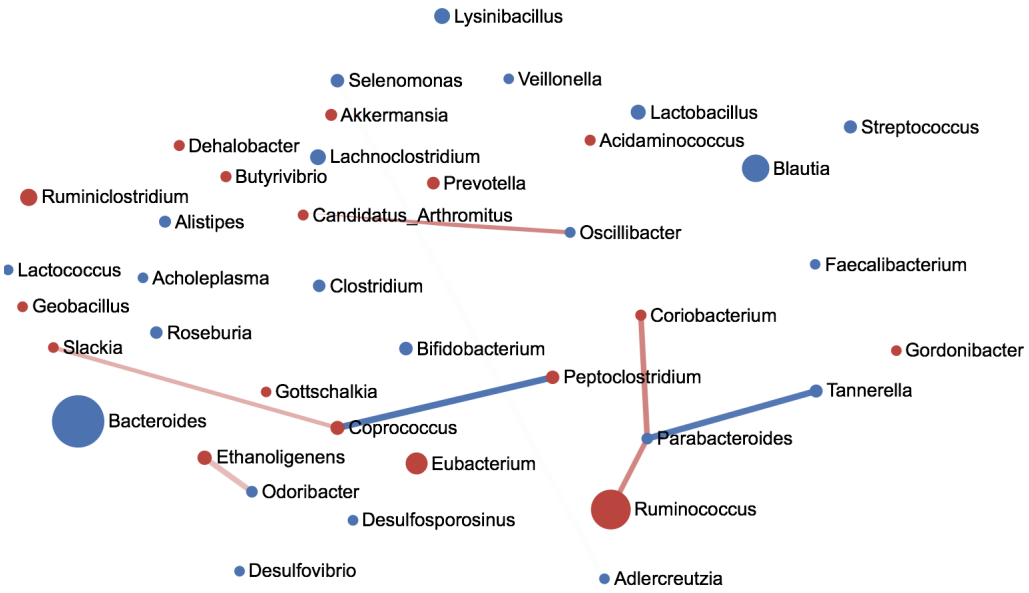


Figure 4.5: **Healthy/UC differential network:** *Bootstrap = Healthy; Subsampling = UC; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in UC patients; Blue nodes = Increase in abundance in UC patients. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in UC patients; Blue edges = Increase in ϕ (decrease in association) in UC patients; The darker the edge, the greater the change in ϕ.*

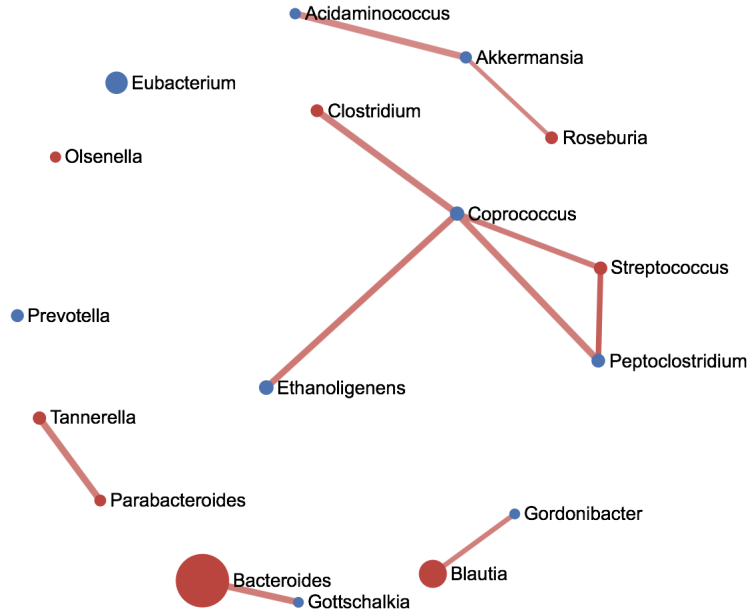


Figure 4.6: **UC/Healthy differential network:** *Bootstrap = UC; Subsampling = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in healthy individuals; Blue nodes = Increase in abundance in healthy individuals. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in healthy individuals; Blue edges = Increase in ϕ (decrease in association) in healthy individuals; The darker the edge, the greater the change in ϕ.*

This stratification is important for maintaining a barrier between bacterial cells and the host's immune system. It is known that the microbiota of the top mucus layer differs from that of the faecal matter, and that the structure of the mucus layer and mucus microbiota differs in IBD cases when compared to controls (Derrien *et al.*, 2010; Tailford *et al.*, 2015; Png *et al.*, 2010). For example, it is also known that ulcerative colitis (UC) patients with active disease have a depleted mucus membrane with bacterial cells in contact with the epithelium, unlike healthy controls (Jakobsson *et al.*, 2015; Wenzel *et al.*, 2015). Furthermore, these protective mucus sheets are constructed using large amounts of the MUC2 mucin, and Muc2-deficient mice have been shown to develop severe colitis, thereby indicating the importance of this protective barrier (Wenzel *et al.*, 2015).

IBD is characterised by a high load of mucosal bacteria with an increased proportion of mucin degraders when compared to healthy controls. It has been suggested that the high proportion of mucin degraders gives rise to increased mucin degradation and, therefore, increased substrate provision for certain types of non-degrading mucosal bacteria (Tailford *et al.*, 2015). In the differential networks shown above for all three forms of IBD (Figures 4.1 to 4.6), the *Bacteroides* genus shows a large increase in abundance in disease cases when compared to controls. Given that the *Bacteroides* genus contains key mucin degraders such as *Bacteroides thetaiotaomicron*, this finding agrees with the previous observations of increased load of mucosal bacteria in the guts of IBD cases.

It has been shown that the extracellular vesicles produced by mucin-degrading *Akkermansia muciniphila* have a protective effect against the development of IBD (Kang *et al.*, 2013) and, unsurprisingly, the proportion of *A. muciniphila* is known to be lower in IBD patients when compared to controls (Tailford *et al.*, 2015). This is also the case for this study — *Akkermansia* is present within the iCD and UC networks and is shown to have decreased in abundance.

However, it has also been shown that mice exhibiting intestinal inflammation induced by colonisation of *Salmonella enterica* Typhimurium show increased inflammation with the addition of *A. muciniphila* (Ganesh *et al.*, 2013), thereby suggesting that the effects of different bacteria

upon the host are contextual, and are dependent upon the other constituents of the microbiota, as well as the host's immune system and mucosal barriers. This serves to highlight the importance of expanding research beyond the existing abundance measures in order to understand the ecological mechanisms at play within these communities, such as competition and symbioses.

Some progress has already been made into the understanding of metabolic symbioses occurring within the gut microbiota. For example, we already know that degradation of mucin can give rise to free sialic acid and fucose within the lumen of the gut, which then provide energy, nitrogen and carbon to other constituents of the microbiome. Sialic acid metabolism has been studied using wet lab techniques and gene sequences, and it is becoming increasingly apparent that different types of bacteria are mutually dependent on one another for complete utilisation of this particular growth medium (Tailford *et al.*, 2015). A case in point includes the cooperation of *Bacteroides thetaiotaomicron* (a well-characterised mucin degrader known to rely on host-derived glycans for colonisation) with either *Salmonella enterica* Typhimurium or *Clostridium difficile* (also known as *Peptoclostridium difficile*). The former encodes a sialidase to release sialic acid from mucin and the latter two both encode the Nan operon that enables them to utilise the free monosaccharide, but none of these species have both the sialidase and Nan operon in co-occurrence (Tailford *et al.*, 2015; Ng *et al.*, 2013). Furthermore, it has been shown that by using sialidase-deficient *B. thetaiotaomicron* to colonise gnotobiotic mice, the levels of free sialic acid levels in the lumen are reduced, thereby resulting in the down-regulation of the sialic acid catabolic pathway in *C. difficile* and its reduced growth rate. The effects of the sialidase-deficient *B. thetaiotaomicron* upon *C. difficile* growth can be reversed via the exogenous addition of sialic acid, further supporting the case for bacterial metabolic symbioses within the gut microbiota (Tailford *et al.*, 2015; Ng *et al.*, 2013). In the context of the proportionality differential networks shown above, the observed increase in association between *Peptoclostridium* and *Bacteroides* in iCD cases (Figure 4.1) may reflect an increased strength in this metabolic symbiosis between *B. thetaiotaomicron* and *C. difficile* in patients suffering with ileal Crohn's disease, because *C. difficile* has also been reclassified as *Peptoclostridium difficile* (Milani *et al.*, 2016; Yutin & Galperin, 2013).

In the iCD and non-iCD networks (Figures 4.1 to 4.4), mucin-degraders besides *Bacteroides* appear to have increased associations with the *Blautia* genus; *Blautia* is linked to *Prevotella* in iCD and *Akkermansia* in non-iCD. *Prevotella* is known to include strains that degrade mucin (Derrien *et al.*, 2010; Wright *et al.*, 2000; Bergstrom & Xia, 2013) and *Akkermansia* is one of the key mucin degraders of interest in IBD cases due to its protective effect in some instances and destructive effect in others, as highlighted previously (Ganesh *et al.*, 2013).

Prevotella is an opportunistic pathogen and is known to be a causative agent in periodontitis, which involves inflammation of the mucosal surfaces of the oral cavity. Interestingly, periodontitis is known to be more prevalent in IBD cases than in the broader population as a whole (Habashneh *et al.*, 2012). However, here, *Prevotella* appears to show a reduced relative abundance in IBD cases. In Figure 4.1, *Prevotella* is observed to be one of the hubs within the proportionality differential network, showing increased associations with *Bacteroides*, *Blautia*, *Lachnoclostridium* and *Eubacterium*, which could indicate metabolic symbioses that support or cause the change in growth patterns of *Prevotella* in IBD cases.

As well as its increased association with *Prevotella*, *Blautia* also has increased associations with *Parabacteroides* and *Odoribacter* in the iCD differential networks. These observations are interesting because *Prevotella*, *Parabacteroides* and *Odoribacter* are all genera that were once classified under *Bacteroides* but have subsequently been reclassified independently (Rajilić-Stojanović & de Vos, 2014). Given that they used to be classified within the same genus, they may have large similarities in their metabolic functions and requirements — therefore, all three may be associating with *Blautia* in disease cases for similar reasons, such as a shared need for acetate which can be readily produced by *Blautia* species. Acetate is one of the many short chain fatty acids (SCFAs) that is produced by bacteria within the microbiota and is an important precursor to butyrate, which is key to limiting inflammation in the gut.

4.5.2 Butyrate and other short chain fatty acids (SCFAs)

The thick mucosal barrier along the walls of the intestine during good health prevents bacteria within the gut lumen from triggering an inflammatory response at the epithelial surface of the gut. However, this is also complemented by the action of SCFAs within the gut lumen. Acetate, propionate and butyrate are three SCFAs commonly found within the gut and they are the end products of carbohydrate metabolism. Acetate is the SCFA with the highest concentration within the gut lumen and can act as a precursor metabolite in the production of butyrate. The ability of butyrate to aid attenuation of colitis is broadly appreciated and has been demonstrated by a number of studies (Ríos-Covián *et al.*, 2016; Malago & Sangu, 2015; Furusawa *et al.*, 2013).

It has been shown that butyrate can modulate the levels of pro-inflammatory cytokines (such as interleukin-8 (IL-8)) and the associated recruitment of inflammatory cells within the colon (Malago & Sangu, 2015). Butyrate has also been shown to induce the differentiation of regulatory T (T_{reg}) cells in mice via epigenetic effects; these T_{reg} cells can produce interleukin-10 (IL-10), which subsequently attenuates inflammatory responses. It has been shown that administration of intraperitoneal butyrate has a greater anti-inflammatory effect than intrarectal or oral administration. This may be due to the fact that the damaged epithelium in UC is not as effective at absorbing butyrate from the gut, partly due to down-regulated butyrate transporters in the colonic mucosa (Furusawa *et al.*, 2013); systemic administration bypasses the need for uptake in the gut (Malago & Sangu, 2015). Unsurprisingly, as well as the inefficiencies of butyrate uptake caused by the damaged epithelium in IBD patients, it has also been known for some time that the abundance of butyrate-producing bacteria are underrepresented in IBD cases when compared to healthy controls (Ríos-Covián *et al.*, 2016; Furusawa *et al.*, 2013).

As well as its anti-inflammatory effects, butyrate is key for gut health purely as an energy source for colonocytes. Butyrate encourages growth and differentiation of colonocytes, as well as energy production within them; without butyrate, the colonocytes within the epithelium enter autophagy (Donohoe *et al.*, 2011; Dumas, 2011). It is at this point that the link between

the gut epithelium, mucus membrane and SCFAs becomes important. Due to the dependency of the colonocytes upon butyrate, it is beneficial to maintain high levels of butyrate producers and butyrate-precursor producers (such as acetogens) close to the gut epithelium (Koropatkin *et al.*, 2012). Given that *Blautia* is a key acetate (butyrate precursor) producer in the gut, it is unsurprising that it would be found in association with mucus-inhabitors that are situated so close to the epithelium. Therefore, the previously discussed associations between *Blautia* and mucin-degraders (*Prevotella* and *Akkermansia*) in the iCD and non-iCD networks may be beneficial for keeping the acetate supply high for use by butyrate producers near the epithelium, such as *Eubacterium*, which is also shown to have increased associations with mucin degraders *Prevotella* and *Bacteroides*.

In line with the concept of keeping SCFA producers close to the epithelium, the iCD differential networks (Figures 4.1 and 4.2) show an increase in association between *Eubacterium* (known to include butyrate producers) and *Ruminococcus* (including the mucus-degrading *Ruminococcus torques* and *Ruminococcus gnavus*) (Crost *et al.*, 2016; Hoskins *et al.*, 1985). We also see increased associations between *Prevotella* (a mucin degrader) and *Bacteroides*, *Blautia*, *Lachnoclostridium* (acetate producers) and *Eubacterium* (butyrate producers). This common theme around SCFA metabolism may indicate an area of exploration that could provide some insight into the increase in associations between *Prevotella* and these other genera in disease cases.

Following the same logic, it is interesting to note that within the non-iCD network, (Figures 4.3 and 4.4) there are also increased associations between *Faecalibacterium* (a highly important butyrate producer) and *Streptococcus*, which is known to degrade mucins in the oral cavity (Derrien *et al.*, 2010). If some strains are shown to also degrade mucins in the gut, there could be a similar mechanism at play here. Furthermore, an increase in association between the mucus-inhabiting *Neisseria* and butyrate-producing *Butyrivibrio* is observed within the non-iCD network (Barcenilla *et al.*, 2000), again highlighting the importance of relationships between the SCFA producers and mucin degraders at the epithelial surface (Koropatkin *et al.*, 2012).

4.5.3 Natural immunomodulators

Faecalibacterium prausnitzii is one of the main producers and sources of butyrate within the gut, but *F. prausnitzii* has also been shown to have additional anti-inflammatory effects beyond those associated with its production of butyrate. Supernatant from *F. prausnitzzi* cultures have been shown to suppress activation of IL-1β-mediated NF-κB in colon carcinoma (Caco-2) cells (the effect of which was not reproduced with butyrate alone) and decrease secretion of the pro-inflammatory cytokine IL-8 (Sokol *et al.*, 2008). Furthermore, when peripheral blood mononuclear cells (PBMCs) were exposed to *F. prausnitzii* or its supernatant, an anti-inflammatory profile was observed with high levels of IL-10 secreted and very low levels of IL-12, thereby enforcing an anti-inflammatory effect (Sokol *et al.*, 2008). Unsurprisingly, it was also shown that the severity of 2,4,6-trinitrobenzenesulphonic acid (TNBS)-induced colitis was reduced by administration of *F. prausnitzii* and its supernatant (Sokol *et al.*, 2008). These influences upon the immune system are thought to be some of the reasons why depleted levels of *F. prausnitzii* are associated with certain forms of IBD and with increased chances of postoperative recurrence of Crohn's disease (Sokol *et al.*, 2008).

The reduction in levels of *F. prausnitzii* contributes greatly to the reduction in *Firmicutes* observed in patients with ileal Crohn's disease when compared to healthy controls (Sokol *et al.*, 2008; Lopez-Siles *et al.*, 2012). These findings agree with the networks shown above, where the iCD networks show a decrease in levels of *F. prausnitzii* in disease cases, whilst the networks for non-iCD and UC show an increase in cases.

F. prausnitzii is also able to utilise galacturonic acid (a constituent of pectin), even though uronic acid utilisation is usually observed exclusively within the *Bacteroides* genus (Lopez-Siles *et al.*, 2012). *F. prausnitzii* has also been shown to grow on host-derived substrates such as N-acetylglucosamine and its growth is stimulated by acetate, so it is thought that it can switch between host-, microbe- and diet-derived substrates. Whilst this allows for some flexibility, it is also known that growth is severely limited in the absence of carbohydrates (Lopez-Siles *et al.*, 2012). However, its generalist properties may explain why it is decreased in some forms

of IBD (such as iCD) and not others (such as UC and non-iCD), depending on the metabolic capabilities of (and competition with) the other constituents of the dysbiotic community.

4.5.4 Lactate

It is known that lactate can build up in the gut of UC patients, whereas in healthy patients, lactate is utilised by constituents of the microbiome, which gives rise to low levels of lactate in suspension of fecal matter. Therefore, in UC patients, either lactate producers must increase their production of lactate (either by increased rates of production or increased numbers of cells) and/or lactate users must utilise it at a slower rate (again, either through decreased numbers of cells or altered metabolism with slower rates of lactate uptake).

It is thought that the luminal pH of the colon is reduced in UC patients (Nugent *et al.*, 2001) and a study identifying different levels of lactate accumulation at varying levels of pH indicates that the altered pH in UC patients may influence the build up of lactate (Belenguer *et al.*, 2007). It was shown that lactate is converted to acetate, butyrate and propionate in the gut at pH values above 5.9, but utilisation of lactate dropped at pH 5.2, giving rise to lactate accumulation. Therefore, reductions in production of acetate, propionate and butyrate by SCFA producers (such as *Butyrivibrio*, *Eubacterium* and *Faecalibacterium*) would give rise to higher levels of lactate in the gut lumen (Belenguer *et al.*, 2007). This conclusion agrees with the fact that overall abundances of butyrate producers are known to be reduced in UC (Ríos-Covián *et al.*, 2016). Furthermore, an edge within one of the UC differential networks (Figure 4.6) suggests that there is an increase in the association between *Coprococcus* (a butyrate producer that uses lactate (Pryde *et al.*, 2002; Louis & Flint, 2009)) and *Streptococcus* (a lactate producer (Pessione, 2012)) in healthy individuals, suggesting that there could be a loss of symbiosis in UC patients. Furthermore, the fact that *Coprococcus* exhibits a lower abundance and *Streptococcus* exhibits a higher abundance in UC patients than controls may be an example of why lactate concentrations increase within the lumen of UC patients.

4.5.5 Agreement across pairs of networks

The pairs of networks produced to compare each case dataset to the healthy controls (where in one, the controls are bootstrapped, and the other, the cases are bootstrapped) show some edges that agree across the network pairs. For example, within the networks created using the samples from iCD patients, both networks contained altered dependencies between *Prevotella* and *Lachnoclostridium*, *Eubacterium* and *Ruminococcus*, and *Bacteroides* and *Eubacterium*. This suggests that the effect size (i.e. the difference in the median values for the case and control distributions) are large relative to the variances of the distributions. Similarly, the networks for the UC patients show some agreement, as do the networks for the non-iCD patients. The edges which do not show up in both networks for a given comparison are likely to either have small effect sizes (as discussed previously) or distributions of ϕ values that are largely overlapping. These edges are still of interest and worth exploring further.

When comparing the pairs of networks, there appear to be more edges (i.e. non-zero certainty scores) in the networks generated by bootstrapping the controls rather than the cases. The asymmetry across the pairs of networks may either indicate small effect sizes, or that the case and control distributions largely overlap for the edges shown in one network but not the other (as discussed in Section 3.6.5 in the previous chapter). The fact that there are more edges present when controls are bootstrapped may indicate that the ϕ distributions in the cases are broader than the equivalent distributions in controls, thereby suggesting that the bacterial associations are more variable across the disease cases. Perhaps, therefore, the disease phenotype may be caused by various states of dysbiosis, with numerous different profiles of bacterial associations causing the disease phenotype across the samples.

4.6 Comparing proportionality and correlation differential networks

The finished method (as described previously in Section 3.4) was also altered to use Pearson's correlation coefficient (r) in place of ϕ. This enables us to make a direct comparison of the resulting differential networks and show the benefits of using ϕ in place of r.

The finished method using Pearson's r was applied to previously published 16S rRNA gene count data with 27 healthy control samples and 43 samples from ileal Crohn's disease patients (Morgan *et al.*, 2012). Here, all aspects of the correlation and proportionality analyses remained the same except for the fact that Pearson's r was used in place of the asymmetric ϕ values.

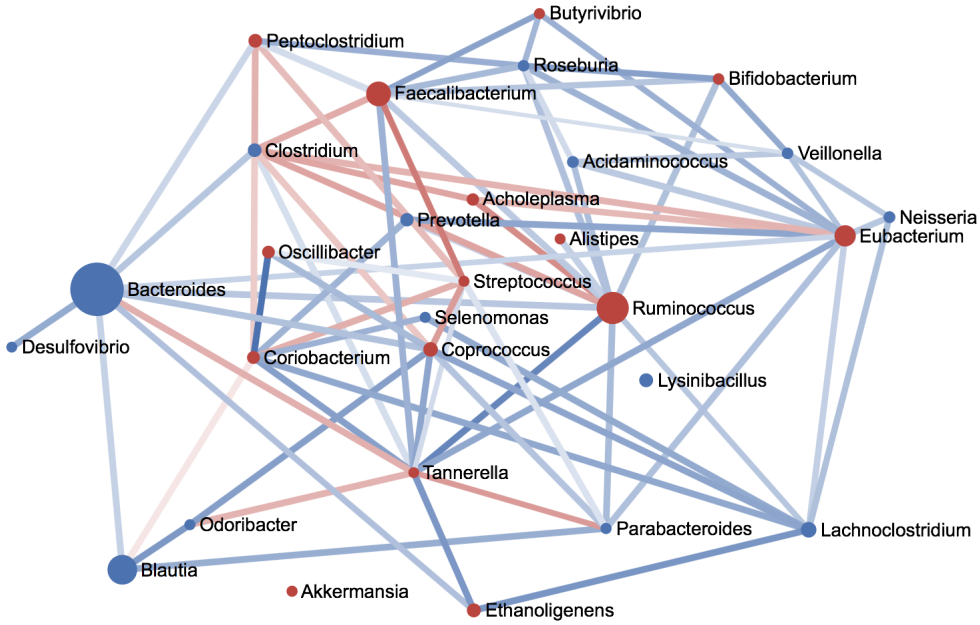


Figure 4.7: **Healthy/Non-iCD differential network using Pearson's correlation:** *Bootstrap = Healthy; Subsampling = Non-iCD; 2500 bootstrap and subsampling iterations; Certainty >0.5. Red nodes = Decrease in abundance in non-iCD patients; Blue nodes = Increase in abundance in non-iCD patients. The larger the node size, the larger the relative change. Red edges = Decrease in r in non-iCD patients; Blue edges = Increase in r in non-iCD patients; The darker the edge, the greater the change in r.*

As expected, the correlation differential network was more dense than the equivalent proportionality differential network. For all certainty scores (s >0), there were 110 edges in the correlation differential network (shown in Figure C.7 within Appendix C) and 55 in the proportionality

network (Figure C.1 within Appendix C), meaning there were twice as many edges when using correlation to measure dependencies between genera. For certainty scores over 0.5 (s >0.5), the difference was over double: there were 77 edges present when using correlation to measure dependency (Figure 4.7), and 30 edges when using the proportionality measure (Figure 4.1). The number of edges in the correlation networks could be reduced further still by increasing the choice of cutoff. As with the cutoff for certainty values of 0.5 and 0, this network could be compared to equivalent proportionality network with the same cutoff value for the certainty score. These results confirm that correlation gives rise to more dense networks. This is most likely explained by high levels of false positives that arise due to correlation being observed in the relative data, which would not be observed in underlying absolute data (Lovell *et al.*, 2015).

4.7 Conclusions and future work

In this chapter, the proportionality differential network method was applied to real biological data taken from inflammatory bowel disease patients and healthy controls. A comparison was also made to an implementation of the method that used a correlation measure in place of ϕ. This test showed that the proportionality differential networks provide sparser networks than the correlation differential networks, thereby showing the benefits of using ϕ in place of correlation in order to reduce the number of false positive results observed in relative data.

However, reconciliation of the pairs of networks produced for each comparison of two datasets is not ideal. It is possible to interpret the two networks together: for example, matching edges across a given pair of networks suggests that a given relationship is very different in cases and controls. Similarly, edges identified in only one network may be partly due to a small effect size and/or just a change in variance for ϕ. In the future, it would be beneficial to develop this method further in order to use a test statistic that does not give rise to different results when we bootstrap one dataset to compare against the values of the other (for example, we could use the difference between the case and control ϕ values as a test statistic, or measure the overlap of case and control ϕ distributions).

Despite these limitations, the resulting proportionality differential networks appear to be largely congruent with the relevant existing literature, which highlights the associations of the microbiota, metabolites and immune system to the inflammatory phenotype observed in disease cases (as discussed in Section 4.5). This gives an element of confidence in the relevance of the method and usefulness of proportionality measures within the context of microbial data. These results suggest that pursuing some of the potential alterations in the method (discussed previously in Chapter 3, Section 3.7) may be a worthwhile endeavour.

Chapter 5

Inferred metanetworks

5.1 Introduction

In Chapters 3 and 4, a new method was presented that allows us to explore potential shifts in linear dependencies between constituents of the microbiota within disease cases when compared to controls. In order to characterise these shifts in dependencies within the context of metabolic symbioses, it could be useful to determine how the overall metabolic capabilities of the microbiota are different in disease cases when compared to controls. The differences in metabolic capabilities may help us to characterise potential causes or outcomes of the shifts in dependencies.

Morgan *et al.* (2012) have discussed and demonstrated the effectiveness of producing inferred metagenomes from 16S rRNA abundance measures from bacterial communities. By linking OTUs to associated sequenced genomes, they were able to infer the metagenomes of case groups of inflammatory bowel disease (IBD) patients and healthy controls (Morgan *et al.*, 2012). They used these inferred metagenomes to identify KEGG modules, KEGG pathways and Gene Ontology (GO) terms associated with cases or controls, by using the OTU abundances to apply averaged weights to these pathways and GO terms in cases and controls (Kanehisa *et al.*, 2016; Ashburner *et al.*, 2000). By identifying areas of metabolism that are different within

the microbiota of cases when compared to controls, we may be able to use this information to perturb the system to return it to a healthy state, either via drug treatment, prebiotic treatment (such as dietary changes) or probiotic treatment (via supplementation of the microbiota).

The use of KEGG modules gives us an insight into the broad differences in metabolism between case and control communities, but it does not allow for finer resolution at the reaction level. It would be preferable if the analysis was not limited or constrained by assigning scores to pre-defined pathways. Instead, it could be beneficial to model the metabolic capabilities of a whole microbial community as a metagenome-scale network, where individual reactions within that network could be identified as being differentially abundant in cases and controls. One of the ways in which we can compare changes in weighted metabolic networks is via the use of AMBIENT, which identifies modules of change within the networks without using predefined sets such as KEGG pathways and KEGG modules (Bryant *et al.*, 2013). However, AMBIENT was designed for use with single-cell models and has previously only been used on genome-scale models, which typically contain only a few thousand reactions. Within a metagenome-scale metabolic model, we would expect to see well over 100,000 reactions, thereby stretching the application of AMBIENT well beyond its usual application and capacity.

In this chapter, we explore the possibility of modelling the metabolic capabilities of the microbiota in this way. We demonstrate that it is possible to map the metabolic capabilities of the microbiome onto a community-scale metabolic model in order to create weighted networks, where the reactions of the network are associated with normalised abundance data. However, we also demonstrate that the distributions of fold-change for the huge numbers of reactions across cases and controls currently prevent AMBIENT from predicting newly-defined modules representing significant change. We go on to discuss the reasons for this and the potential future work that could be applied to compare the case and control networks.

5.2 Methods

16S rRNA gene sequence data from diabetes patients was obtained from the EBI Metagenomes database and split according to phenotype: impaired glucose tolerance (IGT), normal glucose tolerance (NGT) and type II diabetes (T2D). The sequences in each phenotype group were classified using NCBI taxonomy IDs, which represent the different types of bacteria identified within the microbiota; the number of sequences classified under a given taxonomic ID gave rise to the abundance weight for that ID, which was also adjusted according to the copy number of the 16S rRNA gene within the given organism associated with the given ID. These IDs were mapped to KEGG genomes and the associated metabolic reaction lists, before mapping the resulting KEGG reaction IDs to MetaNetX reaction IDs. These MetaNetX IDs are crucial because the MetaNetX database contains information on all known metabolic reactions, thereby making it ideal for building a metagenome-scale metabolic model containing all known metabolic reactions. These MetaNetX reaction IDs associated with each NCBI ID were subsequently weighted with the adjusted NCBI taxonomy abundance data for the three patient phenotypes (IGT, NGT and T2D), which gave rise to three weighted sets of reactions. Finally, pairwise comparisons of these metabolic reaction lists were made by calculating the fold change for each reaction, which were presented as input for AMBIENT. Therefore, these fold changes associated with MetaNetX reaction IDs were mapped onto a metagenome-scale metabolic model, which was generated from the MetaNetX database, giving rise to three comparisons using AMBIENT. All of these steps are presented in detail below, as follows:

- **Section 5.2.1:** Obtaining and processing the data
- **Section 5.2.2:** Mapping NCBI taxonomy IDs to metabolic reactions
- **Section 5.2.3:** Correcting weights according to 16S rRNA gene copy numbers
- **Section 5.2.4:** Weighting the metagenome-scale metabolic model
- **Section 5.2.5:** Comparing the two weighted networks using AMBIENT

5.2.1 Obtaining and processing the data

16S rRNA gene sequence data from a full metagenomics study of diabetes and pre-diabetes in European women was obtained from the EBI Metagenomics database (project number: ERP002469 (Mitchell *et al.*, 2016; Karlsson *et al.*, 2013)). The samples were sorted into their three subgroups – IGT, NGT and T2D – according to their associated metadata. The data groups are shown in Table 5.1.

Table 5.1: *Classes of diabetes samples.*

Class label	**Disease/control classification**
NGT	Healthy individuals.
IGT	Impaired fasting glucose (pre-diabetes).
T2D	Type 2 diabetes.

