School of Chemical Engineering
and Advanced Materials

Human-Microbiota Interactions in Health and Disease: Bioinformatics Analyses of Gut Microbiome Datasets

Matthew Collison

Submitted for the degree of Doctor of Engineering in the School of Chemical Engineering and Advanced Materials, Newcastle University

24.11.2016

Acknowledgements

I would like to thank Professor Robert P. Hirt from Newcastle University for his guidance and supervision throughout my research and in the production of this thesis. Robert's passion for science is contageous and his support throughout has been nothing short of brilliant. I would also like to thank Professor Anil Wipat from the School of Computing Newcastle University for his supervision and team leadership. Being part of the integrative bioinformatics research group shaped my values as a researcher. Also thanks to Dr. James R. Brown and Dr. Philippe Sanseau GlaxoSmithKline for their guidance on the industrial applications of the work. I would like to thank the staff at the BBTC Chemical Engineering for coordinating the Engineering Doctorate course.

I would next like to thank my peers at Newcastle University. Throughout the Engineering Doctorate I have had the pleasure of colleagues who made fantastic friends and exceptional peers. The Systems Modelling and Analysis (SysMA) group in the School of Chemical Engineering and Advanced Materials provided a stimulating research environment during the early years of my research and the School of Computing had a major influence on my academic development, particularly Michael Bell, Jennifer Warrender and Matt Forshaw.

Finally I would like to thank Penny, my friends, family and my late grandma for their unconditional support.

Declaration

I delare that this thesis, apart from the help recognised, is my own work. No part of this thesis has previously been submitted for a degree or any other qualification at Newcastle University or any other institute.

Matthew Collison

November 2016

Publications Arising from this Thesis

Matthew Collison, Robert P. Hirt, Anil Wipat, Sirintra Nakjang, Phillipe Sanseau and James R. Brown (2012) 'Data Mining the Human Gut Microbiome for therapeutic Targets', *Briefings in Bioinformatics*. PMID: 22445903.

Michael J. Bell, **Matthew Collison** and Phillip Lord (2013), 'Can inferred provenance and its visualisation be used to detect erroneous annotation? A case study using UniProtKB', *PLoS One*. PMID: 24143170.

Matthew Collison, Anil Wipat, Robert P. Hirt *(in preparation)* 'The Role of Proteases in Protein and Peptide Associated Glycan Metabolism in the Microbiota'.

Conference proceedings

M Collison, JR Brown, P Sanseau, A Wipat, RP Hirt 'A Data Integartion Platform for Studying the Microbiota: A Network Approach' Poster presentation - International Human Microbiome Conference, Paris 2012.

M Collison, JR Brown, P Sanseau, A Wipat, RP Hirt 'Human-Microbiota Interactions in Health and Disease: Bioinformatics Analyses of Microbiome Datasets' - Oral and poster presentation, Biopharmaceutical process development conference Newcastle 2013.

M Collison, JR Brown, P Sanseau, A Wipat, RP Hirt 'Computational approaches to drug targeting in the gut microbiota' - Oral presentation, Biopharmaceutical process development conference Newcastle 2012.

M Collison, JR Brown, P Sanseau, A Wipat, RP Hirt 'The Role of the Gut Microbiome in Health and Disease' - Poster presentation, Biopharmaceutical process development conference Newcastle 2011.

Abstract

The human gut harbours a vast diversity of microbial cells, collectively known as the gut microbiota, that are crucial for human health and dysfunctional in many of the most prevalent chronic diseases. Until recently culture dependent methods limited our ability to study the microbiota in depth including the collective genomes of the microbiota, the microbiome. Advances in culture independent metagenomic sequencing technologies have since provided new insights into the microbiome and lead to a rapid expansion of data rich resources for microbiome research. These high throughput sequencing methods and large datasets provide new opportunities for research with an emphasis on bioinformatics analyses and a novel field for drug discovery through data mining.

In this thesis I explore a range of metagenomics analyses to extract insights from metagenomics data and inform drug discovery in the microbiota. Firstly I survey the existing technologies and data sources available for data mining therapeutic targets. Then I analyse 16S metagenomics data combined with metabolite data from mice to investigate the treatment model of a proposed antibiotic treatment targetting the microbiota. Then I investigate the occurence frequency and diversity of proteases in metagenomics data in order to inform understanding of host-microbiota-diet interactions through protein and peptide associated glycan degradation by the gut microbiota. Finally I develop a system to facilitate the process of integrating metagenomics data for gene annotations.

One of the main challenges in leveraging the scale of data availability in microbiome research is managing the data resources from microbiome studies. Through a series of analytical studies I used metagenomics data to identify community trends, to demonstrate therapeutic interventions and to do a wide scale screen for proteases that are central to human-microbiota interactions. These studies articulated the requirement for a computational framework to integrate and access metagenomics data in a reproducible way using a scalable data store. The thesis concludes explaining how data integration in microbiome research is needed to provide the insights into metagenomics data that are required for drug discovery.

Contents

List of Figures

List of Tables

Acronyms

ALT alanine transaminase

ANGPTL4 Angiopoietin-like 4

ANOVA analysis of variance

AST aspartate transaminase

BT Bacteroides thetaiotaomicron

bHBA beta-hydroxybutyrate

cazymes carbohydrate targetting enzymes

CBM carbohydrate binding motif

CK creatine kinase

DIO diet induced obesity

GIP gastric inhibitroy polypeptide

GIT gastro-intestinal tract

GLP-1 glucagon-like peptide-1

GSK GlaxoSmithKline PLC

GWAS Genome Wide Association Studies

HDL-C high-density lipoproteins carrying cholesterol

HPC High Performance Compute

IBD inflammatory bowel disease

IBS irritable bowel syndrome

HMP Human Microbiome Project

InsR insulin receptor

JVM Java Virtual Machine

JRE Java Runtime Environment

LGT lateral gene transfer

LPS Lipopolysaccharides

MetaHIT Metagenomics of the Human Intestinal Tract

NIH National Institute of Health

NEFA non-esterified fatty acids

OFS Oligo-fructo saccharide

ORF Open Reading Frame

OTU operational taxonomic unit

PCoA principal coordinate analysis

PPr pancreatic protein

PUL polysaccharide utilisation locus

PYY peptide YY

RDF Resource Description Framework

T2DM Type II Diabetes Mellitus

tBil total bilirubin

XML eXtensible Markup Language

16S 16S small subunit ribosomal RNA

Chapter 1

Introduction

Studies of the human genome have lead to important advances in biomedical sciences and drug discovery (Lander, 2011; Lander *et al.*, 2001; Venter *et al.*, 2001). However, it is becoming increasingly apparent that determinants of our health are not solely controlled by our own genomes. Rather, many disease pathologies involve the interplay between the human body, the external environment and the complex communities of microorganisms residing on the mucosal surfaces of our respiratory tract (Kiley, 2011), urogenital tract (Danielsson *et al.*, 2011), gastro-intestinal tract (GIT) (Sekirov *et al.*, 2010) and skin (Grice & Segre, 2011).

This complement of microbial cells co-inhabiting any given individual, their ***microbiota***, exceeds at least ten-fold the number of human origin cells that make up that individual (Ley *et al.*, 2006a; Savage, 1977). Furthermore, the gene collective of this residing microbial community, their ***microbiome***, exceeds by at least 150 times the complement of unique genes present in their own human nuclear genome (Qin *et al.*, 2010). While knowledge of the health implications of many isolated bacterial, viral and eukaryotic pathogens is well-established, the roles of these complex non-pathogenic microbial communities are a relatively recent development in our understanding of health and disease pathologies.

Major public funding initiatives such as the American National Institute of Health (NIH) Human Microbiome Project (HMP), initiated in 2007 (Nelson *et al.*, 2010), and the European Metagenomics of the Human Intestinal Tract (MetaHIT) project, initiated in 2008 (Ehrlich & Consortium, 2011), have invested heavily in the characterisation of microbiomes from thousands of individuals of different ages, geographical, dietary and disease backgrounds. This is reflected in the growing number of publications using the terms ‘microbiota’ or ‘microbiome’ in the title or abstract, Figure 1.1, since the technique for culture independent metagenomics sequencing was published in 1991 (Schmidt *et al.*, 1991).

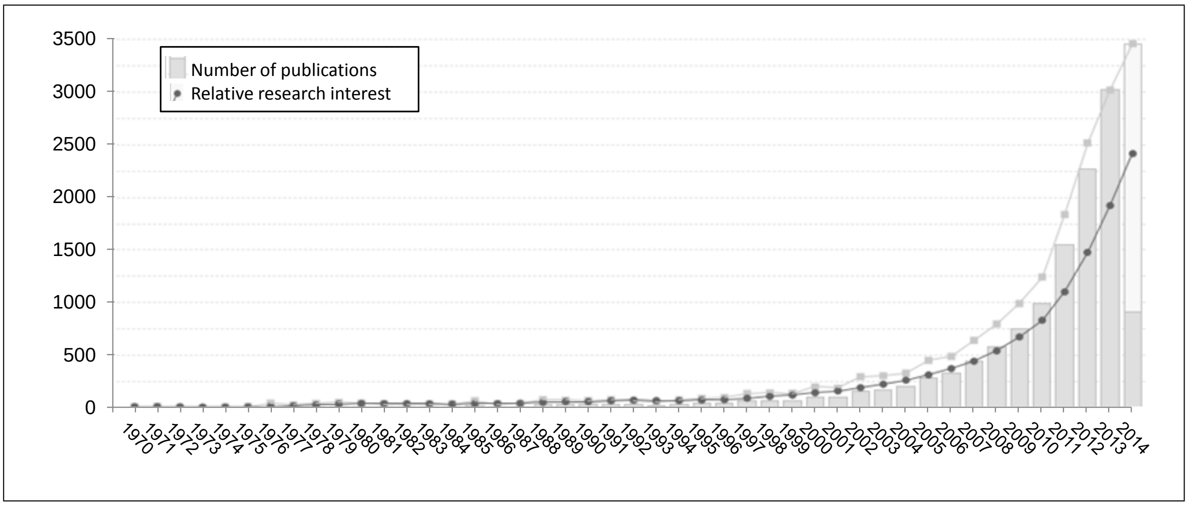


Figure 1.1: Historical emergence of the terms 'microbiome' and 'microbiota' in the title and abstract of publications. Figure generated using GoPubMed. Figure updated from Collison *et al.* (2012)

Together these technologies, investments and new insights have lead to an expansion of the data available for microbiome research. However, there are currently significant challenges in maximising the value of datasets to improve biological understanding and being able to use the data produced to identify new drug discovery opportunities.

1.1 Drug discovery in the microbiota

Historically chronic and infectious diseases have been distinct and non-overlapping areas of drug discovery and research development (Collison *et al.*, 2012). The traditional translational development of research in chronic disease has an established target space known as the druggable genome (Hopkins & Groom, 2002). That is the accessible targets for small molecule drugs in the human cells. Over recent decades the solution space for molecular targets in the druggable genome has started to become saturated and the number of new molecular entities approved by the FDA has been declining (Federsel, 2010; Research, 2014). Therefore the pharmaceutical industry is requiring more novelty in drug discovery for chronic diseases. At the same time in-depth diagnostics are offering added value in the pharmaceutical industry by predicting how a patient will respond to a treatment based on their habits and physiological measurements increasing the opportunity for secondary treatments.

Rapidly emerging science in human-microbe interactions is opening the possibility for a new therapeutic paradigm where understanding of the microbiota contribution to the progression and pathology of chronic diseases is opening treatment options that include anti-bacterial or anti-viral therapeutics, potentially extending the druggable solution space, Figure 1.2, and the treatment of side effects. This is a paradigm that has also been proposed to work in both directions; viral, bacterial and parasitic infections are treated via human host gene targets (Brown *et al.*, 2011). The microbiota is part of a complex pathology in many diseases and can affect many factors of health and disease, including the efficacy of other treatments. Therefore the microbiota represents not only a new target space for disease treatment but also a likely target for coadministered drugs to alleviate side effects for treatments and an important feature of personalised medicine (Angelakis *et al.*, 2014).

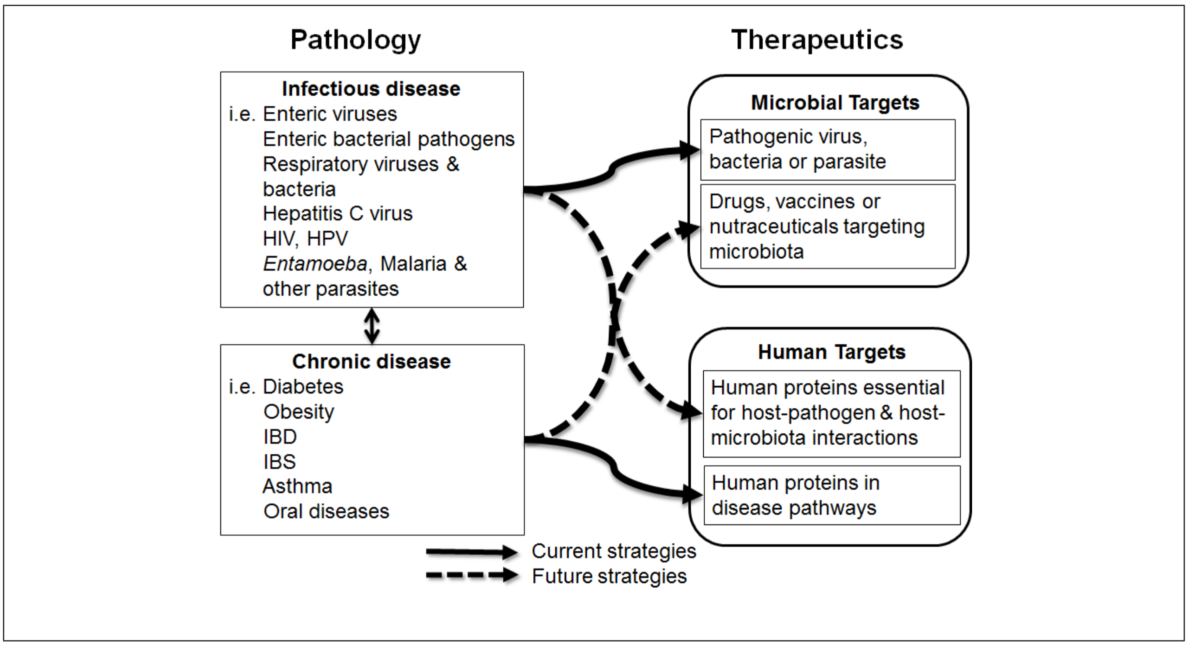


Figure 1.2: Shows the traditional segmentation of research and drug discovery in chronic diseases and infectious diseases and the potential cross over offering opportunities for novel drug discovery. Figure adapted from Collison *et al.* (2012)

1.2 Metagenomics data analyses

Metagenomics developments have undoubtedly provided new insights into the microbiota but also changed the challenges faced by researchers in this field. Particularly the scale and complexity of sequence data from microbiome metagenomics studies puts more emphasis on the bioinformatics analyses (Prakash & Taylor, 2012). A number of metagenomics annotation pipelines have been developed, mimicking genomic annotation pipelines, that complete the assembly and annotation of sequence reads typically allowing species and functions to be identified from existing databases (Meyer *et al.*, 2008). However, due to the high proportion of uncharacterised genes and the variability in metagenomics samples there are major challenges in extracting low level insights from metagenomics data for translational studies in drug discovery.

1.3 Aims and objectives

The link between human genotype and phenotype is a complex relationship. Genomics, transcriptomics and proteomics can provide a detailed picture of internal human biology and metagenomics can give a detailed snapshot of the microbiome. However, host-microbiota interactions that affect the host phenotype are often not clear from standard annotation pipelines and isolated datasets (Viney, 2014). To gain insights into these critical host-microbiota interactions additional analyses are required to disect the data rich metagenomics resources.

1.3.1 Aim

> Explore the biology of host-microbiota interactions through secondary bioinformatics analyses to improve opportunities for new therapeutic strategies involving the microbiota.

1.3.2 Objectives

1. Evaluate the opportunities for drug discovery from existing microbiome analysis pipelines
2. Analyse the role of host-microbiome interactions in health, disease and pharmaceutical interventions
3. Provide computational tools to facilitate microbiome analyses and improve access to microbiome data for drug discovery

1.4 Thesis structure

- Chapter two, the background chapter, investigates computational tools and pipelines used in metagenomics analyses to evaluate existing approaches to identify drug targets in the microbiome using sample datasets from existing studies.
- Chapter three documents the bioinformatics analyses of an experimental dataset provided by GlaxoSmithKline PLC (GSK). A mouse model was used to test the efficacy of a micorbiota targeted antibiotic intervention for type 2 diabetes. Analyses cover 16S metagenomics and metabolite markers for physiological systems and the data show a significant improvements

in the phenotype of the high dose Gram negative antibiotic treatment group and construct a disease model involving the microbiota and a complementry treatment model.

- Chapter four is an exploratory study into the role of microbial proteases in protein and peptide associated glycan degradation in the gut. In this chapter the protease complement of the microbiome in healthy individuals is characterised and a group of novel genes of interest are identified that contain putative CBM-protease pairs thought to be important in glycan metabolism.

- Chapter five documents the development of a computational framework to enable effective data integration and promote reproducible secondary analyses across integrative bioinformatics at the scale of metagenomics studies.

- Chapter six concludes the thesis by evaluating the existing understanding of host-microbiota interactions, the tools available to study host-microbiota interactions and the scope for translational studies that will develop new therapeutic strategies involving the microbiota.

Chapter 2

Background

2.1 The gut microbiota in health and disease

With over 200m^2 of mucosal surface area and a nutrient rich environment, the GIT hosts the majority of the human microbiota (Ley *et al.*, 2006a; Whitman *et al.*, 1998), Figure 2.1. Absorption in the distal gut where the microbiota reside results in approximately 10% of the metabolites in the host systemic blood flow being products of bacterial metabolism (Wikoff *et al.*, 2009). Changes in gut microbiota can therefore be linked with numerous GIT diseases and other systemic diseases (Sekirov *et al.*, 2010). The broad implications of GIT microbiota on human physiology combined with sample accessibility, a large proportion of faecal material being microbial biomass (Sekirov *et al.*, 2010; O'Hara & Shanahan, 2006), has facilitated the exploration of the human GIT microbial ecosystem in human health and disease.

In healthy individuals, changes in the GIT microbiota have been associated with host genetics (Spor *et al.*, 2011), aging (Biagi *et al.*, 2010) and dietary patterns (Wu *et al.*, 2011; Turnbaugh *et al.*, 2009). A cohort study involving 39 individuals from different cultures, geographical locations, races, as well as disease patients, found that the GIT microbial communities of these individuals could be segregated into three statistically robust clusters (Arumugam *et al.*, 2011). These clusters, known as enterotypes, were also found to be consistent across other larger cohort studies, including 85 European (Qin *et al.*, 2010) and 154 American (Turnbaugh *et al.*, 2009) individuals. Enterotypes were identified using from latent variables of a Principal Coordinate analysis (PCoA) based on biodiversity profile. The weighting of individual variables shows the clusters are mainly differentiated by the relative abundances of three bacterial genera, Bacteroides, Prevotella (both of the phylum Bacteroidetes) and Ruminococcus (of the phylum Firmicute) (Arumugam *et al.*, 2011). The stratified

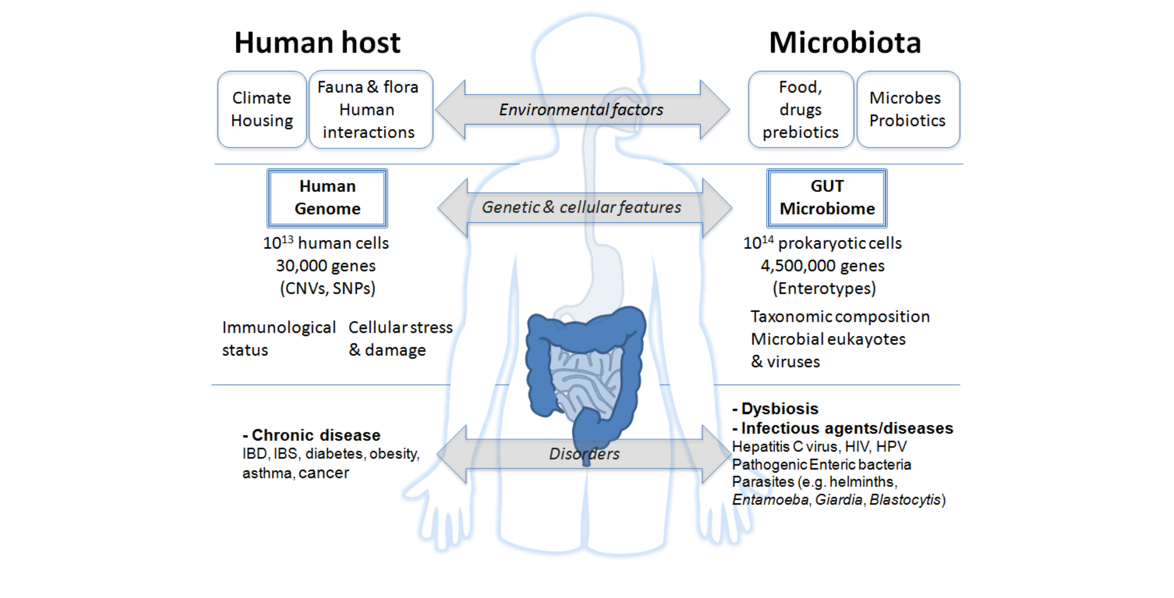


Figure 2.1: The image represents the microbiota on the human body with the depth of colour (transparency percentage) approximately representing density of the microbiome in the corresponding area. The information represents a comparison between the microbiome and the human genome, including environmental factors that affect the respective genetic collection at the top, the size and complexity of the genetic collection in the middle and the implications of each on disease at the bottom. Figure adapted from Collison *et al.* (2012)

nature of the enterotype data indicates that the microbial ecology in each individual's gut establishes a stable and structured biodiversity that is independent of nationality, continent, sex, age or body mass index.

Other studies suggest that long-term diet is important in shaping GIT microbial communities. Comparisons between children from Europe on a typical Western diet, high in animal protein and fat, with children from Burkina Faso in Africa, on a low animal protein and high carbohydrate diet, found the Bacteroides enterotype higher in Europeans while the Prevotella enterotype predominated in African children (De Filippo *et al.*, 2010). Another study also found that animal fat and high protein versus carbohydrate rich diets correlated with the Bacteroides and Prevotella enterotypes, respectively (Wu *et al.*, 2011). In the same study, controlled feeding of 10 subjects with high fat, low fiber versus low fat, high fiber diets produced detectable changes in their microbiome within 24 hours. However, overall the enterotypes of individuals remained stable for the duration of the 10-day study, suggesting that long-term rather than transient dietary trends determine the ecological structure of the gut microbial communities.

Other factors such as environmental exposure also influence biodiversity and the gene content of the GIT microbiota. Babies delivered naturally have microbial communities most similar to that of the mother's vagina while the microbiota from neonates born by Caesarean section closer resembles the mother's skin bacteria (Dominguez-Bello *et al.*, 2010). A fascinating case of lateral gene transfer (LGT) from marine bacteria genomes to the human microbiome involved genes encoding carbohydrate-active enzymes acting on marine algal polysaccharides being frequently detected in the microbiome of Japanese individuals who have such seaweeds in their diets (Hehemann *et al.*, 2010). Further understanding of the microbiota variation across human populations, as well as the environmental and genetic factors shaping it, will be important in designing future therapeutic strategies.

One central aspect of the human-microbiota symbiosis is the dialogue between the microbiota and the immune system (Hooper & Macpherson, 2010; Wells *et al.*, 2011). The microbiota contributes to the development of both the mucosal and systemic immune systems. It is now appreciated that loss of homeostasis in the GIT immune system plays a central role in numerous disease conditions, for example inflammatory bowel disease (IBD) (Sekirov *et al.*, 2010; Abraham & Medzhitov, 2011; Willing *et al.*, 2010). Homeostasis of the mucosal immune system requires the development of both tolerance to the residential microbiota and, at the same time, regulation to avoid overgrowth and invasion of internal tissues. Both the microbiota and the innate and adaptive immune systems contribute to the establishment of an optimal equilibrium whereas disturbances can lead to dysbiosis (Round & Mazmanian, 2009) and disease states through the development of intestinal inflammation (Willing *et al.*, 2010; Abraham & Medzhitov, 2011). This is reflected on the host side by genetic predispositions to IBD, which point to the importance of the immune system and microbial sensing. For example mutations in human genes NOD2, ATG16L1 and those encoding defensins are known predispositions to IBD (Willing *et al.*, 2010; Abraham & Medzhitov, 2011).

The homeostasis of the GIT is also affected by the presence of eukaryotic parasites. Many parasitic protists, for example Entamoeba, Giardia and Blastocystis, and parasitic animals, for example helminth worms, have specialised in colonising the human gut and affect the immune status of the GIT. Parasitic protists may contribute to the disease state (Boorom, 2007). In the case of helminth worms, the human GIT interactions can positively influence immune homeostasis (reflecting host-parasite adaptations) by modulating the immune system towards an optimal anti-inflammatory status (Elliott & Weinstock, 2009).

Dramatic changes over the last 50 years in antibiotic use, hygiene (especially the eradication of helminth worms in most of the developed world) and diet (increase proportions of meat and animal fat) are all potential contributing factors to disturb this balance established over millennia. The hygiene hypothesis suggests that the rapidly changing human lifestyle throughout the developed world has resulted in predisposition to multiple currently prevalent diseases due to imbalances in immune-microbiota coupling (Elliott & Weinstock, 2009). A better understanding of the contribution of the different partners in dysbiosis, and the cross-talk taking place between them, represents important opportunities to develop new strategies to treat chronic diseases, such as IBD and asthma, that have reached epidemic proportions in some populations (Elliott & Weinstock, 2009; Ley *et al.*, 2006b).

While the natural variation of the human microbiota has yet to be fully determined, significant changes in gut microbial communities have been associated with several diseases, including type II diabetes (Burcelin *et al.*, 2011), obesity (Ley *et al.*, 2006b), fatty liver disease (Dumas *et al.*, 2006), irritable bowel syndrome (IBS) (Saulnier *et al.*, 2011) and inflammatory bowel diseases (IBD) (Frank *et al.*, 2011). Complementary to these associations based on trends in populations, obesity studies in animal models, including reciprocal transplants of faecal contents from obese to germ-free mice, confirmed experimentally the role of the gut microbiota in controlling body weight and energy homeostasis (Turnbaugh *et al.*, 2006). Furthermore, in animal obesity models links between microbiota dysbiosis, low grade inflammation and diet induced type II diabetes have also been confirmed experimentally (Cani & Delzenne, 2011). This combination of population based observations and experimental modification of aspects of host-microbiota interactions can be very informative. Patient cohorts receiving pharmaceutical interventions can also help in this approach.

In human patients with infective endocarditis, treatment with vancomycin has shown an increased weight gain which is thought to be caused by dysbiosis due to colonisation of Lactobacillus species that are intrinsically resistant to this antibiotic (Thuny *et al.*, 2010). Clinically, antibiotic usage has also been associated with increased incidence of Crohn's disease and ulcerative colitis in both adults (Shaw *et al.*, 2011) and children (Hviid *et al.*, 2011). The status of gut microbiome is increasingly being studied in a wide variety of further clinical conditions such as post small bowel transplantation (Hartman *et al.*, 2009), colorectal cancers (Marchesi *et al.*, 2011) and malnutrition-immunity imbalances (Kau *et al.*, 2011). Other areas of interest are the potential roles of the GIT microbiota in diseases of distal organs, including asthma (McLoughlin & Mills, 2011) and behavioural disorders (Lyte, 2011). For these diseases and others, the GIT microbiota is an intriguing therapeutic target because of its potential to dramatically expand the human druggable genome, presently estimated to be around 20% of the human proteome sensu stricto (Hopkins & Groom, 2002; Plewczynski & Rychlewski, 2009). Here, a key contribution can be made through computational approaches to link specific changes in microbial communities with physiological changes in the host that reflect disease phenotypes.

2.1.1 Drugs and the microbiome

Targeted therapeutics for manipulating the microbiota are still nascent and rudimentary. Prebiotics and probiotics are the most commonly marketed generic supplements for GIT disorders (Preidis & Versalovic, 2009). However, our understanding of their mechanisms of symptom relief is limited. Probiotics are usually supplements of single bacterial strains which integrate into the broader microbiota with limited global GIT effect (Shanahan, 2010; Quigley, 2011), unless the microbiota is temporarily significantly depleted such as after antibiotic treatments (Hickson *et al.*, 2007). Prebiotics, nutrients aimed at stimulating the growth of specific microbial species, have shown greater potential for manipulating the environmental pressures which shape the microbial ecology, especially in the developmental stages of life (Arslanoglu *et al.*, 2007), although the effects of these supplements are short term and can be overshadowed by the overall diet (Preidis & Versalovic, 2009). It is widely accepted that a full complement of biodiversity is important for a healthy microbiota and single species supplements or substitutions have little effect on the long-term host microbiota-phenotype as determined by underlying biology (Sekirov *et al.*, 2010; Arumugam *et al.*, 2011). Responsive probiotics are the closest the field has come to targeted microbial therapies (Bailey *et al.*, 2011). By introducing iron responsive probiotics, the microbiota can be prepared for trauma and internal bleeding, such as prior to surgery.

Other more extreme therapies are intentional infections by parasitic helminth worm (Weinstock & Elliott, 2009) and faecal transplants (Landy *et al.*, 2011) which have been tested as alternatives to invasive surgery in IBD patients. In preliminary clinical trials, these approaches seem to somewhat regulate the host gut inflammatory response (Landy *et al.*, 2011), although more thorough controls are desirable and the duration of relief is unclear.