For each sample within each subgroup (NGT, IGT, T2D), every 16S rRNA gene sequence read was classified using the Kraken classification tool with the output in the unmapped (non-MPA) format, such that each sequence is assigned directly to a NCBI taxonomy ID (Wood & Salzberg, 2014; NCBI Resource Coordinators, 2015). This was done using a k-mer database that was built from the NCBI database release in July 2014 (NCBI Resource Coordinators, 2015).

5.2.2 Mapping NCBI taxonomy IDs to metabolic reactions

The NCBI taxonomy IDs that represent the constituents of the microbiome were mapped to bacterial genomes and, subsequently, to reaction IDs in order to determine the metabolic capabilities of the microbiota. This was done for the NCBI taxonomy IDs within the NGT, IGT and T2D datasets via the following three stages described below:

- Mapping NCBI taxonomy IDs to KEGG genomes
- Mapping KEGG genomes to MetaNetX reactions
- Creating a reaction matrix from the reaction vectors for each NCBI taxonomy ID.

Mapping NCBI taxonomy IDs to KEGG genomes

The union set of all NCBI taxonomy IDs identified within the the sample groups (NGT, IGT and T2D) were mapped to their associated KEGG genome(s) via navigation of the NCBI taxonomy tree and by using the REST-style KEGG API (Kanehisa *et al.*, 2016). For the NCBI taxonomy IDs that map one-to-one with a particular KEGG genome, this mapping process is straightforward because a single genome is used to represent that organism (represented in Figure 5.1 a)). If a direct match to the original NCBI taxonomy ID is not found within the KEGG genomes database, the NCBI taxonomy tree is used to determine which KEGG genomes would best represent that classification ID, which is discussed fully below. This mapping process means that each NCBI taxonomy ID is ultimately represented by one or more KEGG genomes.

Many of the taxonomy IDs will not map directly to a single KEGG genome because the KEGG Genomes database is not exhaustive, and a number of the NCBI taxonomy IDs identified within the microbiota actually represent higher taxonomic levels (i.e. many bacterial species). Therefore, there are many more bacterial NCBI taxonomy IDs than there are genomes in the database. Any NCBI taxonomy IDs that do not map directly onto a singular KEGG genome are comprised of the KEGG genomes that map to their nearest relations. We begin by searching for their descendants, which are identified via navigation through the NCBI taxonomy tree. If any of these descendants map directly onto any KEGG genomes, these KEGG genomes are used to represent the original NCBI taxonomy ID (represented in Figure 5.1 b)).

If there are no descendants in the tree or the existing descendants do not map onto any KEGG genomes (shown as grey circles in Figure 5.1), the tree is navigated again in order to identify the parent of the original node (shown as a purple circle in Figure 5.1 c)) and the full set of its descendants is identified. If these descendants map onto any KEGG genomes, these genomes are used to represent the original NCBI taxonomy ID — refer to Figure 5.1 c).

If this also yields no positive mapping results, we continue to sequentially climb the tree to the level above (i.e. the grandparent, great-grandparent etc., shown in Figure 5.1 d) as an orange circle) and identify all of their children and the corresponding KEGG IDs (represented

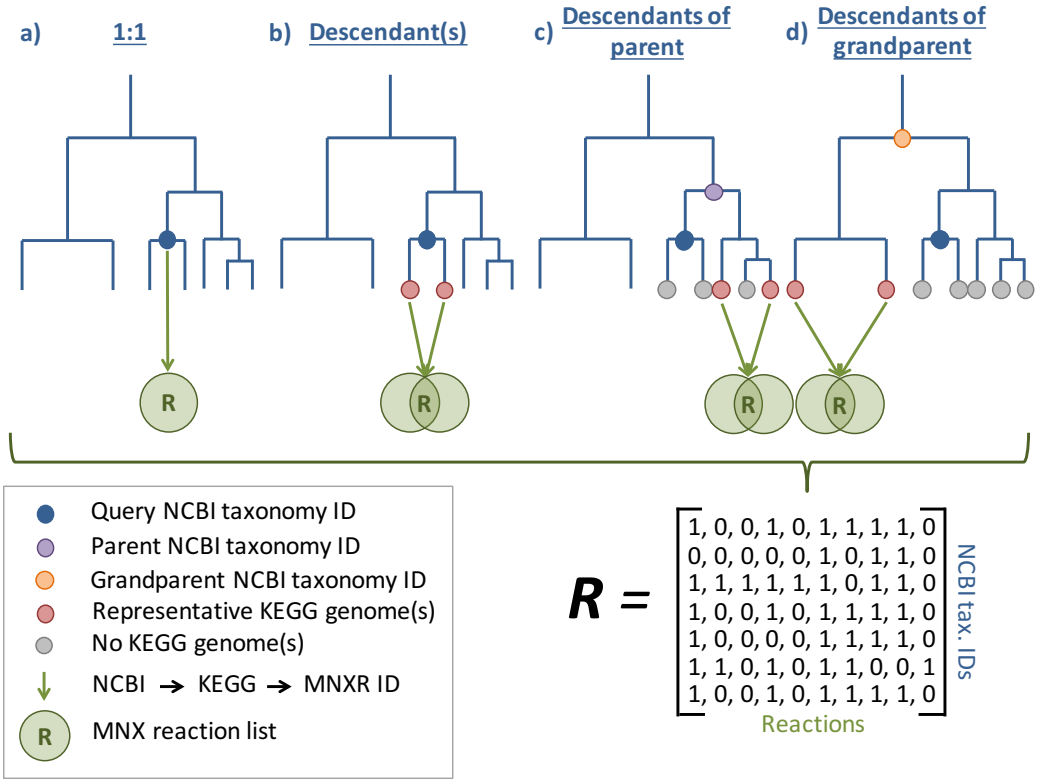


Figure 5.1: **a)** *When a 1:1 mapping between the NCBI ID and a KEGG genome exists, the metabolic reaction set (green circle) for that genome is used to represent the metabolic capabilities of the NCBI ID. The metabolic reaction set is determined by mapping from the NCBI ID to the KEGG genome, KEGG reactions and then the MetaNetX reaction IDs (shown by the green arrows).* **b)** *When 1:1 mapping between the NCBI ID and a KEGG genome is not possible, we search for the descendants of the query NCBI ID (shown as red nodes). If the genome of at least one descendant is present within the KEGG genomes database, the query NCBI ID is mapped to the available genome(s) and reaction list(s) of the descendants (green arrow) and the intersection of the available reaction lists (green circles) is used to represent the metabolic capabilities of the query NCBI ID.* **c)** *If the NCBI ID does not map directly to a KEGG genome and genomes of its descendants are not available (shown by the grey nodes), the parent NCBI ID is identified (purple node). The descendants of the parent node are identified and, if they can map to at least one KEGG genome (red nodes), the available genome(s) are used to map the NCBI ID to the reaction lists associated with the available genome(s) (green arrows). The intersection of the available reaction lists (green circles) are used to represent the query NCBI taxonomy ID.* **d)** *If the NCBI ID does not map directly to a KEGG genome and no genes are identified within the KEGG database for its descendants are the descendants of its parent, the tree is navigated again to the next level up — in this case, the grandparent node is identified (shown by the orange node). The descendants of this node are identified, as was done previously for the parent node, and the same mapping process is used to determine the reactions associated with the query ID. If this yields no representative genomes, tree navigation continues upwards (great-grandparent etc.) until the closest representative are found within the KEGG genomes database.* **Matrix (R):** *Each NCBI taxonomy ID is finally represented by a set of reactions which can be represented within a reaction matrix,* **R**.

in Figure 5.1 d)). These genomes are used to represent the original NCBI taxonomy ID.

Mapping KEGG genomes to MetaNetX reactions

Each KEGG genome has a set of metabolic reactions associated with it within the KEGG database, and these can be retrieved via the REST-style KEGG API. For the NCBI taxonomy IDs that mapped directly to a single KEGG genome, we simply take the reaction list associated with that genome. In the cases where there are multiple KEGG genomes representing a given NCBI taxonomy ID, we take the intersection of the reactions found within those genomes, as shown by the Venn diagrams in Figure 5.1 b), c) and d). This is based on the assumption that the reactions present in all of the mapped descendants are likely to be present in the most recent common ancestor (MRCA) of the bacteria represented by the NCBI taxonomy ID. Once all of the KEGG reactions have been identified, these can be converted to MetaNetX reaction (MNXR) IDs using the namespace mapping flat files available for download from the publicly available MetaNetX database (Ganter *et al.*, 2013).

Matrix representation of reaction vectors for each NCBI taxonomy ID

After identifying which reactions are associated with each NCBI taxonomy ID, this information can be represented by a reaction matrix, $\mathbf{R}$. This matrix is a binary matrix where each NCBI taxonomy ID is represented by a different row and each column represents a different reaction — refer to Figure 5.1. Values of 1 represent a particular reaction being present in the bacteria represented by the given NCBI taxonomy ID, whereas 0 represents absence of that reaction in that bacteria.

5.2.3 Correcting weights according to 16S rRNA gene copy numbers

We can assign a weight to each NCBI taxonomy ID based on the number of reads classified with that ID within the microbiota samples for the three phenotype groups (NGT, IGT and T2D).

These weights are related to the relative abundances of each of these taxonomy IDs within each phenotype group. However, these numbers do not directly represent the relative abundances of the different bacteria due to the fact that the 16S rRNA gene is present with varying copy number across different bacteria. In order to correct for this, we need to take into account the 16S copy number for each NCBI taxonomy ID and use these copy numbers to scale the weights accordingly. This was done by creating a database to map NCBI taxonomy IDs to their 16S rRNA gene copy numbers. The methods used to create and query the database are discussed below.

Creating a MySQL database to map GG IDs to NCBI taxonomy IDs

In order to map the NCBI taxonomy IDs to the best approximations of their 16S rRNA gene copy numbers, we must use databases beyond those available from NCBI. One such database is *rrnDB*, but the flat files available to the public via the *rrnDB* File Transfer Protocol (FTP) do not include mapping files for NCBI taxonomy IDs (Klappenbach *et al.*, 2001). However, the PICRUSt tool that Morgan *et al.* (2012) used for their analysis includes a file that maps GreenGenes IDs to their associated, pre-calculated 16S rRNA copy numbers (Langille *et al.*, 2013). GreenGenes is a publicly available 16S rRNA gene sequence alignment that can be used for taxonomic classification of 16S rRNA gene sequence reads (DeSantis *et al.*, 2006). Therefore, it is possible to map the GreenGenes IDs (which represent different bacterial operational taxonomic units, OTUs) to NCBI IDs using their taxonomic classifications (i.e. scientific names) and taxonomic levels, otherwise known as ranks (e.g. species, genus, etc.). This mapping was done by creating a MySQL database with three tables, and a subsequent view with joins, such that the 16S rRNA copy numbers could be efficiently retrieved via queries to this database. The schema for this database is shown in Figure 5.2.

The *NCBI_CLASSN* table was created using data from the NCBI taxonomy database, to allow subsequent mapping to another table containing the equivalent GreenGenes taxonomic data. The data for the *NCBI_CLASSN* table was obtained by processing the *nodes.dmp* and

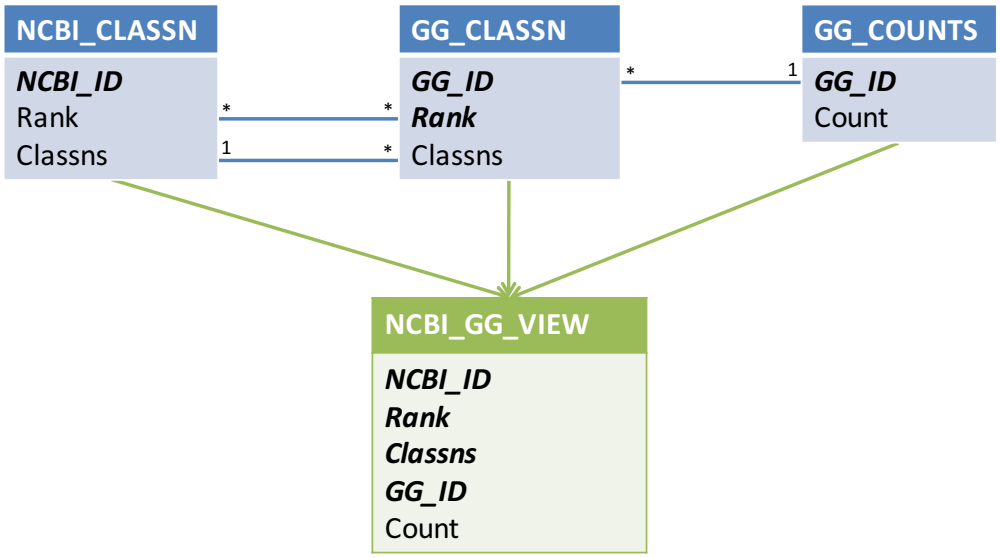


Figure 5.2: *The three database tables are shown in blue and the resulting view is shown in green.* **NCBI_CLASSN:** *This table was built using data from the NCBI taxonomy database. The NCBI taxonomy ID acts as the primary key and is linked to its corresponding taxonomic rank and scientific name.* **GG_CLASSN:** *This table was built using GreenGenes data files that map OTUs to their scientific name and rank. In this table, the GreenGenes ID and rank form a composite primary key because each GreenGenes ID is present independently with every rank. The rank and classification entries can be mapped to those within the NCBI_CLASSN table.* **GG_COUNTS:** *This table was created from the PICRUSt flat file that maps the GreenGenes IDs to 16S rRNA gene copy numbers. The GreenGenes IDs act as the primary key for this table. These IDs can be mapped directly to the GreenGenes IDs present within the GG_CLASSN table.* **NCBI_GG_VIEW:** *This integrated view was created by joining the three tables. NCBI_CLASSN and GG_CLASSN are joined via Rank and Classns. The mappings via Rank are many-to-many because many entries in both tables will share the same rank status (e.g. species, genus etc.). The joins via Classns will be one to many because the NCBI_CLASSN table will only have one entry with a given classification (scientific name), while the GG_CLASSN table will have many entries with a given classification (scientific name) due to the fact that it contains an entry for every rank of the given classification, unlike the NCBI_CLASSN table, which only contains the lowest rank for a given entry (i.e. if* Escherischia coli *was entered with its NCBI ID as a species, it would not also be entered with its NCBI ID and its genus rank of* Escherichia*, unlike the GG_CLASSN table, which would contain the genus rank as a separate entry with the same GG ID. The resulting view can be queried using NCBI taxonomy IDs. It is only by mapping both Rank and Classns together that we can get a one-to-one mapping of NCBI_ID to GG_ID. GG_CLASSN and GG_COUNTS tables are joined via the GG_ID, which arises many times in the GG_CLASSN table but only once in the GG_COUNTS table. The mappings between the three tables give rise to the NCBI_GG_VIEW, which allows us to map the gene copy number from the GreenGenes data ('Count') to the NCBI IDs ('NCBI_ID').*

names.dmp files from the NCBI Taxonomy database FTP, such that each NCBI taxonomy ID (*NCBI_ID*) was mapped to its rank (*Rank*, which is a single letter representing rank, e.g. species = s, genus = g etc.) and its classification (*Classns*), which was determined by the *'scientific*

name' category in the *names.dmp* file. The *NCBI_ID* is the primary key for this table.

The *GG_CLASSN* table was created using data obtained from the GreenGenes website where they provide the taxonomy files such that each GreenGenes ID (*GG_ID*) is mapped to its classification (*Classns*) across all ranks (*Rank*), where the classification is the scientific name for that rank. The GreenGenes IDs were saved within the *GG_CLASSN* table such that there is an entry for each *GG_ID* with each *Rank*. Therefore, the *GG_ID* and *Rank* form a composite primary key for this table. Therefore, this table contains data equivalent to the *NCBI_CLASSN* table, but is based on the GreenGenes IDs rather than the NCBI IDs. Note that the GreenGenes IDs represent different OTUs, which means that different GreenGenes IDs may map to the same taxonomic classification labels (i.e. many GreenGenes IDs may map to a single NCBI ID). This is due to the way that OTUs are formed via sequence-similarity clustering.

The *GG_COUNTS* table is the most straight-forward table. It was generated using the tab-delimted pre-calculated 16S rRNA gene copy number file available from PICRUSt and was downloaded directly from their website (Langille *et al.*, 2013). The *GG_ID* forms the primary key in this table.

In order to map the NCBI taxonomy IDs to the 16S rRNA gene copy numbers, they have to be mapped via the GreenGenes classifications. This was done such that *GG_COUNTS* is joined to *GG_CLASSN*, which is also joined to *NCBI_CLASSN* via *Rank* and *Classns*. This forms the integrated view, *NCBI_GG_VIEW*.

Querying the MySQL database using NCBI taxonomy IDs

In order to determine the best approximation for the 16S rRNA gene copy number for each of the NCBI taxonomy IDs identified within the IBD data, a non-redundant list of NCBI IDs was generated from all three data sets (NGT, IGT, T2D) and any uninformative classifications (refer to Table 5.2) removed before addressing each ID in turn.

For each *id*, the view was initially queried as follows to find any direct matches:

Table 5.2: *Uninformative classifications removed from the set of NCBI IDs.*

NCBI ID	**Classification**	**Reason for removal**
0	Root	This would return all entries in the database because it is the root of the taxonomic tree. This classification simply means that any sequences returning this ID are unclassified.
2	Bacteria	This would return all bacterial entries in the database, which is not a high enough level of classification to be informative.
131567	Cellular organisms	This would return all cellular organism entries in the database, which is not a high enough level of classification to be informative.

SELECT * from NCBI_GG_VIEW

WHERE NCBI_ID =*id*

which returns all classifications and ranks for the GG_IDs associated with each query NCBI taxonomy ID (*id*) in turn, along with their 16S rRNA copy numbers. The copy number assigned to a given ID is deemed to be the median of all values returned.

For some IDs, there will be no direct matches in the GreenGenes database due to the imperfect nature of bioinformatics databases. If there are no matches in the GreenGenes database, the ID is checked against the NCBI taxonomy tree. If it is present within the tree, the descendants of the original NCBI ID are selected and used to query the database. The copy number assigned to a given ID is deemed to be the median of all 16S copy number values returned.

If, after searching for descendants, there is still no match, the parent of the original node is identified with its ID and is used to query the database. This returns the copy numbers for this ID, including all of its descendants, and therefore the sister clades of the original ID. In this case, the copy number assigned to the original query ID is deemed to be the median of all 16S copy number values returned.

The reason for checking if the NCBI taxonomy ID is found within the tree is because sometimes IDs are merged during updates of the NCBI taxonomy database, giving rise to the loss of an ID. On the rare occasions that the ID is not found within the tree, this ID is considered to be

unclassified and cannot have a copy number assigned to it.

5.2.4 Weighting the metagenome-scale metabolic model

If we represent the NCBI taxonomy ID counts as a vector, ($\mathbf{t}$), and the reciprocals of their corresponding 16S rRNA gene copy numbers in a vector of the same length, ($\mathbf{n}$), we can scale the abundance values via entry wise multiplication of these two vectors, as shown in Equation 5.1:

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} n_1 t_1 \\ n_2 t_2 \\ \vdots \end{pmatrix}, \tag{5.1}$$

where t_i and n_i are the abundance value and reciprocal of the copy number associated with the i^{th} NCBI ID, respectively.

The resulting vector, $\mathbf{c}$, contains all of the corrected relative abundance values for all of the NCBI taxonomy IDs. This can then be used to apply corrected abundance weights to the reaction matrix, $\mathbf{R}$, to create the final weighted reaction vector, $\mathbf{w}$:

$$w_j = \sum_i c_i R_{ij} \tag{5.2}$$

where i is the i^{th} NCBI ID and j is the j^{th} reaction in the reaction matrix.

This process is carried out for cases and controls separately, such that you end up with two weighted reaction vectors, $\mathbf{w}_{case}$ and $\mathbf{w}_{control}$. Each entry in these vectors represents the normalised weight for a single reaction present within the MetaNetX database. These weights can then be associated with their corresponding reactions in the previously discussed metagenome-scale metabolic model (which was created from the MetaNetX database (Ganter *et al.*, 2013)) to create two weighted networks — one for cases and one for controls. The resulting weighted networks for cases and controls can then be compared.

5.2.5 Comparing the two weighted networks using AMBIENT

The two weighted reaction vectors, $\mathbf{w}_{case}$ and $\mathbf{w}_{control}$, must be compared in order to identify how the metabolic capabilities of the microbiota in cases differs to those in the controls. This can be done using AMBIENT (Active Modules for Bipartite Networks), which uses machine learning methods to identify modules (subnetworks) within metabolic networks that demonstrate a coherent and significant change in different conditions (Bryant *et al.*, 2013). In the case of the inferred metanetworks, it can be used in conjunction with a community scale metabolic network generated from the MetaNetX database in order to identify parts of this network that show significantly different weighting in cases when compared to controls.

AMBIENT uses a simulated annealing algorithm to identify metabolic modules that change coherently between two conditions, such as in the disease state and healthy control. These coherent changes are identified using fold change values associated with the reactions in the metabolic network. The input required by AMBIENT includes the full metabolic network (in the case of this project, this is a metagenome scale metabolic network) and scores for the reactions for which there is data.

The aim of the algorithm is to find a set of non-overlapping modules with the highest total score, where the score of a given module takes into account:

- the total number of nodes in the module (i.e. the number of reactions and metabolites),
- the fold change scores associated with the metabolites, which are determined by the degree of the metabolites across the whole metabolic network,
- the weights associated with the metabolites, which are determines by the degree of the metabolites across the whole metabolic network,
- an overall weighting term that balances the reaction scores against the metabolite weights. This is to help balance the effects of common metabolites (e.g. H_2O) to ensure that:
    1. modules are not restricted by metabolite weights,

2. but also do not expand across the network to create modules that are not useful.

The process of identifying the non-overlapping modules with the highest total score is NP-hard, hence requiring a sampling approach such as simulated annealing, where the solution space is searched. This is done by toggling a random set of edges (i.e. sampling a set of edges and adding them to the test set) in each step. The total score of the resulting modules is then considered within the context of the current acceptance threshold from the threshold schedule, at which point the toggled edges are either accepted or rejected. This sampling and comparison protocol is repeated while governed by the schedule of decreasing acceptance thresholds, thereby gradually improving scores of the identified set of modules until the threshold comparison set is exhausted. The p-values associated with the final modules are determined empirically by randomly selecting subnets of the same size from within the full metabolic network P times.

Metagenome-scale metabolic model

Given that the MetaNetX reaction database contains information about every biochemical reaction that could occur in a given bacterial cell, it is possible to generate a metagenome-scale metabolic model in XML format containing all of these reactions. This was done by Dr Will Bryant (unpublished work, Imperial College London), so here we use this metagenome/community-scale metabolic model as the input model for AMBIENT.

The fold change in cases compared to controls was calculated using the two weighted reaction vectors, $\mathbf{w}_{case}$ and $\mathbf{w}_{control}$, in order to generate a suitable input file for use by the AMBIENT tool. This input file and the community-scale metabolic model were run in AMBIENT with the parameters set to those shown in Table 5.3.

Table 5.3: *Table of parameters used to run AMBIENT*

Parameter	**Value**
-m (input metabolic network in SBML format)	metanetx.xml
-s (input scores from a tsv, using reaction IDs)	IGT_to_T2D_AMBIENT.txt NGT_to_IGT_AMBIENT.txt NGT_to_T2D_AMBIENT.txt
-e (output name for results files)	IGT_T2D_pos_run NGT_IGT_pos_run NGT_T2D_pos_run
-N (number of edge toggles)	100,000
-P (number of tests for empirical significance testing)	1,000

5.3 Results and discussion

The proportion of successfully classified sequences are presented below, followed by the results of mapping the taxonomic IDs to their reaction lists. The results of the copy number normalisation are presented with the total number of reactions with non-zero weights presented for each dataset (NGT, IGT and T2D). The failure of AMBIENT to identify differentially abundant modules is discussed alongside some ideas for future work.

5.3.1 Classifying 16S rRNA sequences

The number of sequences within each dataset and the number that were successfully classified are shown in Table 5.4.

Table 5.4: *Summary of sequence classification data for the diabetes datasets*

Dataset	**Number of sequences**	**Number classified**	**Number of unique taxonomy IDs**	**Number of bacterial taxonomy IDs**
NGT	1,074,691	1,039,249	1,366	1,347
IGT	1,449,745	1,405,013	1,393	1,378
T2D	1,713,572	1,664,160	1,411	1,387

5.3.2 Mapping sequence classifications to MNXR IDs

Across all three groups of samples, a total of 1,387 bacterial NCBI taxonomy IDs were identified: of these, 941 NCBI taxonomy IDs mapped directly to a KEGG genome, and 446 did not. This resulted in 1,730 KEGG genomes being used to represent the 1,387 NCBI taxonomy IDs. These extra genomes arise from the fact that any NCBI taxonomy IDs that did not map directly will be associated with numerous KEGG genomes, as described previously in Section 5.2.2. Non-bacterial taxonomy IDs were omitted.

Of the 1,730 KEGG genomes used to represent the NCBI taxonomy IDs, 23 did not have any KEGG reactions associated with them. Therefore, these 23 genomes were removed from the analysis when creating the list of reaction intersections for each NCBI taxonomy ID. This meant that only 17 NCBI taxonomy IDs were not associated with any KEGG reaction IDs, compared to >100 when these genomes were not removed. These 17 IDs are shown in Table D.1 within Appendix D. The NCBI taxnomy IDs that were mapped successfully to KEGG reaction IDs were stored in a dictionary with their KEGG reaction lists.

Each NCBI taxonomy ID from the taxonomy abundance files was used to extract the corresponding reaction information from the previously discussed dictionary, and the KEGG reaction IDs were converted to MetaNetX reaction IDs. 8 bacterial NCBI taxonomy IDs could not be associated with any MetaNetX reactions because they were not in the dictionary discussed above. 2 were general classifications (0 = unclassified, 131567 = cellular organisms) which were removed from the matching process used to map NCBI taxonomy IDs and KEGG reactions. The remaining 6 IDs were old IDs that have since been merged with others within the NCBI taxonomy database. These 6 merged IDs were replaced with the updated IDs, and subsequently were successfully mapped to MetaNetX reactions.

5.3.3 Correcting weights according to 16S rRNA gene copy numbers

All NCBI taxonomy IDs were queried, regardless of whether they were bacterial or not (i.e. including archaea and viral sequences), meaning that 1682 IDs were used to query the copy number database. Only 7 NCBI taxonomy IDs could not have a copy number assigned; of these, 6 were the merged IDs discussed previously and 1 represented the Simbu virus, which is excluded from the rest of the analysis because it is not bacterial. The maximum copy number identified within the set of assigned copy numbers was 12, and the most common copy number was 3.

The copy numbers were successfully used to normalise the reaction weights for each taxonomic assignment. After applying this normalisation step and summing the reaction weights across all taxonomic assignments, the three datasets had differing numbers of reactions with non-zero weights. Of the 135,995 reactions present within the MetaNetX database, NGT contained 3108 with non-zero weights, IGT contained 3153 with non-zero weights and T2D contained 3120 with non-zero weights.

5.3.4 Comparing the two weighted networks

The number of reactions for which fold-change could be calculated for the comparisons of the NGT, IGT and T2D reaction weights are shown in Table 5.5.

Table 5.5: *Numbers of reactions with numerical fold change values when comparing the weighted networks for the three diabetes groups.*

Comparison	**Reactions with non-NA values**
NGT/IGT	3050
NGT/T2D	3031
IGT/T2D	3057

The assessment using AMBIENT revealed no differentially abundant modules in any of the three comparisons made. When the distributions of fold change scores were assessed, their distributions were very different to those provided with the example model for AMBIENT;

specifically, they displayed a skewed distribution rather than a symmetrical normal distribution, unlike the AMBIENT example model. The highest fold change value observed in the example was 3.88, compared to 294.13 for the diabetes data. It may be that this difference in distributions of scores cannot provide predictions of differentially abundant modules. Figure 5.3 shows the fold change values from the comparison of NGT to IGT alongside the fold changes observed within the example model provided with AMBIENT. The 12 fold change values within the NGT/IGT diabetes data that were greater than 20 were omitted for clarity.

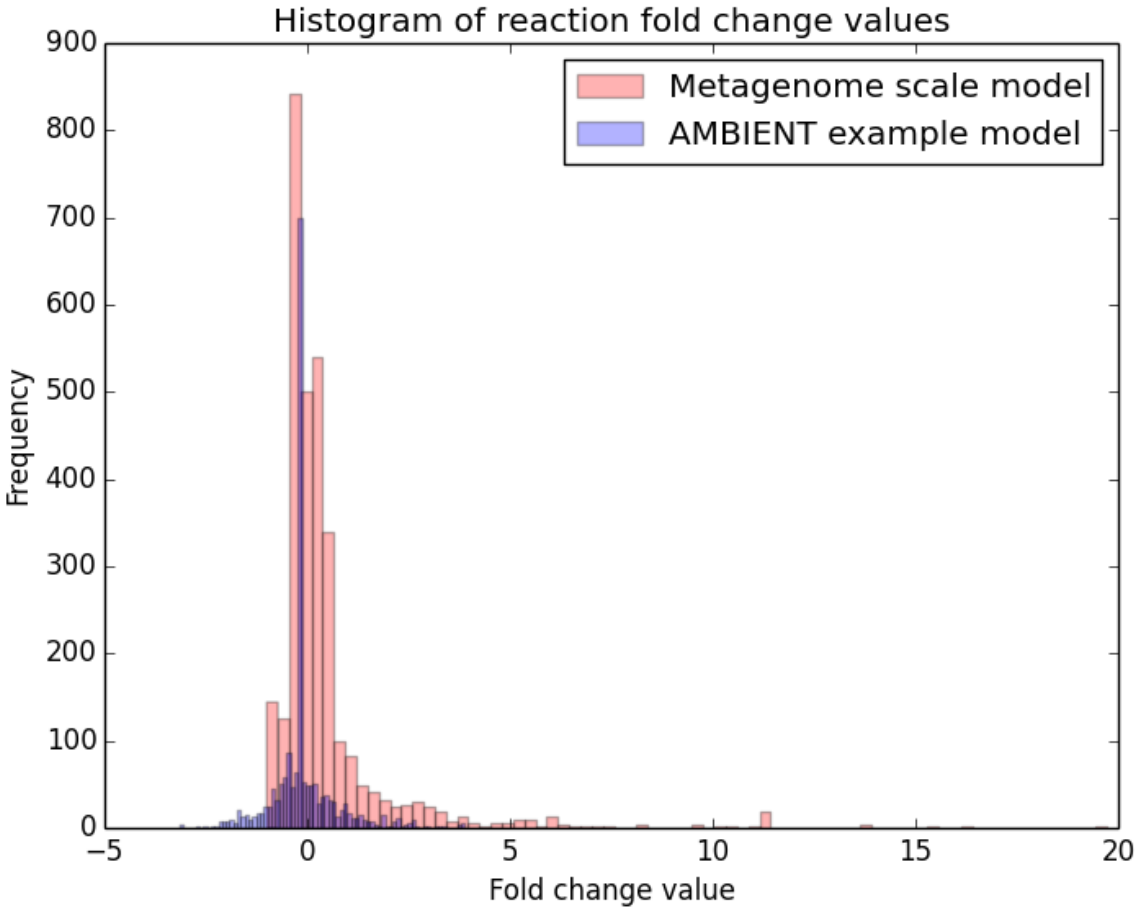


Figure 5.3: *The fold change values when comparing NGT to IGT (shown in red) are compared to those observed in the example model supplied with AMBIENT (shown in blue). The diabetes data (NGT/IGT) shows skewed distributions with more extreme positive values; in contrast, the example data is more symmetrically distributed.*

For future work, it would be beneficial to explore a number of ways to compare the pairs of NGT, IGT and T2D networks. Firstly, we may be able to transform the fold change data or use

a different distance measure, such that AMBIENT could still be used to identify differentially abundant modules within the networks. The log of the fold changes cannot be taken because a large number of values are negative, but a different measure of change may be used in place of fold change, such as measuring the ratios of reaction weights ($NGT_{weights}$, $IGT_{weights}$, $T2D_{weights}$) between two modules:

$$\frac{IGT_{weights}}{NGT_{weights}} \tag{5.3}$$

Secondly, it is important to note that AMBIENT has been used here on a much larger scale model than ever before — this application has stretched AMBIENT beyond the means for which it was built. Therefore, there may be ways in which we can change this analysis to bring it closer to the scale of the genome-scale analyses that the tool was designed for. For example, the reactions that were assigned a fold change value of N/A (of which there were approximately 130,000 for each comparison) are assigned the median value by AMBIENT, but these reactions could be removed from the metagenome-scale metabolic network in order to only assess the reactions for which we have numerical values. Given that most of these reactions are absent from all three datasets (NGT, IGT and T2D), most of the reactions will not be relevant to the comparisons and can therefore probably be removed from the comparison networks. It would be worth trying to remove any nodes that are not in the case or control metagenome-scale networks, if they are also not direct neighbours of non-zero weighted reactions (i.e. remove unreached/unused parts of the network). This will change the distributions of scores by removing superfluous data where AMBIENT has automatically assigned median weight scores, and make the metagenome scale model much smaller, thereby bringing the scale of the network down and bringing it closer to the scale of the genome-scale networks that AMBIENT was designed for.

Thirdly, the huge metagenome-scale network could benefit from using more edge toggles in each search round, thereby searching more of the network at once. Similarly, because the larger network has a larger search space, it could be worth increasing the number of search iterations (i.e. changing the schedule of testing thresholds) to accommodate the larger network.

Finally, it may be possible to develop a new method besides AMBIENT. It would be interesting

to run an analysis that is similar to a gene set enrichment analysis but using reaction sets instead. These reaction sets could be defined by their links in the network, such as shortest paths between nodes. This means that they do not have to be defined by pathways defined in KEGG but instead can take into account the whole network and connectivity between metabolites.

5.4 Conclusions and future work

In this chapter, we have shown that it is possible to create an inferred metagenome and present it as a weighted metagenome-scale metabolic model rather than a collection of weighted KEGG pathways. This creates an opportunity to analyse the metabolic capabilities of the microbiota in a different way to previous studies, with the aim of removing the constraints posed by applying single weights to whole pathways.

The weighted metagenome-scale metabolic networks are a new resource that could be compared to draw conclusions about differences in the metabolic capabilities of different microbial communities. Here, we showed that AMBIENT could not identify differentially abundant modules in these huge metageome-scale networks when using raw fold change values, but it would be beneficial to try other distance measures, transformation of the data, reducing the network, or changing the search parameters used by AMBIENT. All of these are avenues worth pursuing in order to stretch the application of AMBIENT to this new scale of analysis.

Furthermore, it may be possible to develop new comparison methods. The structure of the metagenome-scale metabolic network can impart information about connectivity of metabolites and reactions, which could be exploited alongside the reaction weights to determine reaction sets that can be compared across pairs of models.

The previously existing problem of mapping 16S rRNA gene copy numbers to NCBI taxonomy IDs has been solved via the production of a MySQL database. This database was used successfully in this study to normalise 16S rRNA gene counts. This resource will also be useful

for future analyses that require 16S rRNA gene copy numbers to be mapped via the scientific name of an organism. It could be useful to generate an online front-end for this database such that it could be queried by others for future research projects.

The concept of inferred metanetworks has been explored in this chapter, but there is still a way to go before we can increase the resolution beyond those that use weighted predefined pathways from KEGG.

Chapter 6

ROOMM documentation

6.1 Introduction

Genome-scale metabolic models can now be generated semi-automatically using information in online databases, such as KEGG and Model SEED (Kanehisa *et al.*, 2016; Overbeek *et al.*, 2005). The metagenome-scale model used in Chapter 5 was generated from information available in the MetaNetX database (Ganter *et al.*, 2013). However, there are a number of issues with these semi-automatically generated models, which cause certain types of quantitative analyses to be inaccurate, such as flux analyses. One key limitation is that the biomass reaction included in each of these models contains incorrect stoichiometries, so it is left to the user to curate this reaction based on current knowledge of the cell type being modelled. Flux analyses of these models (such as flux balance analysis, FBA, which was introduced in Chapter 2) are dependent upon accurate biomass reactions and metabolic pathways in order to produce meaningful results. The more accurate the reactions and pathways are within a given model, the more accurate the flux analyses will be.

The process of curating a metabolic model is time consuming, and challenging for those who are not accustomed to editing files in the SBML format. To overcome this problem, I have created a desktop application, *Reaction Operations On Metabolic Models* (ROOMM), which

provides a graphical user interface that allows easy curation of reactions within genome-scale metabolic models. A key motivation for developing the application is to allow fast and accurate transfer of information between models that use different namespaces for their metabolite and reaction IDs. By enabling semi-automated transfer of information from trusted, curated models to new models undergoing improvement, the accuracy of metabolite mapping across models is improved and much time is saved when compared to carrying out the process manually. This application allows experts in the field of molecular and cellular biology to curate existing metabolic models without needing to directly access the SBML files, thereby aiding the development of more accurate genome-scale metabolic models. ROOMM can be used to create more accurate quantitative models for use with methods such as FBA in order to better understand cellular metabolism. Ultimately, the ability to create accurate models of bacterial symbiosis (by combining accurate genome-scale models of potentially symbiotic species) will be immensely useful for understanding metabolic symbioses in bacterial communities, but this cannot be achieved to a high standard until the existing single-cell models are of a better standard. ROOMM is one step towards this goal.

In this chapter, I will discuss the methods that I used to build and implement ROOMM and present a summary of the resulting software. In Chapters 7 and 8 I present two different examples of how ROOMM can be used to guide model curation.

The desktop application has two main functions:

1. **Transfer:** The user can transfer information from a reaction in a curated SBML model (known from hereon as the *template model*) to the draft SBML model requiring curation (known from hereon as the *target model*).

2. **Edit:** The user can manually edit the stoichiometries of metabolites within a reaction within the target model.

Table 6.1 gives definitions of terminology specific to the desktop application.

Table 6.1: *Summary of ROOMM terminology.*

Term	**Definition**
Transfer mode	The mode of the software in which metabolite information can be transferred from a curated template model to a target model. Metabolite information can be converted across all namespaces used within the MetaNetX databases.
Edit mode	The mode of the software in which the metabolite stoichiometric values can be altered to take any value.
Target model	The model requiring curation.
Template model	The curated model with accurate reaction information that will be transferred to the target model.

The *Transfer mode* allows the user to transfer reaction information from a curated template model to the target model being curated (see Table 6.1 for terminology). The chosen reaction in the template model should be one that is also thought to be in the target model. For example, when transferring biomass reaction information from one model to another, the template model should be a curated model of a closely-related cell type that is thought to have a similar biomass composition to that of the target model. Once both models are loaded into the application, the user can transfer information about the metabolites involved in a selected template model reaction to that of a new or corresponding reaction in the target model. This process is enabled using mapping files and chemical classification files available from the online MetaNetX and ChEBI database, respectively (Ganter *et al.*, 2013; Hastings *et al.*, 2013). The MetaNetX IDs aid universal mapping of IDs across namespaces, while ChEBI enables the classification of metabolites into groups such as proteins, lipids or other metabolite groups.

The *Edit mode* allows the user to edit the stoichiometries of metabolites within the selected reaction of choice in their target model. It has been provided for use by those who would like to avoid manually editing the SBML file, but it should be noted that the user cannot add metabolites to the reaction in this mode, unlike in the *Transfer mode*. If the user feels comfortable and confident that they can successfully edit metabolite stoichiometries in the SBML files without introducing errors, they can do this directly in place of using this editing mode.

6.2 Implementation

The application was implemented in Python, and makes use of the *libSBML*, *wxPython*, *matplotlib* and *numpy* packages as well as two of the standard Python library modules, *os* and *re* (van Rossum & F.L., 2001; Bornstein *et al.*, 2008; Rappin & Dunn, 2006; Hunter, 2007; Walt *et al.*, 2011). Refer to Table 6.2 for a brief summary of how these pre-existing Python packages were used to create ROOMM — a fuller explanation is provided in Sections 6.2.1 to 6.2.4 below.

Table 6.2: *Summary of the packages used to implement the graphical user interface.*

Package	**Contribution to GUI functionality**
libSBML	Used for parsing the SBML files, manipulating the models and saving them.
wxPython	Used to create the graphical user interface.
matplotlib	Used to generate the mass composition graphics.
os	Used for file navigation.
re	Used for identifying database IDs using regular expressions based upon ID formats.
numpy	Used in the production of the mass composition graphics.

6.2.1 Implementation of the Transfer and Edit panels

For the *Transfer mode*, the *wxPython* package was used to create two nested panels — one for the template model and one for the target model. *wxPython* button objects, drop-down list objects and navigation box objects were used to provide access to the user's file system and for the selection of model attributes, such as reactions and metabolites. A screenshot of the *Transfer mode* can be seen in Figure 6.1.

For the *Edit mode*, the *wxPython* package was used to create a single nested panel for the target model. As for the transfer panels, *wxPython* button objects, drop-down list objects and navigation box objects were used to provide access to the user's file system and for the selection of model attributes. A screenshot of the *Edit mode* can be seen in Figure 6.2.

6.2.2 Loading, parsing and saving SBML files

The *os* package was used to allow file navigation and the *libSBML* package was used to allow parsing of the input (template and target) SBML models and the writing of the altered target SBML model to a new file. Overwriting limits were applied here to ensure that the altered model could not be saved over any existing files.

6.2.3 Mapping metabolite IDs across namespaces

The *re* package was used to create regular expressions that are used to identify IDs from different namespaces (such as KEGG IDs, MetaNetX IDs and BiGG IDs) and allow mapping from one namespace to another (Kanehisa *et al.*, 2016; Ganter *et al.*, 2013; King *et al.*, 2016). The

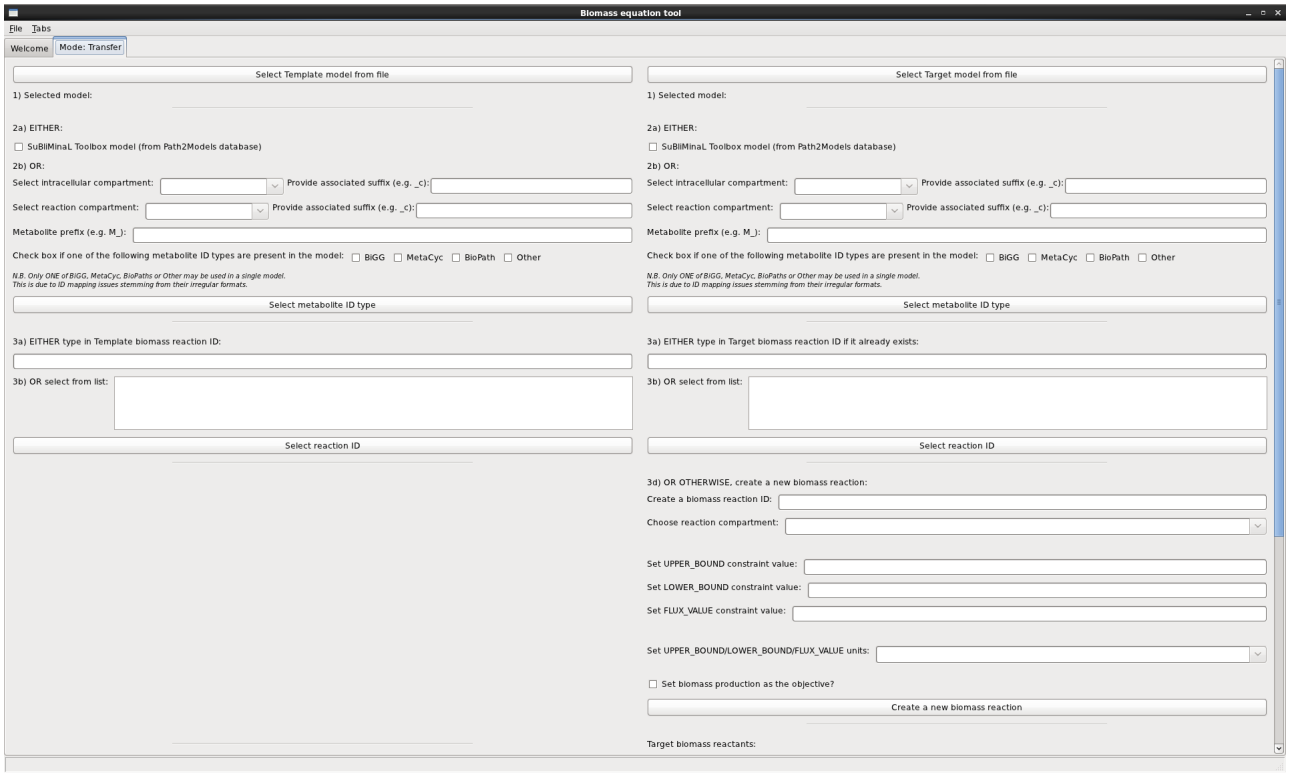


Figure 6.1: *Screenshot of the* Transfer mode *showing two panels: the user loads the curated template model on the left and the target model undergoing curation on the right. The text boxes and buttons allow the user to select reactions and transfer information from the template to the target model.*

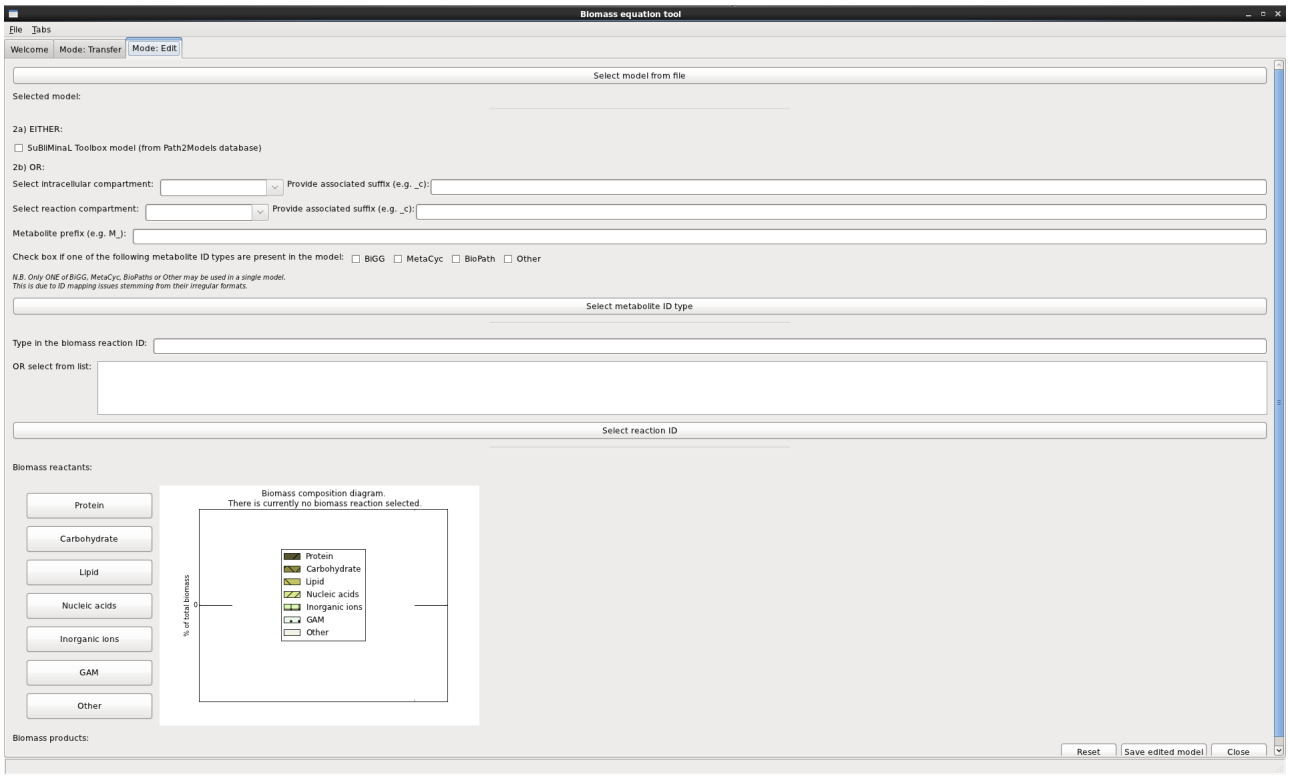


Figure 6.2: *Screenshot of the* Edit mode *showing a single panel. Here, the user loads the target model and selects a reaction in order to modify reactant and product stoichiometries directly.*

mapping of metabolites across from template to target model is facilitated by the MetaNetX metabolite ID mapping files that were downloaded from the MetaNetX database (Ganter *et al.*, 2013).

6.2.4 Mass composition figures

The ChEBI database was used to classify metabolites in the selected reactions of template and target models into different categories, such as proteins, nucleic acids or inorganic ions (Hastings *et al.*, 2013). The relative proportions of these categories are shown in a stacked bar chart for each model (as shown in Figure 6.3) and is updated whenever the selected reactions are updated. The *matplotlib* package was used to create these figures (Hunter, 2007).

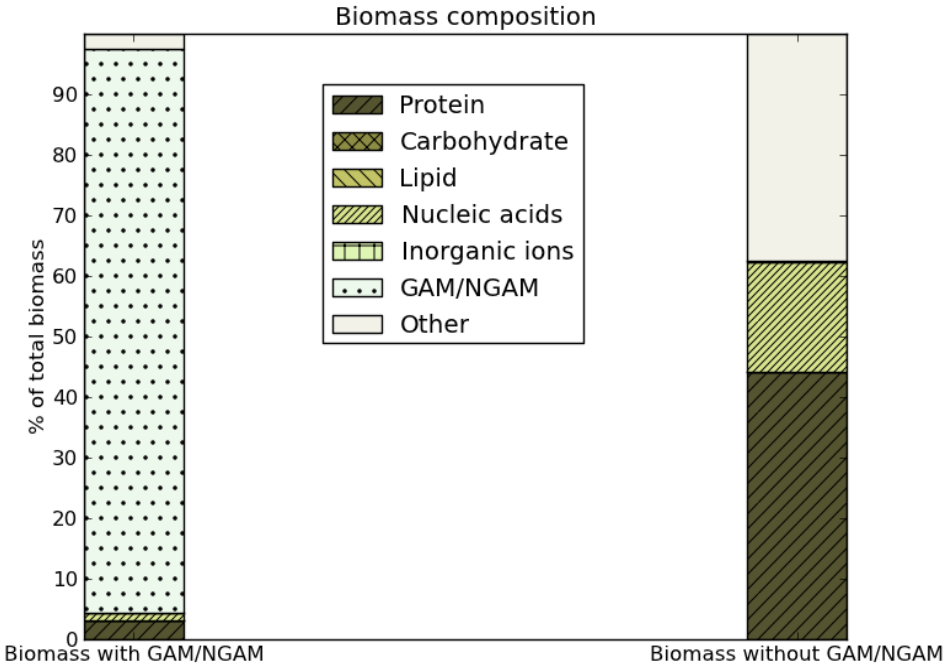


Figure 6.3: *Example of the mass composition diagram for a selected reaction. Here, both bars show the relative proportion of each of the ChEBI categories (shown in different green shades and patterns) within the reactants of the selected reaction. However, the bar on the left includes the GAM/NGAM category (i.e. metabolites classified as energy metabolites) whilst the bar on the right excludes this category. This is because the GAM/NGAM category (shown in light green with dots) will always have a high proportion so more information about the other metabolites can be seen when this category is excluded.*

6.3 User instructions

The software application, ROOMM, resulting from the above implementation is able to carry out two main functions, referred to as the *Transfer mode* and *Edit mode*. Here, the relevant user guidelines and description of purpose are provided for each of the following elements of the resulting software application:

- Transfer mode
- Edit mode
- User input
- Output

6.3.1 Transfer mode

The *Transfer mode* is used for transferring metabolite information from a selected reaction within a curated template model to that of a draft target model requiring curation. The selection criteria for the template model requires that the reaction of interest must be similar to that of the target model organism (e.g. the reaction is transferred from a closely related organism). This is because the process of transferring reaction information from one model to another is based on the assumption that the metabolic capabilites of the two organisms are likely to be similar due to homology.

When the user has selected which metabolites to transfer from the template reaction, ROOMM identifies the corresponding metabolites in the target model and amends the existing metabolite stoichiometries within the target reaction. If a given metabolite is not already present in the target reaction, it adds the metabolite to the reaction along with its associated stoichiometry. Furthermore, as well as being able to transfer information to existing reactions within the target

model, it is also possible to use ROOMM to create a new target reaction, to which metabolite and stoichiometric information can subsequently be transferred.

Classifying metabolites using ChEBI categories

The default mode for transferring metabolite information will copy all of the metabolic information from the template reaction to the target reaction. However, in some cases (e.g. when transferring to biomass reactions), it may be useful to only transfer certain reactants and products. Therefore, the user has the option to select a subset of metabolites to transfer from the template model. This process is helped by organising the reactants into seven different categories (*Protein*, *Carbohydrates*, *Lipids*, *Nucleic acids*, *Inorganic ions*, *Energy* and *Other*) according to the ChEBI ontology for classification of chemical entities (Hastings *et al.*, 2013). The reactants in each of these categories can be viewed within a pop-up window, as shown in Figure 6.4. The user selects which metabolites they want to transfer within each of these categories. The broad compositions of the template and target reactions can be compared by referring to their mass composition figures, as shown previously in Figure 6.3.

Mapping metabolites

The mapping of metabolites across from template to target models is facilitated by MetaNetX metabolite ID mapping files downloaded from the MetaNetX database, so any models utilising IDs with namespaces used within the MetaNetX database can be handled by ROOMM (Ganter *et al.*, 2013). This was done with the aid of regular expressions to identify IDs from different namespaces.

The BiGG, MetaCyc and BioPath IDs can contain any number of any characters and are therefore indistinguishable from each other — these namespaces are poorly-defined. Therefore, only one of these namespaces may be used in a single given model and the user must specify (using the provided drop-down box) if one of these poorly-defined ID types is used. However, these ID types are distinguishable from the other well-defined namespaces with standardised IDs (e.g. KEGG IDs are distinguished as `C` followed by 5 digits). Therefore, any other combination of standardised ID types is accepted by ROOMM, as they can all be easily deciphered from

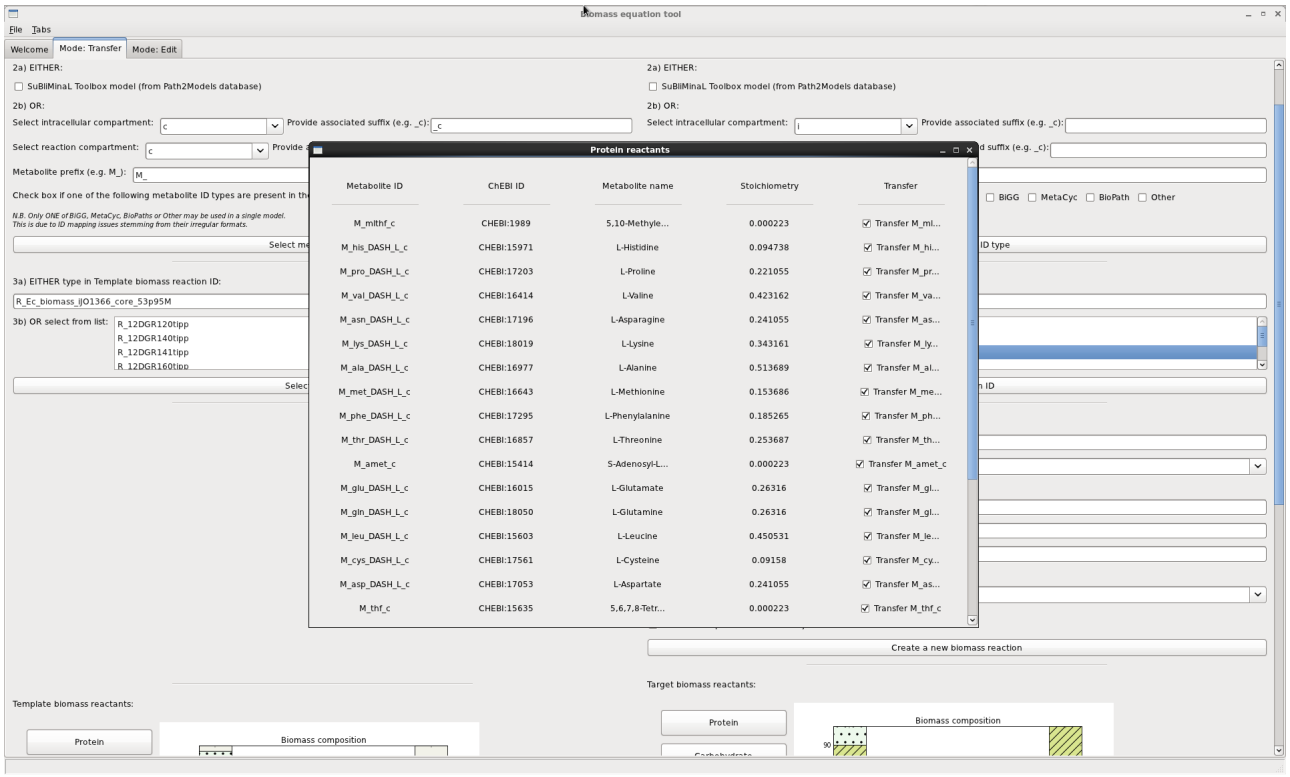


Figure 6.4: *The reactants in each of the seven ChEBI categories can be viewed within pop-up windows, as shown here.*

one another. The standard formats of these well- and poorly-defined namespaces are shown in Table 6.3.

Applying model-specific prefixes/suffixes

ID strings in genome-scale metabolic models are commonly given prefixes and suffixes within models to denote whether they are metabolites or reactions, and to denote which cellular compartment they belong to. For example, the prefix `M_` is often used to denote metabolites whilst a the suffix `_c` is often used to denote the intracellular compartment. Therefore, text control boxes are provided for the user to enter any universal prefixes or suffixes applied to the metabolites in their input models.

Mixed ID types and alterations to standard IDs

The ability to use multiple ID types with custom prefixes and suffixes provides a lot of flexibility

Table 6.3: *Summaries of the namespace formats that can be expressed as regular expressions. Those shown in italics are the poorly-defined namespaces that cannot be distinguished from one another.*

Namespace	**ID format and regular expression**
BiGG	*Variable-length character string.*
MetaCyc	*Variable-length character string.*
BioPath	*Variable-length character string.*
MetaNetX	`MNXM` followed by any number of digits.
KEGG	`C` or `D` or `?` followed by five digits.
ChEBI	Any number of digits.
The Seed	`cpd` followed by any number of digits.

for ID identification and mapping. This allows ROOMM to transfer to and from the semi-automatically generated models (generated via a tool called the SuBliMinaL Toolbox) available from the Path2Models database, which use a combination of BiGG IDs and MetaNetX IDs (King *et al.*, 2016; Ganter *et al.*, 2013; Büchel *et al.*, 2013; Chelliah *et al.*, 2013; Swainston *et al.*, 2011). The prefix and suffix settings used in these SuBliMinaL Toolbox models have been pre-saved in ROOMM so that the user can simply specify their model as being a SuBliMinaL Toolbox model taken from the Path2Models database — this is due to the fact that the BiGG IDs utilised within these SuBliMinaL Toolbox models have had an extra (i.e. non-universal) prefix added. Furthermore, the software can recognise BiGG IDs in various other formats found in many curated models, including a hyphen replacement with a `_DASH_` string and other similar string substitutions within the IDs. These formatting issues with BiGG IDs have been taken into account to allow for a wider source of curated models to be used. Furthermore, the user can load an input model utilising novel IDs if they provide a flat-file for converting those IDs to ChEBI IDs. Whilst this is not recommended due to the work involved in creating the associated flat file, it also means that ultimately any model may be used as a curated template model or a target model, if necessary.

Creating a new reaction

If the relevant reaction does not exist in the target model, the user can create a new target reaction by using the text boxes within the software application to enter a unique reaction ID, flux upper bound, flux lower bound and a flux value. There are extra drop-boxes associated with each of these parameters through which the user can select the correct units for each

parameter. The user can also select a tick-box to set the objective coefficient attribute of the new reaction to 1. This is for cases where the new reaction will be the objective function in FBA models. If this objective coefficient box is selected, all other reactions will then have their objective coefficients set to 0 because there should only be one objective function per FBA model.

Summary of transfer mode

The ID mapping and information transfer carried out by ROOMM is far more efficient than if the user were to manually map each ID across the template and target models individually. Furthermore, the possibility to utilise non-standard IDs if the user provides a conversion flat-file means that ROOMM can be used with any model combinations, if required.

6.3.2 Edit mode

This functionality was provided for use by those who wish to avoid the error-prone process of manually altering a SBML model file and to allow the user to visualise the relative contributions of the different ChEBI categories to the reaction being altered.

Editing metabolite stoichiometries

Once the target model has been loaded, the user can select the reaction to be edited and edit the stoichiometries of the metabolites within that reaction. The reactants are split according to the same ChEBI classifications as those used in the *Transfer mode*. As with the *Transfer mode*, the reactant composition of the target reaction can be visualised by referring to the associated stacked bar plot, which shows the relative contribution of each of the ChEBI categories to the reaction, as shown in Figure 6.3.