Traditionally, antibiotics have been considered the most common course of action against infectious disease, microbial disorders and inflammation. However, growing evidence for microbial contribution to health and advanced understanding of complex microbial diseases has resulted in a re-evaluation of some antibiotic (Dethlefsen *et al.*, 2008; Blaser, 2011) and immunosuppressant treatments (Proal *et al.*, 2011). Antibiotics have been considered as a poor choice for GIT microbiota modification because of tolerance issues associated with long term dosing and the lack of bacterial species specificity. However, antibiotics have the potential to positively modulate chronic disease conditions, such as diabetes and obesity, at least in rodent models (Kootte *et al.*, 2011). Desirable pharmacological properties for a potential GIT microbiota modulator would be selective bacterial species activity and high bioavailability in the gut. Intriguingly, these are precisely the type of molecules which are considered to be failed candidates in antibiotic drug development (Payne *et al.*, 2007). The fact that over 80% of the gut microbial species cannot be cultured using conventional laboratory methods restricts the use of high throughput compound screening campaigns to discover anti-microbiota compounds (Eckburg *et al.*, 2005). However, the development of in vitro human gut models (Feria-Gervasio *et al.*, 2011) as well as using bacteria phylum specific antibiotics, for example against Gram-positive Firmicutes, might accelerate the development of narrow spectrum drugs for microbiota modulation.

Another avenue for potential therapies is targeting host genes involved with microbiota cross-talk. Our knowledge of human receptors engaged in the maintenance of the GIT microbial community balance is growing but far from complete (Zaneveld *et al.*, 2008). Toll-like receptors (TLRs) are responsible for cellular responses against bacterial infections, initiation of inflammation, production of antimicrobial peptides, maturation of antigen-presenting cells and activation of cellular repair and survival pathways (Saleh & Trinchieri, 2011). While TLR2 and TLR4 are primary sensors of pathogenic bacteria, they are also important in maintaining bacterial gut flora homeostasis. Disruptions of TLR and NOD (nucleotide-binding oligomerisation domain) pathways have been associated with colorectal cancer (Saleh & Trinchieri, 2011), IBD (Knight *et al.*, 2008) and other intestinal diseases (Round & Mazmanian, 2009). Microorganisms synthesise a wide range of bioactive signalling low molecular weight molecules and metabolites, many of which are similar to human or eukaryotic produced metabolites (Shenderov, 2011). Using computational meta-analysis of data sources such as human gene expression, bacterial metagenomics and metabolomics, molecules responsible for human-host bacterial cross-talk could be discovered and form the basis for future drug design.

Besides being a potential therapeutic target for chronic disease, the vast metabolic potential of the gut microbiota also plays a key role in the metabolism or biotransformation of xenobiotics, including many marketed drugs. Over 30 drugs are known substrates of bacterial enzymes in the GIT (Sousa *et al.*, 2008) which can have considerable impact on drug development. A tragic example is the reported deaths of several patients co-prescribed a new antiviral drug, sorivudine, along with an oral 5-fluorouracil. The deaths were attributed to toxicity of secondary drug metabolites generated by gut flora (Okuda *et al.*, 1998). Contrasting to the previous case, selective inhibition of gut microbiota enzymes can be potentially used to improve drug efficacy and safety. The colon cancer chemotherapeutic CPT-11 has a dose-limiting side-effect of severe diarrhea caused by reactivation of the prodrug by gut bacteria producing the enzyme beta-glucuronidase. In rodent models, Wallace et al. introduced an inhibitor of bacterial beta-glucuronidase which allowed for higher CPT-11 dosing (Wallace *et al.*, 2010). Interestingly, beta-glucuronidase is not essential for bacterial viability, so inhibition of this enzyme blocked the drug metabolism function while minimising disturbance of the gut microbial community. The development of selective modulators of bacterial enzymes or species responsible for drug biotransformation could be an intriguing strategy for improving the clinical efficacy and safety profiles of particular drugs. The efficacies of many widely used drugs, including statins, are likely determined by both microbiota and host genetic factors (Kaddurah-Daouk *et al.*, 2011), which further prompts integration of microbiome, human genetics and metabolomics into personalised medicine initiatives (Nicholson *et al.*, 2011; Wilson, 2009) and in the identification of therapeutic targets in the microbiome.

2.2 Computational approaches to studying the microbiome

The ‘target’ for therapeutic intervention in the microbiome is a broad concept ranging from molecular entities, such as specific genes or proteins, to pathogenic species or community level modulation (Round & Mazmanian, 2009). Furthermore, the complex pathology often associated with microbiota-related diseases means multiple mechanisms can be linked and several different ‘targets’ can be isolated through studies of the same condition.

The international community is tackling human microbiome research through two complementary sequencing programmes. First, a large number of reference genomes from selected microbial taxa found in different human body sites are being cultured (where possible) and sequenced in their entirety (Nelson *et al.*, 2010; Peterson *et al.*, 2009). Second, bacterial culture-independent metagenomic sequencing is being used to investigate natural microbial ecological adaptations and enterotypes that characterize a human anatomical site or condition. Having both a rich collection of reference genomes as well as broad metagenomic survey data from controlled population studies are crucial for determining microbiota and disease associations.

Historically, clinical microbiology has relied upon culture-dependent methods to seek evidence in support of one of Koch’s key postulates (Falkow, 1988) that one pathogen or pathogenic genetic feature results in a single disease disorder. However, a chronic disease might be the result of interactions involving subtle changes across many different microbiota species and communities. Thus, combined reference genome and culture-independent metagenomic analyses are necessary to study complex microbial ecologies at both the molecular and community levels in order to illuminate linkages between microbiome and specific disease phenotypes.

Various commercial next-generation sequencing (NGS) platforms have been used for metagenomics surveys. Several well-developed, publicly available software tools, such as QIIME (Caporaso *et al.*, 2010b) and MG-RAST (Meyer et al., 2008), provide the means for identifying species and genes from metagenomic data as well as quantifying biodiversity both within and between samples. A thorough discussion of the usage and parameterization of primary analysis software tools (for assembly and initial annotation) is provided elsewhere (Hamady & Knight, 2009). Here I will concentrate on secondary computational analysis geared towards the characterization of microbiome phenotypes and comparative metagenomics.

2.2.1 Reference genomes and comparative genomics

The Human Microbiome Project (Nelson et al., 2010) was one of the first major initiatives with a goal of sequencing genomes from the microbiota to expand reference genome resources. With multiple sequencing centres across America early efforts aimed to sequence 1,000 genomes from the microbiota (Nation Institute of Health, n.d.). By the end of the first phase of the Human Microbiome Project in 2012 there were over 3,000 new genomes published from the microbiota and culturing methods were substantially more advanced to enable previously unculturable species to be sequenced (Markowitz et al., 2012). Furthermore, other national microbiome projects had started across the world that were also contributing to the expanding resource of reference genomes (Dubilier et al., 2015). These included the European metaHIT project (Ehrlich & Consortium, 2011), Canadian Microbiome Project (Government of Canada, 2009), Brazilian Microbiome Project (Pylro et al., 2014) and the Japanese Microbiome Consortium. Since 2015 the national project structure has also been complemented with a Unified Micorbiome Initiative (Dubilier et al., 2015) where the Earth microbiome Project (Alivisatos et al., 2015; Jansson et al., n.d.) aims to facilitate collaboration and coordinate efforts to exapnd the reference resources for microbiome analyses.

Determining the taxonomic content and gene function of a metagenomic dataset primarily depends on the availability and annotation of reference genome sequences. Ideally, the reference genome catalogue for the human microbiota would provide an exact functional homologue of every metagenomic read in a dataset. Therefore, given the associated ecological conditions, human anatomical regions, hygiene status and medical history, each mapped metagenomic sequence read could be assigned some physiological context and molecular function based on the functional annotation of the reference gene (Ellrott et al., 2010; O'Sullivan et al., 2009; Lee et al., 2008). However, most genes do not have complete functional annotations from reference genomes and inferring function and taxonomy from sequence similarity is one of the major challenges in this field. Even when a perfect match is found often genes are annotated as hypothetical proteins. Thus, to discover more about these genes indirect bioinformatics methods for predicting function must be employed.

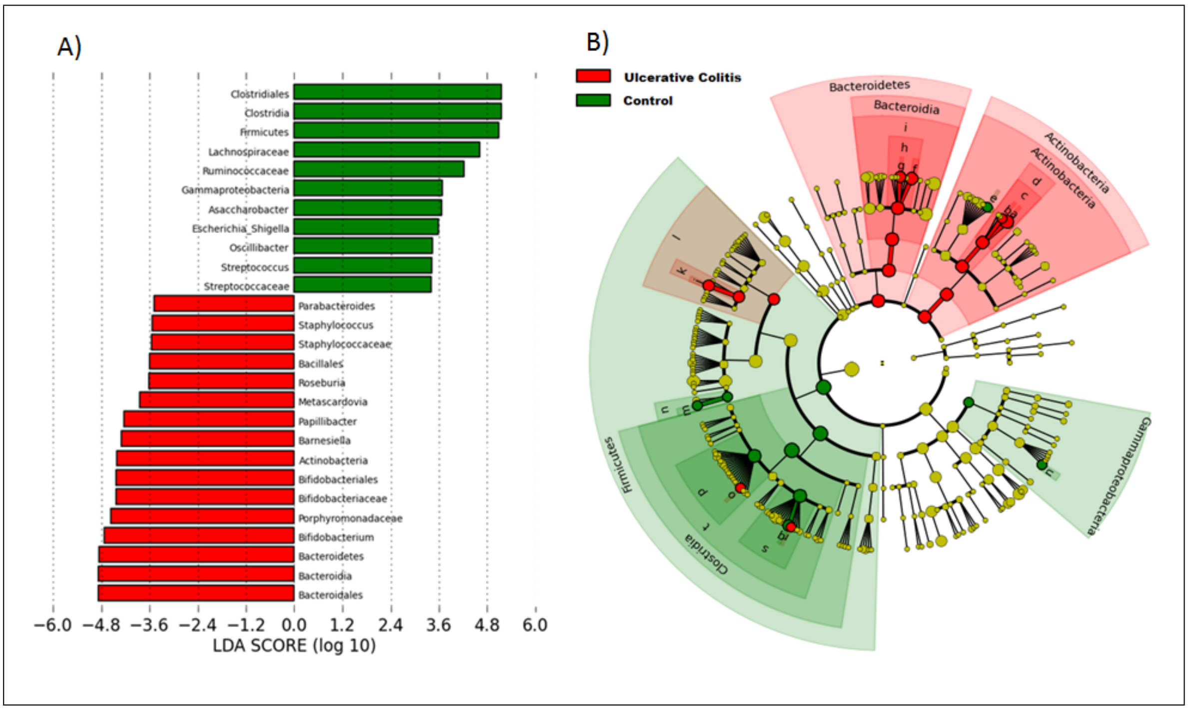


Figure 2.4: Biomarker discovery across a population of healthy and ulcerative colitis model mice, data from Garrett *et al.* (2010). (A) Shows microbial species and their statistical significance as biomarkers for the microbiota phenotype of healthy mice (green) and mice with ulcerative colitis (red). (B) Shows a cladogram indicating the phylogenetic distribution of microbes associated with healthy mice (green) and mice with ulcerative colitis (red). Figures generated in LEfSe (Segata *et al.*, 2011).

pan-genome for microbes (Medini *et al.*, 2005; Frese *et al.*, 2011; Xu *et al.*, 2007). Metagenomics data with its typically low sequence coverage leading to short contigs and numerous singletons is not ideal to investigate individual LGT as the full genomic context of a gene is required to investigate its origin in detail. However, metagenomics data is the only way to get a direct snapshot of the entire metabolic capacity of a microbial community given the exact strains present at the time.

As demonstrated in Figure 2.5, the relative abundance of phyla can vary widely across individual samples. However, the relative abundance of major functional categories remains almost equal, which implies that common gene functions are roughly equally distributed across the microbiota in each individual. This represents a broad adaptation to the environmental niche for the entire microbial population. Despite this high level consistency the functional adaptations to the environment at a finer level are thought to represent the defining characteristics of the microbiota that determine the pathophysiology of dysbiosis and disease. Hence, identifying functional variation at a finer granulation could identify important features corresponding to valuable drug targets, such as proteases and other enzymes degrading xenobiotics, biochemical stressors and toxins (Steck *et al.*, 2011).

Functional diversity is usually visualized in two ways with multiple levels of abstraction; functional

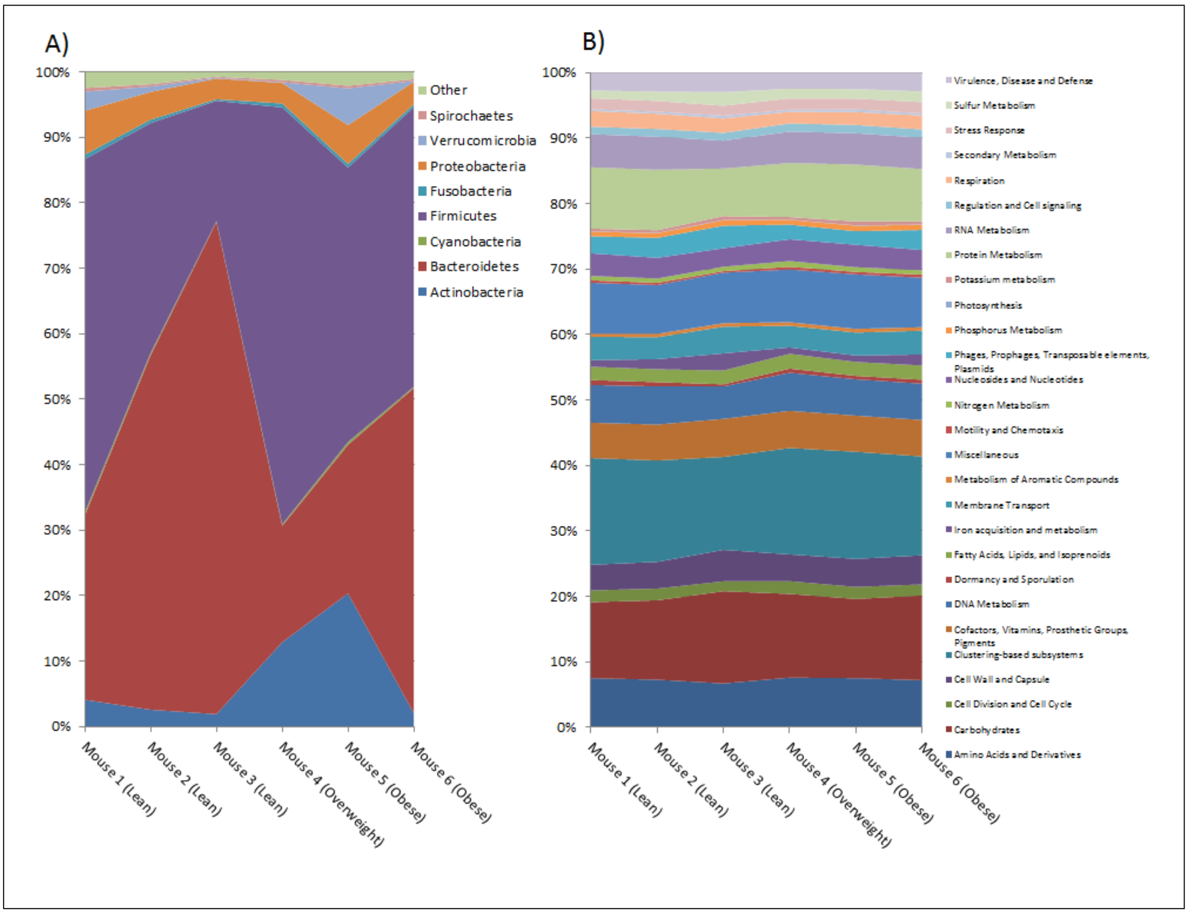


Figure 2.5: Comparison of taxonomic variability against functional variability in six human microbiota samples, a subset of data Turnbaugh *et al.* (2006). (A) Shows phylum-level bacterial variability by percentage abundance. (B) Shows gene function variability across the major functional categories by percentage abundance.

categorization, which reflect many of the same traits as taxonomic visualisation due to the hierarchical structure of the categories, and metabolic pathway reconstruction, Figure 2.6.

In metabolic pathway reconstruction, metagenomic reads are mapped onto known pathways, which can then be compared across the classifications within the data to see which pathways are over represented in which condition. Functional categorization maps each gene to a functional hierarchy and quantifies frequency of genes observed at each level, which again can be compared across classifications in the dataset. These functional analyses have been used successfully to identify gene functions important to a particular habitat (Ellrott *et al.*, 2010). Analysing the distribution of genes and gene functions across different habitats helps generate hypotheses for the level importance of a gene set in a particular environment. This same approach has also been used to identify features of disease vs control microbiomes but mainly seems to identify microbial sub-community biases rather than molecular functions related to disease pathology.

Similar to taxonomic classification, the differential representation of genes and gene functions can

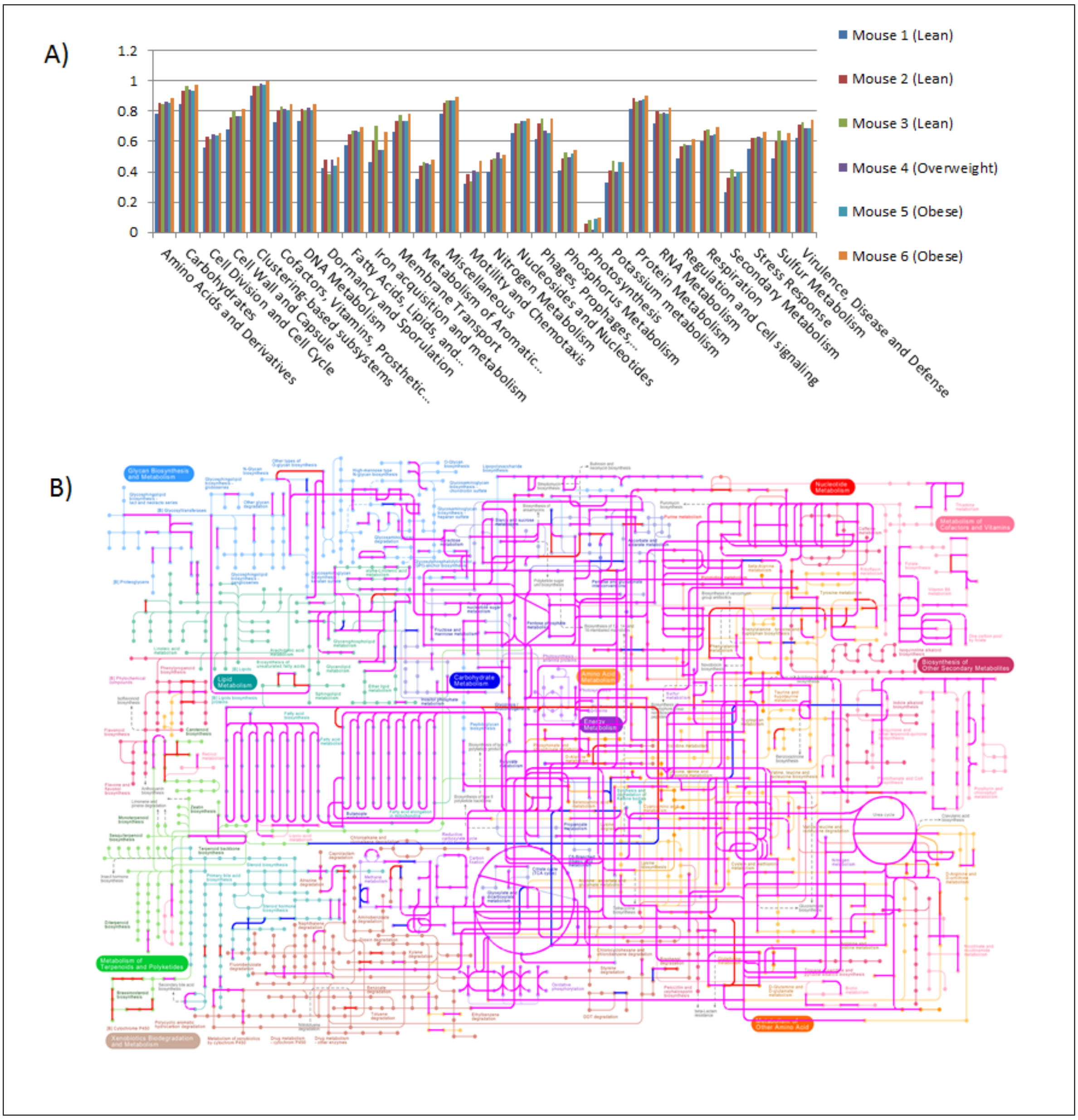


Figure 2.6: Visualization of functional metagenomics across six mice microbiota (three lean, one overweight and two obese), a subset of data from Turnbaugh *et al.* (2009). (A) Shows comparative abundance of the major functional categories (abundance has been normalized to compensate for dataset sizes). (B) Shows comparative metabolic pathways mapped onto the KEGG atlas in the background; purple lines indicate metabolic pathways present in both populations, blue lines indicate metabolic pathways only present in the lean population and red lines indicate metabolic pathways present in only the overweight and obese population. Figure generated in MG-RAST (Meyer *et al.*, 2008).

also be established using the software Metastats (White *et al.*, 2009) and LEfSe (Segata *et al.*, 2011). The scale of gene observations is, however, approximately 3,000 times greater than that of the number of non redundant taxa observations. Therefore, clustering of genes by functional pathways or categorical function can allow for more powerful analyses. MetaPath is customised for statistical identification of over-represented metabolic pathways by using a greedy search algorithm to iterate over the Metastats differential abundance results (Liu & Pop, 2011). Predictive relative metabolic turnover (PRMT) is an extension of this approach that takes into account the metabolic transformations of the pathways as well as sequence abundance. This can be used for approximate system modelling, thereby predicting metabolite usage based on metagenomics datasets (Larsen *et al.*, 2011). Equally, this approach can be reversed to predict microbial functions based on the metabolites present. This level of metagenomics-metabolomics coupling has the potential to be a powerful tool for modelling the host-microbiota metabolite systems.

System modelling is highly dependent on robust assignments of gene function across any particular microbiome, which is not possible at present. In one study in 2011 25% of all predicted Open Reading Frame (ORF)s from human gut metagenomic data were shown to have no significant similarity to any known protein coding sequences from existing complete genomes including hypothetical proteins (Kurokawa *et al.*, 2007). Similarly, in another study over 50% of ORFs identified in human gut metagenomic samples could not be assigned to any known conserved functional region in the Pfam database (Ellrott *et al.*, 2010). In the first large scale metagenomics cohort study approximately 30% of the microbiome predicted genes were unmapped to any known genes and an additional 30% mapped to hypothetical genes (Qin *et al.*, 2010). Furthermore, the majority of mapped functional annotations would have been house keeping genes and a proportion of the genes mapped to known functions would also have been inaccurate and inconsistent (Valencia, 2005; Bell *et al.*, 2013). This highlights the limitations of functional inference when only a section of the datasets have some known molecular assignment; hence, boosting the taxonomic coverage of reference genomes and enriching their functional annotations are primary and fundamental tasks underpinning fruitful metagenomic analysis (Nelson *et al.*, 2010).

In addition to relying on pair-wise sequence similarity to reference genome sequences, profile-based approaches to identify known conserved functional regions in various protein databases, such as Pfam, Interpro and PRIAM, can also be used to assign potential functions to metagenomic reads. Such analyses include highly sensitive profile-profile-based searches (Hildebrand *et al.*, 2009; Kalev & Habeck, 2011) used to investigate divergent protein domains or families identified from metagenomic data (Ellrott *et al.*, 2010) that otherwise reveal no similarity in pair-wise comparison or simple profile analyses.

Notably, protein domain databases are originally derived from the analyses of reference genomes and are continually updated as new sequences from both complete genome and metagenomic sequences become available (Ellrott *et al.*, 2010; Godzik, 2011; Kurokawa *et al.*, 2007). Thus, periodic re-analysis of published metagenomic datasets could illuminate new bacterial species or gene functions in those study conditions as reference genome and protein databases mature over time.

Most metagenomics analysis software packages provide references to multiple databases for annotation. For example, the CoMet web server (Lingner *et al.*, 2011) is an easy-to-use comparative metagenomic package for analysing such datasets. The tool provides a complete pipeline for ORF finding, functional annotation (Pfam, FIGfam and Gene Ontology) and comparative statistical analysis of multiple metagenomic data. However, the majority of these databases are only sparsely populated with environment- specific genes, and the divergent nature of gut microbes provides significant challenges in mapping reads to known-function genes.

Comparative studies of the microbiome typically involve a large number of protein sequences and thus will require a significant amount of computing power. The tremendous amount of sequence data from reference genomes (across all habitats) and metagenomic surveys of the human microbiome necessitates the development of efficient tools for comparative sequence analysis. Several high-throughput computational frameworks employing Grid and Cloud computing have been developed, e.g. QIIME (Caporaso *et al.*, 2010b) and Microbase (Sun *et al.*, 2005). However, the results of these pipelines are often massively reduced incompatible representations of the analyses and require reprocessing for integration. A standard for representing derivative data and a data integration framework capable of handling scalable data would be ideal.

2.3 Moving beyond sequence data

Beyond microbial genomics and metagenomics, other studies have investigated human-host genomics, epigenetics, proteomics, and metabolomics as well as microbiota metabolomics (Kellermayer *et al.*, 2011; Holmes *et al.*, 2008).

Strategies to optimize the integration of the microbiome, host-derived datasets and metabolomics to investigate associations of host-microbe genetic traits with disease conditions will allow a more global understanding of host-microbiota interactions in human health and disease (Holmes *et al.*, 2008). These multi-dimensional human-host-microbial datasets will allow us to investigate the microbiome-encoded metabolic pathways and their relationships with human physiological and environmental variables (Holmes *et al.*, 2008). This implies a predictive nature between the parameters and could allow, in the future, a wider inference from simple sample analysis to extrapolate much broader fields. For example, measurements from multiple areas of the intestine and host tissues coupled with metagenomics sampling and predictive models could be used to develop detailed behavioural analysis of microbial trends and phenotypic behaviour in response to stimuli such as drugs. Ultimately, we would expect this type of analysis to highlight multiple targets for intervention which can be matched up with the compound repositories in silico to search for drug repurposing opportunities (Cockell *et al.*, 2010).

2.3.1 Data integration approaches

For certain diseases, there has been a great deal of activity in the development of tools for the integrative analysis of therapeutics and their targets (Cockell *et al.*, 2010; Gao & Dahlman-Wright, 2011). Integration across the previously described datasets related to microbiome effects on human health will be especially challenging yet critical for the advancement of microbiome targeted therapeutics. Effective strategies for controlling and manipulating our gut microbiota could benefit enormously by taking into account temporal changes in human and microbiome gene expression that could be correlated with other factors such as the host's immune status and genotype as well as environmental conditions. Thus, we can expect to see integrative microbiome models that include information derived from global strategies for measuring gene expression such as RNAseq. These models may, in turn, lead to the identification of biomarkers and expression profiles that are indicative of the microbiota status in terms of disease involvement and responsiveness to therapeutic agents.

Data integration, not a new problem in bioinformatics, is already an active research area in the field of molecular and medical informatics. Conventional approaches such as data warehouses coupled with tools for multi-dimensional statistical analyses still have the potential to provide rich integrative environments that will allow biologists to explore the interdependencies between many factors simultaneously. More recently, graph-based approaches to representing biological data have started to appear and are valuable for the systems level analysis (Smoot *et al.*, 2011; Weile *et al.*, 2011). These graphs can be of different levels of abstraction from functional networks (Lee *et al.*, 2004) to highly annotated and computationally amenable semantic graphs (Ondex) (Weile *et al.*, 2011).

In the future, dynamic systems modelling may play a large part in helping us to understand the complexity of the microbiome. Efforts to integrate data necessary for microbiome analysis may start to reveal deficiencies in upstream processes such as DNA sequence annotation. For example, the clinical status of a patient from where a particular microbiome was sampled might be used as an annotation. This would require not only new sequence annotation formats but also careful blinding of patient data in order to maintain confidentiality. As such, we may find that new standards are required to capture the metadata required for microbiome analysis. In the meantime a first step is mapping metagenomics data to a single representation and providing a computational framework capable of managing this resource.

2.4 Implementation technologies

Throughout the thesis the software development tasks can generally be separated into two categories; statistical scripting and systems development. My local development environment ran programs on Microsoft Windows with up to 8GB RAM. When further compute was needed the software was ported to either a local server machine with a 48 cores and 96GB RAM or a range of Amazon Web Services EC2 instances. Extensive systems administration was also involved in using specific tools to generate analyses and visualisations for the thesis although these are specific to projects and documented individually in each chapter.

MATLAB version R2012B (The MathWorks, n.d.b) was used for all statistical analyses and many of the data visualisation plots. MATLAB is a software package including an integrated development and execution environment that interfaces with the weakly typed MATLAB interpreted programming language. MATLAB was chosen as a familiar solution that was effective for developing single use scripts to leverage the statistics toolbox packages (The MathWorks, n.d.a).

Java (Oracle, n.d.a) was used for systems development. Java is a strongly typed, object oriented, 3rd generation programming language that compiles to javabytes and is executed in the Java Virtual Machine (JVM). The advantage of Java is that it can be executed on any platform where the Java Runtime Environment (JRE) is installed. One of the reasons I chose to use Java over other strongly typed languages is the tool chain. The software lifecycle is well supported with IDEs for source code development, effective build tools that coordinate complex build processes and can be used to form portable execuatable endpoints, and full unit and integration testing suites. I used Netbeans version 8.1 (Oracle, n.d.b), Maven (Apache, n.d.) version 3 and the junit framework version 4. I also used the shade plugin to compile self contained executable jar files which was important for porting programs between the local development environment and remote High Performance Compute (HPC) servers. The Oracle implementation of Java (Oracle, n.d.a) was used throughout and the latest version used for development at the time of writing is version 1.8.151.