6.3.3 User input

Below, the user input required for each aspect of the *Transfer mode* and *Edit mode* are outlined.

Loading the models

In the *Transfer mode*, one file must be supplied containing the template model and another must be supplied containing the target model. In the *edit mode*, only the target model needs to be supplied. Each input file must contain one metabolic model in SBML format and can be loaded by selection in a file navigation window.

Selecting reactions

When a model is loaded, the reaction IDs are parsed from the model and displayed within a list box for the user to select the template/target reaction.

Metabolite selection

The reactants of the biomass reaction are classified according to the ChEBI ontology and are assigned to one of the seven ChEBI groups described previously. Reactants within a given category may be visualised by clicking on the associated category button. This brings up a pop-up window displaying their metabolite information for all of the reactants within that category, including the model-specific metabolite ID and the corresponding ChEBI ID which was used to classify the metabolite type. Alongside the metabolite information for each reactant in the selected ChEBI category, there is a tick box. The user can select/deselect metabolites to transfer/not transfer them to the target model. For a summary of the other metabolite details shown for each metabolite, refer to Table 6.4.

Table 6.4: *Metabolite information displayed in ChEBI category pop-up windows.*

Element	**Definition**
Model ID	The metabolite IDs, as specified within the model.
ChEBI ID	The corresponding ChEBI ID(s), identified by using an ID conversion dictionary.
Metabolite name	The name of the metabolite, as specified within the model.
Stoichiometry	The metabolite stoichiometric value within the selected reaction. (In the *Edit mode*, this is a text box.)
Transfer tick-box	Allows the user to select a given metabolite to be transferred from the template to the target model. (Template pop-ups only.)

Converting IDs

The *Transfer mode* uses ID conversion dictionaries to match up the ID types within the template

and target models. These conversion dictionaries are built in and hidden from the user, and they allow the software application to convert the standard ID namespaces present within the MetaNetX mapping database. If a user wishes to use a SBML model using non-standard IDs, the user must provide flat-files for mapping any non-standard IDs to ChEBI IDs. There is a box provided to the user for the purpose of uploading any conversion files they wish to use. The user must also provide any universal metabolite prefixes or suffixes from within their model(s) — text boxes are provided for this purpose. Check-boxes are also provided in order for the user to specify if either BiGG, MetaCyc or BioPath IDs (i.e. any of the poorly-defined ID namespaces discussed previously) have been used in the model.

Saving the output

Once the user has finished altering a target model (either in the *Edit mode* or *Transfer mode*), the altered model can be saved to a new file in SBML format via a pop-up file navigation window, where the user can provide their chosen file name.

6.3.4 Output

The output can be split into two parts: the transient output within the graphical user interface, and the saved output for the altered model.

Transient output

The biomass composition diagram (refer to figure 6.3) is updated whenever a new reaction is selected or changes are made to an existing reaction when using either the *Transfer mode* or *Edit mode*.

Saved output

Both the *Transfer mode* and *Edit mode* save the altered target model to a new output SBML file. This model will contain all of the changes made to the original (raw) target model. The saved output file can be loaded into ROOMM for further changes at a later date or, if the model curation process is complete, it can be used in constraint-based analyses such as flux balance

analysis.

6.4 Conclusions

The existing bottleneck in the curation of semi-automatically generated genome-scale metabolic models often means that their use in quantitative analyses (such as FBA) is limited. The curation process for these models and the subsequent accuracy of flux-based analyses depend upon the inclusion of a biologically accurate biomass reaction within these models. The process of curation beginning with the biomass reaction is time consuming and challenging for those who are not accustomed to editing files in SBML format. However, 'Reaction Operations on Metabolic Models' (ROOMM) is a desktop application that provides a graphical user interface that allows easy curation of any reactions within genome-scale metabolic models.

Here, I have discussed the methods that I used to implement a model curation tool, ROOMM, as well as presented a summary of the resulting software application. This software application allows experts in the field of molecular and cellular biology to curate existing metabolic models without needing to directly access the SBML files, thereby aiding the development of more accurate genome-scale metabolic models. In Chapters 7 and 8, I discuss two applications of this software: one is a simple example using two models of *Escherichia coli* whilst the other is a more complex example involving models of two *Salmonella enterica* serovars.

Chapter 7

ROOMM: Simple curation of a metabolic model of *Escherichia coli*

7.1 Introduction

To illustrate the usefulness of ROOMM, I have used it to curate two draft genome-scale metabolic models — *Escherichia coli* (this chapter) and *Salmonella enterica* Typhi (Chapter 8). Both of these draft models have been semi-automatically generated by the SuBliMinaL Toolbox and obtained from the BioModels database (Swainston *et al.*, 2011; Chelliah *et al.*, 2013).

E. coli is a well characterised organism due to its prolific use as a model within the wet lab, so models of *E. coli* are used to provide a simple example in this chapter. The target model is the previously-mentioned semi-automatically generated model obtained from BioModels, and the template model is a curated *E. coli* model published by Orth *et al.* (2011), from which we can transfer a quantitatively accurate biomass reaction (Chelliah *et al.*, 2013; Orth *et al.*, 2011). Both models represent the same substrain, K-12 MG1655; in most cases, we would not curate a draft model when a curated version is already available for the given organism, but this example is provided as a simple proof of principle. In the next chapter, we provide a more complex and

realistic application, where information is transferred between two *Salmonella enterica* serovars that are known to cause different, distinct disease symptoms.

7.2 Methods

The curation process begins with the transfer of information from the template model to the target model, followed by fine tuning via stoichiometric coefficient editing; both of these processes are carried out using ROOMM. Once the biomass reaction is as accurate as possible according to the template biomass reaction, FBA is used to determine whether any of the metabolites block flux through the newly altered biomass reaction. Any blocked metabolites are removed from the reaction prior to essential gene predictions using FBA. The essential genes are compared between the raw and altered target models, demonstrating the benefit of curating the biomass reaction and subsequently validating the use of ROOMM to do this. The details of each of these steps are described below.

7.2.1 Transferring the biomass reaction

The template model is a previously published curated model of *E. coli* strain K-12, substrain MG1655 (*i*JO1366, (Orth *et al.*, 2011)). The target model of the same strain of *E. coli* was downloaded from the BioModels database prior to the update made in October 2013 (BMID000000140222, (Chelliah *et al.*, 2013)), so it is a SuBliMinaL Toolbox model that uses IDs in the KEGG format (Kanehisa *et al.*, 2016). All example models are provided with the software.

The target model contains only one biomass reaction with *biomass_reaction* as its ID. The template model has two biomass reactions — the reaction with the following ID was utilised to focus on core metabolism: *R_Ec_biomass_iJO1366_WT_53p95M*. The user input required to run this example is summarised in Table 7.1. Using the *Transfer mode* in ROOMM, all metabolite

and stoichiometric information was selected and transferred from the biomass reaction of the curated template model to the biomass reaction of the target model.

Table 7.1: *The user input supplied in the E. coli example application.*

Input	**Template *E. coli***	**Target *E. coli***
Filename	iJO1366.xml	BMID0000001440222.xml
Reaction compartment	c	i
Reaction suffix	_c	(None)
Intracellular compartment	c	i
Intracellular suffix	_c	(None)
Metabolite prefix	M_	(None)
Namespace	Other	KEGG
Reaction ID	R_Ec_biomass_iJO1366_core_53p95M	biomass_reaction

7.2.2 Editing the biomass reaction

The metabolite representing the lipid content of the target biomass reaction is a unique metabolite with a non-standard ID (ID = _a8dc8dc1_b8b0_4206_b692_6f114244bbb8). This represents a generic lipid formed from various lipid compounds found within the target model. Therefore, due to the non-standard ID, it is not possible to map lipids from the template reaction directly to the target reaction. The total sum of the stoichiometries of the lipid components from the template biomass reaction (shown in Table 7.2) was used as the input for the stoichiometry of the lipid metabolite in the *Edit mode*.

Table 7.2: *Lipids identified within the biomass reaction of the E. coli template model and their stoichiometries. All of these lipids had been characterised within the 'Other' ChEBI category due to the absence of corresponding ChEBI IDs for the conversion flat-file.*

Template lipid ID	**Lipid type**	**Stoichiometry**
M_2ohph_c	Octoprenyl hydroxyphenol	0.000223
M_pe161_c	Phosphatidylethanolamine (**cytoplasmic** dihexadec-9-enoyl, n-C16:1)	0.054154
M_pe161_p	Phosphatidylethanolamine (**periplasmic** dihexadec-9-enoyl, n-C16:1)	0.02106
M_pe160_c	Phosphatidylethanolamine (**cytoplasmic** dihexadecanoyl, n-C16:0)	0.017868
M_pe160_p	Phosphatidylethanolamine (**periplasmic** dihexadecanoyl, n-C16:0)	0.045946

7.2.3 Defining the growth medium for essential gene predictions

A good way to assess the improvement in accuracy of a model is to assess the prediction of essential genes under specific growth conditions; the more accurate the model, the more accurate its essential gene predictions should be. The accuracy of the essential gene predictions can be measured by observing their agreement with experimental results from the wet lab. However, in order to run these simulations, the model must be set up to utilise the same metabolites that are used in the growth medium of the wet lab experiments. Therefore, the target model was simulated to grow on glucose minimal medium, the composition of which was determined by the Model SEED database of growth media, and is shown in Table E.2 (Overbeek *et al.*, 2005). This was done by constraining the upper and lower flux bounds to zero for all import reactions for metabolites other than those within the growth medium. In contrast, the upper bounds for all metabolites within the growth medium were set to 1000, which is the maximum upper flux bound value used within the model.

7.2.4 Identifying and removing flux-blocking metabolites

Some automatically generated metabolic models contain holes in pathways associated with poorly characterised areas of metabolism. Therefore, it is possible that the target model may be unable to fully support the production of all biomass reactants due to insufficient curation in the rest of the model. However, in order to gain predictions of essential genes under specific conditions, such as growth on glucose minimal medium, the cell must be able to pass flux through the biomass reaction when running FBA, otherwise all genes in the model are predicted as essential. Thus, once the target biomass reaction has been curated and the growth medium set, it is compulsory to determine which metabolites (if any) block flux through the biomass reaction.

In order to ensure that the target metabolic model can support the production of its biomass reactants, each of the reactants must be tested separately. This was done by selecting each

biomass reactant in turn and setting all other reactant stoichiometries to 0 within the biomass reaction before running FBA. In each of these analyses, if the cell could grow (i.e. flux was carried through the biomass reaction), the selected metabolite would remain in the biomass reaction; conversely, if the model had a growth rate of 0 and flux could not be carried through the reaction, the selected metabolite would be removed from it. The metabolites that are shown to block flux can either be left out of the analysis or, if necessary, they can be added back in one at a time in order to allow the user to identify where holes lie in the metabolic network. Therefore, these metabolites can be used to guide hole-filling and model curation; this has been done in the next chapter for the complex example using models of *Salmonella enterica* serovars.

7.2.5 Essential gene predictions

To validate the method of information transfer, we have assessed the essential gene predictions of the raw (pre-transfer) and altered (post-transfer) target model. The bacterial strain in the target model is identical to that of the curated template model. Therefore, the changes in essential gene predictions observed in the altered target model were validated using the previously published essential gene datasets that were also used by Orth *et al.* (2011) to validate the curated template model (Baba *et al.*, 2006; Yamamoto *et al.*, 2009). These essential gene data can be found in Table 13 of the Supplementary materials in the publication by Orth *et al.* (2011), showing that of the 433 genes known to be essential for growth on glucose, 248 are contained within the curated iJO1366 model.

As stated previously, the models were simulated to grow on glucose minimal medium to mimic the wet lab experiments. The raw model was set to grow on glucose minimal medium in the same way as the altered model, which is discussed and defined above. The essential genes were predicted for both models using the COBRApy package in Python (Ebrahim *et al.*, 2013).

This works by sequentially selecting each gene declared in the model and blocking flux through any reactions associated with (i.e. dependent upon) the given gene. If the gene is essential, it will block the production of biomass reactants, and consequently block flux through the

biomass reaction (see Figure 7.1, top). Conversely, if there are alternative pathways that allow the reactants of the biomass reaction to be produced and therefore allow flux through the biomass reaction, the gene is considered non-essential (See Figure 7.1, bottom).

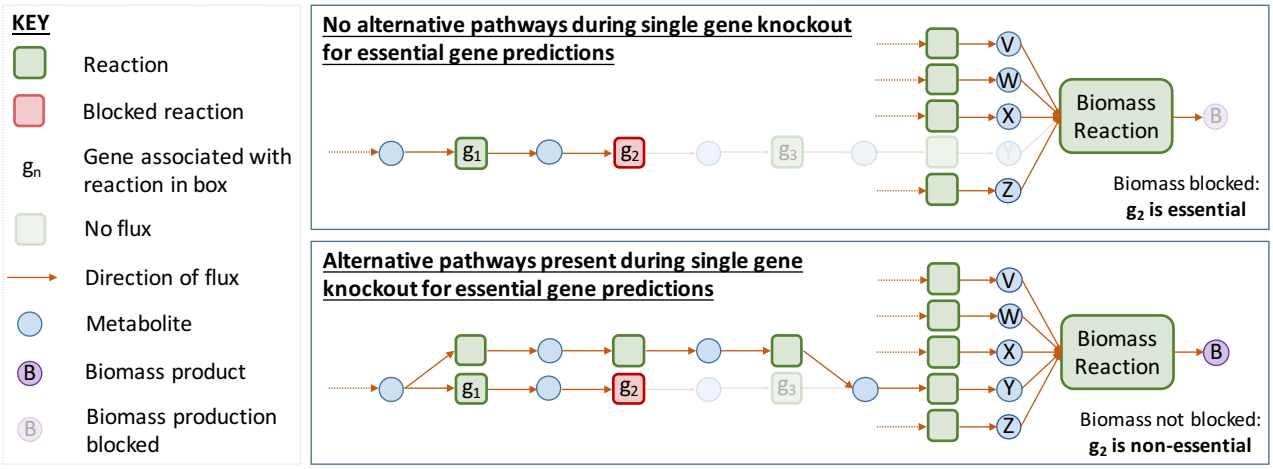


Figure 7.1: *When testing if* Gene 2 *(g_2) is essential, all reactions (green boxes) associated with that gene are blocked (red boxes) by setting its associated flux constraints to 0.* **Top box:** *If the given edge (g_2) is essential its downstream metabolites (blue nodes) cannot be produced (downstream pathway is shown as watermark), thereby blocking production of the biomass reactant, Y. This results in no simulated growth (shown as a watermarked biomass node).* **Bottom box:** *The given gene (g_2) is not essential when alternative pathways can produce the downstream biomass reactant, Y, and therefore bypass the blocked reaction to produce biomass (purple node).*

The sensitivity, specificity and F-measure (F) of both models were calculated for comparison, defined in Equations 7.1 to 7.3 (Lalkhen & McCluskey, 2008; Powers, 2011):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7.1}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7.2}$$

$$F = \frac{2TP}{2TP + FP + FN} \tag{7.3}$$

where FP = False positives, FN = False negatives, TP = True positives, and TN = True negatives.

7.3 Results and discussion

The results from the transfer and editing process are presented and discussed below with the list of blocking metabolites. Finally, the essential gene prediction results are compared for the raw and altered target models, demonstrating the improvement in the target model following the transfer and editing processes using ROOMM.

7.3.1 Transferring the biomass reaction

For the *E. coli* example, all 24 metabolites within the *Protein* category were successfully transferred. The metabolite within the *Carbohydrate* category was successfully transferred, as were all eleven metabolites within the *Nucleic acids* category. Of the eight *Inorganic ion* metabolites, five were not transferred to the target model because these ions are not contained within the model and can therefore be neglected without adverse effects on the flux through other parts of the metabolic network. Three of the four metabolites within the *Energy* category were transferred successfully. The metabolite that could not be transferred here was thiamine diphosphate (known by its BiGG ID, *thmpp*, in the template model and its KEGG ID, *C00068*, in the target model), due to the inability of the application to map between these IDs using the conversion dictionaries built from the MetaNetX database. All products were transferred successfully.

Of the 20 metabolites within the *Other* category, 3 were transferred successfully. The 17 that could not be transferred are shown in Appendix E (Table E.1) alongside the reasons for unsuccessful transfer from the template model to the target model. These reasons include the absence of certain metabolites from the target model, and the absence of mapping IDs in the MetaNetX mapping files (Ganter *et al.*, 2013).

There were no reactants to transfer within the *Lipids* category, but upon further inspection, five of the metabolites within the *Other* category are shown to be lipids that could not be classified due to the fact that they are not contained within the ChEBI database, which is used

to classify metabolites into groups (Hastings *et al.*, 2013). These lipids are the five lipids that were used to generate a stoichiometric coefficient for the generic lipid metabolite within the target model, as shown previously in Table 7.2.

7.3.2 Editing the biomass reaction

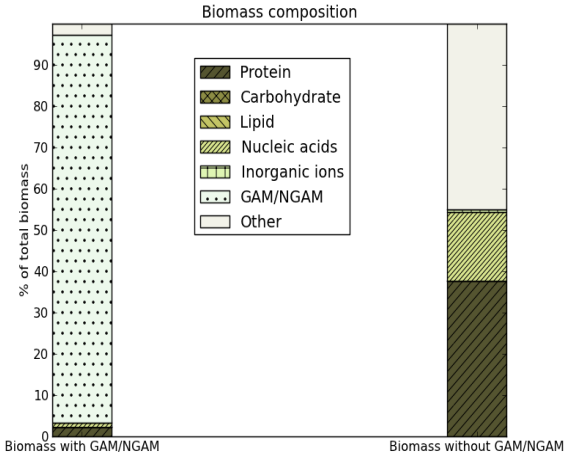
The summed value of the lipid components identified within the template model were successfully applied to the generic lipid reactant within the target biomass reaction. The final altered biomass composition (following the use of the *Transfer mode* and *Edit mode*) more closely represents that of the template model than the original raw target model — the corresponding biomass composition diagrams are shown in Figure 7.2.

7.3.3 Defining the growth medium for essential gene predictions

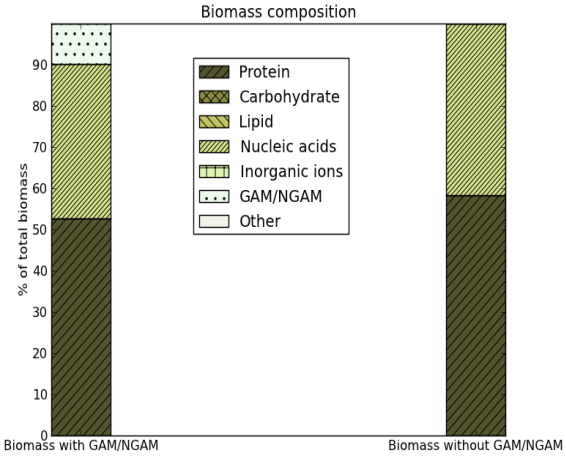
Out of the 18 growth medium components, nine were not contained within the target model and could therefore be neglected. Out of the other nine components, two (Fe^{2+} and Fe^{3+}) did not already have import reactions within the model, so these two import reactions were added. Details of all metabolites within the growth medium and their inclusion in the target model can be found in Table E.2 of Appendix E.

7.3.4 Identifying and removing flux-blocking metabolites

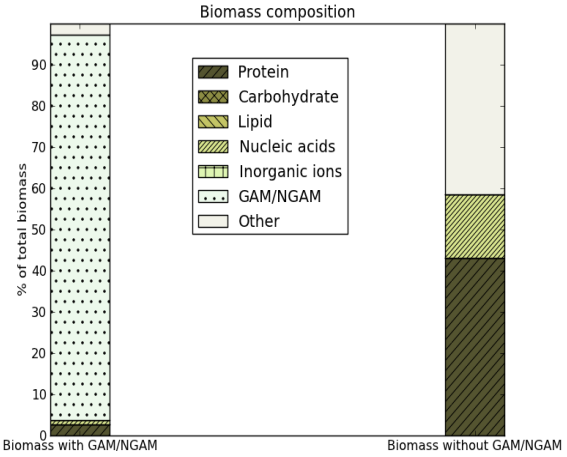
Six metabolites were found to be blocking flux through the biomass reaction, as shown in Table 7.3. Five of these were newly added to the biomass reaction, whilst one was in the original generic biomass reaction of the raw target model. All six were removed from the biomass reaction in order to allow flux through it, and therefore allow for the prediction of essential genes.



(a)



(b)



(c)

Figure 7.2: Figure 7.2 (a) shows the composition of the biomass reaction selected for use within the template model (Orth *et al.*, 2011). Figure 7.2 (b) shows the composition of the biomass reaction within the target model before transferring information from the template model (Chelliah *et al.*, 2013). Figure 7.2 (c) shows the composition of the biomass reaction within the target model after transferring the information from the template model — as expected, the new composition of the target biomass reaction is far more similar to that of the template model.

Table 7.3: *Metabolites removed from the target E. coli biomass reaction due to flux blockage (i.e. preventing growth). Most of these metabolites were newly added to the target biomass reaction due to their presence in the template reaction.*

Metabolite name	**KEGG ID**	**Newly added to target reaction**
L-leucine	C00123	✗
FAD	C00016	✓
Pyridoxal phosphate	C00018	✓
KD02 lipid	C06025	✓
Riboflavin	C00255	✓
Biotin	C00120	✓

7.3.5 Essential gene predictions

The predicted essential genes from the raw and altered target *E. coli* model were compared to experimental results discussed previously (Orth *et al.*, 2011; Baba *et al.*, 2006; Yamamoto *et al.*, 2009). The raw model predicted 30 essential genes with sensitivity of 0.11, specificity of 0.99 and F-measure of 0.20; the altered model predicted 69 essential genes with sensitivity of 0.24, specificity of 0.98 and an F-measure of 0.37. The sensitivity and F-measure have greatly increased in the altered model compared to the raw model, with only a minor decrease in specificity, thereby demonstrating the importance of increased accuracy of the biomass reaction and the ability of ROOMM to aid this process of reaction curation.

7.4 Conclusions and further work

More accurate metabolic models are required in order to understand the metabolic capability of bacterial cells. This is of particular importance for pathogenic organisms in order that drug targets for treatment may be found. Furthermore, metabolic interactions between populations within the microbiota are of interest, given that it has been shown that the microbiota plays a large part in dictating the health of its host. However, before great progress can be made in creating models of symbiosis, quantitative models of metabolic networks must become more accurate. ROOMM aids the integration of expert knowledge of cellular and molecular biologists

with that of systems biologists to enable more accurate modelling of metabolic systems via curation of automatically-generated models. This will allow for better use of these models.

Here we have provided a simple example of how the transfer of biomass information can immediately make semi-automatically generated metabolic models more accurate. However, transfer of biomass information is only the beginning of the curation process. Further curation of the target *E. coli* model would include the investigation of the blocking metabolites shown in Table 7.3, which would guide hole-filling in the network, as discussed in Section 7.2.4. In the next chapter, we provide a more complex example where curation of the biomass reaction alone does not generate a useful model; instead we use the biomass reaction to inform further curation of the wider metabolic model, such as identifying holes in pathways and incorrect flux directions.

Chapter 8

ROOMM: Complex curation of a metabolic model of *Salmonella enterica*

8.1 Introduction

In the previous chapter, two different models of an *E. coli* strain were used to demonstrate that draft metabolic models can be improved by using ROOMM to transfer information from curated models. This example served as a proof of principle: curating a model of a strain that already has an accurate, published model available is largely redundant. In this chapter, I present a more complex example involving two different *Salmonella enterica* serovars, and provide an example of how biomass reaction curation can give an insight into the quality of the model and subsequently guide model curation.

The example presented here demonstrates how a curated model of *Salmonella enterica* Typhimurium was used to improve a draft model of *Salmonella enterica* Typhi (Thiele *et al.*, 2011). This example is more complex than the *E. coli* example for a number of reasons. Firstly, the draft model of *Salmonella enterica* Typhi was downloaded from the BioModels database in

August 2014 (Chelliah *et al.*, 2015); due to an update in the BioModels database, this meant that the draft model was created in the new SuBliMinaL Toolbox format (Swainston *et al.*, 2011), unlike the draft model of *E. coli*. The main differences are the namespaces used for reaction and metabolite IDs, and the accuracy of reaction direction definitions.

Another complexity within this example is due to the growth medium that is used in order to match the model to wet-lab experiments that identified *S.* Typhi essential genes. The lysogeny broth (LB) growth medium commonly used to grow *S.* Typhi within the wet-lab is more complex than the glucose minimal medium used for *E. coli*. The use of a rich growth medium immediately reduces the number of essential genes observed within the cell, and does not highlight all holes within the metabolic network. For this reason, growth was predicted on both the rich and minimal media, thereby adding extra steps to the curation process.

After curating the biomass reaction and attempting to simulate growth of the draft *Salmonella enterica* Typhi model cell using FBA, it became clear that there were many holes within the model due to the number of crucially important metabolites blocking flux through the biomass reaction. The blocking metabolites indicated which pathways needed to be traced and curated, so missing reactions were subsequently identified and filled. However, production of some metabolites was still blocked. After checking reaction directions, it became apparent that the altered methods used by the updated SuBliMinaL Toolbox incorporated many reactions with incorrect flux directions, thereby causing blockages within crucial metabolic pathways.

Given the sequential nature of the curation process outlined above, it is easiest to present and discuss the methods and results chronologically. For this reason, the methods are outlined alongside their associated results below. They are presented in the order in which they were applied:

- **Section 8.2:** We begin with the curation of the biomass reaction using ROOMM
- **Section 8.3:** The curation steps based on LB growth medium are presented
- **Section 8.4:** The curation steps based on glucose minimal medium are presented

Many of the curation steps have a direct impact on the biomass reaction (refer to the top three boxes in Figure 8.1) or are influenced by the biomass reaction (see bottom right box of Figure 8.1). These processes demonstrate how ROOMM can be used to guide and aid multiple steps during model curation.

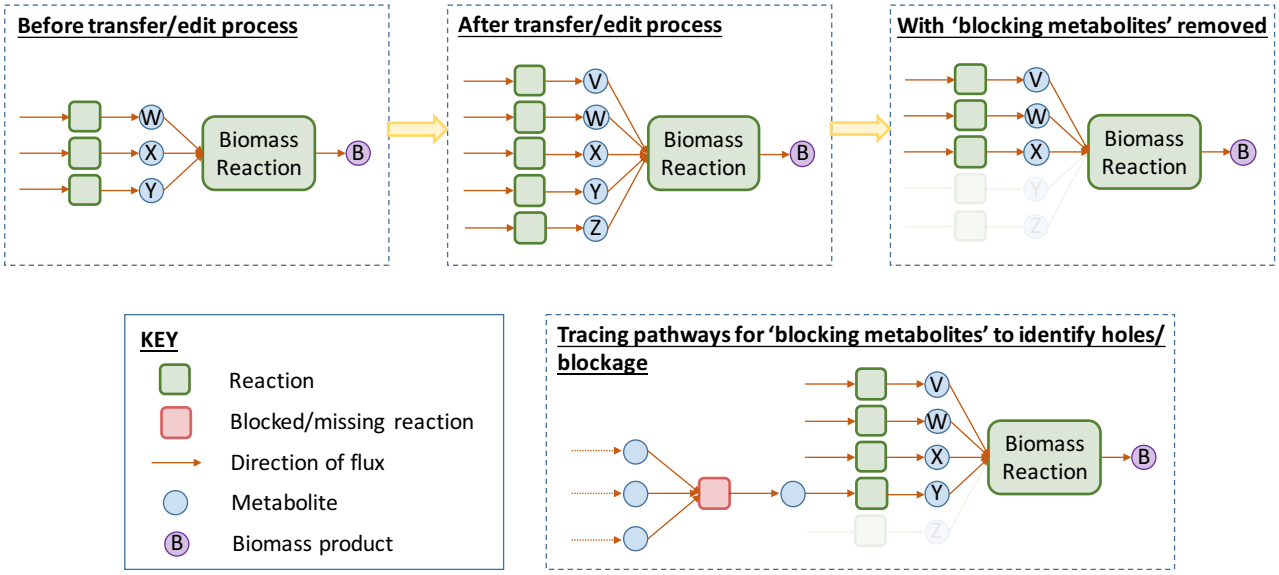


Figure 8.1: **Before transfer/edit process:** *The generic biomass reaction (green box) produced by the SuBliMinaL Toolbox is not curated and contains a large list of reactants (blue circles) that may or may not be present within a realistic biomass reaction for the given cell.* **After transfer/edit process:** *Any metabolites in the biomass reaction of the curated model are added to that of the draft template model, and all stoichiometries are set to match the template model.* **With 'blocking metabolites' removed:** *After transferring metabolite information, the target model may not be able to produce biomass due to pathway holes in the wider model. This provides information about which areas of the model require further curation. In order to run initial essential gene predictions, these blocking metabolites are removed from the reaction. However, they are sequentially returned to the reaction for further curation processes.* **Tracing pathways for 'blocking metabolites' to identify holes/blockage:** *Once a blocking metabolite is returned to the biomass reaction, it is possible to trace back through the pathways that are supposed to produce this metabolite. By doing this, it is possible to determine where pathway holes and/or incorrect flux constraints (red box) cause blockages in flux.*

8.2 Transferring and editing the biomass reaction

The biomass reactant and product stoichiometric information was transferred from the curated template model of *Salmonella enterica* Typhimurium (Thiele *et al.*, 2011) to the draft model of *Salmonella enterica* Typhi using the *Transfer mode* within ROOMM. The user input provided to ROOMM for this example is shown in Table 8.1. For further information about this process and the ChEBI categories referred to below, the reader is referred back to Chapter 6.

Table 8.1: *The user input supplied in the Salmonella enterica example application of ROOMM.*

Input	**Template** ***S. enterica*** **Typhimurium**	**Target** ***S. enterica*** **Typhi**
Filename	STM_v1.0.xml	BMID000000141741.xml
Reaction compartment	c	bm
Reaction suffix	_c	_bm
Intracellular compartment	c	i
Intracellular suffix	_c	_i
Metabolite prefix	M_	(None)
Namespace	BiGG	SuBliMinaL Toolbox (BiGG & MetaNetX)
Reaction ID	R_biomass_iRR1083	BIOMASS_REACTION

All 21 metabolites within the *Protein* ChEBI category were transferred successfully with their stoichiometric coefficients, as were all 15 reactants within the *Nucleic acid* category and the single metabolite within the *Energy* category (Hastings *et al.*, 2013). There were no reactants classified within the *Lipid* or *Inorganic ion* categories. The only reactant classified within the *Carbohydrate* category was glycogen, which could not be transferred. This is because this metabolite is present in the target model with a MetaNetX ID that is no longer mapped within their database. Of the 13 metabolites classified as *Other*, five were successfully transferred; those that were not transferred are involved in lipid metabolism and have non-standard IDs that cannot be converted. All four products were transferred successfully.