Chapter 3

Antibiotic Remodelling of the Gut Microbiota in a Type II Diabetic Mouse Model

3.1 Introduction

T2DM is a major healthcare issue that is increasing at epidemic rates across the world (Smyth & Heron, 2006; Shen *et al.*, 2012; Zhao *et al.*, 2012). The International Diabetes Federation estimated that in 2014 over 387 million people worldwide (8.3% of the global population) were living with diabetes and more than 612 billion USD (11% of global healthcare expenditure) was spent on healthcare for diabetic patients, of which 90% suffered from T2DM. Furthermore, due to the complex and heterogeneous nature of T2DM pathophysiology, an estimated 179 million of the sufferers were undiagnosed (International Diabetes Federation., 2013).

Historically, lifestyle and genetics were thought to be the main factors determining T2DM onset (Scheen, 2003). However, recent developments have shown that the gut microbiota also contributes to the condition (Larsen *et al.*, 2010), opening the possibility for microbiota-targeted therapies to complement established treatments (Cani & Delzenne, 2011).

3.1.1 Type II diabetes mellitus pathophysiology

T2DM is a metabolic disease characterised by a lack of blood glucose regulation which causes chronically high blood glucose levels and over time can cause heart disease, stroke, diabetic retinopathy, kidney failure and circulatory problems (Choices, 2012; Diabetes UK, 2014). This chronic blood glucose elevation is a result of the hyperglycaemic threshold being increased in the direct glucose feedback loop, Figure 3.1. In healthy subjects, insulin is secreted from the pancreas into the blood during periods of high blood glucose concentration and this insulin promotes glucose uptake and utilisation by bodily tissues which reduces blood glucose. In T2DM patients, despite an increase in levels of circulating insulin, blood glucose levels remain high for longer periods. This disease mechanism is known as *insulin resistance*.

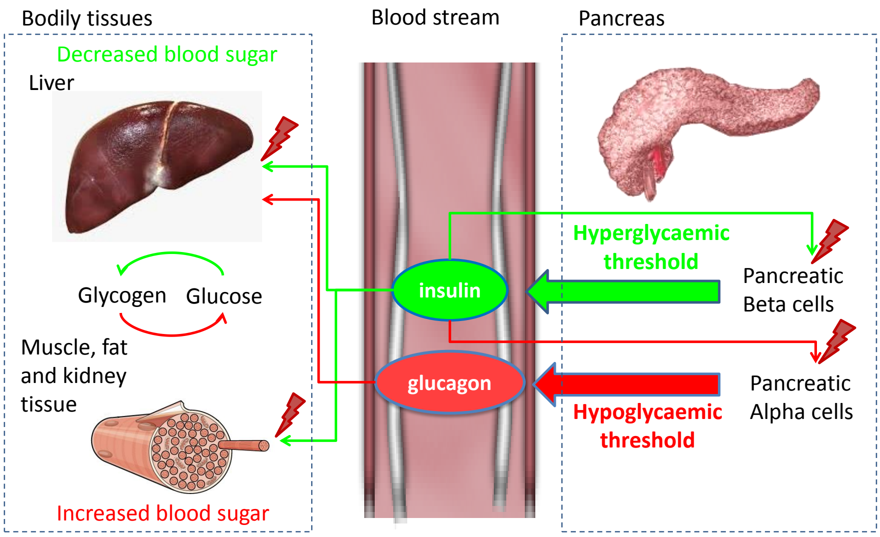


Figure 3.1: **Direct blood glucose feedback loop.** Wide arrows indicate the hormones insulin and glucagon are secreted under hyperglycaemic and hypoglycaemic conditions, respectively, from the pancreas into the blood stream. Each hormone is then transported to target tissues indicated by narrow arrows. Red arrows show target tissues that are activated in low blood glucose conditions and result in an increase in blood glucose, green arrows show target tissues that are activated in high blood glucose conditions and result in a decrease in blood glucose. Red lightning bolts show where the insulin dependent mechanisms are implicated in T2DM.

The main factors that determine an individual's insulin sensitivity and that can contribute to insulin resistance are metabolic status and inflammatory status (Hu, 2011). Each of these factors has genetic and environmental determinants as well as a variable acute status that is the product of a network of endocrine signalling pathways (Lazar, 2005). This complexity means insulin sensitivity is highly variable over time and between individuals and therefore predisposition to insulin resistance and T2DM can be difficult to predict (Stumvoll *et al.*, 2005).

At a population level, concordance studies show that if one monozygotic twin has been diagnosed with T2DM, the other twin has a 76% chance of developing the condition (Medici *et al.*, 1999; Newman *et al.*, 1987); dizygotic twins have a 38% chance (Newman *et al.*, 1987). Genome Wide Association Studies (GWAS) identified 56 loci that vary linked to T2DM (Morris *et al.*, 2012; Cho *et al.*, 2012; Zeggini *et al.*, 2008; Scott *et al.*, 2007). These genes include monogenic genes, that independently increase the risk of T2DM, and networks of polygenic genes that contribute through complex systems. The biological processes repeatedly implicated in the GWAS studies are cell cycle regulation, adipocytokine signalling and CREBBP-related transcription factor activity (Morris *et al.*, 2012). The combination of all of the genes identified are estimated to only accounted for 10% of the heritability of T2DM (Morris *et al.*, 2012). It is clear from these population based studies that familial traits can lead to strong predispositions, but genetics has minimal direct impact on risk and predisposition to T2DM. Lifestyle trends and environmental factors are central to T2DM onset in most cases (Reue & Donkor, 2007).

There are multiple environmental factors known to increase the risk of T2DM, these include unbalanced diets, sedentary lifestyles, stress, chronic immuno-supressant treatments, lack of sleep, alcoholism and motherhood (American Diabetes Association, 2012). Despite the range of potential contributing environmental factors epidemiological comorbidity trends between obesity and T2DM show that diet and a sedentary lifestyle are the main environmental risk factors causing the high prevalence of T2DM in modern society (Wild *et al.*, 2004; Smyth & Heron, 2006; World Health Organization, 2013; Hu, 2011). The epidemiological trends in comorbidity for obesity and T2DM are repeated but variable across the world, 30% comorbidity in China and Japan, 60-80% comorbidity in Europe and 90% comorbidity Pima Indians and Pacific Islanders (Wild *et al.*, 2004; Hu, 2011). Although it is unclear why there are inconsistencies in the comorbidity rates between obesity and T2DM across the world the molecular mechanisms linking obesity and T2DM have been well characterised.

3.1.1.1 Molecular mechanisms of diet-induced obesity and type II diabetes mellitus

The main pathway responsible for obesity driven T2DM is through inflammation caused by increased adiposity (Taylor, 2013). High levels of adiposity cause low grade inflammation and this inflammation increases insulin resistance, Figure 3.2. Molecular mechanisms underlying adiposity driven inflammation revolve around leptin and Angiopoietin-like 4 (ANGPTL4) secretion (Brännmark *et al.*, 2013). The molecular mechanisms underlying inflammation driven insulin resistance are a combination of selective lack of insulin receptor (InsR) expression, intracellular pathway activation shifts and pancreatic β-cell function (Pickup, 2004).

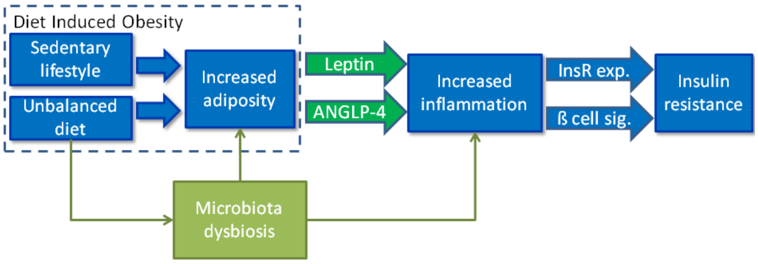


Figure 3.2: **Diet-induced obesity and T2DM disease model.** The boxes and arrows represent the established disease progression linking obesity and T2DM. Blue arrows indicate predictable steps in the progression towards insulin resistance. Green arrows indicate a step that leads to variable outcomes across human populations.

This obesity driven T2DM disease model has been largely developed in mice as genetics and environment can be controlled. The disease model is commonly referred to as diet-induced obesity and T2DM (Surwit *et al.*, 1988b; Winzell & Ahren, 2004). However, in humans there is major variability in the outcomes of this model as shown in the comorbidity trends (Wild *et al.*, 2004; Hu, 2011). This variability is due to the vast variability in environmental factors, particularly in dietary variation and cultural differences, that can contribute to low grade inflammation through mechanisms other than through leptin and ANGPTL4 secretion. One theory is that the microbiota is the central link between diet, environment and inflammation.

3.1.2 Type II diabetes mellitus and the microbiota

Recent studies have begun to provide insights into the role of the microbiota in dietary and inflammatory regulation and mediation. There is also emerging research on how obesity affects the microbiota and how the microbiota can contribute to obesity, however, the implications for T2DM remain unclear.

In humans and mice, obesity and an increase of fat in the diet are associated with increases in *Firmicute* communities and decreases in *Bacteroidetes* communities in the gut microbiota (Turnbaugh *et al.*, 2009; De Filippo *et al.*, 2010). Furthermore, reciprocal transplants of the microbiota from healthy mice and obese mice cause reversal of the obese and healthy phenotypes showing a somewhat causative link (Million *et al.*, 2013; Ridaura *et al.*, 2013). Recent studies have also associated the distinction between *Firmicute* and *Bacteroidetes* dominated communities with increased levels of low grade inflammation and other early complications associated with T2DM (Cani & Delzenne, 2011; Qin *et al.*, 2012; Tilg & Kaser, 2011; Ley *et al.*, 2005). These studies show the microbiota in mice is affected by the diet and the microbiota is capable of contributing to T2DM.

Obesity linked microbiota communities have a physiological impact on multiple disease associated functions such as; energy balance (Cani *et al.*, 2012), gut permeability (Muccioli *et al.*, 2010; Cani *et al.*, 2009), metabolic endotoxemia through lipopolysaccharides (Cani *et al.*, 2007), and metabolic inflammation (Cani *et al.*, 2008; Muccioli *et al.*, 2010; Cani *et al.*, 2009, 2007), Figure 3.3. These Firmicute dominant microbiota-disease associations have been tested in studies with clear null hypotheses, however, their relative contribution to diabetes onset remains unclear. The microbiota dysbiosis categorised in these studies is often based on the *Firmicute* dominated microbiota communities identified in previous studies (Turnbaugh *et al.*, 2009; De Filippo *et al.*, 2010). The relative importance of sub communities in these microbiota is also unclear for taxonomic groups of different levels. Moreover, the communities shared between humans and mice remains unclear beyond the divide in *Firmicute* dominated microbiota and *Bacteroidetes* dominated microbiota.

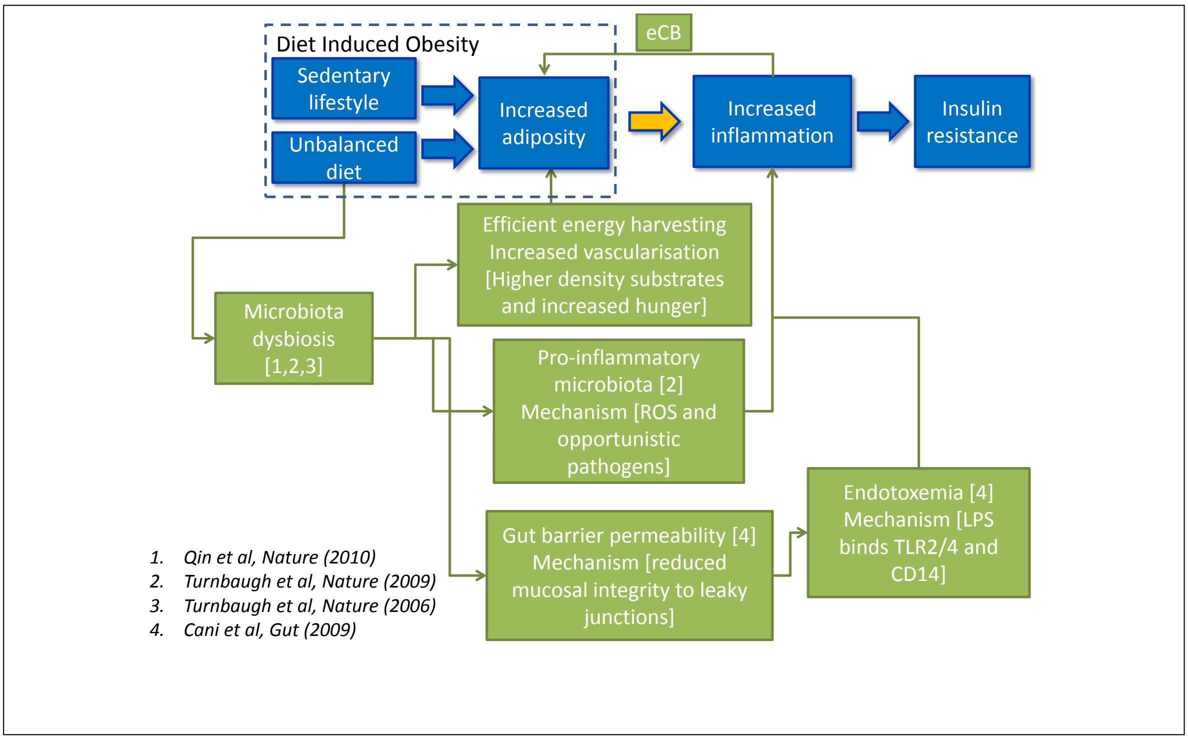


Figure 3.3: **T2DM disease model and the microbiota mechanisms.** The blue boxes represent the classic understanding of disease progression. Blue arrows indicate well characterised steps in the progression towards insulin resistance. The yellow arrow indicates a step that is variable in human populations. Green boxes represent contributions of microbiota dysbiosis to the disease progression. eCB = endocannabinoid system.

This study seeks to investigate a disease model of T2DM including a role for high level microbial communities and therefore a treatment model in which modifying the broad diversity can benefit the phenotype of a mouse model.

3.1.3 Mouse model of obesity induced T2DM

A common model for insulin resistance, and the model used in this study, is the diet induced obesity (DIO) mouse model (Surwit *et al.*, 1988b; Winzell & Ahren, 2004). Compared to human T2DM the disease progression in DIO mice is much more predictable and the physiological conditions can be closer controlled. One of the challenges in the DIO mouse model is that there is no diagnostic threshold for T2DM. To provide an accurate indication of insulin sensitivity, and an indication of T2DM phenotype, the blood glucose and insulin are measured from and the insulin:glucose ratio is calculated. This ratio provides a measure of insulin sensitivity partially accounting for fluctuations in glucose and insulin levels caused by dietary intake, digestion time and molecular half lives of insulin and glucose (Surwit *et al.*, 1988b; Winzell & Ahren, 2004). Another indirect marker for insulin sensitivity is glucagon-like peptide-1 (GLP-1). GLP-1 is a hormone known to directly affect insulin sensitivity by promoting expression of insR receptors (Drucker, 2003). Together these markers give a measure of blood glucose regulation state. A range of other biomarkers and measurements, including body weight, metabolic biomarkers, endocrine signalling molecules and endogenous enzymes are also measured to monitor the physiological state of each study mouse and study the physiological context to generate a disease treatment model.

Weight loss is widely known to change the metabolic status and subcutaneous fat in adipose tissues has a key role in endocrine signalling in T2DM. Weight loss is an important control to test for the effect of any treatment beyond the confounding beneficial effects of weight loss. OFS is a food sweetener and an additive often used to induce weight loss in mice through increased satiety (Macfarlane *et al.*, 2008). OFS is added to a high fat chow diets used to induce weight loss in many mouse model studies (Winzell & Ahren, 2004). Additionally OFS works as a sweetener due to its resistance to digestion in the stomach but this also makes it an effective prebiotic as it is digested by the microbiota in the gut.

The liver is a key organ in the blood glucose feedback loop and often in metabolising antibiotics and other xenobiotic metabolites (Sahin *et al.*, 2006). The markers measured for paranchymal cell integrity in the liver were aspartate transaminase (AST) and alanine transaminase (ALT). If liver function is disrupted these are found in high volumes and the measured enzymatic activity would be increased. Bilirubin is a protein produced by the liver and total bilirubin (tBil) is an indicator of liver function related to hydrophobic molecule transportation. If tBil is high this indicates the liver may not be processing albumin effectively. Albumin is made exclusively by the liver and has a half life of around 2 weeks. Albumin levels can be used as a direct general marker for long term liver functionality and will be dramatically lower if the liver is damaged. Albumin levels are can also be inflated by dehydration.

Blood glucose levels are closely related to blood pressure and osmolarity of the blood which are controlled mainly by the kidneys. Kidney function was monitored by measuring creatinine and creatine kinase (CK). If the kidneys are damaged both of these markers are found in high volumes (Ward, 1988).

Beta cells in the pancreas are the central cells controlling glucose regulation. Clearly pancreas integrity is an important consideration in the measurement of mechanisms affecting onset of T2DM. The markers used for pancreatic function were amylase and lipase. These measurements should be zero in healthy mice as the enzymes should be contained in the pancreas and not leaked into the blood.

Energy substrates and endocrine signalling molecules can be used to investigate metabolic status of the host. Cholesterol, triglycerides and high-density lipoproteins carrying cholesterol (HDL-C) can be increased as a result of dietary digestion and catabolic states, where the host is consuming more energy than is being used. Conversely, when the host consumes less energy than is expended beta-hydroxybutyrate (bHBA) can be produced by vital organs when glucose is critically low reflecting an anabolic metabolic status. Additionally measurements for amylin, leptin, ghrelin, gastric inhibitroy polypeptide (GIP), peptide YY (PYY), pancreatic protein (PPr) and GLP-1 show how the body is responding to satiety and a more long term metabolic status.

3.1.4 Hypothesis

This study was carried out to investigate the effects of a broad spectrum Gram-positive selective antibiotic (antibiotic A) and a broad spectrum Gram-negative selective antibiotic (antibiotic B) on the microbiota and glucose regulatory system in a mouse model of diet-induced obesity and T2DM. This chapter aims to analyse biomarkers of T2DM that may provide information as to the underlying value, if any, of the antibiotic treatments targeting the microbiota. Furthermore to investigate if the microbiota can be included in the T2DM disease model and be manipulated to generate a treatment model to help clarify the opportunities for drug discovery.

The hypothesis is that Gram-positive or Gram-negative antibiotics can improve the glucose regulatory system in a diet-induced obesity mouse model of T2DM by modifying the micorbiota.

3.2 Methods

The experimental design and laboratory work to generate FASTQ format sequence data and metabolite measurements were carried out by collaborators at GSK, under the direction of Dr. James R. Brown. These methods are listed under Section 3.2.1.

Metagenomics data, FASTQ format sequence data from a 454 sequencer, and metabolite measurements were supplied for subsequent data analysis at Newcastle University. These methods are listed under Section 3.2.2. All scripts for reproducing figures and statistical tests, excluding commercially sensitive data, are available on request[1].

3.2.1 Experimental Design

The experimental method involved 90 mature mice (Taconic C57 black 6) (Surwit *et al.*, 1988b) living in shared facilities, separated into nine treatment groups including three dosage sub-groups per antibiotic treatment. For simplicity throughout the report I will refer to the common treatment groups as defined below.

- L - Lean control group
- O - Obese control group
- A_{500} - High dose Gram-positive antibiotic group
- A_{150} - Medium dose Gram-positive antibiotic group
- A_{50} - Low dose Gram-positive antibiotic group
- B_{500} - High dose Gram-negative antibiotic group
- B_{150} - Medium dose Gram-negative antibiotic group
- B_{50} - Low dose Gram-negative antibiotic group
- $D_{10\%}$ - OFS supplement group

Each group followed a 10 week dietary regime, followed by a 14 day treatment period, Figure 3.4. Following this treatment period a selected metabolite profile was measured from blood serum and faecal samples were collected and prepared for 16S metagenomic sequencing as described in Section 3.2.1.2.

[1] https://github.com/mcollison/diabetes-disease-and-treatment-model

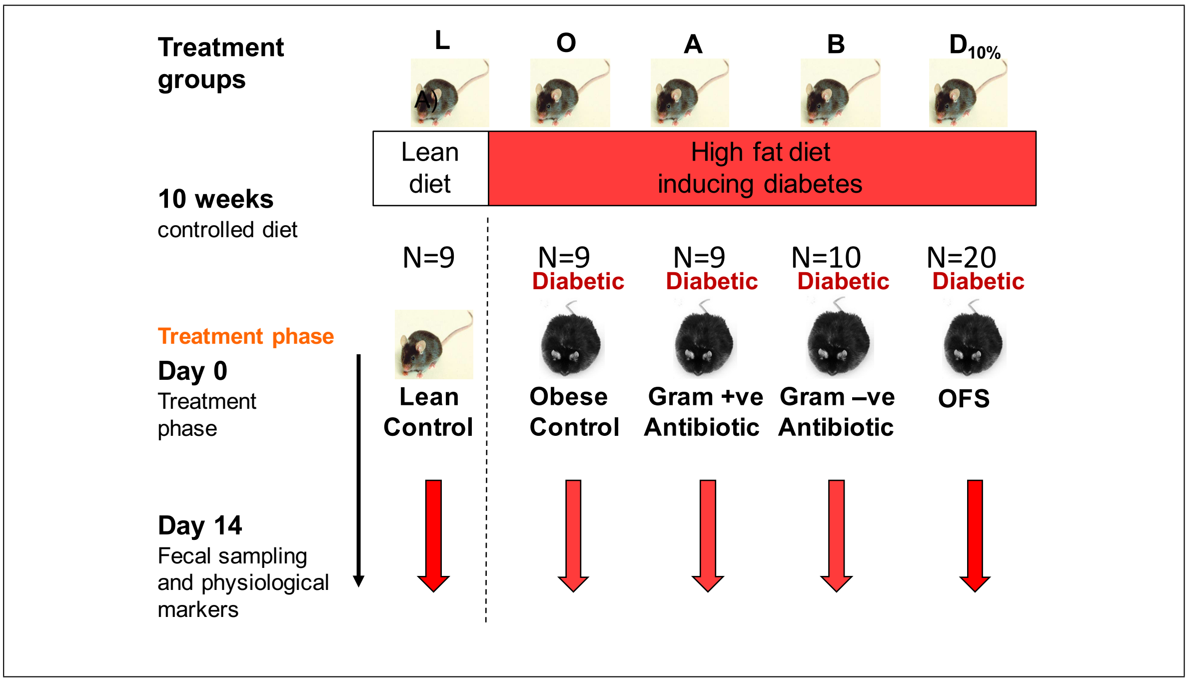


Figure 3.4: **Overview of the experimental design.** There were nine treatment conditions L, O, A_{500}, A_{150}, A_{50}, B_{500}, B_{150}, B_{50} and $D_{10\%}$, as defined in the text above. Each group followed a 10 week dietary regime; L standard chow with 10% calories from fat and O, A, B and D contained high fat content chow with 45% calories from fat. This diet was followed by a 14 day treatment period where the previous diet was continued with additional treatments listed in treatment phase. After the 14 day treatment period faecal samples were collected and physiological markers were measured from blood samples. Images in the figure are not images of mice from the study and are included for illustrative purposes only to demonstrate the expected impact of the stated dietary conditions. OFS=Oligo-fructosaccharide supplement.

The 10 week dietary regime for the lean control group (L) was an ad-libitum standard chow with 10% calorific content from fat. The dietary regime for the diet-induced obesity groups (O, A, B and $D_{10\%}$) was an ad-libitum high fat content chow with 45% calorific content from fat (Surwit *et al.*, 1988a; Winzell & Ahren, 2004). The treatment period for the OFS supplement group ($D_{10\%}$) was an ad-libitum high fat content chow with 45% calorific content from fat mixed with an OFS supplement constituting an overall 10% calorific content. Mice selection for each group was pseudo-randomised according to pre-experimental percentage body fat and all antibiotic treatments were given as food supplements as part of the evening feed each day. Both antibiotics were dosed at 500mgkg^{-1}, 150mgkg^{-1} and 50mgkg^{-1}. As experimental procedures became more complex only selected subjects' specimens were used for certain experimental techniques. This decision was made by personel at GSK to reduce costs and wasn't considered to comprimise the fundamental study design. Table 3.1 summarises the number of subjects from each treatment group that were successfully analysed using each technique. A more detailed per subject breakdown is provided in Appendix I Table 6.1.

Table 3.1: A summary of the number of mice represented in the datasets from each treatment group in each experimental technique.

Condition	number of mice assayed		
	Weight	**metabolite profiling**	**16S metagenomics**
L	10	10	9
O	10	10	9
A_{500}	9	9	9
A_{150}	10	0	0
A_{50}	9	0	0
B_{500}	10	10	7
B_{150}	9	0	0
B_{50}	10	10	6
$D_{10\%}$	10	10	9

3.2.1.1 Metabolite Profiling

The metabolite measurements were obtained through undisclosed assays at GSK. The marker name, physiological function, units and number of subjects per group is summarised in Table 3.2.

3.2.1.2 Metagenomic Sequencing

Faecal specimens from each subject were collected and homogenised before an 8ml sample was extracted and frozen in a solution of DNA stabiliser. In preparation for sequencing each sample was thawed and a 1.4ml subsample was taken for DNA extraction, which was carried out following the steps from the PSP spin stool DNA plus kit (molecular, 2012). Transmission spectrophotometry was then used to quantify the amount of DNA in each sample.

For 16S metagenomic sequencing, extracted DNA from each stool sample was diluted to normalise pre-amplification concentration across samples based on spectrophotometry measurements. Reciprocal barcode primers were ligated to the 16S small subunit ribosomal RNA (16S) primers and were subsequently used to amplify the DNA in triplicate. The 16S bacterial specific primers targeted regions V1 to V3 on the 16S gene. The primers were 27F (5-AGAGTTTGATCCTGGCTCAG-3) (Weisburg *et al.*, 1991) and 534R (5-ATTACCGCGGCTGCTGG-3) (Muyzer *et al.*, 1993), designed to suit an average 454 read length of over 500 base pairs and for high coverage of the Bacteria domain (Baker *et al.*, 2003; Klindworth *et al.*, 2012). This primer pair is also widely used by the HMP (Caporaso *et al.*, 2011). Multiple samples were then pooled and simultaneously sequenced using GS FLX Titanium series reagents on a Roche 454 sequencer. The multiplex barcode was later used in bioinformatics analyses to reassign sequences to their experimental origin.

Table 3.2: Subject numbers for each metabolite measurement. High density lipoprotein = HDL, Asp=aspartate, Ala=Alanine.

Metabolite	Physiological function	units	L	O	A_{50}	B_{50}	B_{500}	$D_{10\%}$
Glucose	energy substrate	$mgdL^{-1}$	10	10	9	10	10	10
Non-esterified fatty acids	energy substrate	$mEqL^{-1}$	10	10	9	10	10	10
Triglycerides	energy substrate	$mgdL^{-1}$	10	10	9	10	10	10
Cholesterol	energy substrate	$mgdL^{-1}$	10	10	9	10	10	10
HDL(cholesterol)	energy substrate	$mgdL^{-1}$	10	10	9	10	10	10
glycerol	energy substrate	$mgdL^{-1}$	10	10	9	10	10	10
beta-hydroxybutyric acid	energy substrate	$mgdL^{-1}$	10	10	9	9	10	9
Total bilirubin	kidney function	$mgdL^{-1}$	10	10	9	10	10	10
Creatinine	kidney function	UL^{-1}	10	10	9	9	10	9
Creatine kinase	kidney function	UL^{-1}	10	10	9	10	10	10
Asp aminotransferase	liver function	UL^{-1}	10	10	9	10	10	10
Ala aminotransferase	liver function	UL^{-1}	10	10	9	10	10	10
Amylase	pancreatic function	UL^{-1}	10	10	9	9	10	10
Lipase	pancreatic function	UL^{-1}	9	10	9	10	10	10
Total proteins	general health	gdL^{-1}	10	10	9	9	10	9
Albumin	inflammatory status	gdL^{-1}	10	10	9	9	10	9
Albumin/Globulin ratio	inflammatory status	-	10	10	9	9	10	9
Insulin	endocrine signalling	*ND	8	9	9	8	10	9
Amylin	endocrine signalling	*ND	8	9	9	8	10	9
Leptin	endocrine signalling	*ND	8	9	9	8	10	9
Ghrelin	endocrine signalling	*ND	8	9	8	8	10	9
Gastric-inhibitory peptide	endocrine signalling	*ND	8	9	9	8	10	9
Peptide YY	endocrine signalling	*ND	8	9	9	8	10	9
pancreatic protein	endocrine signalling	*ND	7	9	8	8	9	9
active GLP-1	endocrine signalling	*ND	8	9	9	8	10	9
total GLP-1	endocrine signalling	*ND	8	9	9	8	10	9

*ND = not disclosed

3.2.2 Bioinformatics Analyses

3.2.2.1 Metabolite Markers

Throughout this report, population statistics are tested using analysis of variance (ANOVA), Kruskal-Wallis tests, paired t-tests, unpaired t-tests, Wilcoxon tests and one-sample Kolmogorov-Smirnov tests from the MATLAB R2013b distribution with the statistics package R2013b.