After using the *Transfer mode*, the *Edit mode* was used to make the final alterations to the biomass reaction. Firstly, glycogen (which was not successfully transferred due to absence from the mapping file) was identified in the target biomass reaction as MNXM876_bm. This reactant was edited to have a stoichiometric coefficient (0.0274762) to match the template model exactly.

Secondly, a metabolite labelled as MNXM18_bm in the target biomass reaction still had the default stoichiometry value. Upon closer inspection, it became apparent that MNXM18_bm was another ID used to represent glutamate in the target model, but bigg_glu_L_bm had been added to the biomass reaction during the transfer process. Therefore, during the edit phase, MNXM18_bm was removed from the biomass reaction by setting the stoichiometry coefficient to 0. However, this incident also indicates that there may be more than one ID used within the model to represent the same metabolite, which could cause problems in the model, such as pathway redundancies.

Finally, there were seven other metabolites (nucleic acids) within the target biomass reaction that still had the default stoichiometry value after the transfer took place. These nucleic acids were extra metabolites that were not present within the curated template model but are always included in the generic biomass reactions produced by the SuBliMinaL Toolbox (Swainston *et al.*, 2011). These seven metabolites were edited to have stoichiometric coefficients of 0 because they were not present within the curated template model (i.e. there was no evidence that they should be included and therefore there were no stoichiometric estimates), thereby removing them from the target biomass reaction.

8.3 Simulating growth on lysogeny broth (LB)

It is possible to assess the accuracy of a draft model by assessing the essential gene predictions under specific growth conditions, as discussed previously for the *E. coli* example in Chapter 7. The more accurate the model, the more accurate the essential gene predictions should be when compared to the essential genes observed within the wet-lab. Here, growth was simulated

using the raw (pre-transfer) and altered (post-transfer) versions of the target model, with LB as its simulated growth medium, in order to compare the accuracy of the raw and altered target models.

8.3.1 Defining the growth medium

The LB growth medium used in these simulations was chosen to match the growth medium used by Langridge *et al.* (2009) in the wet-lab. The contents of LB were determined using the Model SEED database, which revealed 65 metabolites to be present in LB. After mapping the 65 Model SEED IDs to the BiGG IDs used to label metabolites in the draft model, eight of the metabolites were not present within the target model and could therefore be omitted from the growth medium: protoheme IX, Mn^{2+}, Zn^{2+}, Cu^{2+}, Hg^{2+}, Cd^{2+}, N(2)-acetyl-L-ornithine zwitterion and chromate. All other components of the growth medium had existing import reactions within the target model, which were set to allow import with an upper bound of 1000, which matches the highest upper bounds used in the model.

8.3.2 Identifying metabolites blocking flux

As with the *E. coli* model, once the target biomass reaction had been curated and the growth medium set, it was necessary to determine which metabolites (if any) within the biomass reaction block flux and consequently inhibit cell growth in the model. This is because many automatically generated metabolic models contain holes in less well characterised areas of metabolism, as well as incorrect reaction direction assignment.

In order to determine which metabolite(s) block flux through the biomass reaction, each one was tested separately. This was done by selecting each biomass reactant in turn and setting all other biomass reactant stoichiometric values to 0 before running FBA. For each of these FBA analyses, if the cell could grow, the selected metabolite would be included in the curated biomass reaction; conversely, if the model had a growth rate of 0, the selected metabolite would

not be included in the biomass reaction. This process is outlined in Figure 8.2. The seven reactants that blocked flux through the biomass reaction (and were therefore excluded from the reaction) are listed in Table 8.2. All of these were newly added to the biomass reaction during the transfer process.

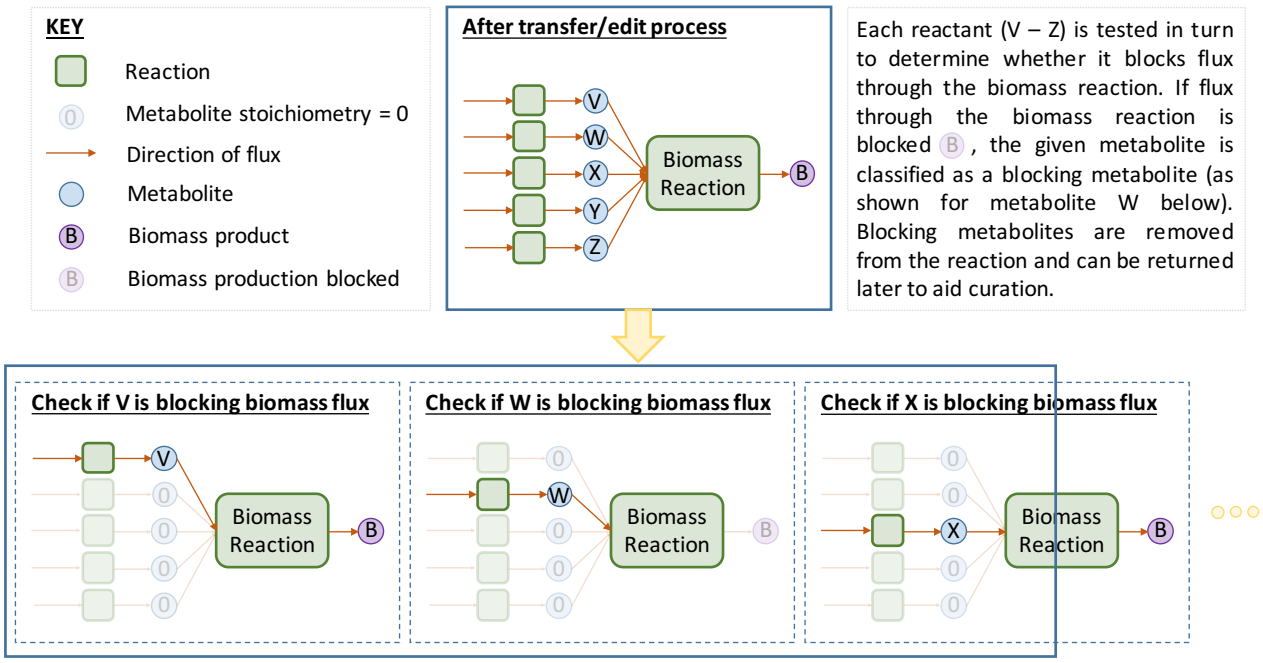


Figure 8.2: **Top box:** *After the transfer/edit process, the biomass reaction (large green box) contains the metabolite (blue circles) and stoichiometric information that allows for FBA. The production of biomass is dependent upon its reactants being produced via upstream reactions (green boxes). If any of the required upstream reactions are missing, or their flux bounds (i.e. direction) are incorrectly defined, these biomass reactants cannot be produced and will therefore block flux through the biomass reaction.* **Bottom boxes:** *Each dashed box shows how a single biomass reactant (V – X) is tested to check if it is blocking flux through the biomass reaction. This is done by temporarily setting all other reactant stoichiometries to 0 and running FBA. If the model exhibited a growth rate of 0, the selected metabolite is a blocking metabolite.*

As with the *E. coli* example above, the blocking metabolites can either be left out of the analysis, or added back in one at a time in order to determine where there are holes or flux blockages within in the network. In this more complex example, we begin by omitting these metabolites to generate essential gene predictions, before adding them back in to identify holes and blockages within the wider model.

Table 8.2: *Metabolites removed from the target Salmonella enterica Typhi biomass reaction due to flux blockage (i.e. preventing growth). All of these metabolites were newly added to the target biomass reaction due to their presence in the template reaction.*

Metabolite name	**BiGG ID**	**Newly added to target reaction**
Coenzyme A (CoA)	coa	Yes
Acetyl-CoA	accoa	Yes
Succinyl-CoA	succoa	Yes
Spermidine	spmd	Yes
FAD	fad	Yes
NADP	nadp	Yes
NADPH	nadph	Yes

8.3.3 Essential gene predictions

The essential genes were predicted for the raw and altered model using the COBRApy package in Python (Ebrahim *et al.*, 2013). This is done by sequentially selecting each gene declared in the model and blocking flux through any reactions associated with (i.e. dependent upon) the given gene. If the gene is essential, it will block the production of biomass reactants, and consequently block flux through the biomass reaction (see Figure 8.3, top). Alternatively, if there are alternative pathways that allow the reactants of the biomass reaction to be produced and therefore allow flux through the biomass reaction, the gene is considered non-essential (See Figure 8.3, bottom).

The raw target model was set to grow on LB in the same way as the altered target model (discussed previously in Section 8.3.1) in order to compare the essential gene predictions before and after altering the biomass reaction.

The raw model predicts two essential genes, both of which are absent from the essential gene lists published by Langridge *et al.* (2009):

- MetaCyc_GH90_4382_MONOMER_i
- STY1000_i

The first is *glgB* (an enzyme that adds branches to glycogen molecules) and the second is

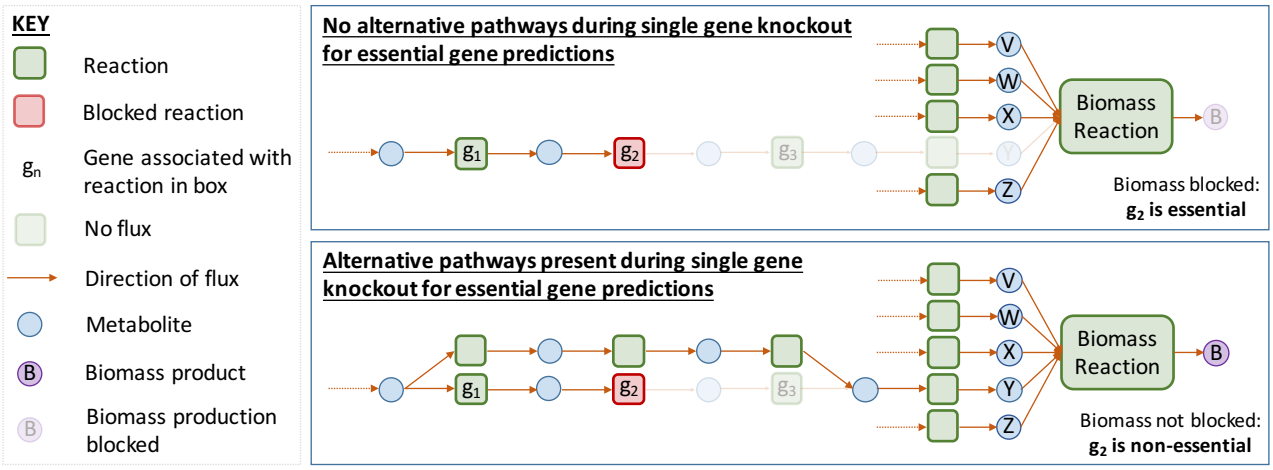


Figure 8.3: *When testing if* Gene 2 *(g_2) is essential, all reactions (green boxes) associated with that gene are blocked (red boxes) by setting its associated flux constraints to 0.* **Top box:** *If the given edge (g_2) is essential its downstream metabolites (blue nodes) cannot be produced (downstream pathway is shown as watermark), thereby blocking production of the biomass reactant, Y. This results in no simulated growth (shown as a watermarked biomass node).* **Bottom box:** *The given gene (g_2) is not essential when alternative pathways can produce the downstream biomass reactant, Y, and therefore bypass the blocked reaction to produce biomass (purple node).*

aspC (aspartate aminotransferase). The altered target model predicts only the former gene as essential, so no improvement is seen in the essential gene predictions as a consequence of transferring biomass reaction information from the template to the target model. This result is unsurprising given that crucial metabolites (such as FAD, NADP and Acetyl-CoA) were blocking biomass production. It is clear that the pathways representing central metabolism within the model are in need of further curation. We can begin to assess these pathways by using the blocking metabolites in the biomass reaction.

8.3.4 Unblocking metabolic pathways

Unlike the target *E. coli* model, the *Salmonella enterica* Typhi model did not show improved essential gene predictions after culling the blocking metabolites. These poor predictions are due to a number of factors:

- The blocked metabolites are crucially important to the cell and are usually well characterised in curated models. Therefore, they cannot be realistically omitted from the model with the expectation of an accurate growth simulation or flux predictions.

- There may be more holes/blocked pathways in the model that are not immediately evident due to the fact that growth has to be simulated on the rich LB medium in order to match the wet-lab experiments.

- The target model contains approximately twice as many reactions when compared to the curated model of the *Salmonella enterica* Typhimurium strain, which indicates that there may be a number of redundancies and repeated reactions in the model.

- The accuracy of the essential gene predictions is dependent on the quality of the gene associations incorporated in the model. For example, by comparing the *S.* Typhi genome to the draft metabolic model, it is clear that many genes are absent from the model. Those that are absent cannot be included within the essentiality tests and therefore cannot be predicted as essential.

Given the inaccuracy of the essential gene predictions and the potential reasons for their inaccuracy, the seven blocking metabolites were returned to the biomass reaction and used to guide further curation of the model. The blocking metabolites can be arranged into groups: CoA metabolites, NADP metabolites, FAD and spermidine. Therefore, the model requires curation within these four areas of metabolism. The process of model curation was focused on these key areas of the network, and the methods used to curate these areas of metabolism are discussed below.

Adding/transferring reactions

For each of the blocked metabolite groups, the metabolite IDs were identified within the target model and used to search through the model and trace the related reactions via the reactant and product definitions. The identified reactions were compared to the defined KEGG pathways

available in the KEGG database (Kanehisa *et al.*, 2016). The reactions within these related pathways (i.e. not just those with the blocking metabolites as direct reactants or products) were identified as either present or absent within the target model.

When checking for absent reactions in the model through manual searching in the XML file, 18 reactions within the KEGG CoA, FAD and NADP pathways were identified as absent from the target model. The genes from the template model and their associated missing reactions in the target model are shown in Table 8.3. Of these reactions, 9 could not be included in the target model because one or more metabolites involved in those reactions were not present within the model. This is unsurprising because many models focus on central metabolism and do not include peripheral areas of metabolism.

The remaining 9 missing reactions were identified within the template model and transferred to new reactions within the target model using the ROOMM *Transfer mode*. After transferring the reactions to the template model, FBA was used to determine whether these metabolites were still blocking simulated growth of the cell. This revealed that the transfer process did not fix the blocks in these pathways, which meant that the direction of flux within parts of these pathways must be incorrect and blocking the production of these metabolites.

With respect to the missing metabolites, it could be argued that a recursive approach could be taken here to add the missing metabolites (and associated reactions) to the model. However, it must be noted that the KEGG pathways used here are not organism-specific and therefore include some metabolic elements that would be inappropriate for the cell type being modelled. For example, many bacteria produce antimicrobial substances that are very specific to the producer, and it would therefore be inappropriate to introduce these substances into the metabolic network of other bacteria. Therefore, if a recursive approach is taken, some inappropriate metabolites and reactions may be incorporated in the model. Whilst taking a recursive approach to build in new metabolites and reactions may be appropriate in some cases, it would be a lengthy approach that would be difficult to automate. Furthermore, one could argue that if this approach was automated, perhaps there is a chance that all networks would end up as

an expansive metagenome-scale model with no realistic representation of the species that is supposed to be modelled. Therefore, if a recursive approach was to be applied, it would need to be done manually in order to ensure that no inappropriate substances and reactions were introduced into the model.

Table 8.3: *Gene/reaction associations and reaction transfer status for the missing genes that were expected to be present in the target model.*

Template gene IDs	**Reaction ID**	**Metabolite**	**Successfully transferred**
STM0586	FEENTERR1	FAD	✗
STM0586	FEDHBZS3R1	FAD	✗
STM0071	CTBTCAL2	CoA	✗
STM0970	OBTFL	CoA	✓
STM0472	GLCATr	CoA	✗
STM0472	MALTATr	CoA	✓
STM3045	PFOR	CoA	✓
STM0071	CRNDCAL2	CoA	✗
STM0071	CRNCAL2	CoA	✗
STM1196	MCOATA	CoA	✓
STM1582	ACANTHAT	CoA	✗
STM2649/STM3915	TRDR	NADP	✓
STM3680	ALDD3y	NADP	✓
STM3680	ALDD2y	NADP	✓
STM3045/STM4085	FLDR	NADP	✗
STM2556	NODOy	NADP	✓
STM0255	ALR2	NADP	✗
STM0255	DKGLCNR1	NADP	✓

Comparing reaction directions between template and target models

The direction of flux through each reaction associated with these blocked pathways was compared to the direction defined within the KEGG pathways and the curated template model. Upon inspecting the flux constraints in this way, it became clear that the flux directions in the target model did not agree with those in the curated template model and KEGG pathways.

Reaction directions within the CoA pathway that did not agree with the reaction directions in the template model and KEGG pathways were altered to agree with both of these. For cases where the curated model and KEGG pathways disagreed, the direction was made reversible.

This method of tracking and matching directions was also applied to the pathways leading to FAD and spermidine production, which were also blocking growth. Given that NAD was not blocking growth but NADP and NADPH were, the same method was also applied to the NAD/NADP conversion reactions. In total, 10 of the 24 reactions identified and compared within these pathways required their flux constraints to be amended.

After amending flux constraints for CoA, it also allowed for the production of succinyl-CoA and acetyl-CoA. Similarly, once flux constraints had been amended for the production of NADP, it also allowed for the production of NADPH. After amending these flux directions, the model could simulate growth on LB growth medium with no blocked metabolites. However, the essential gene predictions remained the same as those mentioned previously.

Given that the model was growing on a rich medium, it is not possible to identify all of the holes within the model, due to the fact that a rich medium does not force certain areas of metabolism to be active within the simulated cell. Therefore, the curation process guided by the biomass reaction was extended by modelling growth on glucose minimal medium.

8.4 Simulating growth on glucose minimal medium

LB is a rich medium and therefore does not employ certain parts of the network during growth simulations. Therefore, simulations using LB growth medium may be concealing other areas of metabolism that are not well-characterised within the model. However, *Salmonella enterica* is capable of growing on glucose minimal medium within the wet-lab, so growth was simulated on this medium in order to guide further curation of the model (Hartman *et al.*, 2014).

8.4.1 Defining the growth medium

The growth medium was set using the same methods as previously described for LB, but with glucose minimal medium metabolites, as described previously for the *E. coli* model. All of the

metabolites were present within the model and had associated import reactions, which were constrained using their upper and lower flux bounds to allow metabolite import.

8.4.2 Identify metabolites blocking flux

In order to determine which metabolite(s) block flux through the biomass reaction, each one was tested separately. This was done by selecting each biomass reactant in turn and setting all other biomass reactant stoichiometric values to 0 before running FBA, as described previously in Section 7.2.4. In each of these FBA analyses, if the cell could grow, the selected metabolite would remain in the biomass reaction; conversely, if the model had a growth rate of 0, the selected metabolite would be removed from the biomass reaction via a stoichiometric assignment of 0.

Ten metabolites were identified to be blocking growth on glucose minimal medium: L-tyrosine, L-lysine, L-isoleucine, L-arginine, L-histidine, L-methionine, L-tryptophan, putrescine, FAD, and spermidine. This shows that amino acid metabolism is inaccurate within this model, illustrating the extent of the inaccuracies within these semi-automatically generated models. Curation at this level would take too much time to do manually via pathway tracing, so a script was created to enable automation of parts of the curation process.

8.4.3 Comparing reaction directions between template and target models

The reaction directions of equivalent reactions in the target and template models were compared by creating a script to map the target model MNX reaction IDs to the corresponding reaction IDs that are used in the curated template model. This was done by using the flat-files available from the MetaNetX database (Ganter *et al.*, 2013; King *et al.*, 2016). Any target reaction direction that did not match that of the corresponding template reaction was altered to match the template.

Of the 655 reactions that could be mapped between the two models, 90 were made reversible to match the curated model and 167 were flipped to the opposite direction to their original orientation. These altered reactions represent almost a quarter of those that could be mapped between the template and target models. Whilst this process enabled the production of most of the blocked metabolites, the model was still unable to produce tyrosine, tryptophan and FAD. Therefore, these remaining pathways required further assessment, as described below.

8.4.4 Tracing tyrosine and tryptophan metabolism

Given that tyrosine and tryptophan have a common precursor (chorismate), the associated pathways (shown within KEGG map 01063: *'Biosynthesis of alkaloids derived from shikimate pathway'*) were compared to the target model. In order to test whether the block in the pathways leading to tyrosine and tryptophan production was prior to the production of chorismate, the model was altered to allow the import of chorismate; if there are no holes in the pathway after chorismate, tryptophan and tyrosine should be producible when chorismate is provided to the cell. When tested via FBA, chorismate import showed the production of trytophan and tyrosine, indicating that the production of chorismate was causing the block in tryptophan and tyrosine production.

Chorismate is produced via the Shikimate pathway (KEGG map 00400: *'Phenylalanine, tyrosine and tryptophan biosynthesis'*) and is known to include some essential genes. Therefore, chorismate import was removed and replaced with the import of an upstream precursor, erythrose-4p (e4p). The reactions required for transformation of e4p to chorismate were identified manually within the model and were present with the correct flux directions. Therefore, by allowing import of e4p and not chorismate, the essential genes within the e4p-chorismate pathway should be predicted as essential when FBA is applied.

When FBA was applied, the essential gene predictions were still unchanged from those predicted on LB; more importantly, the essential genes that we would expect to observe within the Shikimate pathway were not predicted as essential. The target model has almost twice as

many reactions present when compared to the curated template model of *Salmonella enterica* Typhimurium, indicating that redundancies in the target model may be causing the lack of essential gene predictions. Furthermore, the inaccuracies of flux directions throughout the model would suggest that there may also be too much inaccuracy and/or flexibility in the flux constraints throughout the model.

Following these investigations into the Shikimate pathway and its associated essential genes, the import of e4p was removed from the model to restore it to the original growth medium. Rather than trace back through the preceding reactions to curate that pathway further, the previously acknowledged problem of wide-spread incorrect flux directions was addressed in order to fix the blockage in e4p production. The previous mapping from the template reactions to the target reactions allowed only 655 of the 4357 *S.* Typhi reaction directions to be checked automatically, due to the issue of mapping between the template and target models. However, eQuilibrator is a tool that uses thermodynamic data to predict flux directions of metabolic reactions, which can be applied to any reaction within the model that can be represented using KEGG IDs (Kanehisa *et al.*, 2016; Flamholz *et al.*, 2012). This allows us to predict flux directions for those reactions that could not previously be mapped across the two models.

8.4.5 Applying eQuilibrator

Given the seemingly widespread incidence of reaction directionality issues within the target model, the eQuilibrator tool was used to predict the reaction directions within the target *Salmonella enterica* Typhi model with the aim of reducing flux blockages in pathways (Flamholz *et al.*, 2012). The eQuilibrator input file was generated by converting all MNX reaction IDs to KEGG reaction IDs using the MNX mapping dictionaries (Ganter *et al.*, 2013; Kanehisa *et al.*, 2016). The format of the input file is a tab-delimited file including the reaction IDs, reactants with their stoichiometric coefficients and products with their stoichiometric coefficients, as follows:

```
Reaction ID    C00012 + 2 C01010 = C12111 + C01211
```

Of the 4357 reactions within the target *Salmonella eneterica* Typhi model, 2869 reactions could be converted to KEGG IDs and formatted to create an eQuilibrator input file. This input file was run through a setup of eQuilibrator and the corresponding output file returned a value for each reaction that represents either a forward, reverse or reversible direction. These directions were applied to the target model; as a result, the *S.* Typhi model could grow on glucose minimal medium with no blocking metabolites; tyrosine, tryptophan and FAD no longer blocked biomass production.

8.4.6 Essential gene predictions

After applying eQuilibrator, the essential gene test was run again on glucose minimal medium using the COBRApy package for FBA. The results remained unchanged from the previous tests on LB growth medium.

There are a number of factors that are likely to be contributing to the lack of essential gene predictions observed within the model. Firstly, the essential genes are predicted by running FBA simulations of single gene knock-outs, which is done by blocking flux through the reactions associated with a given gene. If the blocked reactions result in zero flux through the biomass reaction, the gene being knocked out is considered to be essential. However, it may be the case that some of the essential gene knock-out simulations result in negligible non-zero flux through the biomass reaction, whereby the gene is not flagged as essential, but reduces the yield to a growth rate so small that the cells would not be viable if tested *in vivo*. Therefore, close checking of the flux values through the biomass reaction may reveal more essential genes than a simple binary test.

Another potential reason for the lack of essential gene predictions involves the metabolite IDs incorporated within the model. Most of the metabolite IDs are in the BiGG ID format, but a minority are called by their MetaNetX IDs. Given that at least one metabolite is present with both forms of ID (bigg_glu_L_bm/MNXM18_bm), this would suggest that some parts of the model are incorporated twice — one for each namespace. Ideally, we would identify all

metabolites that have been incorporated twice and reconcile the pathways associated with the two metabolite IDs.

Furthermore, there are almost twice as many reactions within the *S.* Typhi model than in the curated *S.* Typhimurium model. Given that the genomes of these two *Salmonella enterica* serovars are known to be 89-99% similar (Garai *et al.*, 2012), it would seem likely that the two models should incorporate a similar number of reactions encoded by their respective genomes. Therefore, this supports the idea that there are likely to be many extra reactions within the draft target model. If these redundant reactions could be easily identified and removed, it may be possible to predict more of the essential genes that we expect to see when FBA is applied to the model.

Incorrect reaction directions may also be present within the remaining reactions that were not able to be converted and included in the eQuilibrator predictions. However, these reactions are likely to be within the redundant pathways given that they did not cause a blockage in biomass production.

As well as the redundancies identified within the model, there are also many holes. For example, some of the missing reactions could not be transferred from the template model to the target model due to the absence of a reactant and/or product within the target model. This implies that whole areas relating to those metabolites are not characterised and included within the model.

8.5 Conclusions and future work

More accurate metabolic models are required in order to understand the metabolic profiles of cells. This is of particular importance for pathogenic organisms in order that drug targets for treatment may be found. Furthermore, metabolic interactions between populations within the microbiota are of interest, given that it has been shown that the microbiota plays a large part in dictating the health of its host. A major bottleneck in the progression of research in this

area is caused by the lengthy curation process involved in creating accurate metabolic models. However, ROOMM can help to integrate the expert knowledge of cellular and molecular biologists with that of systems biologists to enable beneficial collaboration between the communities, ultimately leading to more accurate modelling of metabolic systems via curation of automatically-generated models. This will allow scientists to make the most of these models and the databases available to them, such as KEGG and Model SEED. By providing a graphical user interface and automated mapping between metabolite namespaces, ROOMM can aid model curation and make the process more efficient.

In this chapter, I have presented a complex example demonstrating how ROOMM can be used to guide the curation of a semi-automatically generated metabolic model. ROOMM was used to begin model curation via the transfer and editing of biomass reaction information from the *S.* Typhimurium template model to the *S.* Typhi target model. Subsequently, ROOMM was used to facilitate hole-filling within the target model by using the biomass blocking metabolites to highlight inaccurate pathways within the model. Following these hole-filling steps, inaccuracies in flux directions became apparent, thereby guiding further curation of the model. By speeding up the process of reaction transfer and guiding model curation, ROOMM facilitates the improvement of these metabolic models, which is crucial before these models can be used to model metabolic symbioses within the context of the microbiota.

Chapter 9

Conclusion

The human microbiota changes throughout our lives and is crucial to our health from birth to death. Perturbations to the microbiota can have life-changing implications for its host, but we are still in the early stages of understanding the complexities of these crucial microbial communities and how they affect our health. In this thesis, I present new methods and tools for creating and analysing networks of the human microbiota, including dependency analyses and metabolic networks.

Within Chapter 3 of this thesis, the development of a new method was presented for observing changes in dependencies within the microbiota. This new method uses differential networks and a measure of proportionality to determine how dependencies within the microbiota differ in cases and controls. The method was presented as an alternative to using correlation measures of dependency, due to the issues surrounding the application of correlation measures to relative data. Validation studies justifying the use of the proportionality measure show that it is a valid measure of dependency when datasets are large, and an application to real data allowed for a direct comparison to equivalent networks produced using correlation coefficients in place of the proportionality measure. The resulting proportionality networks were less dense than the corresponding correlation network, which is due to the inappropriate application of correlation to relative data. Furthermore, when the method was applied to real data in Chapter 4, the bio-

logical results were congruent with existing scientific literature. Therefore, the proportionality differential network method may be applied to future microbiota datasets with the aim of identifying changes in dependencies in disease cases when compared to controls. However, there are ways to take this method forward, such as looking at non-linear dependencies and test statistics that would return identical networks no matter whether cases or controls were set as the null distribution. For example, the difference between a given proportionality measure for cases and controls could be used and tested against a suitable null distribution in order to maintain a directional measure of change, or the overlap between the case and control distributions for a given proportionality value could be used.

Following the identification of shifts in dependencies within the microbiota of disease cases, a method for presenting the metabolic capabilities of the microbiota as a weighted metagenome-scale metabolic model was presented in Chapter 5. These models allow us to represent the metabolic capabilities of the microbiota at a higher resolution than we can by weighting whole pathways. In addition, 16S rRNA gene copy numbers were mapped to taxonomic assignments via their scientific name and rank by creating a new MySQL database. This database could be used in future analyses and could be made publicly available as a new and valuable resource to bioinformaticians. Further work is also required in order to develop a successful method for comparing the metagenome-scale metabolic networks across cases and controls.

As well as assessing the metabolic capabilities of the microbiota as a whole, it would be beneficial to model pairs of symbiotic species, such as those highlighted as increasingly dependent in disease cases. However, single cell models are still imperfect and require many years of curation before models of symbiosis can be expected to be useful. For this reason, ROOMM was made in order to speed up the curation of metabolic models. This software was presented in Chapter 6 with a simple example following in Chapter 7, and a more complex example provided in Chapter 8. In the future, it would be beneficial to provide an online front-end for this tool so that it can be easily accessed without requiring local installation for users. It would also be helpful to provide some of the subsequent analyses, such as growth medium application, FBA runs and essential gene predictions as part of the software, so that users can run them without

needing to write their own scripts.

We conclude by emphasising the need to develop these methods and tools further to aid the number and extent of studies exploring the relationships between constituents of the microbiota. It is only through these studies of dependencies and symbioses that we can begin to fully understand how the microbiota functions and how it affects the health of the host. The proportionality differential network method can be used to hypothesise about altered symbiotic relationships, and the inferred metanetworks and metabolic models may be used to begin to explore shifts in metabolism and metabolic symbioses.

Bibliography

Aroniadis, O.C., Brandt, L.J., Greenberg, A., Borody, T., Kelly, C.R., Mellow, M., Surawicz, C., Cagle, L., Neshatian, L., Stollman, N., Giovanelli, A., Ray, A., & Smith, R. (2016) Long-term Follow-up Study of Fecal Microbiota Transplantation for Severe and/or Complicated Clostridium difficile Infection: A Multicenter Experience, *Journal of clinical gastroenterology*, **50**:398–402.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., & Sherlock, G. (2000) Gene ontology: tool for the unification of biology, *Nat Genet*, **25**:25–29.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., & Mori, H. (2006) Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection, *Molecular Systems Biology*, **2**:2006.0008–2006.0008.

Balter, M. (2012) Taking stock of the human microbiome and disease, *Science (New York, N.Y.)*, **336**:1246–1247.

Barcenilla, A., Pryde, S.E., Martin, J.C., Duncan, S.H., Stewart, C.S., Henderson, C., & Flint, H.J. (2000) Phylogenetic relationships of butyrate-producing bacteria from the human gut, *Applied and Environmental Microbiology*, **66**:1654–1661.

Belenguer, A., Duncan, S.H., Holtrop, G., Anderson, S.E., Lobley, G.E., & Flint, H.J. (2007)

Impact of ph on lactate formation and utilization by human fecal microbial communities, *Applied and Environmental Microbiology*, **73**:6526–6533.

Bergstrom, K.S.B. & Xia, L. (2013) Mucin-type o-glycans and their roles in intestinal homeostasis, *Glycobiology*, **23**:1026–1037.

Bordbar, A., Lewis, N.E., Schellenberger, J., Palsson, B.Ø., & Jamshidi, N. (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions, *Molecular systems biology*, **6**:422.

Bornstein, B.J., Keating, S.M., Jouraku, A., & Hucka, M. (2008) LibSBML: an API library for SBML, *Bioinformatics (Oxford, England)*, **24**:880–881.

Brandes, A., Lun, D.S., Ip, K., Zucker, J., Colijn, C., Weiner, B., & Galagan, J.E. (2012) Inferring carbon sources from gene expression profiles using metabolic flux models, *PloS one*, **7**:e36,947–e36,947.

Bryant, W.A., Sternberg, M.J.E., & Pinney, J.W. (2013) AMBIENT: Active Modules for Bipartite Networks–using high-throughput transcriptomic data to dissect metabolic response, *BMC systems biology*, **7**:26.

Buccigrossi, V., Nicastro, E., & Guarino, A. (2013) Functions of intestinal microflora in children, *Current opinion in gastroenterology*, **29**:31–38.

Büchel, F., Rodriguez, N., Swainston, N., Wrzodek, C., Czauderna, T., Keller, R., Mittag, F., Schubert, M., Glont, M., Golebiewski, M., van Iersel, M., Keating, S., Rall, M., Wybrow, M., Hermjakob, H., Hucka, M., Kell, D.B., Müller, W., Mendes, P., Zell, A., *et al.* (2013) Path2models: large-scale generation of computational models from biochemical pathway maps, *BMC Systems Biology*, **7**:1–19.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M.,

Reeder, J., *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data, *Nature methods*, **7**:335–336.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., & Karp, P.D. (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic acids research*, **44**:D471–80.

Chelliah, V., Juty, N., Ajmera, I., Ali, R., Dumousseau, M., Glont, M., Hucka, M., Jalowicki, G., Keating, S., Knight-Schrijver, V., Lloret-Villas, A., Natarajan, K.N., Pettit, J.B., Rodriguez, N., Schubert, M., Wimalaratne, S.M., Zhao, Y., Hermjakob, H., Le Novère, N., & Laibe, C. (2015) BioModels: ten-year anniversary, *Nucleic acids research*, **43**:D542–8.

Chelliah, V., Laibe, C., & Le Novère, N. (2013) BioModels Database: a repository of mathematical models of biological processes, *Methods in molecular biology (Clifton, N.J.)*, **1021**:189–199.

Cho, I. & Blaser, M.J. (2012) The human microbiome: at the interface of health and disease, *Nature reviews. Genetics*, **13**:260–270.

Clooney, A.G., Bernstein, C.N., Leslie, W.D., Vagianos, K., Sargent, M., Laserna-Mendieta, E.J., Claesson, M.J., & Targownik, L.E. (2016) A comparison of the gut microbiome between long-term users and non-users of proton pump inhibitors, *Alimentary pharmacology & therapeutics*, **43**:974–984.

Collins, S.M. & Bercik, P. (2009) The relationship between intestinal microbiota and the central nervous system in normal gastrointestinal function and disease, *Gastroenterology*, **136**:2003–2014.

Costello, S.P., Tucker, E.C., La Brooy, J., Schoeman, M.N., & Andrews, J.M. (2016) Establishing a Fecal Microbiota Transplant Service for the Treatment of Clostridium difficile Infection, *Clinical Infectious Diseases*, **62**:908–914.

Crost, E.H., Tailford, L.E., Monestier, M., Swarbreck, D., Henrissat, B., Crossman, L.C., & Juge, N. (2016) The mucin-degradation strategy of ruminococcus gnavus: The importance of intramolecular trans-sialidases, *Gut Microbes*, **7**:302–312.

Derrien, M., van Passel, M.W., van de Bovenkamp, J.H., Schipper, R.G., de Vos, W.M., & Dekker, J. (2010) Mucin-bacterial interactions in the human oral cavity and digestive tract, *Gut Microbes*, **1**:254–268.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G.L. (2006) Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb, *Applied and Environmental Microbiology*, **72**:5069–5072.

Diaz Heijtz, R., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., Hibberd, M.L., Forssberg, H., & Pettersson, S. (2011) Normal gut microbiota modulates brain development and behavior, *Proceedings of the National Academy of Sciences of the United States of America*, **108**:3047–3052.

Donohoe, D.R., Garge, N., Zhang, X., Sun, W., O'Connell, T.M., Bunger, M.K., & Bultman, S.J. (2011) The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon, *Cell Metabolism*, **13**:517–526.

Dumas, M.E. (2011) The microbial-mammalian metabolic axis: Beyond simple metabolism, *Cell Metabolism*, **13**:489–490.

Ebrahim, A., Lerman, J.A., Palsson, B.O., & Hyduke, D.R. (2013) Cobrapy: Constraints-based reconstruction and analysis for python, *BMC Systems Biology*, **7**:1–6.

Faith, J.J., McNulty, N.P., Rey, F.E., & Gordon, J.I. (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice, *Science (New York, N.Y.)*, **333**:101–104.

Flamholz, A., Noor, E., Bar-Even, A., & Milo, R. (2012) equilibrator—the biochemical thermodynamics calculator, *Nucleic Acids Research*, **40**:D770–D775.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T.A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., Takahashi, M., Fukuda, N.N., Murakami, S., Miyauchi, E., Hino, S., Atarashi, K., Onawa, S., Fujimura, Y., Lockett, T., Clarke, J.M., *et al.* (2013) Commensal microbe-derived butyrate induces the differentiation of colonic regulatory t cells, *Nature*, **504**:446–450.

Ganesh, B.P., Klopfleisch, R., Loh, G., & Blaut, M. (2013) Commensal akkermansia muciniphila exacerbates gut inflammation in salmonella typhimurium-infected gnotobiotic mice, *PLoS ONE*, **8**:e74,963.

Ganter, M., Bernard, T., Moretti, S., Stelling, J., & Pagni, M. (2013) MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks, *Bioinformatics (Oxford, England)*, **29**:815–816.

Garai, P., Gnanadhas, D.P., & Chakravortty, D. (2012) Salmonella enterica serovars typhimurium and typhi as model organisms: Revealing paradigm of host-pathogen interactions, *Virulence*, **3**:377–388.

Greenblum, S., Turnbaugh, P.J., & Borenstein, E. (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease, *Proceedings of the National Academy of Sciences*, **109**:594–599.

Gutiérrez-Díaz, I., Fernández-Navarro, T., Sánchez, B., Margolles, A., & González, S. (2016) Mediterranean diet and faecal microbiota: a transversal study, *Food & function*, **7**:2347–2356.

Habashneh, R.A., Khader, Y.S., Alhumouz, M.K., Jadallah, K., & Ajlouni, Y. (2012) The association between inflammatory bowel disease and periodontitis among jordanians: a casecontrol study, *Journal of Periodontal Research*, **47**:293–298.

Hamady, M. & Knight, R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges, *Genome Research*, **19**:1141–1152.

Hartman, H.B., Fell, D.A., Rossell, S., Jensen, P.R., Woodward, M.J., Thorndahl, L., Jelsbak, L., Olsen, J.E., Raghunathan, A., Daefler, S., & Poolman, M.G. (2014) Identification of

potential drug targets in salmonella enterica sv. typhimurium using metabolic modelling and experimental validation, *Microbiology*, **160**:1252–1266.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., & Steinbeck, C. (2013) The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013, *Nucleic Acids Research*, **41**:D456–D463.

Heinken, A., Sahoo, S., Fleming, R.M.T., & Thiele, I. (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut, *Gut microbes*, **4**:28–40.

Hoskins, L.C., Agustines, M., McKee, W.B., Boulding, E.T., Kriaris, M., & Niedermeyer, G. (1985) Mucin degradation in human colon ecosystems. isolation and properties of fecal strains that degrade abh blood group antigens and oligosaccharides from mucin glycoproteins, *The Journal of Clinical Investigation*, **75**:944–953.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, *Bioinformatics (Oxford, England)*, **19**:524–531.

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

Hunter, J.D. (2007) Matplotlib: A 2d graphics environment, *Computing in Science & Engineering*, **9**:90–95.

Jakobsson, H.E., Rodríguez-Piñeiro, A.M., Schütte, A., Ermund, A., Boysen, P., Bemark, M., Sommer, F., Bäckhed, F., Hansson, G.C., & Johansson, M.E. (2015) The composition of the gut microbiota shapes the colon mucus barrier, *EMBO Reports*, **16**:164–177.

Johnson, R. & Wichern, D., *Applied multivariate statistical analysis* (Pearson, New Jersey, 2007), 6th edition.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016) Kegg as a reference resource for gene and protein annotation, *Nucleic Acids Research*, **44**:D457–D462.

Kang, C.s., Ban, M., Choi, E.J., Moon, H.G., Jeon, J.S., Kim, D.K., Park, S.K., Jeon, S.G., Roh, T.Y., Myung, S.J., Gho, Y.S., Kim, J.G., & Kim, Y.K. (2013) Extracellular vesicles derived from gut microbiota, especially akkermansia muciniphila, protect the progression of dextran sulfate sodium-induced colitis, *PLoS ONE*, **8**:e76,520.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergstrm, G., Behre, C.J., Fagerberg, B., Nielsen, J., & Bckhed, F. (2013) Gut metagenome in european women with normal, impaired and diabetic glucose control, *Nature*, **498**:99103.

King, Z.A., Lu, J., Drger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., & Lewis, N.E. (2016) Bigg models: A platform for integrating, standardizing and sharing genome-scale models, *Nucleic Acids Research*, **44**:D515–D522.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., & Schmidt, T.M. (2001) rrndb: the ribosomal rna operon copy number database, *Nucleic Acids Research*, **29**:181–184.

Koropatkin, N.M., Cameron, E.A., & Martens, E.C. (2012) How glycan metabolism shapes the human gut microbiota, *Nat Rev Micro*, **10**:323–335.

Kumar, A., Suthers, P.F., & Maranas, C.D. (2012) MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases, *BMC bioinformatics*, **13**:6–6.

Lalkhen, A.G. & McCluskey, A. (2008) Clinical tests: sensitivity and specificity, *Continuing Education in Anaesthesia, Critical Care & Pain*, **8**:221–223.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., & Huttenhower, C. (2013) Predictive functional profiling of microbial communities using 16s rrna marker gene sequences, *Nat Biotech*, **31**:814–821.

Langridge, G.C., Phan, M.D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., Wain, J., Parkhill, J., & Turner, A.K. (2009) Simultaneous assay of every salmonella typhi gene using one million transposon mutants, *Genome Research*, **19**:2308–2316.

Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data.

Li, S.S., Zhu, A., Benes, V., Costea, P.I., Hercog, R., Hildebrand, F., Huerta-Cepas, J., Nieuwdorp, M., Salojärvi, J., Voigt, A.Y., Zeller, G., Sunagawa, S., de Vos, W.M., & Bork, P. (2016) Durable coexistence of donor and recipient strains after fecal microbiota transplantation, *Science (New York, N.Y.)*, **352**:586–589.

Lopez-Siles, M., Khan, T.M., Duncan, S.H., Harmsen, H.J.M., Garcia-Gil, L.J., & Flint, H.J. (2012) Cultured representatives of two major phylogroups of human colonic faecalibacterium prausnitzii can utilize pectin, uronic acids, and host-derived substrates for growth, *Applied and Environmental Microbiology*, **78**:420–428.

Louis, P. & Flint, H.J. (2009) Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine, *FEMS Microbiology Letters*, **294**:1–8.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., & Bähler, J. (2015) Proportionality: a valid alternative to correlation for relative data, *PLoS computational biology*, **11**:e1004,075.

Lozupone, C. & Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities, *Applied and environmental microbiology*, **71**:8228–8235.

Lundin, D., Severin, I., Logue, J.B., Ostman, O., Andersson, A.F., & Lindström, E.S. (2012) Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity?, *Environmental microbiology reports*, **4**:367–372.

Malago, J.J. & Sangu, C.L. (2015) Intraperitoneal administration of butyrate prevents the severity of acetic acid colitis in rats, *Journal of Zhejiang University. Science. B*, **16**:224–234.

Milani, C., Ticinesi, A., Gerritsen, J., Nouvenne, A., Lugli, G.A., Mancabelli, L., Turroni, F., Duranti, S., Mangifesta, M., Viappiani, A., Ferrario, C., Maggio, M., Lauretani, F., De Vos, W., van Sinderen, D., Meschi, T., & Ventura, M. (2016) Gut microbiota composition and clostridium difficile infection in hospitalized elderly individuals: a metagenomic study, *Scientific Reports*, **6**.

Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., ten Hoopen, P., Fraser, M., Pesseat, S., Potter, S., Scheremetjew, M., Sterk, P., & Finn, R.D. (2016) EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data, *Nucleic acids research*, **44**:D595–603.

Moore, N.E., Wang, J., Hewitt, J., Croucher, D., Williamson, D.A., Paine, S., Yen, S., Greening, G.E., & Hall, R.J. (2015) Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans, *Journal of clinical microbiology*, **53**:15–21.

Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., Bousvaros, A., Korzenik, J., Sands, B.E., Xavier, R.J., & Huttenhower, C. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment, *Genome biology*, **13**:R79.

NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information, *Nucleic acids research*, **43**:D6–17.

Nelder, J. & Wedderburn, R. (1972) Generalized linear models, *Journal of the Royal Statistical Society*, **135**:370–384.

Ng, K.M., Ferreyra, J.A., Higginbottom, S.K., Lynch, J.B., Kashyap, P.C., Gopinath, S., Naidu, N., Choudhury, B., Weimer, B.C., Monack, D.M., & Sonnenburg, J.L. (2013) Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens, *Nature*, **502**:96–99.

Nugent, S., Kumar, D., Rampton, D., & Evans, D. (2001) Intestinal luminal ph in inflammatory

bowel disease: possible determinants and implications for therapy with aminosalicylates and other drugs, *Gut*, **48**:571–577.

Opgen-Rhein, R. & Strimmer, K. (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC systems biology*, **1**:37.

Orenstein, R., Dubberke, E., Hardi, R., Ray, A., Mullane, K., Pardi, D.S., Ramesh, M.S., & PUNCH CD Investigators (2016) Safety and Durability of RBX2660 (Microbiota Suspension) for Recurrent Clostridium difficile Infection: Results of the PUNCH CD Study, *Clinical Infectious Diseases*, **62**:596–602.

Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., & Palsson, B.Ø. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011, *Molecular systems biology*, **7**:535.

Orth, J.D., Thiele, I., & Palsson, B.Ø. (2010) What is flux balance analysis?, *Nature biotechnology*, **28**:245–248.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic acids research*, **33**:5691–5702.

Parzen, E. (1962) On estimation of a probability density function and mode, *Ann. Math. Statist.*, **33**:1065–1076.

Pessione, E. (2012) Lactic acid bacteria contribution to gut microbiota complexity: lights and shadows, *Frontiers in Cellular and Infection Microbiology*, **2**:86.

Petrof, E.O., Gloor, G.B., Vanner, S.J., Weese, S.J., Carter, D., Daigneault, M.C., Brown, E.M., Schroeter, K., & Allen-Vercoe, E. (2013) Stool substitute transplant therapy for the eradication of Clostridium difficile infection: 'RePOOPulating' the gut, *Microbiome*, **1**:3.

Png, C.W., Linden, S.K., Gilshenan, K.S., Zoetendal, E.G., McSweeney, C.S., Sly, L.I., McGuckin, M.A., & Florin, T.H.J. (2010) Mucolytic Bacteria With Increased Prevalence in IBD Mucosa Augment In Vitro Utilization of Mucin by Other Bacteria, *Am J Gastroenterol*, **105**:2420–2428.

Powers, D.M.W. (2011) Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation, *Journal of Machine Learning Technologies*, **2**:37–63.

Preidis, G.A., Ajami, N.J., Wong, M.C., Bessard, B.C., Conner, M.E., & Petrosino, J.F. (2016) Microbial-Derived Metabolites Reflect an Altered Intestinal Microbiota during Catch-Up Growth in Undernourished Neonatal Mice, *The Journal of nutrition*, **146**:940–948.

Pryde, S.E., Duncan, S.H., Hold, G.L., Stewart, C.S., & Flint, H.J. (2002) The microbiology of butyrate formation in the human colon, *FEMS Microbiology Letters*, **217**:133–139.

Quinn, G. & Keough, M., *Experimental design and data analysis for biologists* (Cambridge University Press, Cambridge, 2002).

Rajilić-Stojanović, M. & de Vos, W.M. (2014) The first 1000 cultured species of the human gastrointestinal microbiota, *Fems Microbiology Reviews*, **38**:996–1047.

Ramette, A. (2007) Multivariate analyses in microbial ecology, *FEMS microbiology ecology*, **62**:142–160.

Rappin, N. & Dunn, R., *wxPython In Action* (Manning Publication Co., 2006).

Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., Muehlbauer, M.J., Ilkayeva, O., Semenkovich, C.F., Funai, K., Hayashi, D.K., Lyle, B.J., Martini, M.C., Ursell, L.K., Clemente, J.C., Van Treuren, W., *et al.* (2013) Gut microbiota from twins discordant for obesity modulate metabolism in mice, *Science (New York, N.Y.)*, **341**:1241,214.

Ríos-Covián, D., Ruas-Madiedo, P., Margolles, A., Gueimonde, M., de los Reyes-Gavilán, C.G.,

& Salazar, N. (2016) Intestinal short chain fatty acids and their link with diet and human health, *Frontiers in Microbiology*, **7**:185.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.*, **27**:832–837.

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., & Palsson, B.Ø. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0., *Nature protocols*, **6**:1290–1307.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., & Weber, C.F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities, *Applied and environmental microbiology*, **75**:7537–7541.

Shankar, V., Agans, R., Holmes, B., Raymer, M., & Paliy, O. (2013) Do gut microbial communities differ in pediatric IBS and health?, *Gut microbes*, **4**:347–352.

Sim, K., Cox, M.J., Wopereis, H., Martin, R., Knol, J., Li, M.S., Cookson, W.O.C.M., Moffatt, M.F., & Kroll, J.S. (2012) Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing, *PloS one*, **7**:e32,543.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.J., Blugeon, S., Bridonneau, C., Furet, J.P., Corthier, G., Grangette, C., Vasquez, N., Pochart, P., Trugnan, G., Thomas, G., Blottière, H., Doré, J., Marteau, P., Seksik, P., & Langella, P. (2008) Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients, *Proceedings of the National Academy of Sciences of the United States of America*, **105**:16,731–16,736.

Storey, J.D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value, *Annals of statistics*, **31**:2013–2035.

Swainston, N., Smallbone, K., Mendes, P., Kell, D., & Paton, N. (2011) The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks, *Journal of integrative bioinformatics*, **8**:186.

Tailford, L.E., Crost, E.H., Kavanaugh, D., & Juge, N. (2015) Mucin glycan foraging in the human gut microbiome, *Frontiers in Genetics*, **6**:81.

Terzer, M., Maynard, N.D., Covert, M.W., & Stelling, J. (2009) Genome-scale metabolic networks, *Wiley interdisciplinary reviews. Systems biology and medicine*, **1**:285–297.

Thiele, I., Hyduke, D.R., Steeb, B., Fankam, G., Allen, D.K., Bazzani, S., Charusanti, P., Chen, F.C., Fleming, R.M., Hsiung, C.A., De Keersmaecker, S.C., Liao, Y.C., Marchal, K., Mo, M.L., Özdemir, E., Raghunathan, A., Reed, J.L., Shin, S.I., Sigurbjörnsdóttir, S., Steinmann, J., *et al.* (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen salmonella typhimurium lt2, *BMC Systems Biology*, **5**:1–9.

Thiele, I. & Palsson, B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction, *Nature protocols*, **5**:93–121.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., & Gordon, J.I. (2007) The Human Microbiome Project, *Nature*, **449**:804–810.

Valcárcel, B., Würtz, P., Seich al Basatena, N.K., Tukiainen, T., Kangas, A.J., Soininen, P., Järvelin, M.R., Ala-Korpela, M., Ebbels, T.M., & de Iorio, M. (2011) A differential network approach to exploring differences between biological states: an application to prediabetes, *PloS one*, **6**:e24,702.

van Rossum, G. & F.L., D. (2001) *Python Reference Manual*, http://www.python.org; Virginia, USA.

Walt, S.v.d., Colbert, S.C., & Varoquaux, G. (2011) The numpy array: A structure for efficient numerical computation, *Computing in Science & Engineering*, **13**:22–30.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., & Knight, R. (2016) Correlation detection strategies in microbial data sets vary widely in sensitivity and precision, *The ISME journal*, **10**:1669–1681.

Wenzel, U.A., Jonstrand, C., Hansson, G.C., & Wick, M.J. (2015) Cd103(+)cd11b(+) dendritic cells induce t(h)17 t cells in muc2-deficient mice with extensively spread colitis, *PLoS ONE*, **10**:e0130,750.

Wood, D.E. & Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome biology*, **15**:R46.

Wright, D.P., Rosendale, D.I., & Roberton, A.M. (2000) Prevotella enzymes involved in mucin oligosaccharide degradation and evidence for a small operon of genes expressed during growth on mucin, *FEMS Microbiology Letters*, **190**:73–79.

Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., Datsenko, K.A., Nakayashiki, T., Tomita, M., Wanner, B.L., & Mori, H. (2009) Update on the keio collection of escherichia coli single-gene deletion mutants, *Molecular Systems Biology*, **5**:335–335.

Yutin, N. & Galperin, M.Y. (2013) A genomic update on clostridial phylogeny: Gram-negative spore-formers and other misplaced clostridia, *Environmental microbiology*, **15**:2631–2641.

Appendix A

Correlation analyses

A.1 Correlation

Calculating correlation coefficients

In order to measure or identify associations between pairs of variables, correlation can be used.

The Pearson correlation coefficient measures the linear relationship between a given pair of variables and is calculated as the covariance of two random variables (X, Y) divided by the product of their standard deviations, as shown in Equations A.1 – A.4 (Quinn & Keough, 2002).

$$S_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{A.1}$$

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{A.2}$$

where x is any observed/measured value of the random variable, X, with mean $\bar{x}$; y is any

observed/measured value of the random variable, Y, with mean $\bar{y}$; n is the number of samples within each distribution ($n = n_y = n_x$) (Quinn & Keough, 2002).

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{A.3}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{A.4}$$

where σ_X is the standard deviation of the random variable X and σ_Y is the standard deviation of the random variable Y (Quinn & Keough, 2002).

A.2 Partial correlation

It is likely that various OTUs within a microbial community interact with one another in a complex manner and affect each others' abundance, meaning that each pairwise correlation will be affected by other OTUs. In such cases as these, it could be argued that it is preferable to use partial correlation in place of standard correlation because partial correlation is the correlation that remains between two variables (e.g. OTUs) after the effects of other variables have been removed by regression (Opgen-Rhein & Strimmer, 2007). In order to define partial correlation, we must first begin with regression.

Regression

Regression models allow us to either predict values for a given response variable based upon the values of a predictor variable, or to evaluate the relative effects of numerous predictor variables upon a given response variable (Johnson & Wichern, 2007). In our case, it would allow us to

look at the predictive value of OTU abundances upon one another.

Based upon the classical linear regression model, let our response variable be Y and let $z_1, z_2, ..., z_r$ be r predictor variables. The linear regression model for n independent observations may be written as follows:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \beta_2 z_{12} + \cdots + \beta_r z_{1r} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 z_{21} + \beta_2 z_{22} + \cdots + \beta_r z_{2r} + \epsilon_2 \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_n &= \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + \cdots + \beta_r z_{nr} + \epsilon_n \end{aligned} \tag{A.5}$$

which may be rewritten as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \dots & z_{1r} \\ 1 & z_{21} & z_{22} & \dots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \dots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_r \end{bmatrix} \tag{A.6}$$

which may be shortened to:

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon} \tag{A.7}$$

where ϵ represents an error term with the following properties:

1. $\mathrm{E}(\epsilon_j) = 0$;
2. $\mathrm{Var}(\epsilon_j) = \sigma^2$;

3. $\text{Cov}(\epsilon_j,\epsilon_k) = 0$, $j \neq k$;

(i.e. the error terms are normally distributed about 0 and have covariance of 0, implying either independence or non-linear dependence) (Johnson & Wichern, 2007). Therefore, the response variable may be represented by the sum of the weighted mean of the predictor variables ($z_1, z_2, ..., z_r$) and the error. Each observation (sample) is represented by a different row and each predictor variable is in a different column of $\boldsymbol{Z}$.

The model is fitted to the observed data to determine the values of the regression coefficients ($\boldsymbol{\beta}$) and the error variance (σ^2). This can be done by using the method of least squares (Johnson & Wichern, 2007). If we let b represent the values for $\boldsymbol{\beta}$ within a vector $\hat{\boldsymbol{\beta}}$, the method of least squares fits $\hat{\boldsymbol{\beta}}$ such that the difference between the observed response value, y_i, and the value of $b_0 + b_1 z_{j1} + \cdots + b_r z_{jr}$ is minimised:

$$\hat{\epsilon}_j = y_i - (b_0 + b_1 z_{j1} + \cdots + b_r z_{jr}) \tag{A.8}$$

$$\begin{aligned} \textit{residual sum of squares} &= \sum_{j=1}^{n} (y_j - (b_0 + b_1 z_{j1} + \cdots + b_r z_{jr}))^2 \\ &= \hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} \end{aligned} \tag{A.9}$$

where the deviations ($\hat{\epsilon}$) are the residuals (Johnson & Wichern, 2007). The least squares estimates, $\hat{\boldsymbol{\beta}}$, and the vector of residuals, $\hat{\boldsymbol{\epsilon}}$, can be obtained from $\boldsymbol{Z}$ and $\boldsymbol{Y}$ using matrix operations (Johnson & Wichern, 2007). The least squared estimates can then be used to determine the relative contributions of the predictor variables, according to confidence interval calculations for each element within $\hat{\boldsymbol{\beta}}$ (Johnson & Wichern, 2007). Furthermore, this classical linear regression model can be generalised to accommodate for variables with discrete distributions or bounded distributions, and in these cases the models are referred to as generalised linear models (GLM) (Nelder & Wedderburn, 1972).

Partial correlation following regression

Partial correlation coefficients are calculated as the correlation between the error terms in a regression model. First we must refer back to the regression models where we had a response variable Y and r predictor variables, $Z_1, Z_2, \ldots, Z_r$ (represented as a vector, $\boldsymbol{Z}$); however, this time we shall suppose that the variables are random and have a joint distribution (Johnson & Wichern, 2007). Using regression, the linear predictor allows us to predict outcomes of Y for any set of $\boldsymbol{Z}$:

$$Y = b_0 + b_1 z_1 + \cdots + b_r z_r = b_0 + \boldsymbol{b'Z} \tag{A.10}$$

As demonstrated in equation A.8, the error term associated with this predictor is given by:

$$\epsilon = Y - b_0 - \boldsymbol{b'Z} \tag{A.11}$$

Consequently, we select b_0 and $\boldsymbol{b}$ so as to identify the optimal linear predictor which, by definition, minimises the mean square error (Johnson & Wichern, 2007), given by:

$$E(Y - b_0 - \boldsymbol{b'Z})^2 \tag{A.12}$$

Now, in the case where $Y, Z_1, Z_2, \ldots, Z_r$ have a joint distribution with a vector of means, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, we can go on to express the optimal linear predictor in terms of these parameters. First, we must partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ into their Y and Z components (Johnson & Wichern, 2007):

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \boldsymbol{\mu_Z} \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{YY} & \boldsymbol{\sigma'_{ZY}} \\ \boldsymbol{\sigma_{ZY}} & \boldsymbol{\Sigma_{ZZ}} \end{bmatrix} \tag{A.13}$$

where $\boldsymbol{\sigma'_{ZY}} = [\sigma_{YZ_1}, \sigma_{YZ_2}, \ldots, \sigma_{YZ_r}]$.

Given that the linear predictor will have minimum mean square values when the following is true (Johnson & Wichern, 2007):

$$\boldsymbol{\beta} = \boldsymbol{\Sigma_{ZZ}^{-1}\sigma_{ZY}} \qquad \beta_0 = \mu_Y - \boldsymbol{\beta'\mu_Z} \tag{A.14}$$

we can substitute these equations for the $\boldsymbol{b}$ and b_0 terms in the definition of the mean square error (equation A.12):

$$E(Y - \beta_0 - \boldsymbol{\beta'Z})^2 = E(Y - \mu_Y - \boldsymbol{\sigma'_{ZY}\Sigma_{ZZ}^{-1}(Z - \mu_Z)})^2 \tag{A.15}$$

Now that we have the mean square error from the optimal linear predictor in terms of the means and covariances of the variables, we can consider the pair of errors we would obtain from the optimal linear predictors of two variables, Y_1 and Y_2:

$$\begin{aligned} \boldsymbol{\epsilon}_1 &= \boldsymbol{Y}_1 - \mu_{\boldsymbol{Y}_1} - \boldsymbol{\Sigma_{Y_1Z}\Sigma_{ZZ}^{-1}(Z - \mu_Z)} \\ \boldsymbol{\epsilon}_2 &= \boldsymbol{Y}_2 - \mu_{\boldsymbol{Y}_2} - \boldsymbol{\Sigma_{Y_2Z}\Sigma_{ZZ}^{-1}(Z - \mu_Z)} \end{aligned} \tag{A.16}$$

The correlation of the error terms for the two variables, Y_1 and Y_2, gives us the partial correlation coefficient. Therefore, the partial correlation is the correlation remaining between Y_1 and Y_2

after we have eliminated the effects of all other variables, $Z_1, Z_2, \ldots, Z_r$, via regression.