Box plots are displayed to show and compare the distributions of sample populations using the box-plot function from the statistics MATLAB package R2013b. The box represents the 25^{th} to 75^{th} percentiles and the red line indicates the median value. The whiskers (dashed line to T bar) show the range excluding outliers, which are shown as red crosses. Points are drawn as outliers if they are larger than $q3+w(q3-q1)$ or smaller than $q1-w(q3-q1)$, where $q1$ and $q3$ are the 25^{th} and 75^{th}

percentiles, respectively, and w equals 1.5. The default of 1.5 corresponds to approximately plus or minus 2.7 standard deviations and 99.3% coverage if the data are normally distributed.

Normal probability plots for each sample population per metabolite have been generated using the normplot function from the statistics package R2013b and are displayed in Appendix 6.2. These plots show the frequency distribution of each subsample without bining data, which can be misleading in such sparse datasets, to give an indication of whether the data follow a normal distribution. Simulated data is superimposed on these plots in red and follows a predicted normal distribution with standardised mean and standard deviation. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. These plots were used in combination with significance values presented in the results to determine which differential tests were suitable for comparing sub populations.

A correlation matrix plot was produced using Pearson's correlation scores generated by the statistics package R2013b corr function and visualised using a customised MATLAB script provided in the code available on request.

3.2.2.2 16S Metagenomics

To analyse the 16S metagenomics data, a bioinformatics pipeline was constructed in QIIMEv1.7 (Caporaso *et al.*, 2010b) with the major steps listed below.

1. Quality control and file integrity
2. Clustering of operational taxonomic unit (OTU)s
3. Taxonomic assignment
4. Diversity and meta-analysis

Parameters for quality control when processing the FASTQ format sequence data with the denoising script (Reeder & Knight, 2010) provided in QIIMEv1.7 remained as the default values; minimum mean average quality score of 25, minimum and maximum lengths of 200 to 1000 base pairs and zero mismatches allowed in the primer sequence. Within this step the multiplexing barcodes in the mapping file were used to assign each sequence from the two sequencing runs to their respective subject sample and then the forward and reverse primers were removed as well as the barcode. The corresponding results justifying threshold cutoffs are presented in the results Section 3.3.8.

OTU clustering was calculated using uclust-v1.2.2 (Edgar, 2010) and a threshold of 97% with no reverse strand mismatches. The clusters were subsequently assigned a representative sequence based on the most popular sequence in the cluster.

PyNAST (Caporaso *et al.*, 2010a) was used in combination with the template alignment from greengenes, file core_set_aligned.fasta.imputed, to align OTUs to the reference taxonomy alignment. The final alignment was written to the Newick tree file format and used for unifrac distance calculations (Lozupone *et al.*, 2011). Taxonomy was inferred using the RDP naïve Bayesian classifier algorithm version 2.2 (Wang *et al.*, 2007) and RDP database with 80% confidence threshold. The results were formatted to populate an OTU occurrence table.

Secondary analyses were then completed. Chao score, Shannon index and species count were calculated as diversity metrics with rarefied and non-rarefied data. Due to large variation in data available for each treatment group non-rarefied data is represented in the diversity plots within this chapter.

Differential statistics were then calculated for classification of the phenotype. Independent t-tests were calculated on each OTU in the OTU table.

3.3 Results

The data generated from the experimental methods described in Section 3.2.1, were produced by collaborators at GSK directed by Dr James R. Brown. All analyses presented here were subsequently completed as part of this project at Newcastle University.

This study tests the hypothesis that antibiotic modulation of the microbiota improves insulin sensitivity in a diet induced mouse model of T2DM as measured using physiological markers, endocrine signalling markers and metagenomics from the gut microbiome.

3.3.1 A Mouse Model for Type II Diabetes Mellitus

All of the blood glucose regulation markers from groups L and O show normal distributions ($P>0.05$) and significant variability ($P<0.05$), Table 3.3. Figure 3.5 shows the population level phenotype shift from healthy to insulin resistant assumed T2DM caused by the 12 week high fat diet (10 weeks pre-treatment period plus 2 week treatment period). These results are consistent with those reported in the literature on this diet-induced obesity and T2DM mouse model (Surwit *et al.*, 1988b; Winzell & Ahren, 2004; Wang & Liao, 2012).

There is one anomaly in the lean group; subjects 1, as shown by their insulin:glucose ratio and insulin levels. These measurements are likely due to a hyper glycaemic stage in the regulation cycle. There is one anomaly in the obese group; subject 11, as shown by their total GLP-1 and active GLP-1 measurements. However, GLP-1 expression is dynamic. For further note obese control subjects 19, 17, 16 and 14 have an insulin:glucose ratio within the range of lean control mice.

This results demonstrates that the diet-induced obesity protocol had a systematic effect on blood glucose, insulin and GLP-1. Importantly, the diet-induced obesity experimental control groups represent the mice pre-treatment and show the mice in groups O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ would have been insulin resistant with a T2DM phenotype directly before the treatment period.

Table 3.3: **Population statistics for blood glucose regulation markers between the lean control group (L) and the obese control group (O).** L and O columns show one-sample Kolmogorov-Smirnov test *p*-values for whether the sample population is normally distributed and the t-test asks if there is variability across the two groups, given the subpopulations are normally distributed.

Blood glucose regulation marker	**L** n=10	**O** n=10	**T-test *p*-value**
Glucose	0.871	0.612	**0.035**
Insulin	0.330	0.906	**0.009**
Active GLP-1	0.982	0.888	**<0.001**
Total GLP-1	0.898	0.402	**0.002**
Insulin:glucose ratio	0.281	0.910	**0.014**

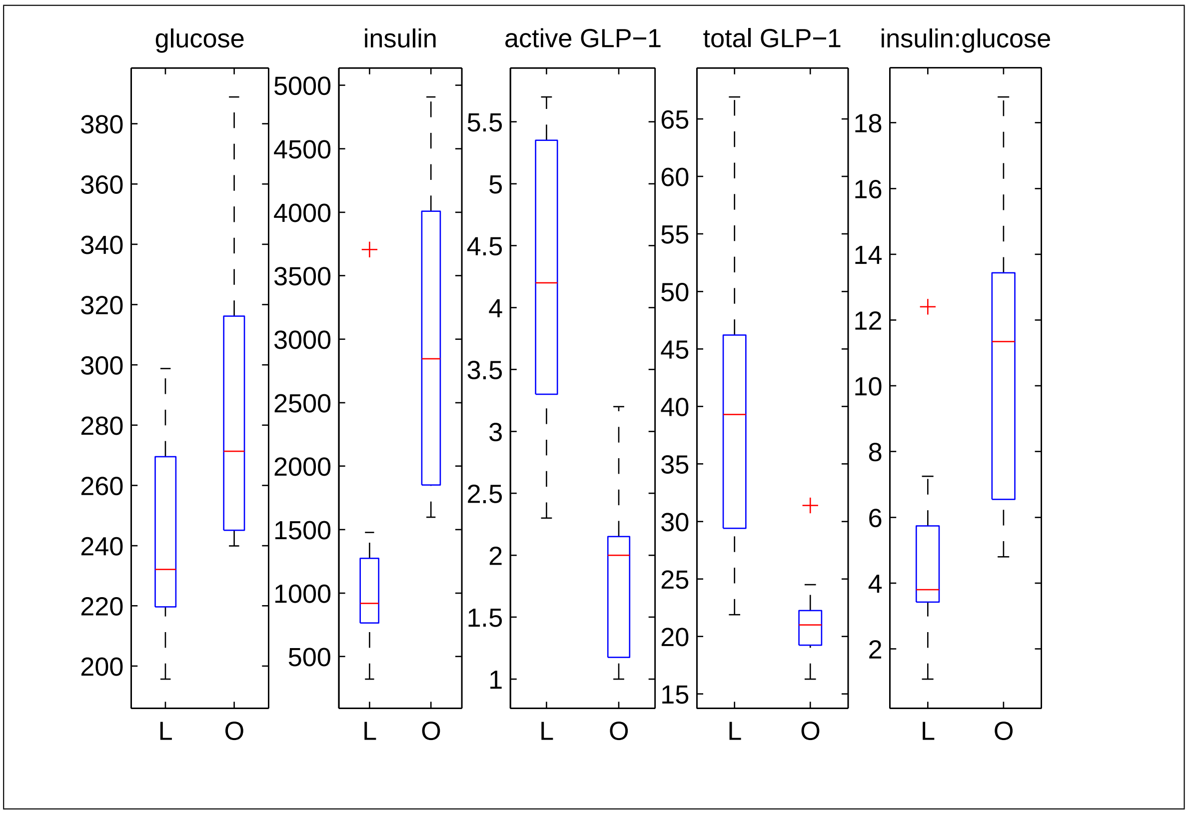


Figure 3.5: **Boxplots showing the effect of a 12 week high fat diet on blood glucose regulation markers in the lean control group (L) and obese control group (O).** Supporting statistics for normalisation testing and differential statistics can be found in Table 3.3. Units, where disclosed, are in Table 3.2.

3.3.2 Treatment Effects on Insulin Sensitivity

T2DM is fundamentally a blood glucose regulation dysfunction. To investigate the effects of the treatment period on blood glucose regulation a series of markers, as listed in the above Results Section 3.3.1, were tested across all treatment groups. In order to test for effects of treatment on pre-existing insulin resistance and an assumed T2DM phenotype, differential statistics were calculated excluding the lean control group.

All markers for blood glucose regulation for all treatment groups showed normal distributions (P>0.05) and significant (P<0.05) variability across the treatment groups, except for the insulin:glucose ratio, Table 3.4. The boxplots in Figure 3.6 show a trend towards improved glucose regulation across all treatment groups, although the main differences across the treatment groups are in the Gram-negative antibiotic treatment groups, particularly group B_{500}. The range of insulin measurements and the range of insulin:glucose ratio measurements for group B_{500} has fallen to within the quartile bounds of group L; an unpaired t-test shows there is no significant difference (P=0.797) between group B_{500} and group L for insulin. All other groups show a significant (P<0.05) difference from group L.

Both active GLP-1 and total GLP-1 levels are dramatically increased in the blood for the Gram-negative antibiotic treatment groups B_{500} and B_{50} far beyond increases caused by other treatments. Of further note, groups A_{500} and $D_{10\%}$ also show significant increases (P<0.05) in both active GLP-1 and total GLP-1 levels as well as non significant (P>0.05) reductions in glucose and insulin levels.

There are three subjects with anomalously high insulin levels, subject one in group L which also shows an anomalous insulin:glucose ratio; subject 22 in the group A_{500} which also shows an anomalous insulin:glucose ratio; and subject 59 in group $D_{10\%}$. There is one anomaly in the group B_{50}, subject 45, with anomalously high glucose levels. These subjects are likely in hyper or hypo glycaemic stages of the glucose regulation cycle. Subjects 11 (group O), 37 (group B_{500}), 41 (group B_{50}) and 49 (group B_{50}) show anomalous results for total GLP-1 and subject 37 also shows an anomalous result for active GLP-1. However, considering the observed distribution for GLP-1 expression outliers are expected.

These results are central to this chapter. The blood glucose regulation markers shows a full recovery to health during a 14 day treatment period in response to treatment with antibiotic B treatment at 500mgkg^{-1}. This is one of the first studies to show acute beneficial effect of antibiotics on T2DM. Additionally, the results show a general improvement in the insulin sensitivity for all treatment groups and a drammatic increase in expression of GLP-1 in response to antibiotic B. The underlying effect of each treatment on broader physiological markers is investigated in the following sections to explore the disease model and propose a model for treatment.

Table 3.4: **Population statistics for blood glucose regulation markers for all treatment groups.** L, O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ columns show one-sample Kolmogorov-Smirnov test *p*-values for whether the sample population is normally distributed. The ANOVA *p*-value tests if there is variability across the treatment groups excluding the lean control group, given the subpopulations are normally distributed.

Blood glucose regulation marker	**L**	**O**	**A_{500}**	**B_{50}**	**B_{500}**	**$D_{10\%}$**	**ANOVA *p*-value**
Glucose	0.871	0.612	0.799	0.980	0.879	0.958	**0.002**
Insulin	0.330	0.906	0.615	0.989	0.880	0.767	**0.038**
Active GLP-1	0.982	0.888	0.606	0.579	0.736	0.837	**<0.001**
Total GLP-1	0.898	0.402	0.951	0.575	0.529	0.843	**<0.001**
Insulin:glucose	0.281	0.910	0.279	0.824	0.999	0.858	**0.159**

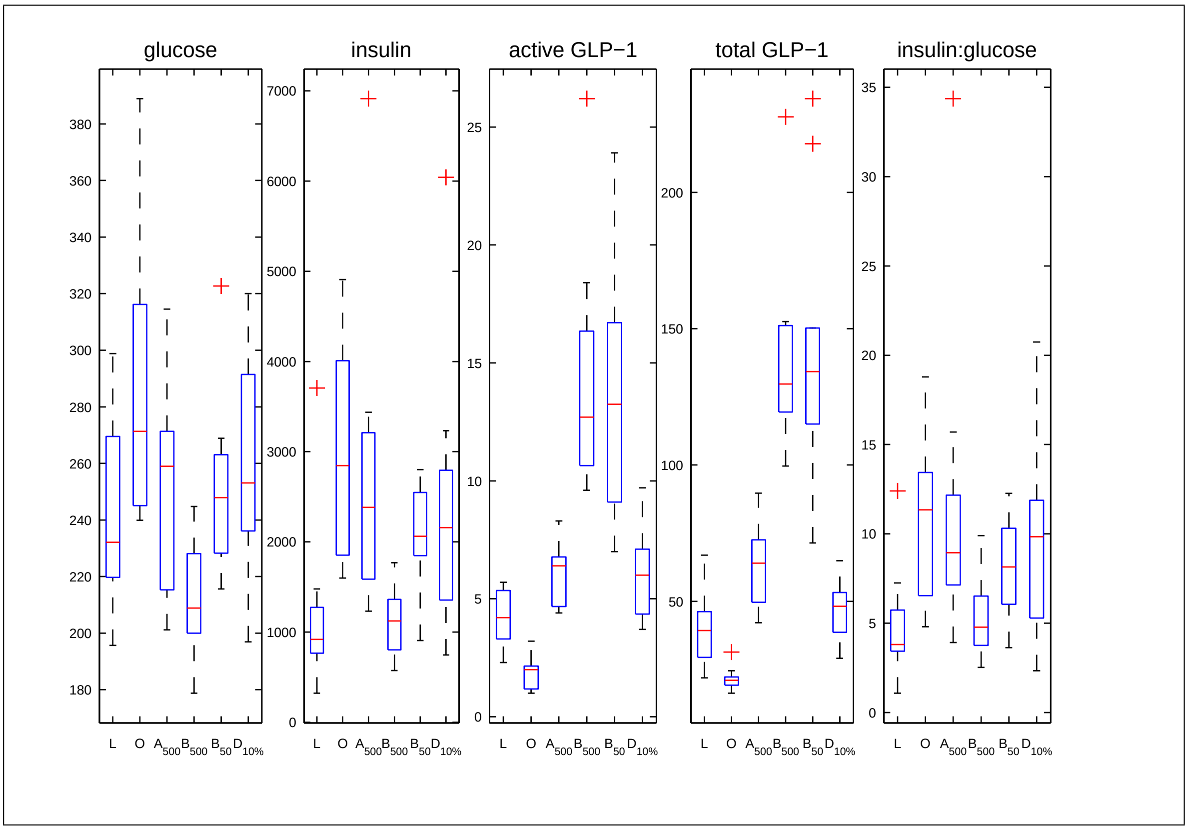


Figure 3.6: **Boxplots showing blood glucose regulation markers, glucose, insulin, total GLP-1, active GLP-1 and insulin:glucose ratio across all groups after the 14 day treatment period.** Supporting statistics for normalisation testing and differential statistics can be found in Table 3.4. Units for markers, where available, are in Table 3.2 and parameters used for drawing the graph are discussed in the methods.

3.3.3 Treatment Effects on Body Weight

Adiposity is a major factor in insulin resistance. Furthermore, weight gain or loss can give an indication of metabolic status when combined with calories consumed. This section investigates whether there were any changes in overall body weight and fat mass over the treatment period and whether there were any differences across the groups.

Overall body weight is significantly (P<0.05) different from pre-treatment to post-treatment in groups L, B_{500}, B_{150} and $D_{10\%}$, Table 3.5. In group L this weight change accounts for healthy growth during this period of aging for mice. Therefore, the average weight change in group L was subtracted from the weight measurements in other treatment groups to correct for natural growth. After correction all groups showed significant (P<0.05) weight loss except group L and group O. The boxplots in Figure 3.7 show the main differences are found in group $D_{10\%}$ (P=<0.001 unpaired t-test) and group B_{500} (P=<0.001 unpaired t-test), although groups A_{150}, B_{150} and B_{50} also show significant (P<0.05 unpaired t-tests) weight loss compared to group O despite largely following the same diet. Interestingly there is a distinction between the dosage groups for antibiotic B. Group B_{500} is significantly (P<0.001) different from the other groups, including B_{50} but not $D_{10\%}$(P=0.507). These results indicate a potential basis for the dosage response differences seen in insulin levels between B_{50} and B_{500} in Results Section 3.3.2.

Importantly for the experimental design, the OFS group show a clear weight loss. OFS as an artificial sweetener increases the satiety of the mice and therefore reduces their desire for food. OFS supplements are a common mechanism for inducing weight loss in mice so in this experimental design represent a natural weight loss control. The implication of this result is that any impact on insulin sentitivity beyond what would be natural from weight loss can be attributed to mechanisms related to antibiotic treatment.

Table 3.5: **Body weight and fat mass measurements before and after the treatment period.** Paired t-test *p*-values show whether there is variability in body weight and fat mass over the treatment period. Baseline correction tests for variability after correcting for weight gain and fat mass gain shown in the lean control group.

Weight (g)	**L**	**O**	**A_{500}**	**A_{150}**	**A_{50}**	**B_{500}**	**B_{150}**	**B_{50}**	**$D_{10\%}$**
Pre-treatment body weight	26.8	40.7	40.3	40.0	40.5	41.0	40.9	40.2	40.5
Post-treatment body weight	28.0	41.4	39.6	39.4	40.3	36.7	39.0	39.0	35.9
Body weight paired t-test	<0.01	0.120	0.312	0.099	0.963	<0.01	<0.01	0.121	<0.01
Baseline corrected t-test	0.941	0.347	0.015	<0.01	0.036	<0.01	<0.01	<0.01	<0.01
Pre-treatment fat mass	3.3	15.6	15.3	15.2	15.5	15.5	15.6	15.3	15.3
Post-treatment fat mass	3.7	16.0	14.1	13.9	14.7	11.6	14.0	14.4	11.2
Fat mass paired t-test	<0.01	0.419	0.081	<0.01	0.096	<0.01	<0.02	0.212	<0.01
Baseline corrected t-test	0.434	1.000	0.023	<0.01	0.015	<0.01	<0.01	0.083	<0.01

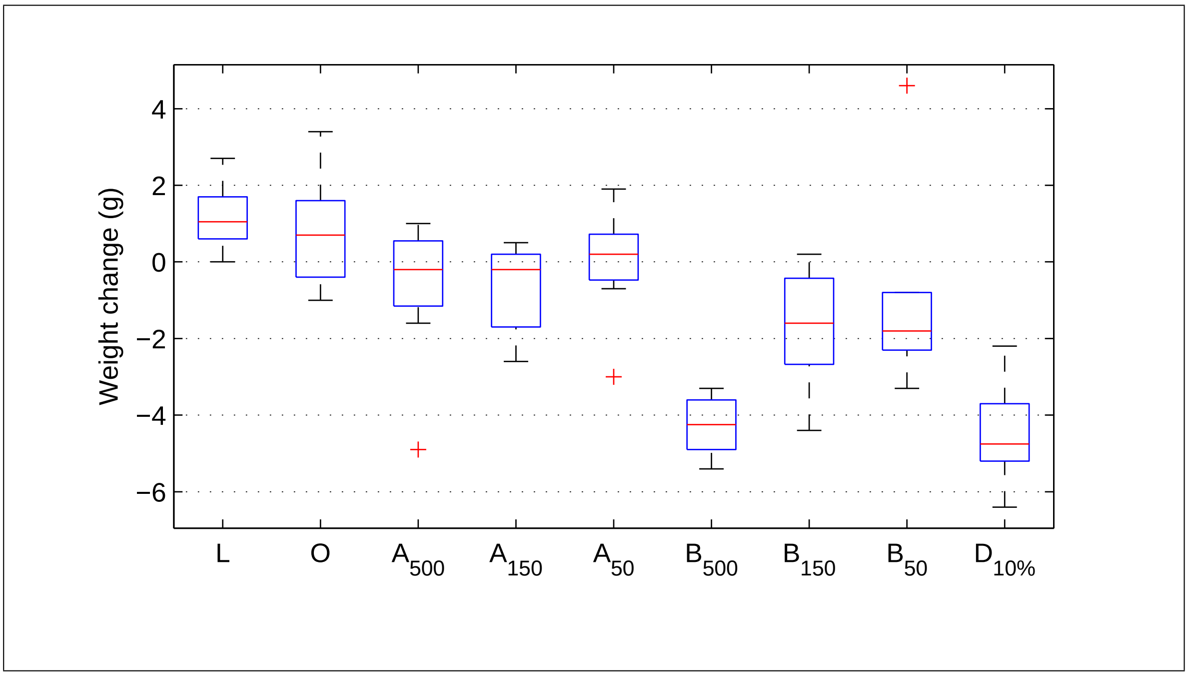


Figure 3.7: **Average body weight change for mice during the 14 day treatment period.** See methods for treatment group labelling.

3.3.4 Treatment Effects on Host Physiology

Blood glucose regulation is complex and involves multiple physiological systems. The next analyses attempt to identify the effect of each treatment on physiological systems in the mice by analysing markers for liver function, kidney function and pancreatic function.

A Kolmogorov-Smirnov test with the CK, amylase and lipase data reject the null hypotheses that the data are normally distributed and show a recurrent trend towards an F distribution. Given these distributions Kruskal-Wallis rank sum tests were used in place of ANOVA tests to test for variability in CK, amylase and lipase.

Differential statistical tests show there was no significant (P>0.05) variability across all of the treatment groups and group L for any of the markers for liver, kidney or pancreatic function except for albumin, Table 3.6. Albumin levels are slightly increased, which suggests dehydration as opposed to reduced liver function which would be also reflected in tBil measurements. This result indicates that liver function, kidney function and pancreatic functions are not damaged across any of the treatment groups.

A closer look at the data reveals there are multiple outliers across all of the groups, as identified by the boxplots, Figure 3.8. These outliers could indicate subjects that are unwell due to reasons outside the experimental procedure or measurement anomalies. Within the measurements identified, none of the subjects showed to be outliers across multiple measurements indicating most are likely to be healthy subjects.

Table 3.6: **Population statistics for marker of liver function, kidney function and pancreatic function across all treatment groups.** L, O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ columns show one-sample Kolmogorov-Smirnov test p-values for whether the sample population is normally distributed. The differential p-value tests if there is variability across the treatment groups including the lean control group, using ANOVA tests where the subpopulations are deemed to be normal distributions and Kruskal-Wallis rank sum tests where the subpopulations are deemed to be not normally distributed.

Physiological markers	L	O	A_{500}	B_{500}	B_{50}	$D_{10\%}$	Differential p-value
AST	0.186	0.039	0.526	0.241	0.589	0.182	**0.897**
ALT	0.147	0.064	0.326	0.981	0.839	0.540	**0.106**
tBil	0.715	0.246	0.350	0.654	0.224	0.302	**0.084**
Albumin	0.312	0.763	0.335	0.216	0.896	0.803	**0.007**
creatinine	0.896	0.244	0.349	0.321	0.790	0.701	**0.392**
CK	0.095	0.071	0.656	0.216	0.012	0.084	**0.101**
amylase	0.015	0.951	0.027	0.088	0.066	0.080	**0.307**
lipase	0.011	0.936	0.025	0.108	0.136	0.019	**0.201**

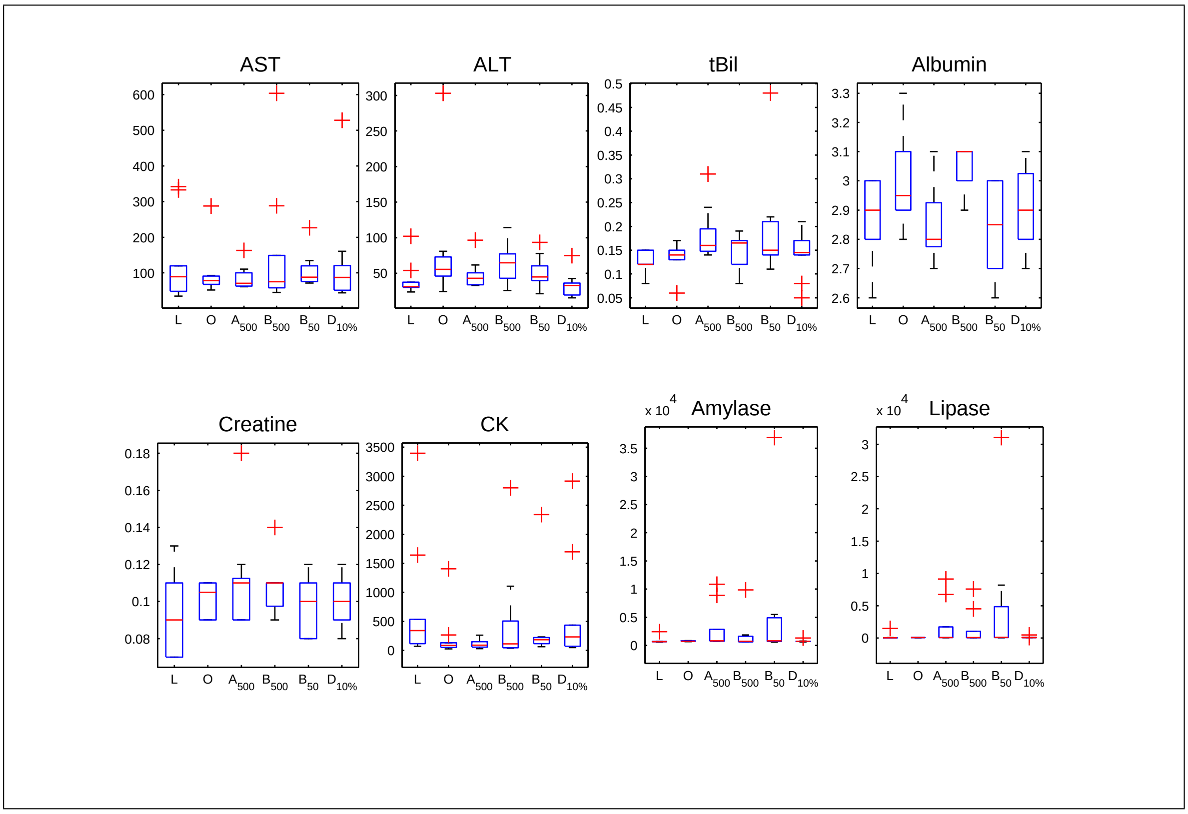


Figure 3.8: **Boxplots showing the treatment effects on liver function, kidney function and pancreatic function.** Supporting statistics for normalisation testing and differential statistics can be found in Table 3.6. Units for markers, where available, are in Table 3.2 and parameters used for drawing the graph are discussed in the methods.

3.3.5 Treatment Effects on Inflammatory Status

Plasma protein markers give an indication of health and inflammation. Total plasma proteins are increased with episodes of inflammation or dehydration. Albumin measurements give a measure of dehydration assuming healthy liver function. Globulin provides a measure of inflammation, although this measure can be affected by dehydration. Globulin is higher when the innate immune system is more active. The albumin:globulin ratio provides an indication of levels of inflammation accounting for dehydration.

All plasma protein marker measurements show normal distributions ($P>0.05$) except group A_{500} for the albumin:globulin ratio. One-way ANOVA tests suggest there is significant ($P<0.01$) variability across the treatment groups including group L for all plasma protein markers. Furthermore one-way ANOVA tests also suggest there is significant variability across the treatment groups excluding group L for total proteins and albumin levels, with *p*-values of 0.020 and 0.006 respectively. Figure 3.9 shows there is a distinction ($P=0.0639$) between group B_{500} and group B_{50} in the albumin:globulin ratio.

These results demonstrate no significant increases in inflammation beyond the bounds of the control groups that would indicate an acute inflammatory in response to illness. This means despite significant variance between the groups the overall inflammatory status for all mice is considered normal given the conditions and can variance can be attributed to the experimental treatment. The variance between the groups is mainly associated to the high fat diet although there is a trend towards reduced inflammation in group B_{500}. However, the measurements are not precise enough to draw any conclusions regarding low grade inflammation. Given the proposed model for the microbiota contributing to the disease model through low grade inflammation this measurement is a significant downfall of the experimental procedure.

Table 3.7: **Population statistics for plasma protein markers across all treatment groups.** L, O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ columns show one-sample Kolmogorov-Smirnov test *p*-values for whether the sample population is normally distributed. The differential *p*-value tests if there is variability across the treatment groups including the lean control group, using ANOVA tests where the subpopulations are deemed to be normal distributions and Kruskal-Wallis rank sum tests where the subpopulations are deemed to be not normally distributed.