In the case of microbiota data, partial correlation may be used to investigate pairwise associations between OTUs with any confounding effects from other OTUs accounted for.

Appendix B

Validation studies for utilising the proportionality measure

B.0.1 Validation studies using simulated data

In order to calibrate the measure of goodness of fit to proportionality (ϕ) against likely features of microbiota data, I created simulated datasets where the relationships between variables were predetermined. The ϕ values for each pair of variables in each dataset were computed using the measure of goodness of fit, as presented previously in Section 2.3.4. The equations for calculating ϕ are repeated here (Equations B.1 and B.2) for convenience:

$$\phi_{(A,B)} = \frac{\mathrm{Var}(\log(A/B))}{\mathrm{Var}(\log A)} \tag{B.1}$$

$$\phi_{(B,A)} = \frac{\mathrm{Var}(\log(B/A))}{\mathrm{Var}(\log B)} \tag{B.2}$$

where A represents the vector of relative counts for genus A, and B represents the vector of relative counts for genus B across all samples within a given group (e.g. cases or controls).

The simulated datasets have been categorised in the following way:

- **Basic simulated data** — This first dataset provides a range of different relationships between variables. This was done by generating data for two variables (A and D) by sampling from defined distributions and subsequently applying varying transformations to these key variables. The transformations applied include:

  - exact scaling of variables (i.e. providing an example of a perfectly proportional relationship),
  - linear transformations with non-zero constant (i.e. providing an example of a linear but not perfectly proportional relationship) and
  - quadratic transformations.

  These transformations reflect potential associations between bacteria, which could equate to symbiosis, competition, coexistence or other ecological relationships.

- **Noisy simulated data** — The second set of simulated data was created to analyse the effect of noise upon ϕ measurements. This was done by generating data for an initial variable by sampling from a defined distribution and then generating the other variables by transforming the original data via the addition of increasing noise terms. These validations are important for understanding how methodical errors and natural variation could affect ϕ measurements.

- **Zero-rich simulated data** — The 16S rRNA gene abundance data used to analyse the microbiota is typically sparse (zero-rich), so it is important to understand how ϕ behaves for zero-rich data, especially given that the presence of zeros directly influences the amount of data available for calculating ϕ. Therefore, data for each variable was generated by sampling from independently defined zero-inflated Poisson (ZIP) distributions with increasing levels of zeros (i.e. increasing sparsity).

The distributions and relationships between variables within these simulated datasets are described below and summarised in Tables B.1 – B.5.

Basic simulated data

The first simulated dataset (Table B.1) was created simply to identify how ϕ values behave for different relationships between different variables. This was done by sampling from known distributions to create two variables (A and D) and then transforming these variables to create new variables with defined relationships to one another. Variable A was determined by randomly sampling from a Normal distribution where $\mu = 200$ and $\sigma^2 = 100^2$; these values were chosen to reflect the count values sometimes observed for the higher abundance bacteria found within the microbiota. Any negative values were replaced with zero values in order to better represent the positive count data generated for microbiota data. The second variable, B, was determined by doubling all values for A in order to act as the control, where B is perfectly proportional to A, so we expect $\phi_{(A,B)} = \phi_{(B,A)} = 0$. Variable C was determined by adding a constant to B to create a variable that is a linear transformation of A that is correlated (but not perfectly proportional to) A and B. Variable D was determined by sampling from another Normal distribution independently of A. This was done in order to generate a variable that should not have any pre-determined relationship to A and therefore should give rise to non-zero values for $\phi_{(A,D)}$ and $\phi_{(D,A)}$. D also has a lower mean and variance than A to better reflect the low-abundance bacteria found within the microbiota. All other variables (E to L) were determined by transforming and/or combining values from A and D in order to determine how ϕ behaves for different known relationships between variables. For exact relationships between variables, refer to Table B.1.

Figure B.1 shows the proportionality values calculated between all pairs of variables within the basic simulated dataset summarised in Table B.1.

As expected, the ϕ values for the AB pair were both zero and the ϕ values for AC were marginally higher, demonstrating the lowered proportionality between A and C compared to A and B; this result is expected due to the application of a negative constant (-10) in the transformation from A to C. Similarly, the ϕ values for BC were matched closely to that for AC.

Variable (i.e. 'OTU')	Parameters
A	$\sim N(\mu = 200, \sigma^2 = 100^2)$; Negative values to 0
B	$2A$
C	$2A - 10$; Negative values to 0
D	$\sim N(\mu = 25, \sigma = 20^2)$; Negative values to 0
E	$2A - D$; Negative values to 0
G	$4A - (43 + D)$; Negative values to 0
H	$2A - 43$; Negative values to 0
I	A^2
J	$\left\lfloor \frac{A^2}{2} \right\rfloor$
K	$43 - D$; Negative values to 0
L	$\left\lfloor \frac{A}{2} \right\rfloor$

Table B.1: Distributions of simulated data with defined relationships between variables from the basic simulated dataset. This set includes varying distributions and transformations in order to demonstrate the behaviour of the ϕ measure for different sorts of relationships between variables. Variables A to L are shown along with their distribution parameters, or their relationship to other variables within the dataset. All negative values were converted to 0 in order to reflect the positive count data that the proportionality method will be applied to in the context of studying the microbiota.

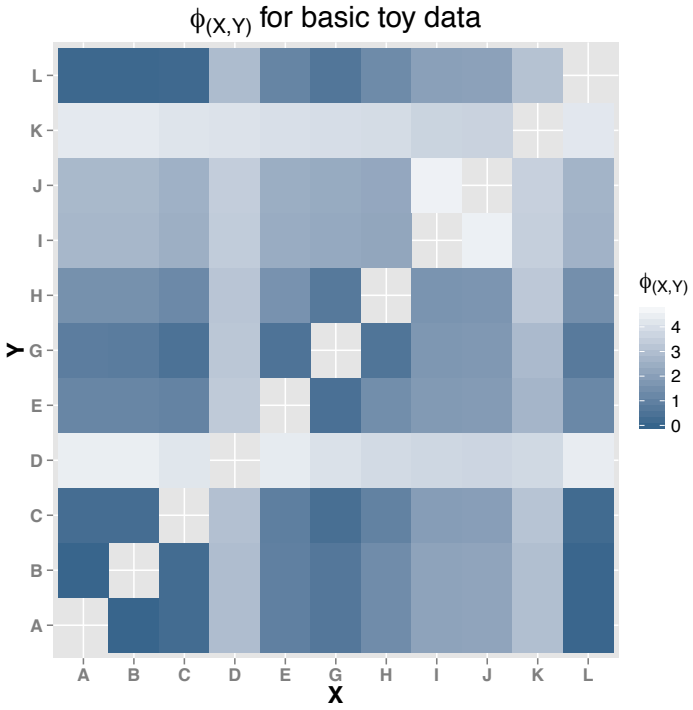


Figure B.1: *Heat map showing the ϕ values for all pairs of variables within the basic simulated dataset. The shading represents increasing ϕ values where 0 (representing perfect proportionality) is shown in dark and the lighter shades represent higher ϕ values.*

Unsurprisingly, we see very little proportionality between A and D due to the fact that the values for D were generated independently of A.

As we compare A, B and C with E through to J, we see a gradual decrease in proportionality (i.e. an increase in ϕ values). This is expected because E through to J become increasingly less proportional to A due to the addition of non-zero constants and non-linear scaling.

K has high ϕ values when paired with all other variables. This is due to the fact that it is independent to variable A (unlike many of the other variables) and, although K is negatively proportional to D, its constant is proportionately large compared to the mean of D, thereby making it non proportional to D.

L shows high proportionality (i.e. low ϕ values) when paired with A to C due to the fact that it is proportional to A with only a small amount of noise introduced by rounding. Hence it also shows a similar pattern to A in its proportionality to variables E through to K. L also shows low proportionality to D (i.e. high ϕ values) due to its independence to D.

These results show how ϕ behaves for different relationships between variables. It illustrates how a low ϕ value demonstrates the existence of a proportional relationship between a given pair of variables, and shows how the ϕ value increases as the linear relationship deviates from a path through the origin (i.e. as the magnitude of an additional constant increases). This suggests (as shown by the relationship between D and K) that for positive count data, this measure would be inappropriate for identifying strong, negative proportional relationships between variables, such as bacterial genera competing with one another. This is due to the fact that the relationship (if negatively proportional) would need to be modelled with a large positve constant in order to give rise to positive count data.

Finally, it is important to note the asymmetry of the heat map along $y = x$. Although the overall pattern is similar on both halves of the heat map, there are subtle differences in the two ϕ values produced for each pair of variables. This is due to the asymmetry of the two ϕ values, which tends to increase as the two ϕ values deviate further from zero.

Noisy simulated data

The second set of simulated data was created to analyse the effect of noise upon ϕ measurements. This is due to the fact that noise is an inherent and unavoidable characteristic of biological data. Here, variable A was determined by randomly sampling from a Normal distribution where $\mu = 200$ and $\sigma^2 = 100^2$; any negative values were replaced with zero values, as above. Variable B was set to match A as a control in order to confirm that $\phi_{(A,B)} = \phi_{(B,A)} = 0$. Data for variables C through to L were determined by taking each value of A and adding a noise term that was proportional to A such that:

$$X = A + \epsilon A; \qquad \epsilon \sim U(-y, y) \tag{B.3}$$

where ϵ is a random variable sampled from a Uniform distribution and y represents 0.1 for C, 0.2 for D, 0.3 for E, and so on. All negative values are replaced with zero values. Refer to Table B.2 for the noise term parameters for all variables in this dataset. By creating this simulated dataset with progressively increasing levels of noise, it is possible to observe the effect of increasing uniform noise on proportionality.

Figure B.2 shows the proportionality values calculated between all pairs of variables within the noisy simulated dataset. As expected, $\phi_{(A,B)} = \phi_{(B,A)} = 0$. We also observe a gradual decrease in proportionality (an increase in ϕ values) as the noise term coefficient increases from C through to L, as expected.

Zero-rich simulated data

The 16S rRNA data that is commonly used to analyse the microbiota is zero-rich and sparse. However, for a given pair of genera, samples can only be used when both genus abundance values are non-zero due to the fact that the proportionality measure is calculated using log transformations and $\log 0$ is undefined. This is discussed in Section 3.3.1 and illustrated in

Variable (i.e. 'OTU')	Parameters
A	$A \sim N(\mu = 200, \sigma^2 = 100^2)$; Negative values to 0
B	A
C	$A + \epsilon A; \epsilon \sim U(-0.1, 0.1)$; Negative values to 0
D	$A + \epsilon A; \epsilon \sim U(-0.2, 0.2)$; Negative values to 0
E	$A + \epsilon A; \epsilon \sim U(-0.3, 0.3)$; Negative values to 0
F	$A + \epsilon A; \epsilon \sim U(-0.4, 0.4)$; Negative values to 0
G	$A + \epsilon A; \epsilon \sim U(-0.5, 0.5)$; Negative values to 0
H	$A + \epsilon A; \epsilon \sim U(-0.6, 0.6)$; Negative values to 0
I	$A + \epsilon A; \epsilon \sim U(-0.7, 0.7)$; Negative values to 0
J	$A + \epsilon A; \epsilon \sim U(-0.8, 0.8)$; Negative values to 0
K	$A + \epsilon A; \epsilon \sim U(-0.9, 0.9)$; Negative values to 0
L	$A + \epsilon A; \epsilon \sim U(-1, 1)$; Negative values to 0

Table B.2: Distributions of simulated data with their relationships between variables from the noisy simulated dataset. This set includes variables that are derived from the first variable (A) with increasing levels of noise applied to variable C through to L. Variables A to L are shown along with their noise distribution parameters. All negative values were converted to 0 in order to reflect the positive count data that the proportionality method will be applied to in the context of studying the microbiota.

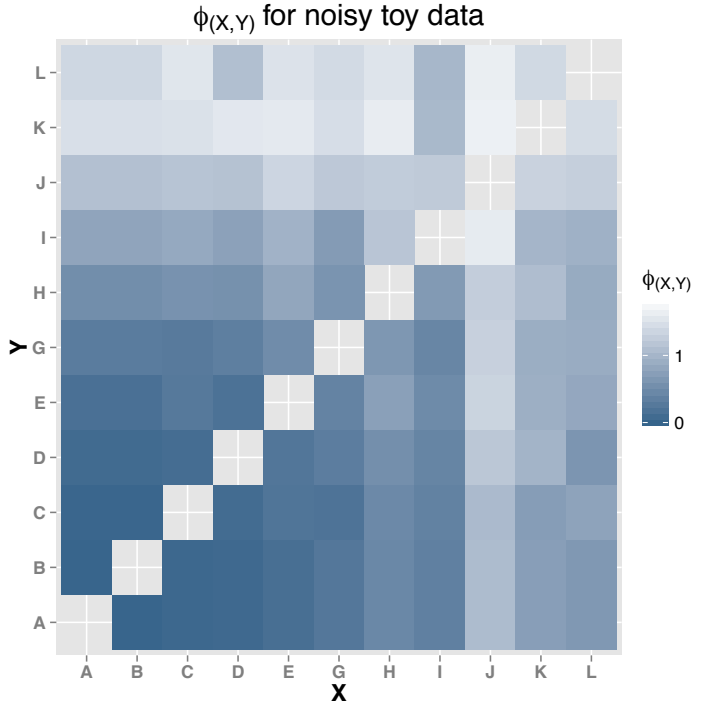


Figure B.2: *Heat map showing the ϕ values for all pairs of variables within the noisy simulated dataset with progressively increased noise. The shading represents increasing ϕ values where 0 (perfect proportionality) is shown in dark and the lighter shades represent higher ϕ values.*

Table 3.7. Therefore, here we explore the effect of zero-rich data upon the calculation of proportionality.

It could be argued that microbial 16S rRNA gene count data could be modelled using a zero-inflated Poisson (ZIP) distribution, which is a mixture distribution between a Poisson distribution and a point mass at 0. Both underlying distributions give rise to zero values, which model the microbiota well because the zero values that arise can be the product of either of two possibilities represented by the mixed distribution:

1. A given zero value is a **true negative** result due to the total absence of that variable within a given sample;
2. A given zero value is a **false negative** result due to primer bias, lack of detection due to low abundance, or experimental error.

Therefore, values for variable A were sampled from a ZIP distribution where $\lambda = 2$ (where λ represents the rate parameter of the Poisson distribution) and $\omega = 0.1$ (where ω represents the probability of a zero value occurring in a single event). This means that A represents a low-abundance genus in our system. The second variable, B, was matched to A in order to act as a control to confirm that $\phi_{(A,B)} = \phi_{(B,A)} = 0$.

Zero-rich data was generated for variables C through to L by increasing the proportion of zero values (increasing the ω parameter) to see how increased levels of zeros in low-abundance data would affect proportionality measures. The number of zeros within the data was progressively increased by 10% for each variable, and all variables were independent of one-another apart from variable B, which was made equal to variable A. Refer to Table B.3 for all parameters.

Figure B.3 shows the proportionality values calculated between all pairs of variables within the zero-rich simulated dataset with varied distribution parameters. Here we see that the ϕ values for the AB pair were both zero, as expected. We also observe the occurrence of undefined ϕ values (shown by grey boxes in the heat map in Figure B.3), which indicates that there were no

Variable (i.e. 'OTU')	Parameters
A	$\sim ZIP(\lambda = 2, \omega = 0.1)$
B	A
C	$\sim ZIP(\lambda = 2, \omega = 0.1)$
D	$\sim ZIP(\lambda = 2, \omega = 0.2)$
E	$\sim ZIP(\lambda = 2, \omega = 0.3)$
G	$\sim ZIP(\lambda = 2, \omega = 0.4)$
H	$\sim ZIP(\lambda = 2, \omega = 0.5)$
I	$\sim ZIP(\lambda = 2, \omega = 0.6)$
J	$\sim ZIP(\lambda = 2, \omega = 0.7)$
K	$\sim ZIP(\lambda = 2, \omega = 0.8)$
L	$\sim ZIP(\lambda = 200, \omega = 0.9)$

Table B.3: Distributions of sparse simulated data (generated using a zero-inflated Poisson (ZIP) distribution) where all variables are independent of one another. Variables A to L are shown along with their associated distribution parameters. This set includes variables that exhibit increasing levels of zeros.

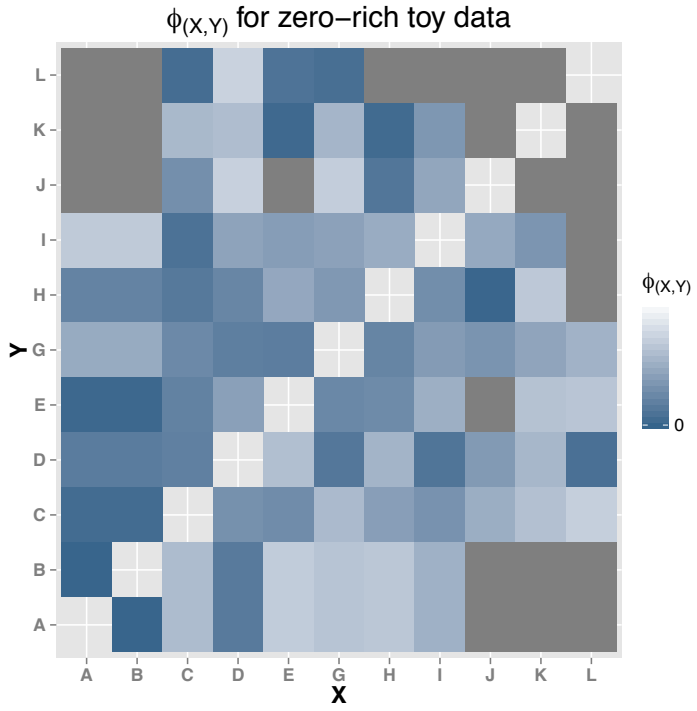


Figure B.3: *Heat map showing the ϕ values for all pairs of variables within the zero-rich simulated dataset. The shading represents increasing ϕ values where 0 (perfect proportionality) is shown in dark and the lighter shades represent higher ϕ values. Grey boxes show where ϕ could not be determined due to a lack of non-zero data for a given pair of variables.*

non-zero pairs that could be used to calculate ϕ for the given pair of variables. Unsurprisingly, the inability to define ϕ becomes more frequent as the proportion of zeros in the data increases; hence, variable L (with 90% zero values) shows very few ϕ values that could be computed with respect to the other variables (A to K). There is no distinct pattern in the distribution of ϕ values produced using this dataset, which is unsurprising because variables C through to L were generated independently of each other and independently of variables A and B.

It is important to note that although the ϕ values show no pattern between the variables, all of the ϕ values were below 1 despite the fact that none of the data is proportional. This is in contrast to the previous datasets, which show that ϕ reaches values above 4 in some of the datasets. These lower ϕ values are generated due to the reduction in usable samples, which is due to the omission of zero pairs. This is demonstrated in the next two validation studies. These studies are presented in Sections B.0.2 and B.0.3, where these biased ϕ values observed in zero-rich data are also considered within the context of the microbiota and differential networks.

B.0.2 Observing distributions of ϕ values for varying levels of sparsity

In order to observe the effect of high levels of zeros upon the distributions of phi measurements calculated from real microbiota data, we can observe the distributions of ϕ values produced by real data with different proportions of zeros in it. To do this, a threshold can be applied such that we begin by including all pairs of genera when we calculate the pairwise ϕ values for a given data set. This can then be repeated including only pairs of genera that have more than 10% of samples providing non-zero values. Again, this can be repeated including only pairs of genera that have more than 20% of samples providing non-zero values, and so on for each 10% interval up to 100% non-zero pairs of data.

Here, this thresholding method was applied to 16S rRNA count data for two previously published datasets from a study of the gut microbiota in inflammatory bowel disease (IBD) (Morgan

et al., 2012). It was applied to 27 samples taken from healthy controls and 43 samples taken from patients with ileal Crohn's disease (iCD).

Figure B.4 shows the distributions of ϕ values between all pairs of genera identified within the 27 healthy individuals (top row) and the 43 patients with ileal Crohn's disease (iCD; bottom row) when we include all data (left column) and when we include only the data with at least 20% non-zero pairs (right column). When we include all pairwise ϕ values for each dataset, we observe more values at the lower end of the range. Therefore, the ϕ values show highly skewed distributions for both datasets (healthy and iCD) such that we appear to have an over-representation of highly proportional pairs of bacterial genera. However, there is also a progressive reduction in low ϕ values as the threshold for the proportion of non-zero data increases; as we eliminate pairs of genera that do not have enough non-zero data, there are fewer values observed at the lower end of the range. These results agree with the observations in the heat map shown above (Figure B.3) where zero-inflated data showed lower ϕ values.

B.0.3 Observing the effects of sample size upon ϕ values

In order to confirm whether the biased ϕ values for zero-rich variables are caused by the severe reduction in usable sample size, a test was carried out on simulated data. Each dataset consisted of variables A to L, which were independently sampled from the same zero-inflated Poisson distribution. However, each of the four datasets had a different sample size and ZIP rate parameters, as shown in Table B.4. Therefore, for each variable with 30 sample values from a ZIP distribution where $\omega = 0.6$ (as is the case for all variables in *Dataset 1* in Table B.4), typically only 12 of the randomly generated values would be non-zero. Similarly, for a variable with 300 sample values from a ZIP distribution where $\omega = 0.6$, typically approximately 120 of the randomly generated values would be non-zero.

In order to ensure that the observed results were not just due to chance because of sampling, this process of data generation and ϕ calculation was repeated 20 times for each of the four datasets to assess the stability of the resulting ϕ values. The results are summarised below

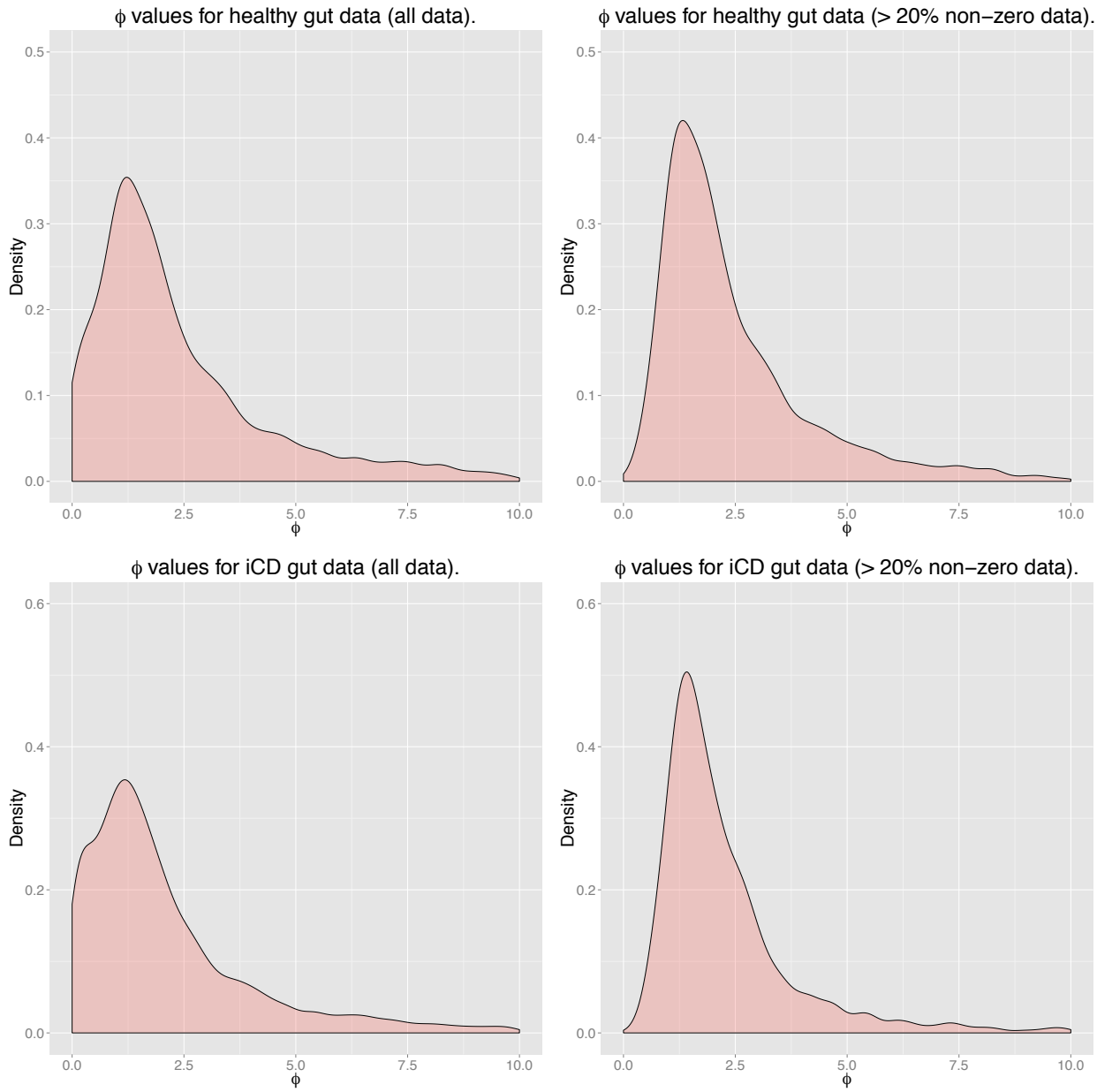


Figure B.4: *Density plots showing the distributions of ϕ values for two datasets.* **Top left:** *ϕ values for all pairs of genera identified within samples taken from healthy individuals. All pairwise ϕ values were included, regardless of the proportion of non-zero data available for the calculation of ϕ.* **Top right:** *ϕ values for pairs of genera identified within samples taken from healthy individuals where at $> 20\%$ of the data available for the calculation of ϕ were non-zero values.* **Bottom left:** *ϕ values for all pairs of genera identified within samples taken from individuals with ileal Crohn's disease (iCD). All pairwise ϕ values were included, regardless of the proportion of non-zero data available for the calculation of ϕ.* **Bottom right:** *ϕ values for pairs of genera identified within samples taken from individuals with iCD where at $> 20\%$ of the data available for the calculation of ϕ were non-zero values.*

in Table B.4, and a heat map taken for a single run for each of the four datasets is shown in Figures B.5 to B.8.

Dataset	Sample size	ZIP parameters	Typical ϕ range	Stability
1	**30**	$\sim ZIP(\lambda = 2, \omega = 0.6)$	$\phi_{(X,Y)max} < 1$	20/20
2	**300**	$\sim ZIP(\lambda = 2, \omega = 0.6)$	$\phi_{(X,Y)max} \geq 3$	20/20
3	**30**	$\sim ZIP(\lambda = 200, \omega = 0.6)$	$\phi_{(X,Y)max} < 1$	17/20
4	**300**	$\sim ZIP(\lambda = 200, \omega = 0.6)$	$\phi_{(X,Y)max} \geq 3$	20/20

Table B.4: Distributions of sparse simulated datasets with varied means and sample sizes (generated using zero-inflated Poisson (ZIP) distributions). Each dataset consists of variables A to L, which are sampled independently from the same distribution. The stability score represents the proportion of times the given dataset parameters caused the maximum ϕ value to fall within the range stated. All of the results were highly stable, where the results were repeated for at least 17/20 times.

These results show that the bias towards low ϕ values in zero-rich data is due to the large reduction in sample size (due to the removal of zero values) and this effect is absent when larger sample sizes are used. Therefore it is important to maximise sample size where possible in order to get an accurate estimate of the ϕ values for a given pair of variables.

The effect of sample sizes upon parameter estimations is a widely known problem, and it is common to determine the ideal minimum sample size for a given study before collecting the samples required. In the case of the estimation of ϕ, the usable sample sizes are smaller than the total number of samples gathered due to the loss of zero-pairs, which is due to sparsity of the data. This subsequently implies that the bias observed in the ϕ values in small samples will depend upon the sparsity of the data (i.e. the number of zero-pairs that have to be omitted from the ϕ calculation) due to its effect on usable sample size. Therefore, it would be interesting to repeat these simulation studies with a varied ω (sparsity) parameter to identify how the sparsity affects the sample size and, hence, the ϕ values. For example, as stated previously, for the simulated datasets of size 30 and 300 with an ω value of 0.6 (shown above), the number of non-zero abundance values for a given bacterium would be approximately 12 and 120, respectively. Therefore, when considering pairs of bacteria for calculating ϕ, the usable pairs of abundance values would be considerably less than 12 and 120 respectively, thanks to

the removal of all pairs of data containing at least one zero value. It would therefore be worth determining the ideal usable sample size for these simulated datasets in light of the sparsity parameter. Using the information from simulation studies such studies, it may be possible to determine appropriate scaling factors for situations where the small sample size and sparsity causes a skew in the ϕ values. Furthermore, by observing the effect of a range of sparsity values and sample sizes upon the ϕ values returned from simulated data, it may be possible to show how the bias in ϕ plateaus as the sample size increases. It may also demonstrate ideal sample sizes for future studies with known levels of sparsity in the data.