Plasma Protein Markers	L	O	A_{500}	B_{500}	B_{50}	$D_{10\%}$	ANOVA *p*-value
total plasma proteins	0.738	0.940	0.480	0.480	0.788	0.538	**<0.001**
albumin	0.312	0.763	0.335	0.216	0.896	0.803	**0.007**
albumin:globulin ratio	0.262	0.535	0.026	0.555	0.574	0.614	**<0.001**

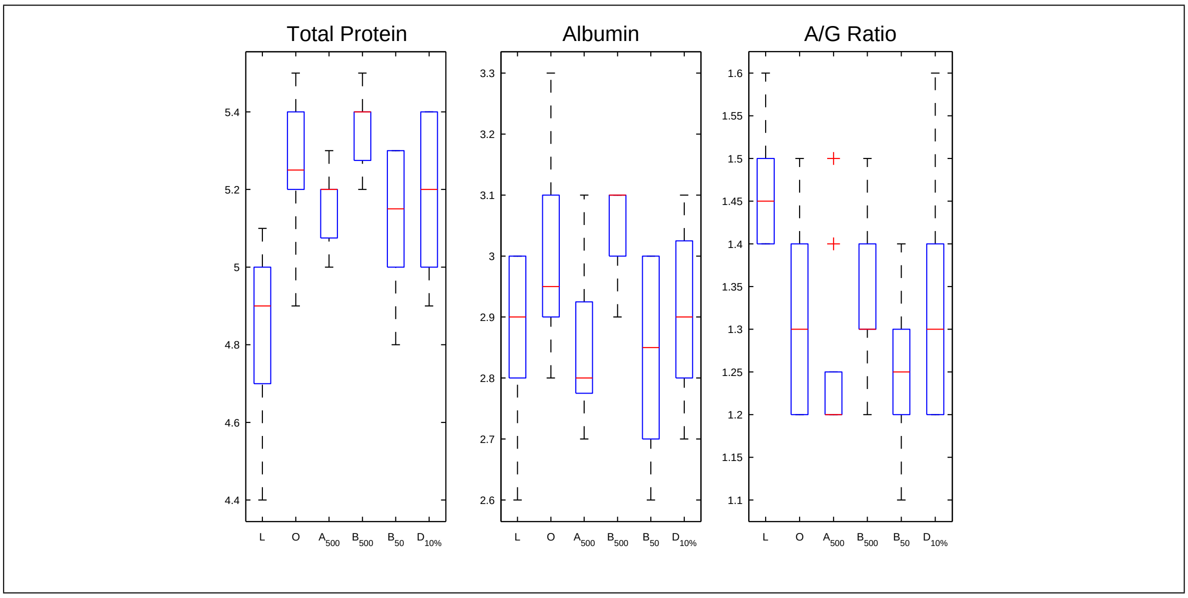


Figure 3.9: **Boxplots showing the treatment effects on plasma proteins.**Supporting statistics for normalisation testing and differential statistics can be found in Table 3.7. Units for markers, where available, are in Table 3.2 and parameters used for drawing the graph are discussed in the methods. A/G ratio = Albumin:Globulin ratio.

3.3.6 Treatment Effects on Metabolic Status

The 10 week high fat diet resulted in high amounts of sub cutaneous fat, as shown in Section 3.3.3, which is known to cause higher levels of glycerols, cholesterol and HDL-C in the blood. One-way ANOVA tests support this with significant (P<0.05) variability across the treatment groups including group L for cholesterol and HDL-C, Table 3.8. However, there is also significant (P<0.05) variability between the treatment groups excluding group L. Interestingly cholesterol and HDL-C show the reverse trend of the weight loss profiles for group B_{500} and group B_{50} Figure 3.10. bHBA varies significantly (P<0.05) across the groups and shows a clear reflection of the weight loss profiles. This is an indication that the weight loss profiles are also a reflection of a healthy metabolic status.

Table 3.8: **Population statistics for metabolic status markers across all treatment groups.** L, O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ columns show one-sample Kolmogorov-Smirnov test *p*-values for whether the sample population is normally distributed. The ANOVA *p*-value tests if there is variability across the treatment groups including the lean control group, using ANOVA tests where the subpopulations are deemed to be normal distributions and Kruskal-Wallis rank sum tests where the subpopulations are deemed to be not normally distributed. NEFA = non-esterified fatty acids.

Energy Substrates	L	O	A_{500}	B_{500}	B_{50}	$D_{10\%}$	ANOVA *p*-value
glucose	0.871	0.612	0.799	0.980	0.879	0.958	**0.003**
NEFA	0.853	0.740	0.833	0.058	0.139	0.984	**0.210**
triglycerides	0.947	0.983	0.827	0.472	0.518	0.948	**<0.001**
cholesterol	0.917	0.975	0.822	0.748	0.421	0.798	**<0.001**
HDL-C	0.834	0.942	0.800	0.925	0.188	0.939	**<0.001**
glycerol	0.935	0.796	0.980	0.349	0.191	0.455	**0.153**
bHBA	0.966	0.214	0.897	0.991	0.983	0.550	**<0.001**

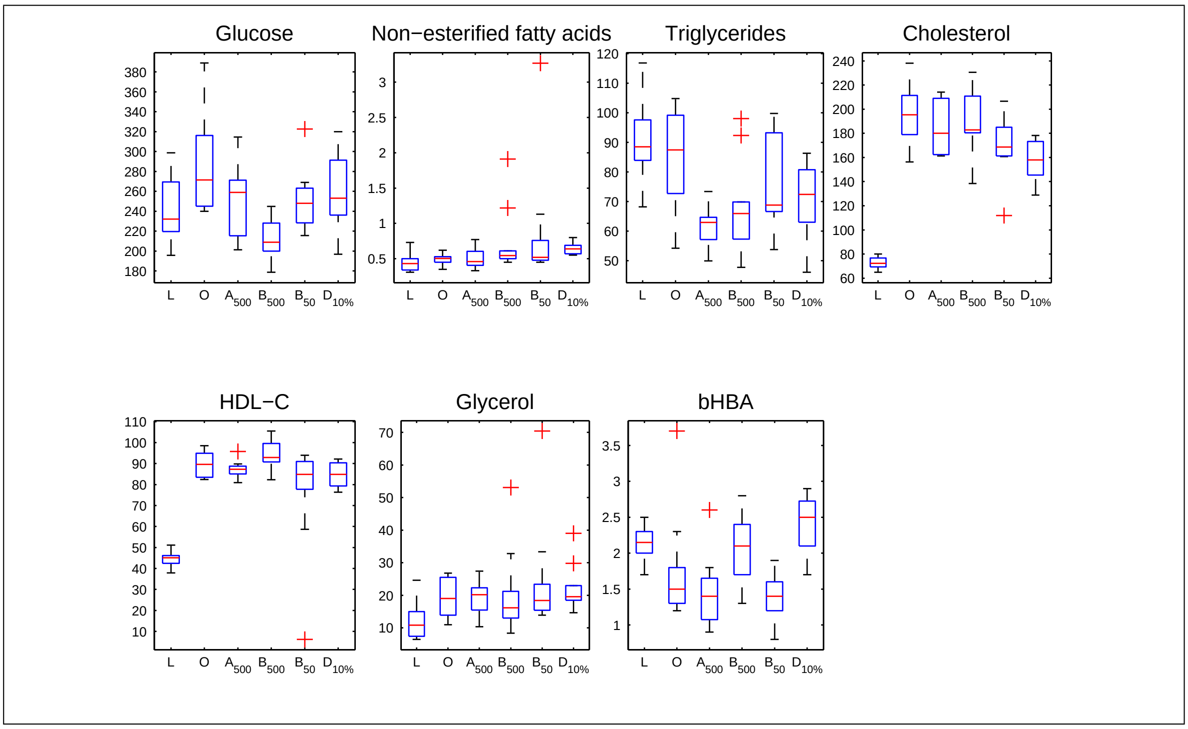


Figure 3.10: **Boxplots showing the treatment effects on energy substrates in the blood.** Supporting statistics for normalisation testing and differential statistics can be found in Table 3.8. Units for markers, where available, are in Table 3.2 and parameters used for drawing the graph are discussed in the methods.

3.3.7 Treatment effects on endocrine signalling

This section investigates whether endocrine signalling was affected by the treatment period. Endocrine signalling can give an indication as to the endocrine pathways contributing to changes in insulin sensitivity.

All endocrine signalling biomarker measurements are normally distributed for all groups (P>0.05). One-way ANOVA tests show there is significant (P<0.05) variability across the groups including group L for insulin, amylin, leptin GIP, PYY, active GLP-1 and total GLP-1. Equally of interest, there was no significant (P>0.05) variability in Ghrelin or PPr. One-way ANOVA tests suggest there is also significant (P<0.05) variability across the treatment groups excluding the lean control group for insulin, amylin, GIP, PYY, active GLP-1 and total GLP-1.

This data shows there were multiple changes in endocrine signalling caused by both the 10 week high fat diet and the treatment period. In particular, Figure 3.11 shows a very similar trend for insulin and amylin, which is expected since they are both dependent on liver insulin sensitivity, and the high dose Gram-negative antibiotic group have reduced levels of insulin. Leptin follows the trend in weight and subcutaneous fat. Ghrelin appears to be affected by the Gram-positive antibiotic but is highly variable across all groups. GIP appears to improve with all treatments. PYY follows the trend of GLP-1 where the Gram-negative treatment groups show much higher levels.

The notable part of these endocrine signalling results are in distinguishing effects of pharmalogical treatment from changes related in weight loss. The most obvious distinction is in GLP-1 marker which does not follow the trends for weight loss. While both Gram-negative antibiotic groups have a drammatic roughly ten fold increase, the Gram-positive antibiotic group also shows a roughly three times increase as well as a significant reduction in GIP.

Table 3.9: **Population statistics for endocrine signalling markers across all treatment groups.** L, O, A_{500}, B_{50}, B_{500} and $D_{10\%}$ columns show one-sample Kolmogorov-Smirnov test *p*-values for whether the sample population is normally distributed. The ANOVA *p*-value tests if there is variability across the treatment groups including the lean control group, using ANOVA tests where the subpopulations are deemed to be normal distributions and Kruskal-Wallis rank sum tests where the subpopulations are deemed to be not normally distributed. NEFA = non-esterified fatty acids.

Diabetic Marker	**L**	**O**	**B**	**A**	**A**	**D**	**ANOVA *p*-value**
insulin	0.330	0.906	0.615	0.989	0.880	0.767	**0.012**
amylin	0.643	0.894	0.946	0.701	0.130	0.783	**0.002**
leptin	0.938	0.054	0.215	0.540	0.180	0.998	**0.013**
ghrelin	0.356	0.980	0.845	0.972	0.767	0.891	**<0.001**
GIP	0.554	0.977	0.616	0.758	0.493	0.632	**<0.001**
PYY	0.513	0.613	0.913	0.632	0.918	0.971	**<0.001**
PPr	0.608	0.799	0.649	0.483	0.162	0.445	**0.794**
active GLP-1	0.982	0.888	0.606	0.579	0.736	0.837	**<0.001**
total GLP-1	0.898	0.402	0.951	0.575	0.529	0.843	**<0.001**

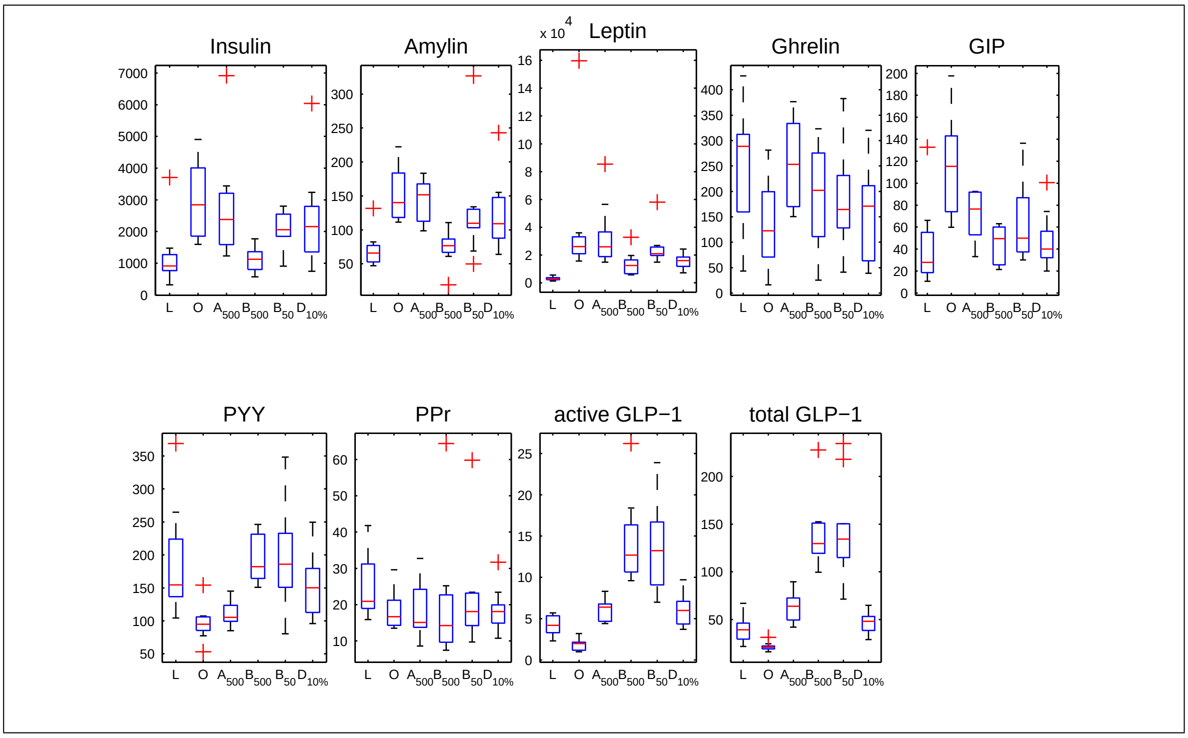


Figure 3.11: **Boxplots showing the treatment effects on endocrine signalling.** Supporting statistics for normalisation testing and differential statistics can be found in Table 3.9. Units for markers, where available, are in Table 3.2 and parameters used for drawing the graph are discussed in the methods.

3.3.8 Variable Correlation

Overall it is clear that the 10 week high fat diet and treatment period had a wide spread impact on the physiology of the mice and the treatment period caused some clear phenotypic distinctions. This section investigates whether there were correlations between the biomarkers measured to look for potential interactions between markers that would facilitate the development of a disease and treatment model.

The correlation matrix shows some strong correlations, mostly between markers linked within the results categories, such as lipase and amylase, AST and ALT, cholesterol and HDL-C, total GLP-1 and active GLP-1. There are also some notable correlations between markers across categories although these appear to be weaker correlations. Amylase and lipase with tBil, non-esterified fatty acids (NEFA) and glycerol. Cholesterol with many of the endocrine signalling markers. PYY and PPr with both GLP-1 markers. These weaker correlations (Pearson's coefficient value between 0.3 and 0.7) provide limited evidence that these markers are physiologically linked. They also support the fact this data is noisy and any conclusions from multivariate analyses would be limited from these measurements.

The hypothesis for this section is that there is a systematic relationship between metabolite markers and insulin that would be indicative of a treatment mechanism. For example, if markers for infammation were part of the treament mechanism I would expect inclusion of the data from those markers to improve the accuracy of a multivariate predictive model. I tested if data from each metabolite was able to improve a multivariate model for predicting insulin levels. The results showed only those metabolites directly linked to the insulin signalling pathways significantly improved the model due to the noise to signal ratio in the other markers.

These results demonstrate a lack of continuous data from the metabolite markers. Many of the markers are measured as dysfunctional markers and show very limited variability in healthy individuals. A more insightful experimental design would have provided a more continuous response curve for the range of healthy individuals.

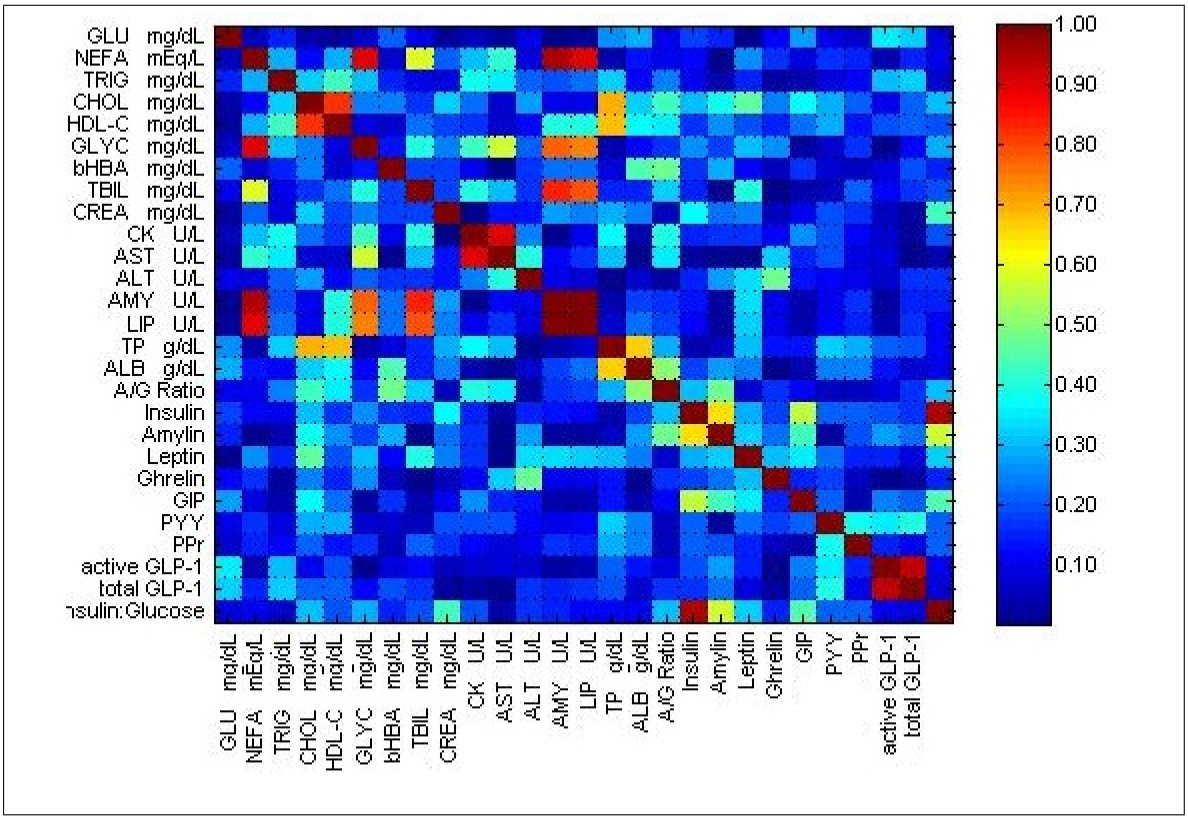


Figure 3.12: **Biomarker correlation matrix** showing the relative Pearson's coefficient between each biomarker. The range starts from zero to one and is represented by a graded colouring from blue to red.

3.3.9 16S Metagenomics

To analyse the microbiome a 16S survey was completed using a 454 sequencer and the data was put through a bioinformatics pipeline as described in Section 3.2.2.2.

269,745 sequence reads were present in the raw data, 21,794 were outside the minimum and maximum length restrictions, 24 contained more than six ambiguous base pairs, 369 had an average quality score lower than twenty five, 2,191 contained a homopolymer run of more than six, 123,118 contained a mismatch in one of the primers or barcodes. This high number of barcode mismatches was due to inclusion of sequencing samples from unknown origins where the barcode could not be included in the analyses. There were 91,174 remaining sequence reads assigned to samples. The minimum, median and maximum number of sequences per sample were 0, 1753.35 and 4478 and distributed as in Figure 3.13. Due to low sequence counts subjects 31, 35, 39, 40, 41, 47, 49 and 53 were not included in later analyses, as identified in Appendix 6.1.

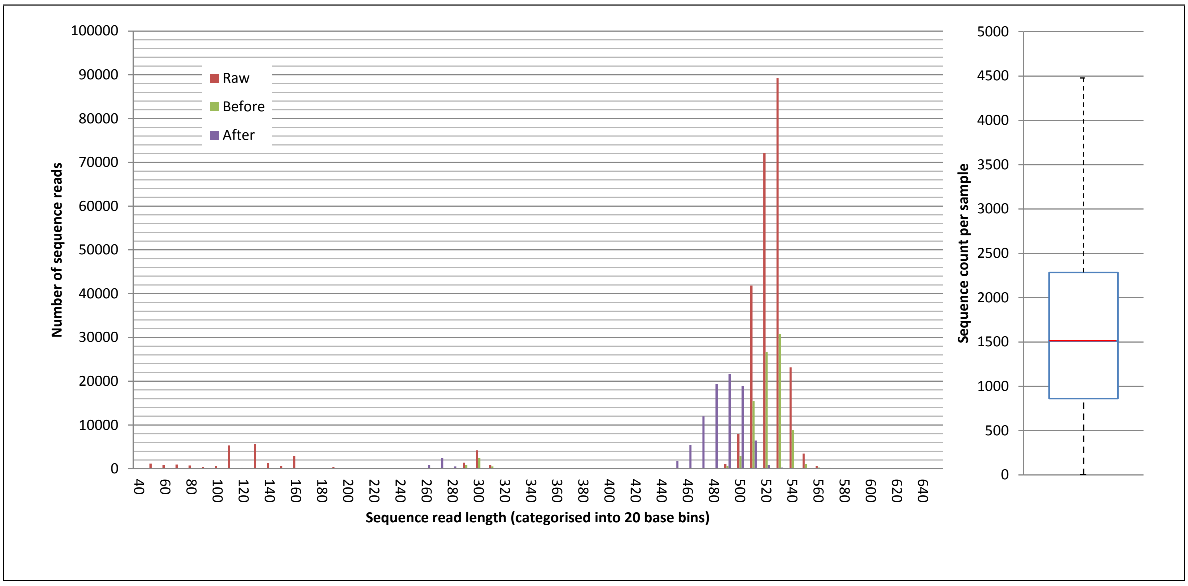


Figure 3.13: Sequence read lengths bined into 20 base intervals and a summary of the range of sequences reads per sample. 'Raw' is red and represents all of the sequence reads generated during sequencing, 'before' is green and represents the sequence reads remaining after initial quality control filtering and 'after' ' represents the remaining sequence reads after quality control and after barcode sequences and amplicon sequences have been removed.

OTUs are terminal nodes in phylogenetic analyses (Blaxter *et al.*, 2005). In this instance they represent a cluster of sequences thought to originate from a single strain or species, exemplified by a consensus sequence. 11,091 OTUs were identified.

Each OTU was assigned a representative sequence that was taxonomically classified. Figure 3.14 shows the proportional representation of each phylum level bacteria from the OTU table taxonomic classification.

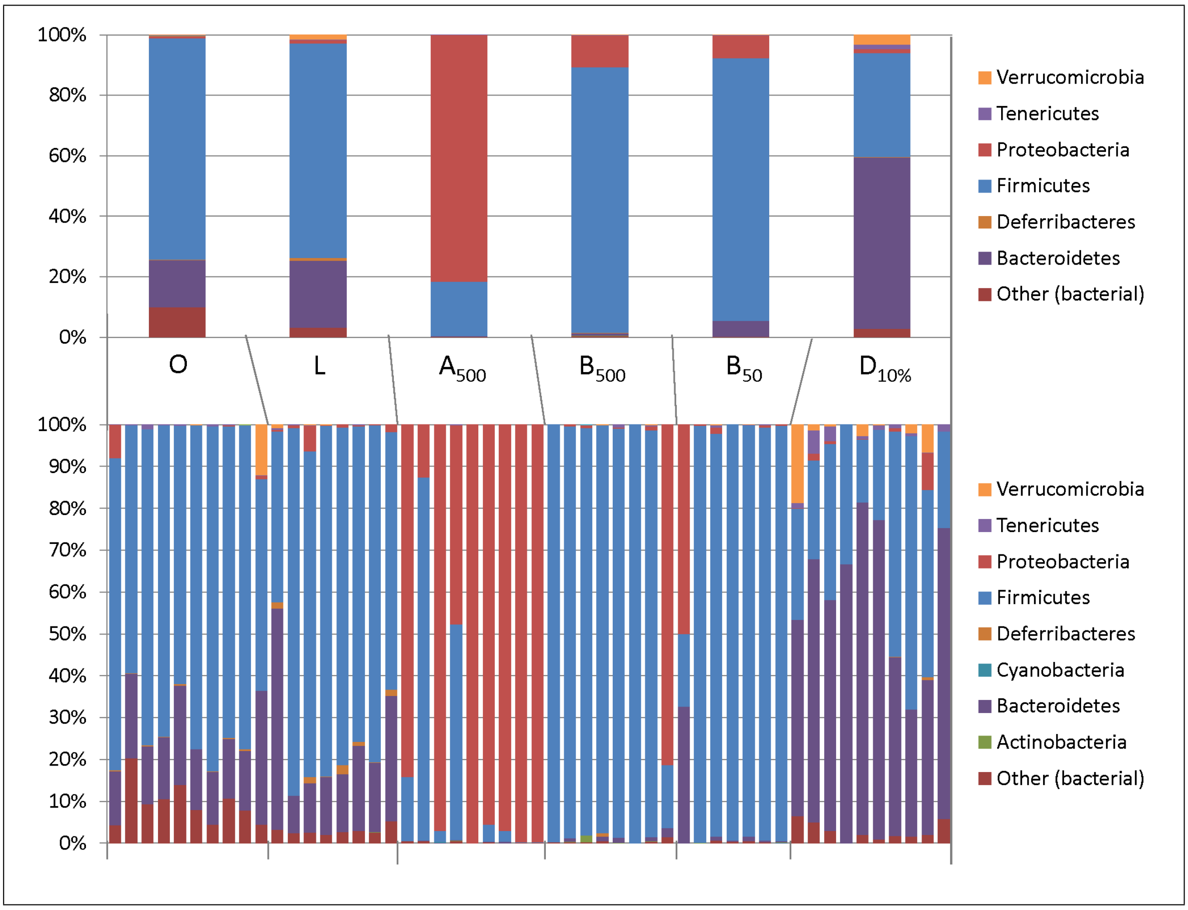


Figure 3.14: Phylum level proportional representation of each microbiomes. Average proportional breakdown is shown at the top and a per subject breakdown of proportional representation is shown directly underneath. Adjoining lines between the top and bottom plots show the division by phenotype of each subject.

What is clear from the initial metagenomic results is that the antibiotics had a broad impact on the microbiome. The Gram-positive antibiotic (antibiotic A) removed almost all of the diversity including the Firmicutes and they were replaced by Proteobacteria. There are two anomalies where a large proportion of Firmicutes remain in the microbiome with around 20% and 40% Proteobacteria. The Gram-negative antibiotic treatments (antibiotic B) also appear to have removed a large portion of the diversity. Although antibiotic B appears to have allowed the Firmicutes to thrive, the dominant species of Firmicute has changed indicating a major remodelling. These remodelling results indicate that both antibiotic treatments had a broad cleansing effect on the microbiota and the resulting population was made up of a previously minor constituent of the microbiota. Despite not having abundance data the implication is that the biomass is dramatically reduced as well (Langdon *et al.*, 2016; Panda *et al.*, 2014).

There are two anomalies that have a large proportion of Proteobacteria, one that also has some Bacteroidetes. The OFS supplements appear to have had a more subtle impact in promoting Bacteroidetes in place of Firmicutes but also allowing other bacteria such as Tenercutes and Verrucomicrobia to thrive in some subjects (Macfarlane *et al.*, 2008).

One way to represent complexity is by quantifying the overall alpha diversity. As well as the total number of species found, the Shannon index, phylogenetic distance (PD_whole_tree) and species richness (Chao1 score) were all measured. The Shannon index reflects the entropy or uncertainty of a system. In this instance, diversity is weighted towards the most common species so less species and skewed abundances would result in a low number and the more species that are more common the higher the number. The phylogenetic distance is calculated based on a cumulative total of the evolutionarily separation of all identified species from the most abundant species using a standard phylogenetic model. The Chao metric of species richness is weighted more heavily on less common species so measures based on how many species are represented by only few sequences.

In order to make the diversity metrics comparable it is necessary to account for experimental coverage of the sample to predict the actual diversity. To do this a rarefaction curve is produced and the alpha diversity is calculated at any given point. To rarify the OTU table a series of randomised sub samples are taken from each sequencing sample, then the average number of OTUs found are plotted on a graph to produce a curve. As the size of the sub sample increases it is expected the diversity measure will get closer to a saturation point (absolute diversity) and plateau. The curve can be extrapolated in a model to predict when the curve will plateau. However, due to the complexity of evolution and adaptation of the microbiome these models are often inaccurate extrapolations.

In Figure 3.15 the species count does not appear to plateau in any of the groups. This shows that there are likely to be multiple species unidentified in the samples and further sequencing would be able to identify further species as well as increasing the number of sequences for currently identified OTUs. Despite this it is clear that all three antibiotic treatments dramatically reduced the diversity, the high fat diet reduces diversity and OFS increases diversity as seen in multiple other studies.

Phylogenetic distance measurements in Figure 3.15 shows a similar trend to species count. This supports the evidence that there are likely to be multiple species unidentified. This graph also shows that the new species identified in higher rarefied fractions of data aren't just close relatives of highly abundant species already identified.

The Shannon index in Figure 3.15 appears to approach a plateau for all groups. This would indicate that the proportion of major constituents is relatively accurate as represented at the phylum level in Figure 3.14. Furthermore this would indicate that the diversity driven by major constituents is more competitive in the lean animals and far less in the antibiotic treated animals. This trend follows that a higher dosage of antibiotics is likely to be less facilitating to a diverse microbiome.

Finally in Figure 3.15 the Chao1 score supports the finding that many minority constituents of the microbiome are likely to be undefined due to low sequence coverage of the community. Importantly the Chao1 score shows the diversity of the minor proportional constituents is dramatically reduced in all antibiotic treatment groups and appears to show the increase in minor constituents is more pronounced.

In summary the antibiotic treatments cause a huge decrease of the diversity and cause a simplified community with few species. A comparison of group $D_{10\%}$ with L would indicate that the higher diversity found in the group $D_{10\%}$ group represents a large number of minor constituents and therefore a more complex community.

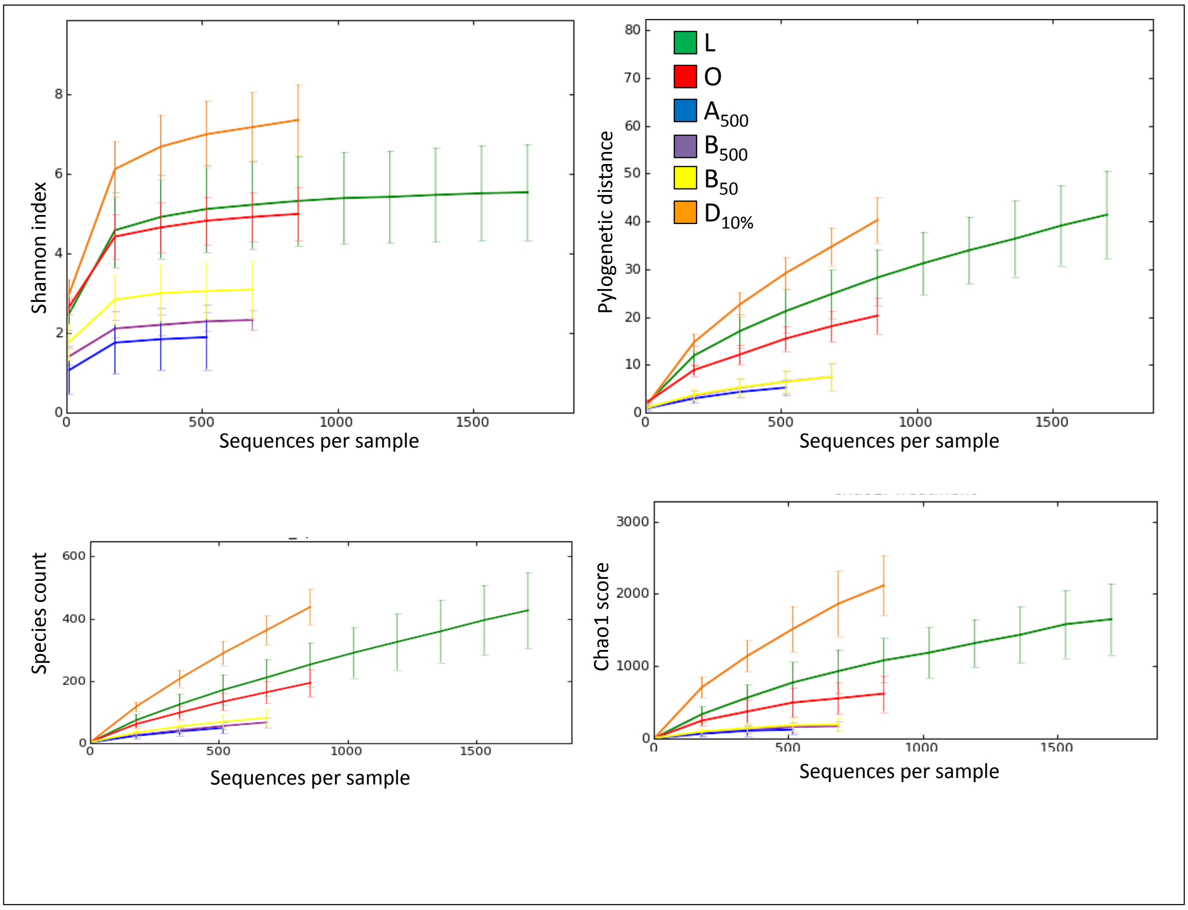


Figure 3.15: **Observed species rarefaction curve.** Green is the lean control group, red is the obese control group, blue is the Gram-positive antibiotic treatment group, purple is low dose Gram-negative antibiotic treatment group, yellow is high dose Gram-negative antibiotic treatment group and orange is the OFS dietary supplement group.

Another approach to summarise the complexity and compare between treatment groups is principal co-ordinate plots. This involves calculating the beta diversity, that is the distance or dissimilarity between each microbiome. Figure 3.16 shows a clear separation in the first principal component between antibiotic treatment and no antibiotic treatment with the majority of the weighting shared among popular species in the non treatment groups. The second principal component shows a distinction between the non treatment groups and is driven by a fraction of the *Firmicute* species. The third principle component separates the groups with a heavier weighting on Bacteroidetes species.

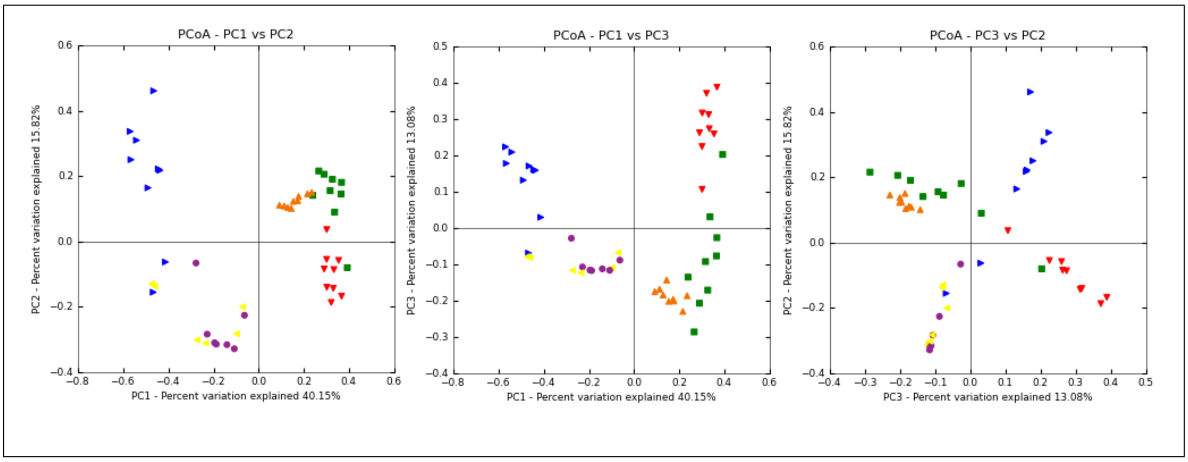


Figure 3.16: This figure shows the first three principle co-ordinates plotted against each other to demonstrate the dissociations between the treatment groups. Green is the lean control group, red is the obese control group, blue is the Gram-positive antibiotic treatment group, purple is high dose Gram-negative antibiotic treatment group, yellow is low dose Gram-negative antibiotic treatment group and orange is the OFS dietary supplement group.

Aside from the drammatic reduction in diversity seen in the antibiotic treatment groups. There is a more subtle observation in the OFS treatment group. As a control group the expectation was that the microbiome would remain the same. Notably the OFS group have an increased representation of *Bacteroidetes* and are the only group that have a consistent representation of *Verrucomicrobia* which are *Akkermansia muciniphila* at the species level and have been considered a marker for a healthy microbiome (Dao *et al.*, 2015; Bland, 2016; Everard *et al.*, 2013).

3.4 Discussion

The hypothesis for this work was that antibiotic modulation of the microbiota would have an impact on glycaemic control in a mouse model of T2DM. This study demonstrates that antibiotics can have a beneficial impact on glycaemic control. Particularly, a Gram-negative antibiotic can cause a full recovery to healthy blood glucose regulation without any notable negative implications to the physiological systems in a mouse model of T2DM. This study is one of the first to demonstrate antibiotics being used as a therapeutic intervention via modulation of the microbiota.

This study also showed trends towards beneficial impacts for all treatment groups, including Gram positive antbiotics treatment, low dosage Gram negative antibiotic treatment and OFS supplements with improvements in weight, changes in endocrine signalling pathways and remodelling of the microbiota. These groups showed some indications of improvements towards a healthy status in glucose, insulin and GLP-1, however, these trends were not supported by a large enough sample size to draw any conclusions.

Host-microbiota interactions in T2DM have been clearly shown to contribute to the disease model for T2DM as changes in the microbiome have caused drammatic changes in diagnostic markers. Turnbaugh *et al.* (2009) previously showed a consistent difference in the Bacteroidetes:Firmicutes ratio between lean and obese subjects. In the Turnbaugh *et al.* (2009) study they found obese subjects had an increased representation of Firmicutes and a reduced diversity which is also observed in the obese group and lean group in this study. Interestingly the OFS supplement also worked as a prebiotic and increased the Bacteroidetes representation and overall diversity at the same time as increasing weight loss. However, the antibiotic treatment shows a contradictory trend. When Firmicutes are removed, which is the case for the majority of the A_{500} group there is limited evidence for improvements. Also in the group B_{500} group the bacterial community is almost exclusively Firmicutes and has very little diversity yet this is the group where we observe a recovery to healthy glycaemic control. This range of outcomes demonstrates that the microbiota has a significant contribution in the disease model and can be a key part of the treatment model but further studies are required to establish the low level mechanisms of host-microbiota interactions that will improve understanding of the disease model so treatments can be effectively designed with a clearer biological foundation.

3.4.1 Treatment model for high dosage antibiotic B treatment

Antibiotic B is a third generation cephalosporin that disrupts cell wall biosynthesis. The antibiotic works by inhibiting cross linking in the cell wall by binding in the place of D-alanine-D-alanine (Hayes & Orr, 1983). This inhibition halts cell division and interupts ongoing functions of organelles.

Group B_{500} showed significant changes from the diet-induced obesity control group in glucose, insulin, GLP-1, weight, body fat, triglycerides, albumin, hHBA, PYY, GIP, amylin and leptin. The most likely mode of action for the treatment given this data is as a result of the microbiota remodelling.

The microbiota has been proposed to contribute to T2DM by increasing the capacity and efficiency of substrate uptake from the gut, increasing the abundance of high calorie substrates, decreasing mucosal integrity and increasing endotoxaemia, Figure 3.3. Endotoxaemia increases inflammation and causes a shift in the metabolic status. This exaggerates low grade inflammation by increasing Lipopolysaccharides (LPS) in the blood which is a substrate derived from the microbiota that is higher during periods of microbiota dysbiosis (Cani *et al.*, 2008; Muccioli *et al.*, 2010; Cani *et al.*, 2009, 2007). PYY and GLP-1 are released from the mucosa and are linked to mucosal integrity. Group B_{500} and group B_{50} showed significantly increased levels of both biomarkers compared to other treatment groups. This may indicate that mucosal integrity was increased by the antibiotic treatment, which in turn would have impacted positively on LPS levels. Therefore the remodelling the microbiota the antibiotic treatment likely had a positive impact on the host diabetic status.

Group B_{50} shared many of the same trends except was significantly different from group B_{500} in glucose, insulin, weight, body fat, triglycerides, amylin, leptin, hBHA, HDL-C, albumin and total protein. This raises the question of why there was a difference in the dose responses. The microbiome of groups B_{50} and B_{500} are essentially identical so if the microbiome modulations was causing improvements it should have been indiscriminant across the groups. Equally the endocrine signalling differences thought to be responsible for increased insulin sensitivity are essentially the same. The main differences between dose response groups are insulin sensitivity markers, weight loss and plasma proteins.

One explanation for dose response differences is a slower recovery in the low dosage group. There are signs that the phenotype of the low dosage group was improving. This may indicate that the low dosage antibiotic remodelled the microbiome at a different rate, which wouldn't have been measured given the single sample metagenomics analysis.

Another explanation could be reduced biomass in the high dosage treatment group. If the high dose antibiotic caused reduced total biomass the proportional measurements shown in this analysis would not have accounted for overall biomass and this could be an important factor. . Either way the underlying mechanisms of the treatment are likely to involve increased mucosal integrity and therefore lower mucosal permeability and a reduction in endotoxaemia.

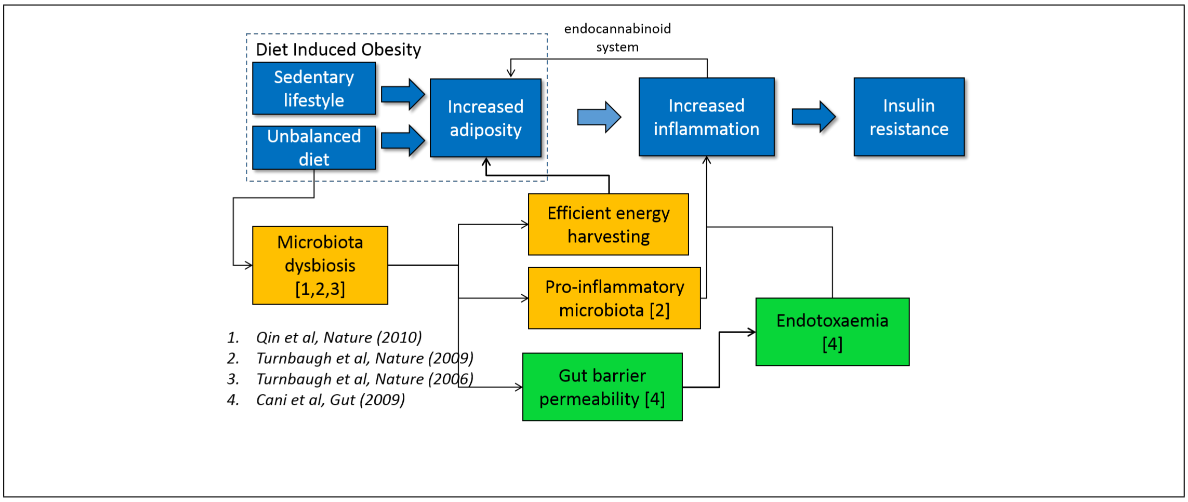


Figure 3.17: **Proposed treatment mechanisms for T2DM.**.

3.4.1.1 Treatment model proposal - Follow up studies

The microbiota is an established contributor to metabolic status through low grade inflammation. This study shows that antibiotics treatments can cause an improvement the metabolic status of a diabetic mouse and modulate the microbiota. However, there is still limited clarity in the exact roles of the microbiota consituents and the host response mechanisms related to microbiota changes in the disease model. To clarify the change of metabolic status caused by antibiotics a robust control for healthy mice with antibiotic treatments and an OFS treatment combined with the Gram negative antibiotic treatment would be ideal. Further, to clarify the inflammation response implied in the microbiota treatment model a measure of LPS in the blood serum could provide a better response curve. This proposal was suggested by reanalysing frozen serum samples although not persued by the study sponsor GSK.

One aim of this work is to bring the empirical models of diabetes closer to the first principal models in our understanding of the condition. The study design largely limitated analytics to binary comparison rather than detailed modelling and limited the robustness of mechanism modelling. GSK are due to run follow up work on this study with rat models although chose not to complete follow up analyses on existing samples from the mouse study.

3.4.2 Treatments for type II diabetes mellitus

The most common treatments for T2DM aim to reduce spikes in blood glucose often by giving insulin supplements that increase glucose uptake by bodily tissues. Some treatments also aim to reduce internal glucose release and increase internal insulin release; sulfonylureas and meglitinides therapies can be used to increase beta cell insulin secretion, metformin and biguanides inhibit hepatic glucose output and pramlintide and other amylin analogues slow digestion of carbohydrates by slowing gastric emptying. These treatments are well established therapies for managing the relatively short term symptoms of T2DM. Recently a series of therapies have also been approved that target insulin sensitivity. GLP-1 analogues such as exendin-4 increase insulin sensitivity which results in increased insulin release and increased amylin release for slower gastric emptying and reduced glucagon release as well as lowering of the hyperglycaemic threshold. Related drugs also inhibit the GLP-1 degradation enzyme DPP-4 to increase the GLP-1 half life.

Treatments for T2DM have improved over the last decade, however, pharmaceutical interventions remain mostly focused on endocrine mimicry in the direct glucose feedback loop. In contrast to pharmaceutical interventions, lifestyle interventions, such as dieting and exercise, are also commonly adopted to treat T2DM symptoms (Taylor, 2013). There is a major challenge in understanding how lifestyle changes are impacting recovery and how lifestyle changes impact on the efficacy of pharmaceutical treatments. This study shows that antibiotic interventions are capable of modifying the insulin sensitivity of a diet induced T2DM mouse model outside of the primary pharmaceutical targets. This demonstrates a target that would be complementary to established treatments and could progress the scientific basis of lifestyle interventions.

3.4.2.1 Microbiota manipulation as a new therapeutic intervention

Microbiome modulation has been studies by four major approaches; probiotics, prebiotics, mixed microbial transplantations, dietary interventions. This study is one of the first studies to target microbiota manipulation using antibiotics as a potential disease treatment.

This chapter clearly shows microbiota-targeted antibiotic interventions are an underexploited therapeutic strategy. With a more detailed understanding of the microbial communities involved in disease it is conceivable that antibiotics could be a valuable addition to the pharmaceutical options for T2DM and therapeutic strategies involving microbiota targeted treatments are a realistic possibility for multiple chronic diseases. However, there are major limitation in translational microbiome studies from mouse microbiota as the gut and inflammatory system is not equal to that of a human. Also it is not clear how long the beneficial effects would last after the antibiotic treatment period or after resistance had evolved in microbiota communities. Furthermore modification of the microbiota may leave the patient susceptible to other diseases and conditions. To address these issues and progress towards an acurate microbiota disease model that would be reliable enough for designing therapeutic strategies, a deeper understanding of the mechanisms underlying host-microbiota interactions is required.

In this chapter I have shown the beneficial effects of anitbiotics modulation of the microbiome and generated a basic disease model involving the microbiota. To develop the disease model to understand the mechanisms of host-microbiota interactions it is crucial to understand what is happening at the mucusal interface which would enable a more precise definition of a disease model and therefore a solid basis for systematic therapuetic interventions.

Chapter 4

Protein and Peptide Associated Glycan Metabolism in the Gut Microbiota

4.1 Introduction

Glycans, glycosidically linked monosaccharide compounds, are large complex molecules that form a major energy source for the gut microbiota and are critical in establishing and maintaining healthy host-microbiota interactions (Koropatkin *et al.*, 2012; Kaoutari *et al.*, 2013; Tailford *et al.*, 2015). Many gut bacteria are adept at foraging glycans so the availability of particular glycan substrates and metabolites in the gut plays an important role in shaping the biodiversity of the microbiota (Pudlo *et al.*, 2015a; den Besten *et al.*, 2013; Pokusaeva *et al.*, 2011; Garrido *et al.*, 2011; Salyers *et al.*, 1977).

Protein and peptide associated glycans are an emerging area of interest due to their complex metabolism and their proposed role in microbial food chains, niche microhabitats and the biogeography of the gut microbiota (Costello *et al.*, 2009; Donaldson *et al.*, 2016; Zhang *et al.*, 2014). Genomics data has provided many insights into the mechanisms involved in processing glycans in the gut through carbohydrate targetting enzymes (cazymes) and polysaccharide utilisation locus (PUL)s which coordinate the systematic digestion of glycans (Tailford *et al.*, 2015; Kaoutari *et al.*, 2013). A recent addition to this understanding emerged from a series of studies that identified three proteases with CBMs that target mucin (Nakjang, 2011; Nakjang *et al.*, 2012; Ndeh, 2013; Navarro-Garcia *et al.*, 2010). These were the first proteases identified with a CBM configuration typically found in cazymes. These studies also suggest that there could be a range of proteases, similar to the ones that were characterised, that can cleave the core peptides and therefore change the model of protein and peptide associated glycan metabolism by breaking down the macromolecular structure instead of digesting the side chains in series and have a broad impact on host-microbiota interactions. The purpose of this chapter is to explore abundance and diversity of proteases targeting protein and peptide associated glycans in metagenomics data and evaluate the scope for metagenomics data to provide a comprehensive view of host-microbiota interactions that could be used to design therapeutic strategies and relate single gene research into microbiome level disease models.

4.1.1 Glycans in the gut microbiota

There are three sources of glycans in the gut; diet derived, microbiota derived and host derived. Each type plays a key role in maintaining gut physiology and each type has different characteristics that can provide selective pressures for specific niches (Koropatkin *et al.*, 2012). Over time glycan availability fluctuates and forms one of the environmental challenges that can disturb the microbiota. The biodiversity impact of glycans is largely a balance of short term adaptations to glycans in the lumen mainly originating from the diet and long term adaptation to microhabitats such as on mucosal surfaces where the majority of glycans are host derived (De Filippo *et al.*, 2010; den Besten *et al.*, 2013; Walter & Ley, 2011; Zhang *et al.*, 2014). These microhabitats are thought to lay the foundations for microbial food chains and provide longevity to microbiota communities.

Dietary glycans make up a large proportion of the microbiota energy source and are highly available. Many glycans are indigestible by human enzymes and can therefore be found in high concentrations in the gut lumen (den Besten *et al.*, 2013). Typically these glycans constitute N-linked glycosylated structures from plant sources and O-linked glycosylated structures from animal sources. Early studies recognised the role of the microbiota in producing the vast array of enzymes required to breakdown the complex dietary glycans into useful substrates for the host (McNeil, 1984). Research has since developed in this area and the diverse roles of cazymes in glycan metabolism and short chain fatty acid production is a highly active and expanding field of research, as demonstrated by the growth of CAZy as one of the most cited databases in bioinformatics literature (den Besten *et al.*, 2013; Lombard *et al.*, 2014; Cantarel *et al.*, 2009; Terrapon *et al.*, 2015).

Microbiota derived glycans mainly consist of peptidogylcans from the cell wall. Peptidoglycans form 90% of the dry weight of Gram positive bacteria and 10% of the dry weight of Gram negative bacteria. They are made up of alternating amino sugar units of N-acetylglucosamine and N-acylmuramic acid cross linked with short peptides (Brown *et al.*, 2015). The regular alternating patterns provide a rigid three dimensional crystal lattice. Cell wall synthesis and degradation can be considered as part of cell division, a competition mechanism and as food source foraging. Peptidoglycan synthesis and degradation in the cell cycle and peptidoglycan degradation as a competition mechanism are areas of research with a long history related to the development of antimicrobial compounds (Aminov, 2010). However, metagenomics has only recently been used to search for novel compounds relating to peptidoglycan degradation (de Castro *et al.*, 2014) and, as far as I am aware, within microbiome research there are only studies investigating antibiotic resistance in this area (Penders *et al.*, 2013; Bengtsson-Palme *et al.*, 2015). Peptidoglycan degradation for food source foraging has shown bacteria can target fungal glycoproteins as a key food source (Temple *et al.*, 2017). However, no studies have investigated the range of bacterial enzymes for foraging bacterial peptidoglycans in metagenomics studies.

Host derived glycans in the gut mainly consist of mucins and other mucosal secretions. 80% of mucin mass is from glycosylated side chains that are attached to a peptide core by O-linked glycosylation (Tailford *et al.*, 2015). The glycan biomass in mucins also contributes to the macromolecular structure of the mucus bilayer which protects the epithelium and is an essential factor in maintaining a healthy gut physiology. Mucins are a dense molecular energy source that facilitates microbiota colonisation in the early stages of life and provides a stable energy source through fast and feast cycles (Mueller *et al.*, 2015). Mucins are one of the key drivers behind the symbiotic host-microbiota relationship and when the host-microbiota interactions breakdown at this interface multiple disease mechanisms can be triggered (Everard *et al.*, 2013).

4.1.2 Glycan metabolism

The capacity for polysaccharide metabolism in the human genome comprises a total of 17 glycosyl hydrolase genes (Kaoutari *et al.*, 2013). Some microbiota organisms alone encodes hundreds of cazymes and the microbiome in total encodes thousands of genes for polysaccharide metabolism (Kaoutari *et al.*, 2013; Lombard *et al.*, 2014). The microbiota also represents a dynamic system that is adaptable to its environment through gene regulation and sub community population size fluctuations. The result of this dynamic adaptability is that the diversity of mechanisms for glycan metabolism is broad and central to microbiota ecology.

Bacteroidetes is by far the most representative Gram negative phylum in the gut (Qin *et al.*, 2012) and has a well characterised strategy for glycan foraging (Koropatkin *et al.*, 2012). Bacteroidetes employ PULs, clusters of carbohydrate active molecules that work together, to target classes of glycans. Typically a PUL is co-regulated and coordinates the capture, transport and degradation of glycan substrates from extracellular space to the intramembranous space to be metabolised before monosaccharides are imported into the cytoplasm (Koropatkin *et al.*, 2012).

Bacteroides thetaiotaomicron (BT) is a particularly well studied Bacteroidetes organism that encodes 88 PULs which comprise 18% of its genome (Terrapon *et al.*, 2015). BT is an example of a 'generalist' that can degrade both diet derived and host derived glycans through selective activation of multiple PULs. BT was first studied as the host of the starch utilisation system (Sus) which is an 8 gene cluster that targets starch and was the first PUL discovered before many more Sus-like systems were characterised (Chaudet & Rose, 2016).

PULs have multiple characteristic markers that make them identifiable in genomics data (Terrapon *et al.*, 2015). Noteably, they contain a set of genes that are physically linked, including adjacent susC-like and susD-like binding proteins as well as cazymes and often regulatory genes on the opposite DNA strand that mark the bounds of the PUL. Carbohydrate active enzymes are obviously critical to PULs although their position within the PUL can vary to facilitate flexibility of regulation. Research into the function of cazymes is stored in CAZy and has rapidly become one of the most valued resources summarising glycoside hydrolases, glycosyltransferases, polysaccharide lyases carbohydrate esterases and carbohydrate binding domains.

Hundreds of PULs have been identified in the microbiome and their mechanisms are central to survival (Tailford *et al.*, 2015). Importantly though there is a high metabolic investment required for producing glycan degrading enzymes and often microbes have a preference for specific substrates. Some 'generalists' are capable of metabolising multiple glycans and can choose their preferred substrate whereas other 'specialists' target only specific glycans. Many microbial species and communities contribute to this system of selective specificity and can form spatial glycan microhabitats (Pokusaeva *et al.*, 2011; Garrido *et al.*, 2011; Salyers *et al.*, 1977). Furthermore, it has been suggested that microhabitats at the interface of host-microbiota interactions are central to mucosal health. Some species, for example *Akkermansia muciniphila* (Everard *et al.*, 2013), are known to digest glycans and have been identified as biomarkers of a healthy microbiota due to their inferred role in maintaining microbial food chains. They can also be used as an indicator of mucosal health and permeability. However, the mechanisms of glycan metabolism that distinguish these healthy microhabitats and health related species remains unclear.

4.1.3 The role of proteases in protein and peptide associated glycan metabolism

Protein and peptide associated glycans are particularly complex molecules that incorporate complexity in their macromolecular structure as a mechanism of generating a stable macromolecule. Due to this macromolecular complexity often only highly specific enzymes can degrade them and these enzymes have mainly been studied through genes identified in isolated strains. A working hypothesis is that due to the central role of protein and peptide associated glycan in host-microbiota interactions a broad range of specific enzymes, exemplified by the PF13402 sequences studied by Nakjang *et al.* (2012), are able to effectively break down the macromolecular structure of protein and peptide associated glycans of various sources including host, microbiota and diet. This study attempts to identify protein and peptide associated glycan targeting protease diversity in metagenomics data.

Mucus is an example protein associated glycan that is highly available in the gut. The molecular structure of mucus is a mesh of mucin molecules, Figure 4.1, which are mainly produced by goblet cells in the epithelium. Mucins come in four varieties in the gut; MUC1, MUC4, MUC2 and MUC5b with MUC 2 being the most prominent (Allen & Pearson, 2000; Pearson & Brownlee, 2010). The mucin molecules are attached to each other by di-sulphide bridges formed between the cysteine rich area at the ends of each core peptide and share gel forming regions due to hydrophobicity in the glycosylated side chains. Some MUC 1 and MUC4 molecules are also have transmembrane domains attached to the mucosa. All mucins are configured with a protein core and multiple glycosylated side chains where the density of the side chains is highest in the middle of the protein core although composition of mucin molecules can vary both in the chemistry at the core O-glycosylation links and peripheral units of the glycosylated side chains and physically in the configuration and density of side chains and polymerisation (Hanisch, 2001; Tailford *et al.*, 2015).

BT is an example of a species that can target mucins and carries 18 PULs that are upregulated in the presence of mucus (Pudlo *et al.*, 2015b). Mucus metabolism can be considered in three sections; reduction of the disulphide bonds, digestion of the side chains and cleaving of the protein core. Reduction of the disulphide bonds can be caused by multiple factors and modern diets regularly lead to increased reducing agents in the gut and an increase in specific sulphate-reducing bacteria (Rey *et al.*, 2013). Digestion of the glycosylated side chains can further be split into at least three enzymatic jobs. Selective removal the peripheral unit from the epitope layer, digestion of the glycosylated backbone and removal of the O-linked GalNac attached to the peptide core (Tailford *et al.*, 2015). Most PULs active to mucins are adept at systematically digesting side chains, however, the terminal unit can require specific enzymes. Until recently the peptide core was considered to be cleaved last due to it being protected by the dense glycosylation. Studies by Ndeh (2013) have now shown BT can access and cleave the protein core before the mucin has been fully degraded with a specific combination of CBMs in the gene configuration. This new model for mucus digestion is a shift in understanding of how the macromolecular structure of mucus is metabolised. However, currently there are only few example of such protease genes.