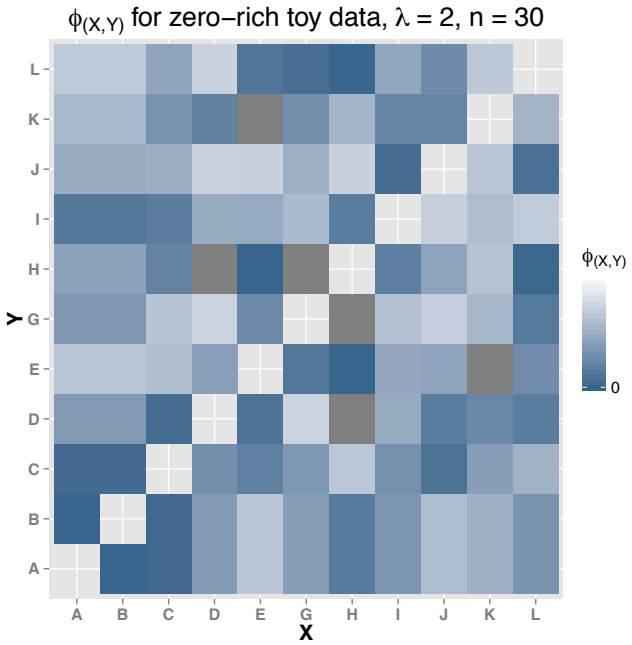


Figure B.5: *Heat map showing the ϕ values for the zero-inflated Poisson toy data with $\lambda = 2, \omega = 0.6$ and sample size of 30. The blue gradient represents progressive ϕ values where 0 is shown in dark and the lighter shades represent higher ϕ values.*

B.0.4 Validating the use of bootstrap empirical distributions

For each pair of genera within the microbiota, a corresponding pair of ϕ values can be calculated to represent the association between the two genera. However, in order to compare these ϕ values between dysbiosis (disease cases) and eubiosis (healthy controls), it is necessary to enforce some criteria upon which to decide whether specific differences in cases and controls are deemed

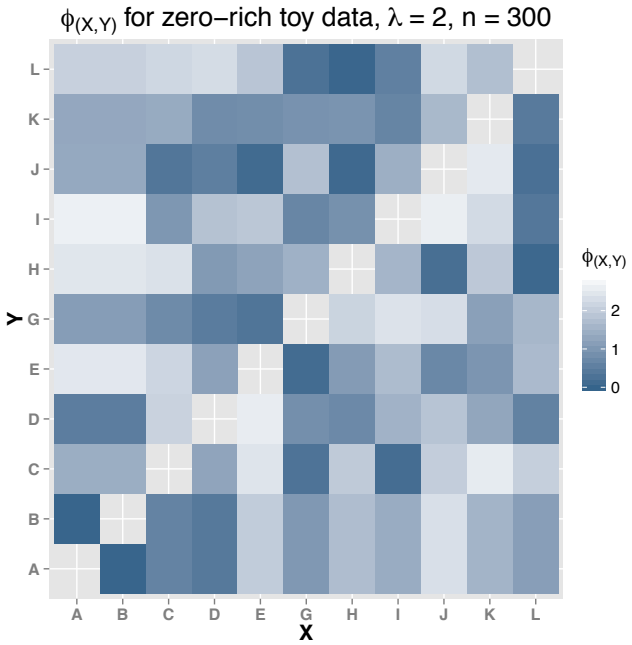


Figure B.6: *Heat map showing the ϕ values for the zero-inflated Poisson toy data with $\lambda = 2, \omega = 0.6$ and sample size of 300. The blue gradient represents progressive ϕ values where 0 is shown in dark and the lighter shades represent higher ϕ values.*

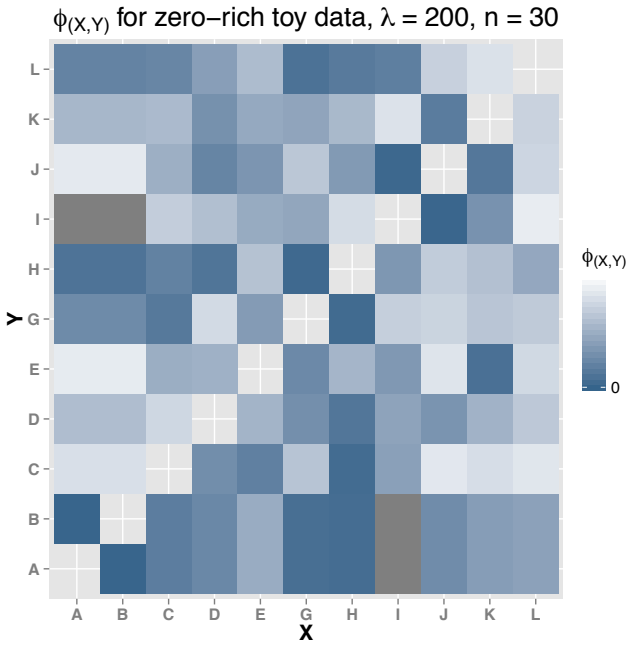


Figure B.7: *Heat map showing the ϕ values for the zero-inflated Poisson toy data with $\lambda = 200, \omega = 0.6$ and sample size of 30. The blue gradient represents progressive ϕ values where 0 is shown in dark and the lighter shades represent higher ϕ values.*

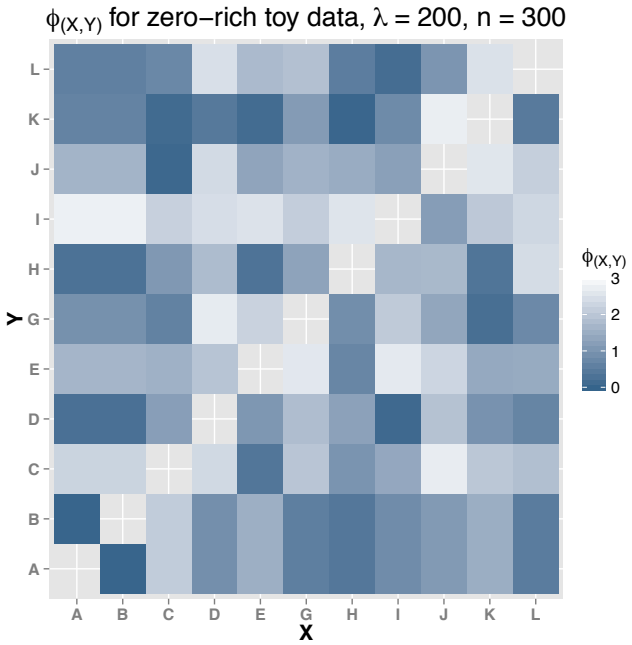


Figure B.8: *Heat map showing the ϕ values for the zero-inflated Poisson toy data with $\lambda = 200, \omega = 0.6$ and sample size of 300. The blue gradient represents progressive ϕ values where 0 is shown in dark and the lighter shades represent higher ϕ values.*

as statistically significant. Here, we do this by generating and using empirical distributions to compare equivalent ϕ values in cases and controls.

An empirical distribution for each ϕ value in cases can be generated by using the bootstrap method to resample the data with replacement and subsequently generate a distribution of ϕ estimators. Here, we demonstrate that the bootstrap distributions do not exhibit a directional bias into the bootstrap ϕ estimators.

Test data was generated by simulating n datasets from a defined sampling distribution and with known relationships between the two variables, A and B. This allowed us to generate a distribution of 'true' ϕ values (ϕ_{true}) for a given sampling distribution and defined relationship between the two variables, A and B (see Table B.5).

In order to generate a corresponding bootstrap distribution of ϕ values ($\phi_{bootstrap}$) for comparison, one of these datasets was randomly selected and bootstrap resampling was applied n times to generate a bootstrap distribution of ϕ estimators.

This process of examination was repeated numerous times with the following variations:

- varying numbers of sampling repeats (i.e. the number of simulated distributions and bootstraps), n,
- varying levels of proportionality via the application of a varied noise coefficient, γ (refer to Table B.5).

For each number of sampling repeats (n) and each level of noise (γ), the median true value ($\tilde{\phi}_{true}$) was compared to the corresponding median bootstrap value $\tilde{\phi}_{bootstrap}$.

Variable (i.e. 'OTU')	**Parameters**
A	$A \sim N(\mu = 200, \sigma^2 = 100^2)$; Negative values to 0
B	$A + \epsilon A; \epsilon \sim U(-\gamma, \gamma)$; Negative values to 0

Table B.5: Distributions of test data and the relationship between variables. Variable B is derived from variable A, where the noise coefficient, γ, is increased for each test.

In order for the bootstrap estimator ($\phi_{bootstrap}$) to be a good match to the distribution of true ϕ values for a given edge (ϕ_{true}), it should not show any directional bias. Using simulated data and bootstrap resampling, we observed that the bootstrap did not show any directional bias.

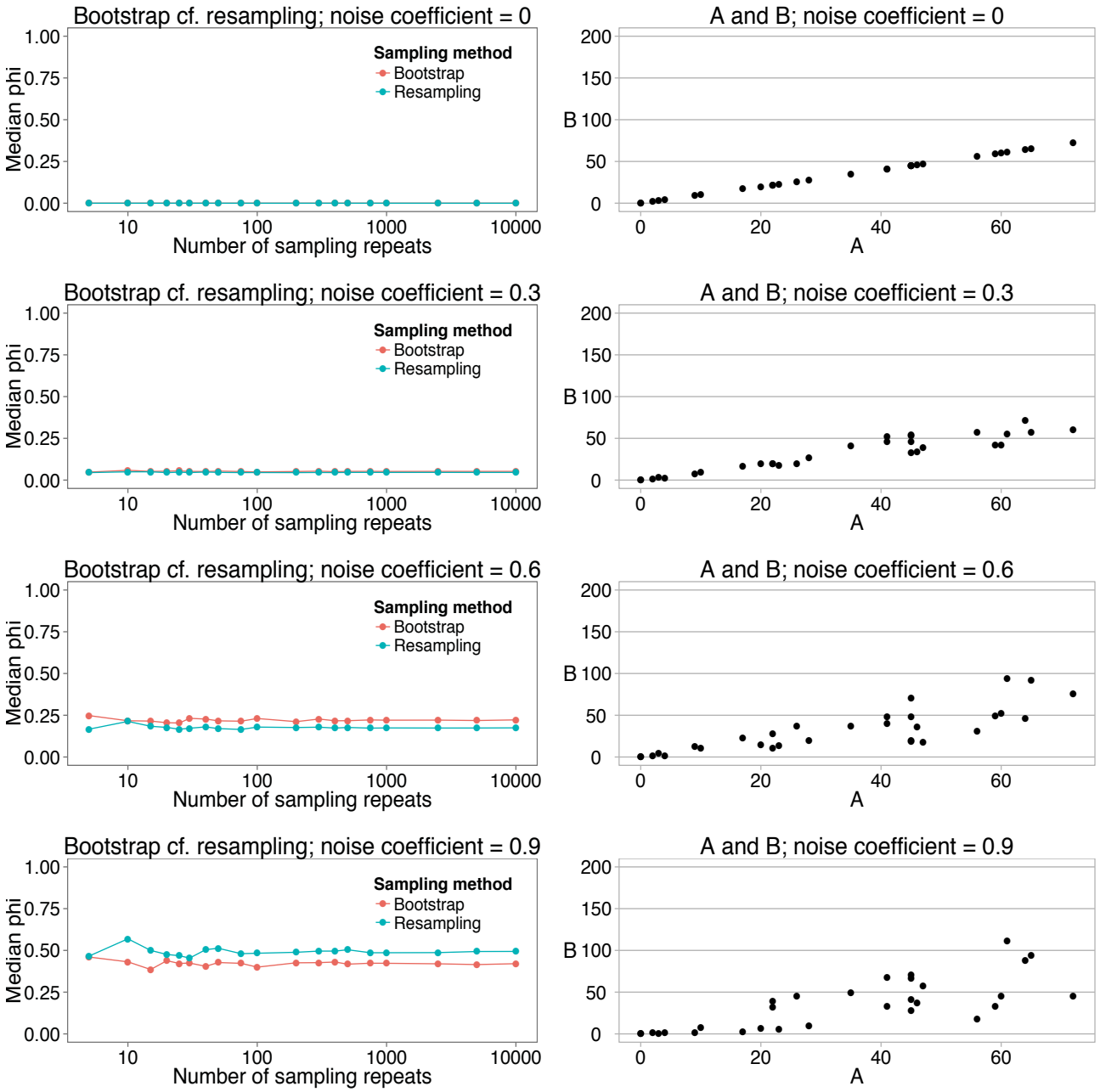


Figure B.9: *A simulated comparison of ϕ values generated by using boostrap subsampling and the equivalent ϕ values generated via repeated subsampling from a known distribution. Each row represents a different dataset where variables A and B become less proportional on each row due to an increasing noise coefficient, γ, as described in Table B.5.* **Left column:** *Plots showing the median of the bootstrap estimators of ϕ ($\phi_{bootstrap}$) in red and the corresponding median of the true ϕ values (ϕ_{true}) determined using simulated datasets from a specified distribution in blue. The x-axis represents the number of bootstrap repeats (i.e. number of ϕ values) used to create the distribution of bootstrap estimators of ϕ and the number of simulated datasets used to calculate the distribution of true ϕ values. The y-axis represents the median of each ϕ distribution.* **Right column:** *Plots showing an example of a single simulated dataset for variables A and B, for which ϕ is calculated. Each dot shows the value of A and B for a single sample in the example simulated dataset.*

Appendix C

Full differential networks for IBD

In Chapter 3, the proportionality differential network method was applied to real microbiota data taken from patients suffering with inflammatory bowel disease (IBD) and healthy controls. Each edge within the resulting network is associated with a certainty score, as discussed in the main chapter. The results presented previously showed all edges with a certainty sore above 0.5. The networks shown here have all edges with a non-zero certainty score so that the reader can observe the full results.

C.1 Proportionality differential networks

C.1.1 Ileal Crohn's disease

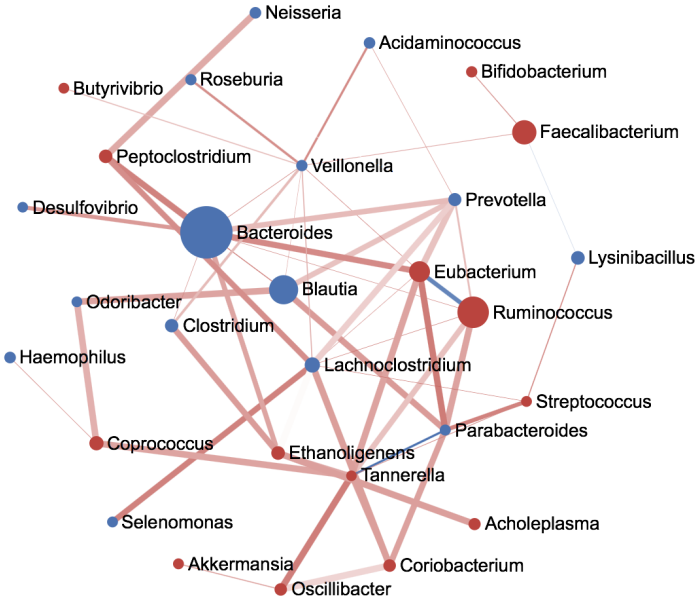


Figure C.1: **Healthy/iCD differential network:** *Controls = Healthy; Cases = iCD; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in phi.*

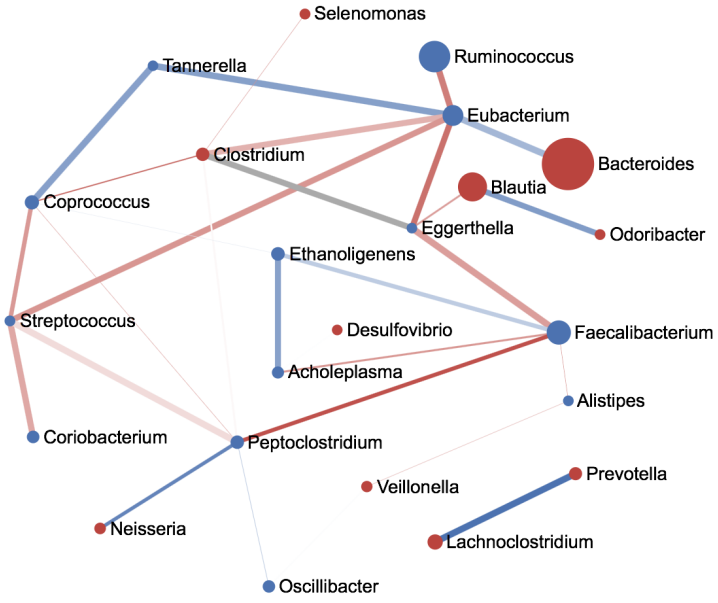


Figure C.2: **iCD/Healthy differential network:** *Controls = iCD; Cases = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in phi.*

C.1.2 Non-ileal Crohn's disease

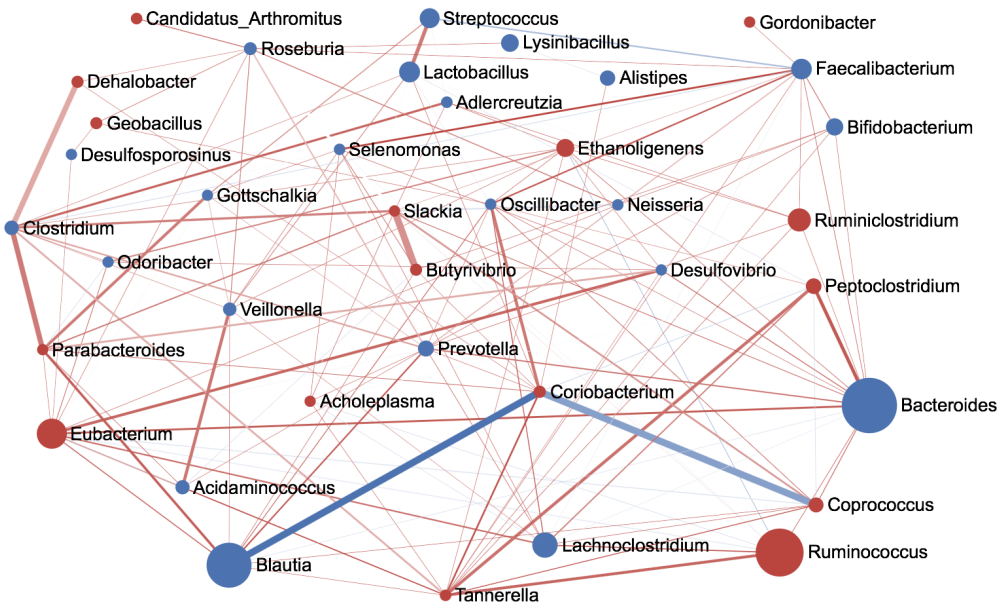


Figure C.3: **Healthy/Non-iCD differential network:** *Controls = Healthy; Cases = Non-iCD; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in phi.*

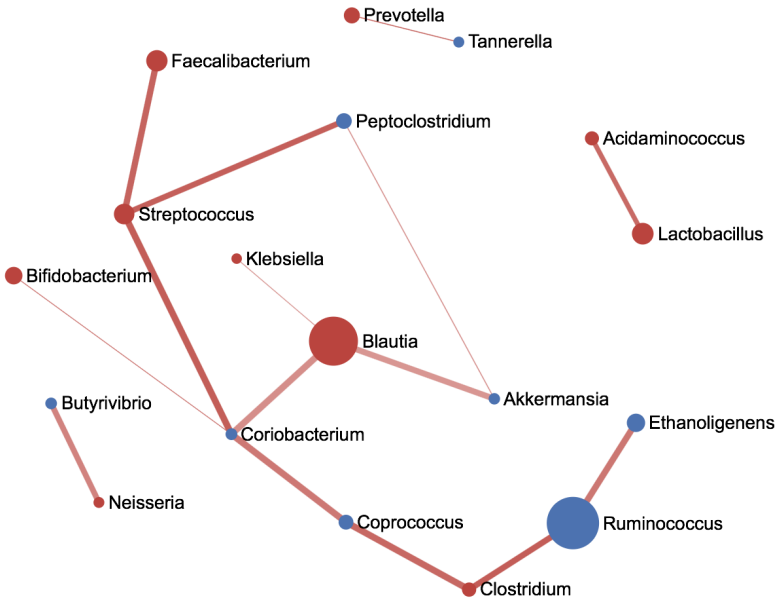


Figure C.4: **Non-iCD/Healthy differential network:** *Controls = Non-iCD; Cases = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in phi.*

C.1.3 Ulcerative colitis

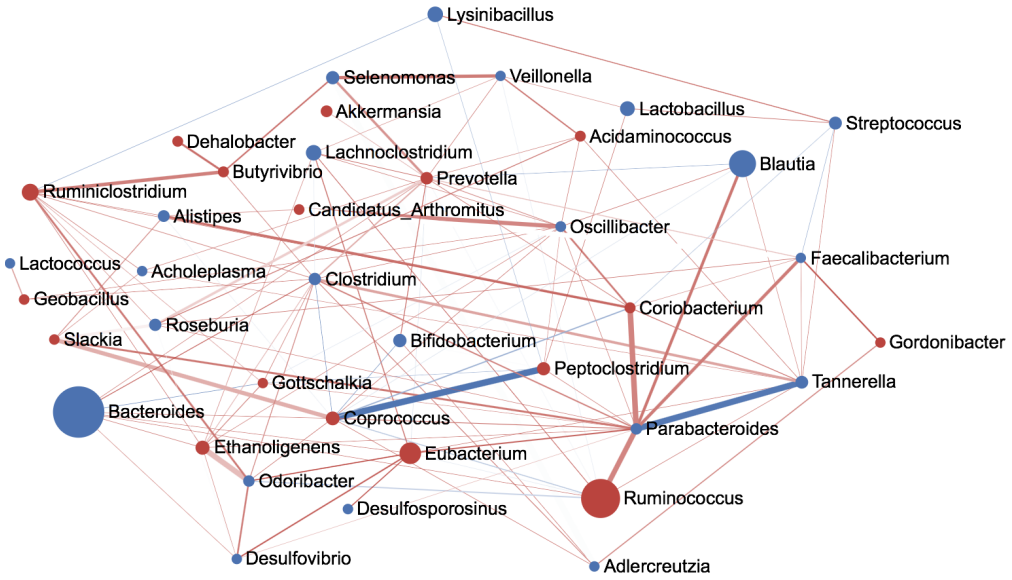


Figure C.5: **Healthy/UC differential network:** *Bootstrap = Healthy; Subsampling = UC; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in ϕ.*

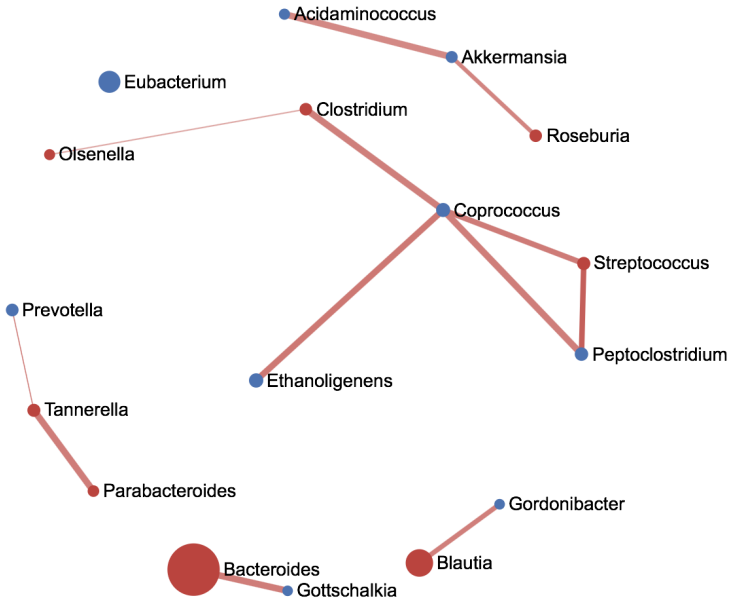


Figure C.6: **UC/Healthy differential network:** *Bootstrap = UC; Subsampling = Healthy; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in cases; Blue nodes = Increase in abundance in cases. The larger the node size, the larger the relative change. Red edges = Decrease in ϕ (increase in association) in cases; Blue edges = Increase in ϕ (decrease in association) in cases; The darker the edge, the greater the change in ϕ.*

C.2 Correlation differential networks

C.2.1 Ileal Crohn's disease

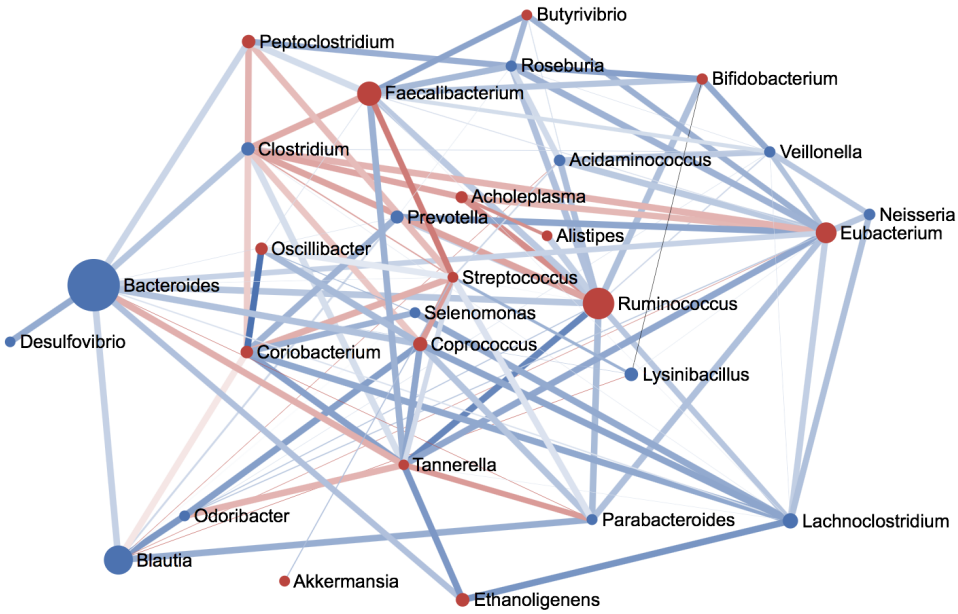


Figure C.7: **Healthy/iCD differential network using Pearson's correlation:** *Bootstrap = Healthy; Subsampling = iCD; 2500 bootstrap and subsampling iterations; Certainty >0. Red nodes = Decrease in abundance in iCD patients; Blue nodes = Increase in abundance in iCD patients. The larger the node size, the larger the relative change. Red edges = Decrease in r in iCD patients; Blue edges = Increase in r in iCD patients; The darker the edge, the greater the change in r.*

Appendix D

Data tables for Inferred Metanetworks

Table D.1: *Table of NCBI taxonomy IDs that were not associated with any KEGG reaction IDs.*

NCBI taxonomy ID	**Name**	**Rank**
2	*Bacteria*	Eubacteria: Superkingdom
131550	*Fibrobacteres/Acidobacteria*	Bacteria
458031	*Acidobacteria*	Bacteria
57723	*Acidobacteria*	Phylum
204434	*Acidobacteriaceae*	Family
91061	*Bacilli*	Class
31	*Myxococcaceae*	Family
1016	*Capnocytophaga*	Genus
1239	*Firmicutes*	Phylum
188711	*Thermodesulfobacteriaceae*	Family
13	*Dictyoglomus*	Genus
392733	*Terriglobus*	Genus
186826	*Lactobacillales*	Order
191394	*Deferribacteraceae*	Family
128827	*Erysipelotrichaceae*	Family
126	*Planctomycetaceae*	Family
118	*Planctomyces*	Genus

Appendix E

Data tables for E. coli model curation and validation

Table E.1: *Of the 20 metabolites within the 'Other' category, 17 could not be transferred to the target model and are shown here. The metabolite ID within the template model (Template ID) is shown in column 1, and the name of the metabolite (Metabolite name) shown in column 2. The MetaNetX ID (MNX ID) that would be used to map from the template model ID to the target model ID is shown in column 3, with the target model ID (Target ID) shown alongside in column 4. The reason why transfer from the template model to the target model is inhibited (Reason for transfer failure) is concluded in column 5.*

Template ID	**Metabolite name**	**MNX ID**	**Target ID**	**Reason for transfer failure**
bmocogdp_c	Bis-molybdopterin guanine dinucleotide	✗	✗	No MNX ID
ca2_c	Calcium	MNXM128	C00076	Not in target model
cobalt2_c	Co^{2+}	MNXM711	C00175	Not in target model
fe2_c	Fe^{2+}	MNXM111	C14818	Unable to map to target model
2fe2s_c	[2Fe-2S] Iron sulfur cluster	✗	✗	No MNX ID
4fe4s_c	[4Fe-4S] Iron sulfur cluster	✗	✗	No MNX ID
h2o_c	H_2O	MNXM2	C00001 D00001	Maps one:many
mg2_c	Magnesium	MNXM653	C00305	Not in target model
mobd_c	Molybdate	MNXM1026	C06232	Not in target model
murein5px4p_p	Two disaccharide linked murein units, pentapeptide crosslinked tetrapeptide (A2pm - D-ala) (middle of chain)	MNXM88338	✗	No KEGG ID
ni2_c	nickel	MNXM3673	C00291 C19609	Not in target model
2ohph_c	2-Octaprenyl-6-hydroxyphenol	MNXM1635	C05811	Unable to map to target model
pe160_c	Phosphatidylethanolamine (dihexadecanoyl, n-C16:0) (cytoplasmic)	MNXM32178	✗	No KEGG ID
pe160_p	Phosphatidylethanolamine (dihexadecanoyl, n-C16:0) (periplasmic)	MNXM32178	✗	No KEGG ID
pe161_c	Phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1) (cytoplasmic)	MNXM1945	✗	No KEGG ID
pe161_p	Phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1) (periplasmic)	MNXM1945	✗	No KEGG ID
sheme_c	Siroheme	MNXM82173	C00748	Unable to map to target model

Table E.2: *Composition of the glucose minimal medium, showing whether the metabolites and import reactions are present in the target E. coli model. Where a given metabolite was present in the model but did not have an associated import reaction, a new import reaction was added (shown by ✓in the 5th column). Where a given metabolite was not present in the model, an associated import reaction was not added (shown by ✗in the 5th column). Where a given metabolite was present in the model and already had an associated import reaction, no action was taken (shown by 'N/A' in the 5th column).*

Metabolite name	**KEGG ID**	**Contained in target model**	**Import reaction in target model**	**Import reaction added**
D-glucose	C00031/C00267/C00293	✓	✓	N/A
Ca^{2+}	C000076	✗	✗	✗
Fe^{3+}	C14819	✓	✗	✓
H^{+}	C00080	✓	✓	N/A
H_2O	C00001/C01328	✓	✓	N/A
K^{+}	C00238	✗	✗	✗
Mg	C00305	✗	✗	✗
Na^{+}	C01330	✗	✗	✗
NH_3	C01342/C00014	✓	✓	N/A
Phosphate	C00009	✓	✓	N/A
Sulfate	C00059	✓	✓	N/A
O_2	C00007	✓	✓	N/A
Cl^{-}	C00698/C01327	✗	✗	✗
Cu^{2+}	C00070	✗	✗	✗
Co^{2+}	C00175	✗	✗	✗
Mn^{2+}	C00034	✗	✗	✗
Zn	C00038	✗	✗	✗
Fe^{2+}	C14818	✓	✗	✓