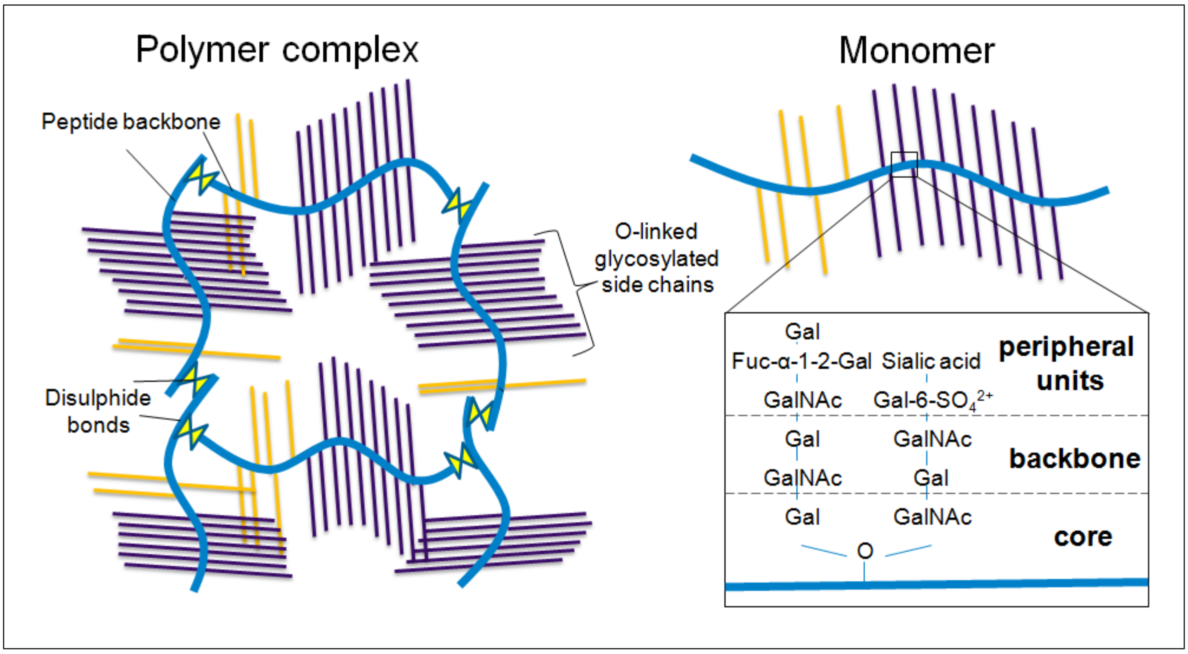


Figure 4.1: Molecular structure of mucin on the right showing the three regions of the glycosylated side chains and the corresponding macromolecular structure of mucus on the left showing the mesh structure created by the disulphide bonds.

4.1.4 M60-like zinc metalloprotease

The M60-like zinc metalloprotease (PF13402) is an enzyme that was initially identified through the process of annotating *Trichomonas vaginalis* (Hirt *et al.*, 2011). The *T. vaginalis* gene was initially uncharacterised although through a process of BLAST alignment and subsequent profile searches homologue were consistently found in the context of mucosal associated bacteria. Through a process of phylogenetic analyses, domain searches and laboratory experiments the functional characteristics of a select few examples of the protease were determined as well as the specificity for mucin degradation when associated with CBMs (Nakjang, 2011; Nakjang *et al.*, 2012; Ndeh, 2013).

Of particular interest to mucus metabolism there are instances of the M60-like zinc metalloprotease enzyme that pair with CBMs (CBM51 or CBM37). This combination is proposed to allow the enzyme to bind to GalNac glycosidic bonds between the glycosylated side chain and mucin peptide core which brings the protease within reach of the peptide core. The endopeptidase protease is then able to cleave protein core of mucin and breakdown the macromolecular structure. The gene is also found with other CBMs configurations that are proposed to allow it to bind variants of glycosidic side chain bonds of mucins and other glycoproteins.

The original study by Nakjang *et al.* (2012) investigated the gene copy number per strain of the mucinase genes in genomics data highlighting the higher copy numbers of the gene in strains isolated from the gut showing a specific adaptation. This was one of the first studies to show a protease with a CBM domain as a specific adaptation for mucus metabolism. A further studies by Ndeh (2013) demonstrated a competitive advantage for BT4255 through this specific adaptation. This study takes a similar approach to using gene occurence and configuration data on a broader scale to explore specific adaptations to microbiota ecology that are central to host-microbiota interactions.

4.1.5 Hypothesis

Microhabitats in the gut are critical to understanding the host-microbiota and microbiota-microbiota interactions. This study investigates the occurrence of bacterial proteases in metagenomics datasets particularly focusing on proteases with specific adaptations to target protein and peptide associated glycans that are thought to form the basis for microhabitats as exemplified by sequences found using PF13402. This study analyses of the presence, diversity, and frequency of CBMs on putative protease genes to identify a series of candidate proteases targeting protein and peptide associayed glycan metabolism.

The hypothesis for this chapter is that a broad range of (uncharacterised) proteases pair with CBMs to target protein and peptide associated glycans.

4.2 Methods

The analyses covered in this chapter explore the protease complement of metagenomics datasets, with a focus on proteases targeting protein and peptide associated glycans. Initially, following methods in Section 4.2.2, a protein family (PF13402) containing a recently characterised mucin targeting protease is searched for using profile and local alignment search algorithms to establish the natural gene frequency and diversity across healthy individuals. Figure 4.2 summarises the data sources and three step analysis pipeline which is described in detail through the methods section. Step 1 following methods in Section 4.2.3, the entire MEROPS collection of protease signatures is then searched for in the set of predicted genes from metagenomics datasets to identify the frequency and diversity of *all* putative protease genes from the MEROPS database in the microbiome. Step 2, following methods in Section 4.2.3, the predicted genes containing hits from the MEROPS database are then used to search for CBMs from the dbCAN database. The outcome is a list of all MEROPS proteases with a CBM domain and a matrix of CBM-protease pair occurrences. Step 3, these final CBM-protease genes of interest are then annotated with gene domain configuration, number of known hits, number of known distinct genes encoding this configuration and known annotations for taxonomy and function.

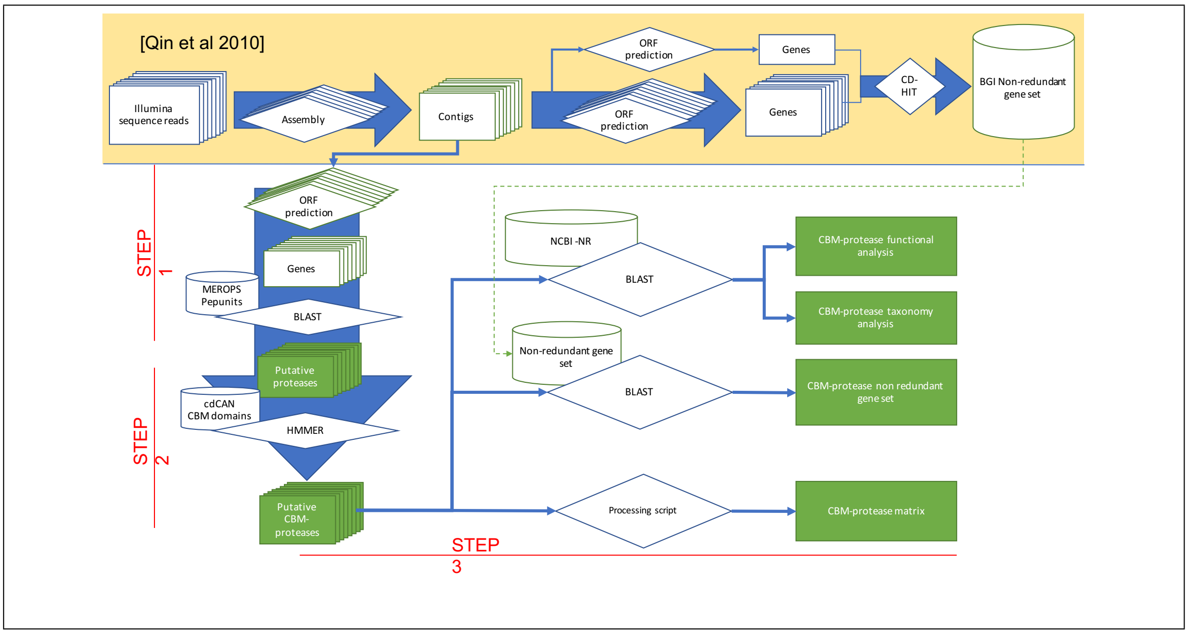


Figure 4.2: The analysis pipeline shows main analysis steps in extracting and annotating the CBM-protease pairs as well as the analysis pipeline from the Qin *et al.* (2010) study that generated the input data for this study and the gene catalog that is used in stage 3 of the analyses. Squares represent flat files in the anlaysis pipeline. Diamonds represent processes and cylinders represent databases. Shaded files represent results files that are available to download on the project webpage. Arrows show the flow data through the process steps and intermediary files.

All code is linked from the text and hosted in the publicly accessible repository linked below. All results files are hosted at the project webpages also linked below and all intermediary files are available on request.

http://github.com/mcollison/CBM-protease-analyses

http://homepages.cs.ncl.ac.uk/matthew.collison/research/CBM-protease-pairs

4.2.1 Metagenomics data source

Metagenomics datasets of the gut microbiome from faecal samples of 85 healthy subjects were taken from the Qin *et al.* (2010) as assembled contigs as shown in Figure 4.2. This data was chosen as one of the largest cohort metagenomics studies at the time of analysis and has been used as a benchmark dataset in many studies since. The data was downloaded as assembled contigs from a web server hosted at EMBL `http://www.bork.embl.de/~arumugam/Qin_et_al_2010/`. The original experiment used the Illumina-GA platform for sequencing and SOAP-denovo (Li *et al.*, 2009) with parameters -m 3 for assembly. For full experimental methods see the primary publication Qin *et al.* (2010).

4.2.2 PF13402: M60-like zinc metalloprotease analyses

PF13402 otherwise known as M60-like zinc metalloprotease was identified as a protease family of interest. Fasta formatted sequences for PF13402 were extracted on 22 May 2013 from the PFAM entry PF13402, `http://pfam.xfam.org/family/PF13402`. Full sequences from the 75% representative proteome were extracted constituting 129 sequences in a fasta format. The sequences were used with both BLAST and HMMER to search for homologs in the metagenomics data.

Firstly, the PF13402 alignment extracted from PFAM was used to generate a HMM profile using HMMER version 3.1b1 for 64 bit Linux. This profile was then used to search the metagenome contigs again using HMMER. BLAST was initially run against the metagenome contigs and Prodigal predicted genes separately. The results showed the same number of hits for the BLAST searches on the Prodigal predicted genes and contigs although the Prodigal predicted genes computed substantially faster and allowed the complete genes to be extracted. Therefore Prodigal genes were used as targets for all BLAST results shown in this chapter. BLAST and HMMER hits were filtered at an *e* value of 0.01, as justified in results Figure 4.3.

The BLAST and HMMER results were parsed, integrated with the metagenomics metadata and output into csv files using custom Java scripts. A spreadsheet was then sorted and visualised as heatmaps using Microsoft Excel.

4.2.3 Protein and peptide associated glycan targeting proteases

This study aims to identify proteases targeting protein and peptide associated glycans in metagenomics studies through a 3 step process.

Step 1: Identifying all putative proteases in the gut microbiome metagenomics samples

metaProdigal version 2.0 (Hyatt *et al.*, 2012) was used to predict genes from the 85 metagenomics datasets described in Sections 4.2.1. A bash script was used to coordinate this process on a local HPC server. The result was a library of predicted genes for each metagenome from 85 healthy individuals.

MEROPS (Rawlings *et al.*, 2008) is a manually curated database for protease genes with entries clustered by clan and by functional family. The database was downloaded in the pep-unit format which lists the active domain of each protease gene in the database. Pep-units were selected for their specificity to reduce false positive hits in alignment searches. These pep-units were used as query sequences to search the libraries of Prodigal predicted genes from the metagenomics datasets. BLAST version 2.2.2 (Altschul *et al.*, 1990) was used to achieve this and a bash script coordinated the analyses to firstly convert the predicted genes into a BLAST database using formatdb and then produce the BLAST report summarising the BLAST hits for each MEROPS pep-unit in each metagenome. The results were parsed with a Java program to extract only the top hit for each predicted gene with an *e*-value less than 0.01 and produce a refined table of filtered BLAST results. The filtered results were then processed the with another Java program to extract and write fasta sequences to file for all of the putative proteases in each of the 85 metagenomes.

Finally a Java program was used to generate a summary csv matrix of each MEROPS entry with a hit documenting the MEROPS ID, the clan ID, the family ID, the corresponding MEROPS description and the number of hits per metagenome. This table was then manipulated as a spreadsheet using Microsoft Excel to group MEROPS families and clans in a master matrix which provides the data for the results section 4.3.2.

Step 2: Identifying CBM domains on putative proteases.

dbCAN (Yin *et al.*, 2012) is a database of profiles for carbohydrate active molecules, largely replicating CAZy, that contains sequence data and profiles for each carbohydrate active domain. Using HMMER version 3.1 (Eddy, 2011) and the hmmpress command, the profiles for carbohydrate active molecules were converted from dbCAN release 4 into HMMER formatted profiles. These profiles were then used in a bash script to search for signatures of carbohydrate active molecules within the set of putative proteases identified in step 1. The HMMER results reports for each metagenome were then converted into dbCAN HMMER hit tables using a another bash script with criteria of the alignment being over 80 amino acids, and *e*-value of less than 1e-5, otherwise use *e*-value less than 1e-3 and check the covered fraction of HMM profile is over 30%.

Next a Java program was developed to use data from the dbCAN HMMER hit tables and the filtered BLAST hit tables to extract each putative protease sequences that contained a CBM. This is the first point that fasta sequences were grouped into files based on the MEROPS family of the protease rather than the metagenomics dataset in which they originated. Also in this step the results were filtered specifically for CBM domains as opposed to including all dbCAN domains.

One artefact of identifying all sequences with a CBM and a MEROPS ID was that sequences with multiple CBMs were recorded multiple time. This redundancy was removed using a small Java program to produce non-replicated fasta results. Annotation details were then added to the fasta files using this Java program to pull coordinates for the CBM domains from the dbCAN hit tables and proteases domains from the filtered BLAST hit tables. The results were a list of all of the annotated fasta sequences for all MEROPS proteases with a CBM. Note all dbCAN domains (including glycoside hydrolases GH, glycosyltransferases GT, polysaccharide lyases PL and carbohydrate-esterases CE) have been annotated in these fasta files when they are present on a sequence that contains a CBM.

Finally a Java program was developed to take data from the dbCAN HMMER hit tables along with the filtered MEROPS BLAST hit tables and record all of the individual occurrences of a CBM and with a MEROPS proteases in each metagenome and wrote them to two column csv files. Another Java program was then used to combine this occurrence data for each metagenome into a master hit matrix. Microsoft Excel to was used to calculate the matrix at the family level which is shown in the results section 4.3.3.

Step 3: Annotating gene configurations, gene function and taxonomy.

The annotated list of all predicted protease genes with a CBM domain identified in the metagenomics datasets contains homologs and partial genes. To remove this redundancy and map partial genes to complete gene BLAST version 2.2.2 (Altschul *et al.*, 1990) was used to convert the BGI non-redundant gene library into a BLAST database and compute alignments to the fasta sequences extracted from the metagenomics data. The resulting BLAST tables map metagenomics genes to the gene catalogue genes.

These BLAST results tables from step 3a were then combined with the annotations from fasta sequences generated in step 2c using a Java program to create a master table for each MEROPS family. Note there was a threshold of over 95% coverage for the BLAST results to be included. For each MEROPS family three pieces of information were included in the master table along with the number of occurrences of each and a spread of occurrences across the metagenomics data; the MEROPS family, all CBM gene configurations that have been observed with that protease family, and the gene catalogue entries that represent each gene configuration. The total number of occurrences and the number of distinct genes from the gene catalogue was recorded for each gene configuration, in files labelled v1 and included in Appendix 6.2, and the full spread of occurrences across the metagenomics samples were recorded in files labelled v2. These files were then modified and concatenated to produce results figures using Microsoft Excel.

Finally another BLAST alignment was run with the fasta sequences extracted from the metagenomics data against NCBI-NR and the BLAST results were processed using a Java program to produce annotation files mapping each putative protease to the gene catalogue gene and any known gene in NCBI NR so each gene taxonomic origin and molecular function could be found.

4.3 Results

The following results explore the protease complement of the gut microbiome in faecal metagenomics datasets from 85 healthy individuals, focusing on proteases targeting protein and peptide associated glycans. Initially the frequency and diversity of the protein family PF13402, known to include a mucin targeting protease gene, was analysed. Then the frequency of occurrence of the entire MEROPS gene set was analysed. Finally CBM profiles were used to identify predicted proteases that bind glycans. These protease-CBM pairs were then grouped into functional families and analysed as genes of interest.

4.3.1 PF13402: M60-like zinc metalloprotease

Figure 4.3 shows the frequency distribution of *e* values for non-redundant BLAST hits. 0.01 was selected as a threshold that would maximise the true positives, potentially at the cost of including false positives, to maximise the possibility of novel gene discovery in subsequent steps of analysis and to account for the divergent evolution of genes in the microbiome environment.

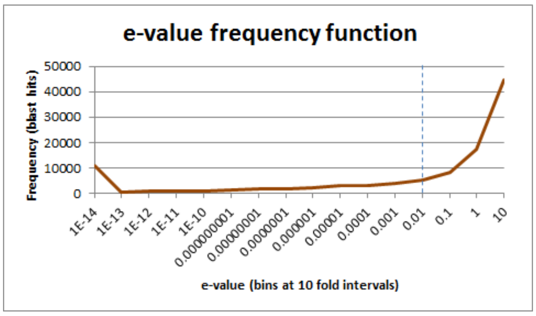


Figure 4.3: Frequency plot showing the frequency of e-value scores for PF13402 BLAST alignments against predicted genes from metagenomes. Data was binned at 10 fold intervals and the axis shows the upper bound of the bin. For example a BLAST hit with e-value of 0.005 would be included in frequency score for the 0.01 bin. A threshold of 0.01 was selected, as shown by the dashed line, to retain as many true positives as reasonably possible.

The results show an average of 11.8 and 10.5 hits for PF13402 per microbiome for BLAST and HMMER respectively, with a standard deviation of 8.8 and 9.5, a maximum of 40 hits per microbiome and a distribution skewed towards zero, Figure 4.4B. The correlation between HMMER and BLAST is shown to be high with a Pearson's coefficient of 0.99, Figure 4.4C. This correlation shows that, within the constraints of the cutoff threshold for filtering hits and given use of the same sequences to produce the HMMER profile as used for BLAST searches, both alignment algorithms captured similar diversity of the PF13402 gene indicating very few, if any, more divergent examples of this gene were identified by HMMER that weren't included in the BLAST search results.

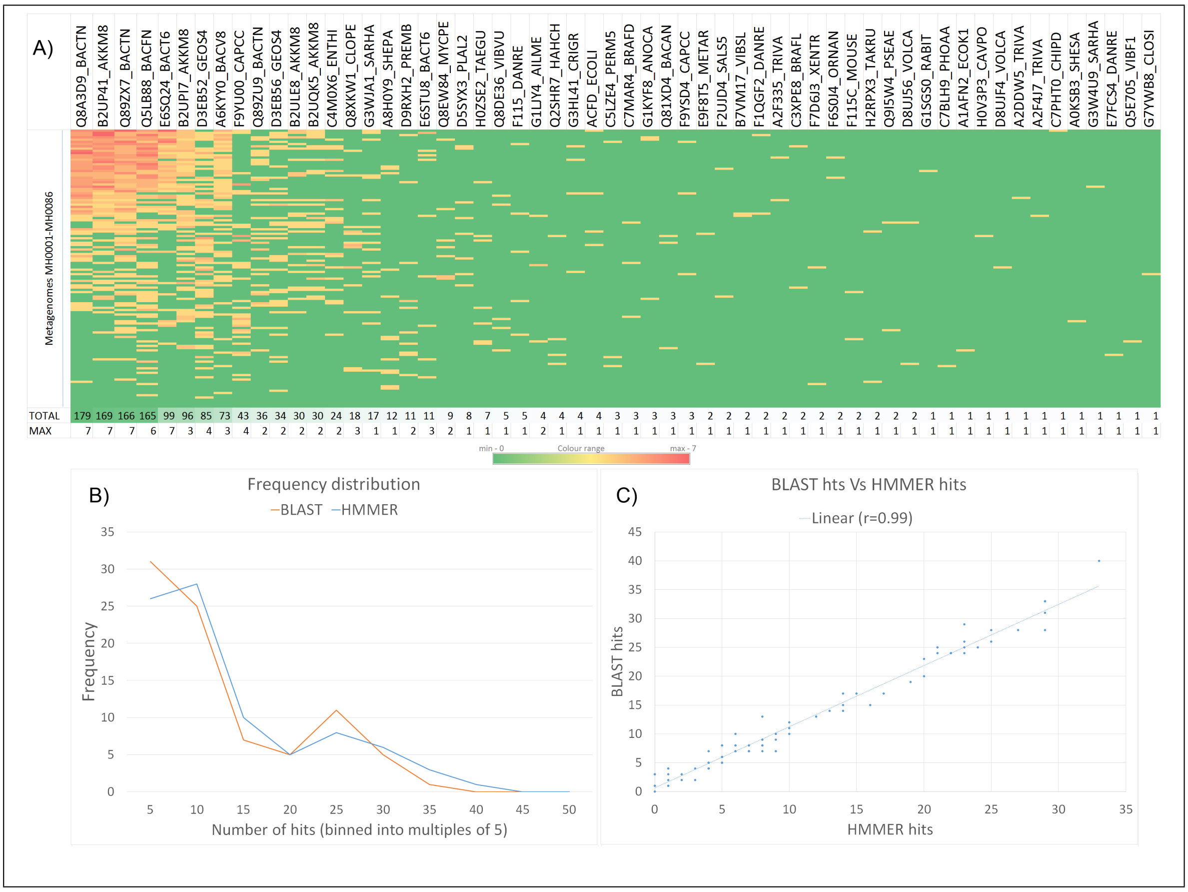


Figure 4.4: BLAST and HMMER hit counts across the metagenomes of 85 healthy individuals for 129 (58 of which had hits) M60-like zinc metalloprotease (PF13402) sequences. A) shows BLAST hits in a heat map for individual sequences. The PF13402 genes are sorted by the most popular left to right and the microbiomes are ordered by those with the most overall hits. B) shows the frequency distribution of the total number hits per metagenome with values binned into multiples of 5. For example the first data point shows the number of metagenomes with less than 5 hits. C) shows the correlation between number of hits found using BLAST searches and HMMER searches.

The majority of BLAST hits are spread across 13 sequences, as shown in Figure 4.4A. The maximum number of hits in a microbiome per individual gene is 7. This shows it is more common for multiple variants of the gene to be encoded, probably encoded by different species, within the same microbiome rather than high numbers of the same species.

One of the most surprising outcomes of these results is the limited number of hits per microbiome. Given the copy number defined in the previous study (Nakjang *et al.*, 2012), up to 15 copies per genome, and the abundance and variety of strains thought to carry this gene it would be reasonable to expect the metagenmomics sample of the entire microbiome to contain substantially more copies of this gene. Clearly the metagenomics method represents a sampling technique (Zoetendal *et al.*, 2002) and sequencing technique (Nayfach & Pollard, 2015) that are incomplete and by their nature although this finding is a clear indication that the fraction represented is minimal and variable. One way to distinguish sampling and sequencing biases is comparing the number of proteases identified against the sequencing coverage.

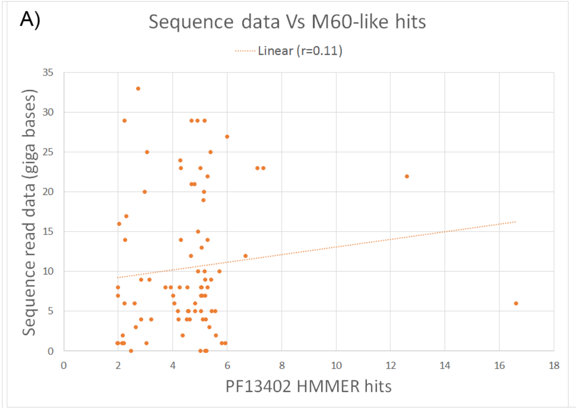
Sequencing coverage, as calculated by the amount of sequence reads multiplied by average sequence read length, and the number of PF13402 hits showed no correlation (r=0.11), Figure 4.5A. This result suggests further sequencing would not have uncovered more PF13402 hits. There could be multiple reasons for lower representation of PF13402 genes including sample collection, DNA extraction, DNA isolation, DNA sequencing and biases in genomic strain sampling and sequencing.

Despite the low numbers of hits there were multiple notably correlations between PF13402 genes including higher correlations between genes of separate species than within species, Figure 4.5B. The five most common genes showed a high and roughly equal representation in the top 30 individuals and a broad reduction in hits in the remaining subjects. Interestingly, notable genes from *Akkermansia* (AKKM8) B2UP41_AKKM8, B2UPI7_AKKM8, B2ULE8_AKKM8 and B2UQK5_AKKM8 showed some correlation but not to the same extent as the most highly scoring five genes that were from different species. This difference in correlation scores could be partially due to correlation metric showing a bias towards higher numbers although the trend seems to persist beyond this limitation. Likely the PF13402 genes are part of the pan genome so not equally represented across all *Akkermansia* strains. The implication being a microbiome having a high number of one *Akkermansia muciniphilia* PF13402 gene does not indicate it will contain a proportional number of the others. Assuming these genes are important or typical of mechanisms for establishing a microhabitat, this also highlights a limitation in the use of *Akkermansia* as a biomarker (Everard *et al.*, 2013; Roopchand *et al.*, 2015).

Interestingly the genes D3EB52_GEOS4 and F9YU00_CAPCC showed an equal spread throughout the metagenomes. In contrast to the highly correlating genes discussed above, these genes would suggest an independent function of the most correlated genes. This poses the question why are the correlated genes captured in some samples and not others.

Publications accompanying large datasets often use rarefaction analysis to demonstrate a significant level of coverage has been achieved in each study. The implication is that if we were to take a first principles approach we can estimate the average genome size and the biomass in the gut to determine the fraction of the microbiome accounted for when the rarefaction curve plateaus and the imply that the remaining gene functions and taxonomy is redundant. Often this analysis is used to demonstrate that the rate at which new species or new genes are being observed and the rate at which the proportions of species or genes is changing has plateaued and therefore additional data would

A)



B)

	Q8A3D9_BACTN	B2UP41_AKKM8	Q89ZX7_BACTN	Q5LB88_BACFN	E6SQ24_BACT6	B2UPI7_AKKM8	D3EB52_GEOS4	A6KYY0_BACV8	F9YU00_CAPCC	Q89ZU9_BACTN	D3EB56_GEOS4	B2ULE8_AKKM8	B2UQK5_AKKM8
Q8A3D9_BACTN	1.00	0.79	0.79	0.72	0.71	0.42	0.09	0.59	-0.15	0.26	0.09	0.14	0.04
B2UP41_AKKM8	0.79	1.00	0.75	0.82	0.71	0.49	-0.02	0.62	-0.14	0.09	0.07	0.33	0.20
Q89ZX7_BACTN	0.79	0.75	1.00	0.67	0.74	0.45	0.03	0.64	-0.09	0.29	0.12	0.24	0.07
Q5LB88_BACFN	0.72	0.82	0.67	1.00	0.67	0.43	0.06	0.52	-0.14	0.12	0.01	0.17	0.04
E6SQ24_BACT6	0.71	0.71	0.74	0.67	1.00	0.40	-0.13	0.62	-0.06	0.00	0.00	0.19	0.03
B2UPI7_AKKM8	0.42	0.49	0.45	0.43	0.40	1.00	-0.01	0.25	0.15	-0.01	0.03	0.40	0.41
D3EB52_GEOS4	0.09	-0.02	0.03	0.06	-0.13	-0.01	1.00	-0.07	-0.15	0.39	0.43	-0.10	0.01
A6KYY0_BACV8	0.59	0.62	0.64	0.52	0.62	0.25	-0.07	1.00	-0.09	0.05	0.02	0.24	0.11
F9YU00_CAPCC	-0.15	-0.14	-0.09	-0.14	-0.06	0.15	-0.15	-0.09	1.00	-0.22	-0.18	-0.03	-0.03
Q89ZU9_BACTN	0.26	0.09	0.29	0.12	0.00	-0.01	0.39	0.05	-0.22	1.00	0.29	-0.10	-0.02
D3EB56_GEOS4	0.09	0.07	0.12	0.01	0.00	0.03	0.43	0.02	-0.18	0.29	1.00	0.12	0.21
B2ULE8_AKKM8	0.14	0.33	0.24	0.17	0.19	0.40	-0.10	0.24	-0.03	-0.10	0.12	1.00	0.59
B2UQK5_AKKM8	0.04	0.20	0.07	0.04	0.03	0.41	0.01	0.11	-0.03	-0.02	0.21	0.59	1.00

Figure 4.5: PF13402 correlation data. A) shows sequence coverage vs number of hits per microbiome for PF13402 HMMER search. B) shows the correlation between the top 13 PF13402 sequences. The remaining sequences had too few hits for any intuitive correlation statistics.

not have any implication on the outcome of the data. However, as shown in the Figure 2.5 from the introduction, high level gene functions are conserved across species common in the microbiome so an early plateau occurs in most diversity metrics for whole microbiome diversity. Additionally we know some of the most physiologically important species are embedded in mucus layers so may be sampled at a different rate, Figure 4.5. This result indicates sampling low level functional diversity (single gene families) does not give an accurate representation of the overall community and current methods of metagenomics analyses only really support robust statistical analysis for more common genes and often those with house keeping functions.

4.3.2 MEROPS protease database

The MEROPS database is a manually curated resource dedicated to proteases (Rawlings *et al.*, 2016, 2008). Searching for MEROPS proteases in metagenomic data will give an indication as to what proteases are prevalent and an indication of what functions are important in the gut.

There were 611,948 hits for 59,193 MEROPS entries across the microbiomes of 85 healthy individuals. Each microbiome had an average of 7,039.91 protease hits and the standard deviation was 3,212.7, showing the observed occurrence of proteases varied broadly (up to ten fold difference) across the microbiomes of individuals in the cohort, Figure 4.6A. The overall number of protease hits did not correlate with the number of hits for PF13402 sequences (r=0.34). The frequency distribution shows the number of hits per MEROPS pepunit was skewed towards zero Figure 4.6C indicating the vast majority of entries had very few hits. There were 31 entries with over 1020 hits (the total number of hits for all PF13402 sequences combined) as summarised in Table 4.1.

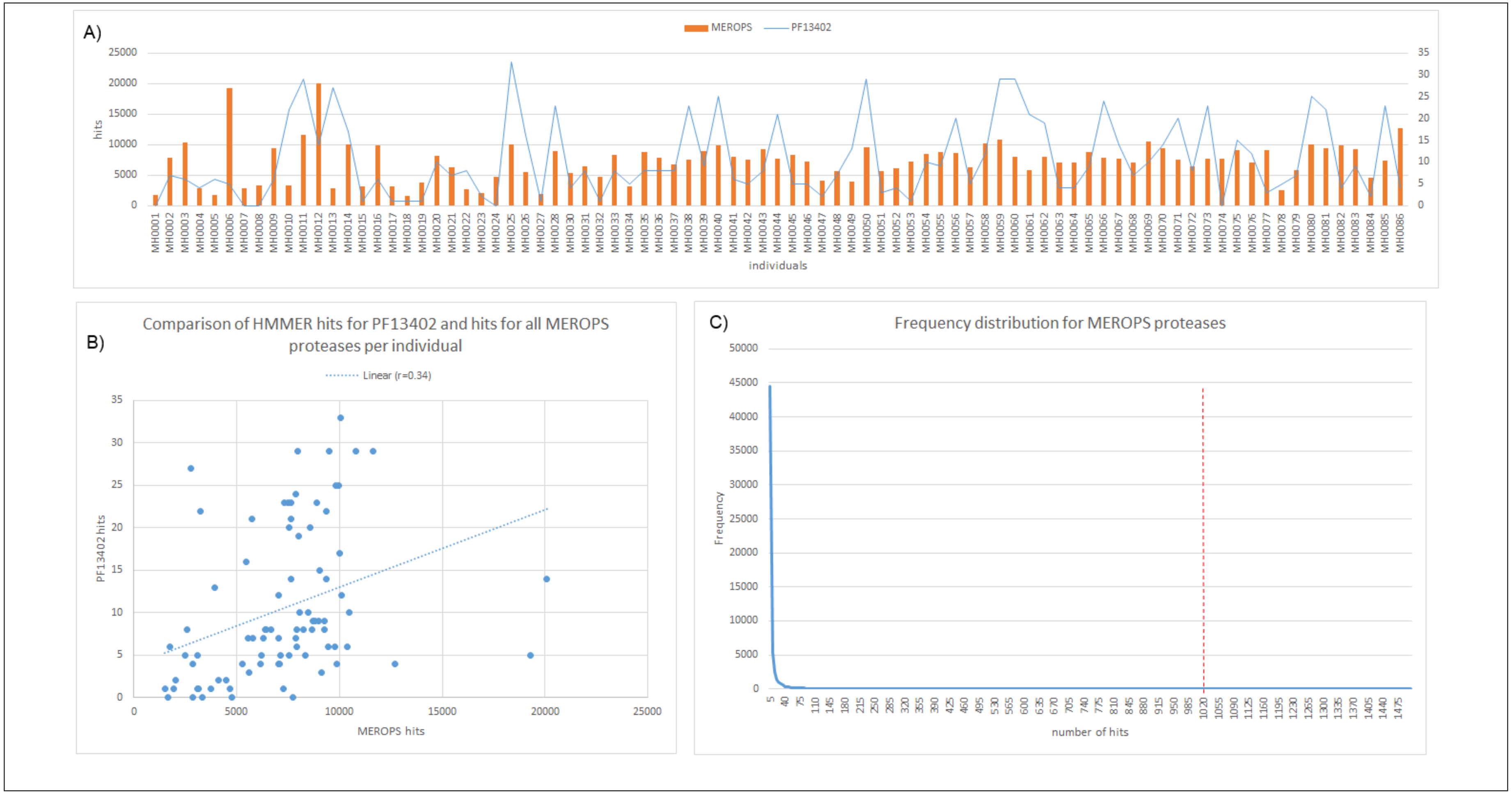


Figure 4.6: Summary of the number of BLAST hits for all MEROPS entries. A) Number of hits for all MEROPS entries per metagenome on left axis and number of hits for PF13402 sequences per metagenome on the right axis. B) Correlation between the number of hits for all MEROPS entries and the number of hits for PF13402 sequences. C) Frequency distribution for all MEROPS entries showing how many entries had how many hits across all metagenomes. The total number of hits for PF13402 sequences is indicated by the red dotted line intercepting the graph at 1020 hits.

Table 4.1: A summary of the MEROPS entries with the most hits across the 85 healthy gut microbiomes from Qin *et al.* (2010). The first three columns show the MEROPS family, MEROPS ID and MEROPS description copied verbatim from the MEROPS database for the corresponding pep unit sequence used in sequence alignment searches. The description contains information about the MEROPS entry functional annotation, taxonomic origins of that sequence and the database index ID. The final column represents the total number of hits in all 85 metagenomes.

No	Family	MER ID	MEROPS description	No of Hits
1	C26	MER260723	carbamyl phosphate synthetase ({Saccharomyces cerevisiae}) (Dasypus novemcinctus) [C26.956]	9077
2	C26	MER077614	family C26 unassigned peptidases (Prochlorococcus marinus) [C26.UPW]	3056
3	C26	MER331802	family C26 unassigned peptidases (Lactobacillus fermentum) [C26.UPW]	2564
4	C26	MER295850	gamma-glutamyl hydrolase (Nematostella vectensis) [C26.001]	2260
5	S16	MER332505	family S16 unassigned peptidases (Desulfotomaculum acetoxidans) [S16.UPW]	2215
6	M23	MER274317	subfamily M23B unassigned peptidases (Roseburia hominis) [M23.UPB]	2000
7	U69	MER255370	family U69 unassigned peptidases (Zobellia galactanivorans) [U69.UPW]	1850
8	I39	MER104267	family I39 unassigned peptidase inhibitor homologues (Spirosoma linguale) [I39.UNW]	1539
9	S33	MER230625	family S33 unassigned peptidases (Micromonospora sp. L5) [S33.UPW]	1527
10	S16	MER281224	family S16 unassigned peptidases (Glaciecola nitratireducens) [S16.UPW]	1499
11	C40	MER027021	PgdS peptidase (Desulfitobacterium hafniense) [C40.005]	1495
12	M41	MER273662	family M41 non-peptidase homologues (Clostridium acetobutylicum) [M41.UNW]	1467
13	I39	MER187397	family I39 unassigned peptidase inhibitors (Sulfurospirillum deleyianum) [I39.UPW]	1391

Table 4.1 – continued from previous page

No	Family	MER ID	MEROPS description	No of Hits
14	M16	MER332268	subfamily M16C unassigned peptidases (Clostridium hathewayi) [M16.UPC]	1388
15	M23	MER287297	subfamily M23B unassigned peptidases (Enterococcus faecium) [M23.UPB]	1355
16	A01	MER282526	subfamily A1A unassigned peptidases (Naumovia dairenensis) [A01.UPA]	1349
17	I87	MER277542	family I87 unassigned peptidase inhibitors (Clostridium bartlettii) [I87.UPW]	1339
18	I39	MER064398	family I39 unassigned peptidase inhibitors (Capnocytophaga ochracea) [I39.UPW]	1307
19	I87	MER331877	family I87 unassigned peptidase inhibitors (Clostridium thermocellum) [I87.UPW]	1259
20	I39	MER196386	family I39 unassigned peptidase inhibitors (Pirellula staleyi) [I39.UPW]	1247
21	C26	MER236171	family C26 unassigned peptidases (Thermus scotoductus) [C26.UPW]	1243
22	S16	MER298341	family S16 unassigned peptidases (Burkholderia mallei) [S16.UPW]	1152
23	I39	MER193949	family I39 unassigned peptidase inhibitors (Chitinophaga pinensis) [I39.UPW]	1147
24	I39	MER129323	family I39 unassigned peptidase inhibitor homologues (Pedobacter heparinus) [I39.UNW]	1112
25	S33	MER209532	family S33 non-peptidase homologues (Strongylocentrotus purpuratus) [S33.UNW]	1107
26	I39	MER133149	family I39 unassigned peptidase inhibitor homologues (Dyadobacter fermentans) [I39.UNW]	1096
27	C44	MER223378	family C44 non-peptidase homologues (Ignisphaera aggregans) [C44.UNW]	1087

From the MEROPS annotations in Table 4.1 the majority of common proteases do not carry assigned functions in MEROPS although their functions can be approximated through the family and subfamily assignments which forms the basis for the following summary of gene funcitons. The four most common protease hits are for the C26 family and are involved in amino acid biosynthesis so can be considered essential house keeping functions present in both Eukaryotes and Prokaryotes. The next top hit is for an unassigned S16 protease involved in degrading unfolded proteins which appears again at number 10 and 22 on the top hits list and is considered to be involved in house keeping internal cell functions. Next we have an unassigned M23B subfamily protease. This is the first hit that is likely to be involved in glycan associated metabolism as it is an endopeptidase that lyses bacterial cell wall peptidogylcans. Next an unassigned U69 family peptidase. This peptide is a known self-processing (autocatalytic) extracellular peptidase. The function is related to excretion rather than extracellular function so the implications could be broad although not specific to a protease activity. I39 family is a broad endopeptidase inhibitor. S33 family is a non essential exopeptidase with a signal peptide. C40 is PgdS peptidase targets cell walls in some instances during sporulation. M41 is a metalloprotease MXXH structure likely targeting membrane proteins.

As shown in Figure 4.6C there is a long tail of individual MEROPS entries with a relatively high number of hits. Each MEROPS entry is assigned to the species it was isolated from so it would be reasonable to expect a species bias for the top individual hits. MEROPS families allow the entries to be grouped into functional clusters. There were 253 MEROPS families with hits and Figure 4.7 summarises the consistency and frequency of hits per MEROPS family across the microbiomes.

Figure 4.7 shows key trends in the MEROPS protease families. The first point to notice is the broad correlation across all of the protease families with the total number of proteases. Noting the proportional consistency, it would seem to suggest that variation in the overall proteases reflect an experimental artifact, such as what would be caused by varied sequence coverage. However, the most popular family is I39, a broad protease inhibitor, is one of the least correlated with any other common proteases indicating the hit count for I39 may be reflective of a biological artifact. After I39, S9 is the next most popular MEROPS family and does not have an individual MEROPS entry in the top hits listed in Table 4.1. S9 is listed as an enzyme targeting biological active molecules.

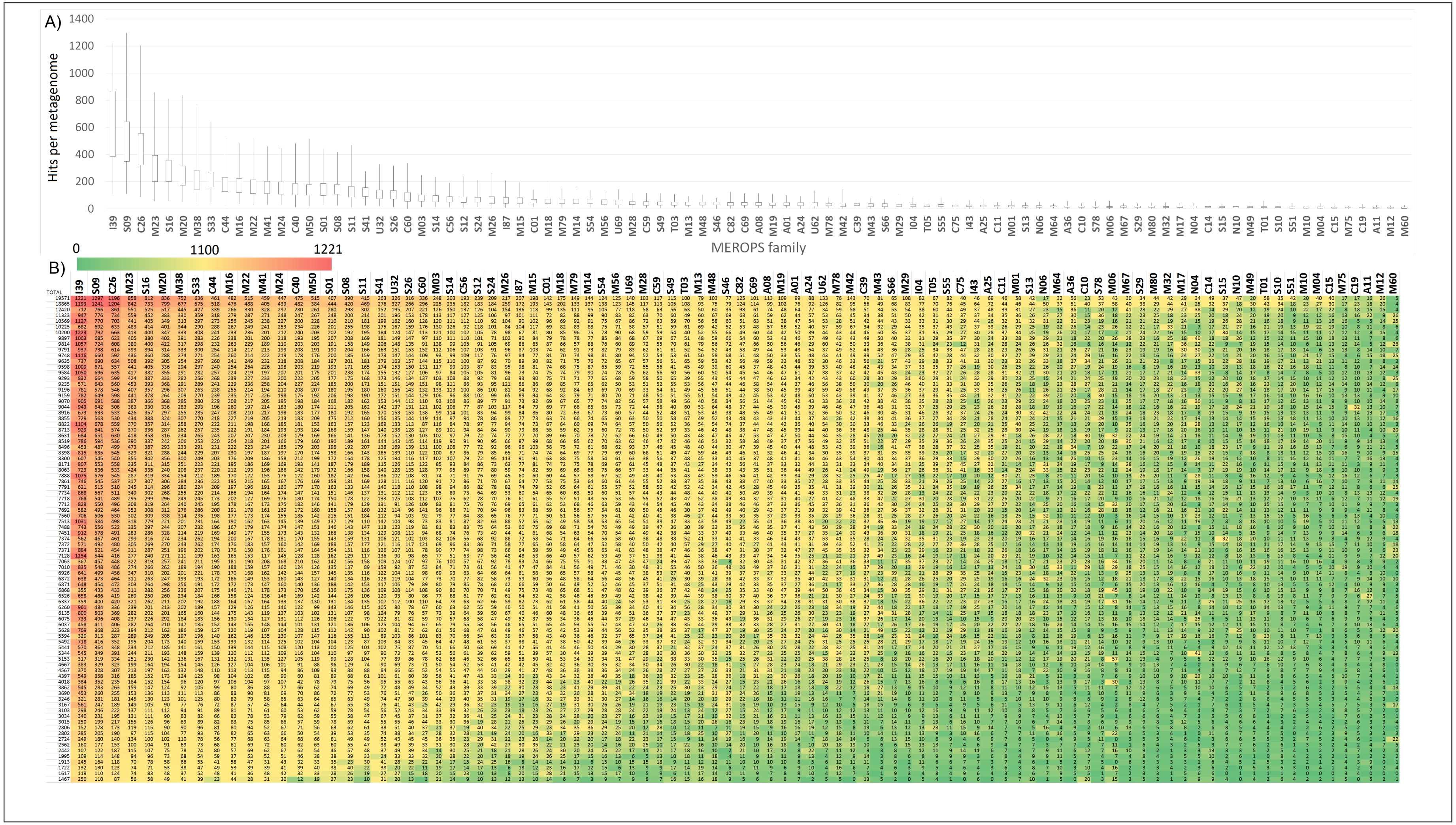


Figure 4.7: Boxplots and heatmap showing how the distribution in the number of hits per MEROPS family across 85 gut metagenomes. A) Boxplots showing the spread in the number of hits for each MEROPS family across the metagenomes. The box represents the upper quartile and lower quartile and the line represents the top and bottom of the range.

Clearly there are multiple MEROPS protease families that are very common in the microbiome and the family annotation gives an indication as to the high level function of that gene. However, the family functional annotation often is not specific enough to distinguish the characteristics of individual proteases. Also as shown in Table 4.1 the individual functional annotations provided by MEROPS database descriptions for many of the most common genes is simply 'unassigned peptidase'. It is common in metagenomics data that a gene will not carry a specific molecular function annotation so it can be useful to use piece together complementary evidence to infer more detail about the individual gene function.

4.3.3 Protein and peptide associated glycan targeting proteases

Looking at the CBM and protease configuration of a gene can provide insights into the target and the function of the proteases. The M60-like zinc metalloprotease was the first protease found to pair with a CBM to target host glycans. The data showed it paired with CBM32, CBM5_12 and CBM51 and that this pairing enabled a functional shift in the ability of the enzyme to target specific protein associated glycans in mucus (Nakjang *et al.*, 2012). Here we investigate CBM configuration at a global scale to gain insights into other protease genes that may target glycans. Combining the configuration data with the number of hits in the microbiome will give an indication as to the diversity and the context of the gene as well as an indication as to the popularity and specificity of gene variants.

CBM profile searches on the putative protease gene set identified a total of 15,914 hits for 1,011 unique CBM-protease pairs (351 unique pairs grouped by MEROPS family) from a total of 56 CBM motifs and 822 proteases (92 MEROPS families). Many of the proteases from the same family share the same function so from here in we will discuss CBM-protease pairs as CBMs linked with one of the 92 MEROPS families. The fasta sequences for all putative CBM-proteases are available on the project website, sequences are grouped into the MEROPS family of the putative protease.

http://homepages.cs.ncl.ac.uk/matthew.collison/research/CBM-protease-pairs

As shown in Figure 4.8 there are 37 MEROPS families that combine with CBMs and accumilated more than 10 hits across the microbiomes. Table 4.2 shows most common MEROPS families, the number of times a CBM is paired with each protease family (No of hits) and the distinct number of CBMs associated with the protease family (NR CBM pairs) as well as the most common CBM domains, the number of times each CBM pairs with a protease and the number of distinct MEROPS families it pairs with (NR MEROPS pairs). Often there are multiple CBM domains on the same gene as a protease, each CBM on a gene is counted as a distinct pairing occurrence.

Of particular interest to the role of proteases in the metabolism of glycans is the association with a range and combination of CBM domains. The range of CBM associations for some proteases is broad whereas other protease associations are dominated by few CBM configurations. To further investigate the nature of CBM-protease pairings a breakdown of the exact gene configurations, the number of hits for each configuration, the number of distinct genes containing that configuration and the species level origin of each distinct genes was computed. Appendix 6.2 summarises this gene configuration and distinct gene data.

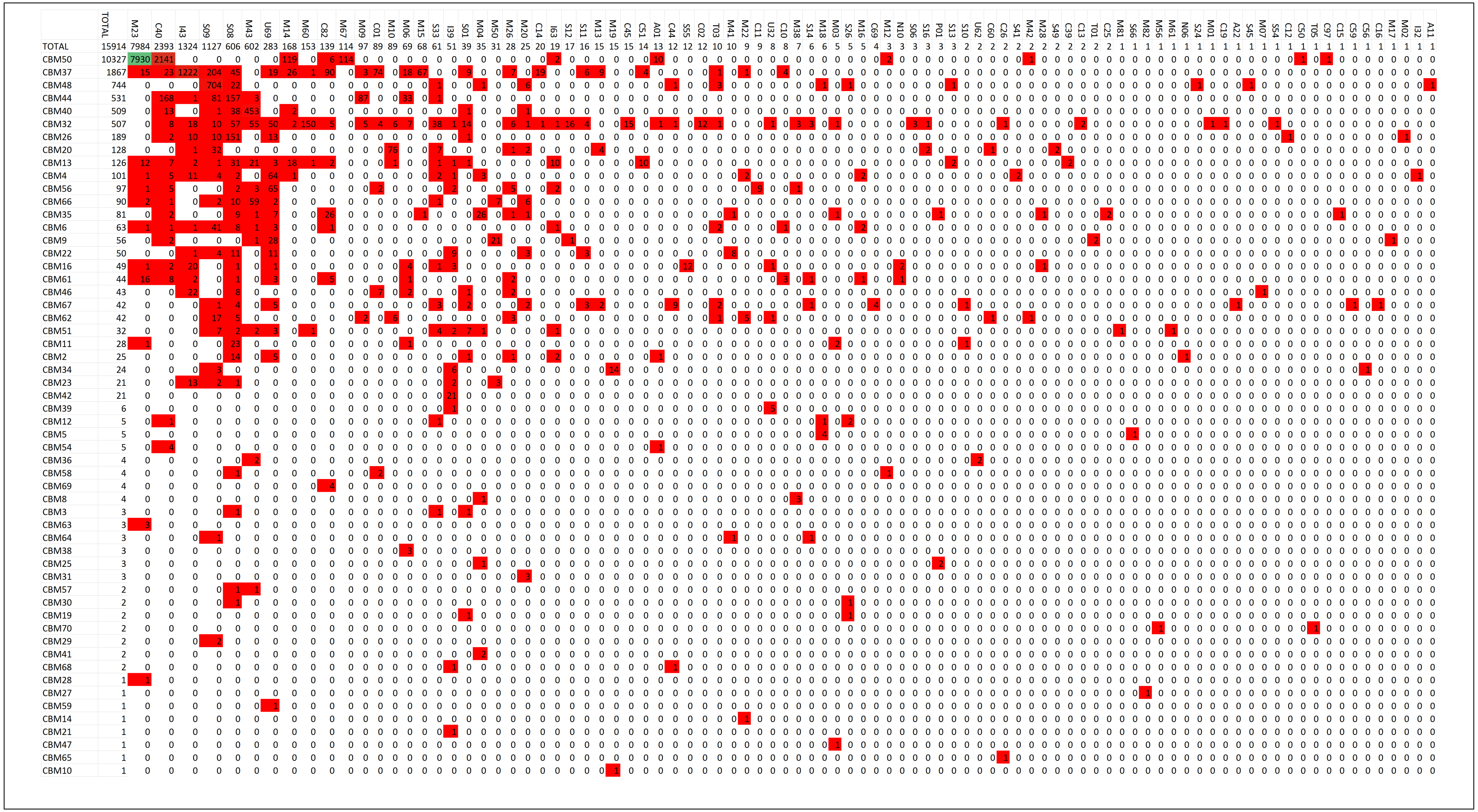


Figure 4.8: **A matrix of the number of hits for combinations of CBMs and MEROPS families.** The rows and columns are sorted vertically and horizontally by the total number of hits per row or column respectively. Each CBM hit registers as a single hit, for example if there are two CBM50 domains with an M23 protease it will add two to the count in the intersection between CBM50 and M23 on the matrix. The colour range was selected to show a distinction the variety of combinations so 0 is white and anything in the range 1-7930 is graded red to green. For the a more focused view of the most common MEROPS families and CBM domains see Figure 4.9.

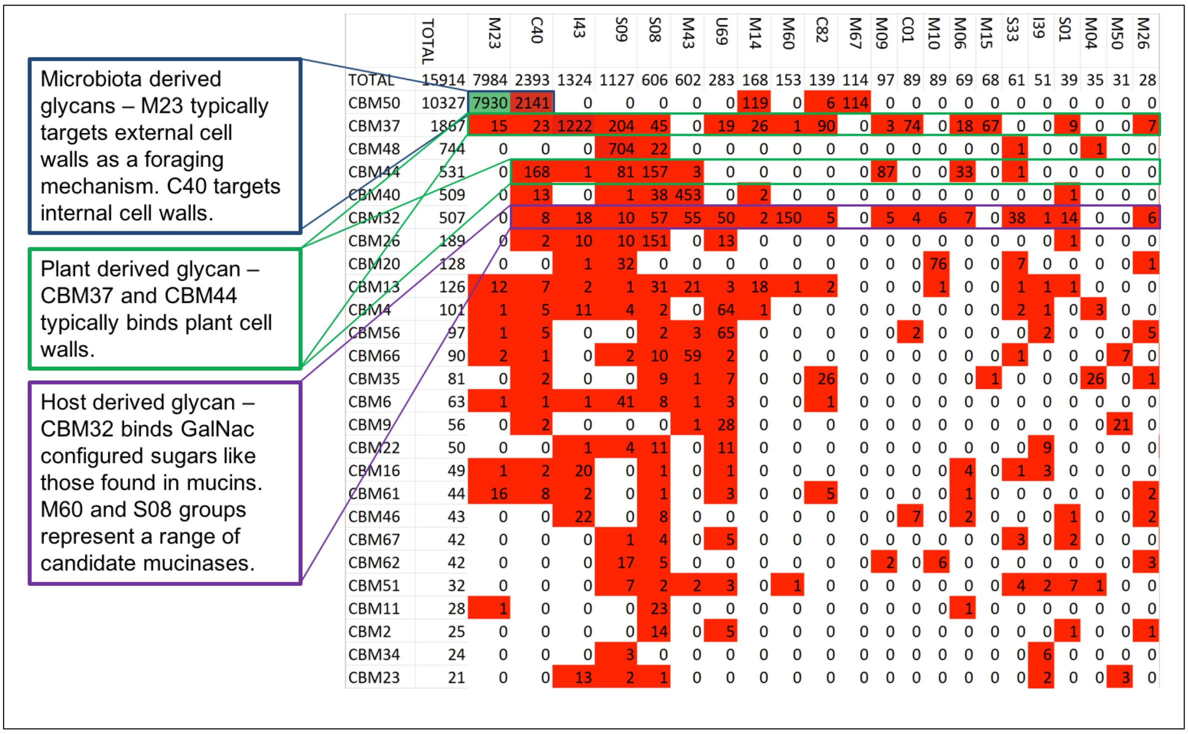


Figure 4.9: **A matrix of the number of hits for combinations of the most common CBMs and MEROPS family protease pairs.** The figure shows the top 26 CBM motifs and top 22 MEROPS families sorted vertically and horizontally by the total number of hits per row or column respectively. Each CBM hit registers as an independent count so duplicates cause an increment of two in the relevant place on the matrix. Colour scheme was selected to show a distinction between the most common combinations and is graded white to red for 0-50 hits and red to green for 50-8000. For the entire matrix including all MEROPS families and all CBM domains see Figure 4.8

Table 4.2: **Column one lists the MEROPS family proteases sorted by the highest number of hits with CBMs. Column three lists the number of different/non-redundant (NR) CBMs each MEROPS family protease paired with. Column 4 lists the CBMs sorted by the highest number of hits with a MEROPS family protease. Column six lists the number of different/non-redundant (NR) MEROPS family proteases each CBM paired with.** Note duplicate CBMs counted multiple times for hits but if a protease family paired with multiple of the same CBM this counted once for the NR count.

MEROPS family	No of Hits	NR CBM pairs	CBM	No of Hits	NR MEROPS pairs
M23	7984	12	CBM50	10327	11
C40	2393	17	CBM37	1867	22
I43	1324	13	CBM48	744	13
S09	1127	19	CBM44	531	8
S08	606	26	CBM40	509	7
M43	602	12	CBM32	507	38
U69	283	17	CBM26	189	8
M14	168	6	CBM20	128	10

Table 4.2 – continued from previous page

MEROPS family	No of Hits	NR CBM pairs	CBM	No of Hits	NR MEROPS pairs
M60	153	4	CBM13	126	18
C82	139	8	CBM4	101	14
M67	114	1	CBM56	97	11
M09	97	4	CBM66	90	9
C01	89	5	CBM35	81	15
M10	89	4	CBM6	63	12
M06	69	8	CBM9	56	7
M15	68	2	CBM22	50	8
S33	61	12	CBM16	49	12
I39	51	13	CBM61	44	12
S01	39	11	CBM46	43	7
M04	35	7	CBM67	42	16
M50	31	3	CBM62	42	10
M26	28	9	CBM51	32	12
M20	25	9	CBM11	28	5
C14	20	2	CBM2	25	7
I63	19	7	CBM34	24	4
S12	17	2	CBM23	21	5
S11	16	4	CBM42	21	1

Figure 4.8 and Table 4.2 show a wide range of pairings but also contains some clear structure. As CBMs have a specific binding target it is possible to break down how different proteases may be using CBMs to target particular substrates at the centre of particular microhabitats. Further we can investigate what species are encoding proteases that exploit each microhabitat. Crucially this level of information would faciliate a more generic overview of host-microbiota interactions and identify how particular perturbations of the microbiota communities may impact on host-microbiota interactions.

The main trends in microbiota, diet and host glycans specificity are idenified in Figure 4.9. Interestingly, some proteases pair with CBMs that target multiple sources of glycan, for example S09 and S08 that have generic functions, however, there are also examples where there are biases that may indicate a functional adaptation to a type of glycan from a particular source.

4.3.4 Proteases targeting microbiota derived protein and peptide associated glycans

The most common CBM-protease pairing in the matrix is M23 and CBM50 with 7930 occurences. CBM50, also known as LysM, is known to be involved in cell wall peptidoglycan processes and has specificty for the N-acetylglucosamine sugars directly linked to the peptide from the cell wall (Ohnuma *et al.*, 2008). LysM is common in single celled organisms, except in Archaea, and LysM hydrolases are also known to be common. The M23 protease family is a well characterised beta-lytic endopeptidase that is known to lyse the cell walls of other bacteria as a defensive mechanism or feeding mechanism. M23 targets either the N-acylmuramoyl-Ala bond between the cell wall peptidoglycan and the cross-linking peptide or a bond within the cross linking peptide, as descibed in the MEROPS entry.

One feature of the M23-CBM50 pairing is the high number of distinct genes. 53 copies was the highest number of a particular M23-CBM50 gene (GL0074796_MH0025_C3650026_1_95_976) found in any micorbiome. Furthermore, as shown in Figure 4.10, the M23 protease has a relatively even taxonomic profile implying that CBM50-M23 genes are present in multiple species in the microbiome. Despite the broad presence of LysM proteases across the microbiota, there is an indication in some instances of specific adaptations through increasing the number of CBM domains to increase avidity and in some instance to create a multimodular domain that provides specificity necessary for function (Mesnage *et al.*, 2014).

LysM (CBM50) has been studied at length in multiple organisms (Buist *et al.*, 2008) and has been used in a range of biotechnology applications, such as cell surface detection and vaccine development (Visweswaran *et al.*, 2014). However, as far as I know, use as a natural microbial molecule has not been explored. Furthermore, this is the first metagenomics study identifying diversity of LysM hydrolases in metagenomics data which identifies areas where novel previously unidentified instances of the gene can be found. As shown in Figure 4.11 there are instances of CBM50-M23 with low similarity to known homologues which could guide novel gene discovery.

MEROPS family C40 has 2388 hits with CBMs, of which 2141 are accountable to CBM50. Similar to MEROPS family M23, this family targets the peptidoglycans in the cell wall, appears to increase avidity or add specificity through increased CBM50 counts and has an even taxonomic spread, Figure 4.12. However, the MEROPS entry suggests C40 is a cytoplasmic enzyme expressed during sporulation so this combination is assumed to have an internal house keeping function. Interestingly C40 also pairs with multiple CBMs that typically are not involved in cell wall processes, such as CBM32 and CBM44.

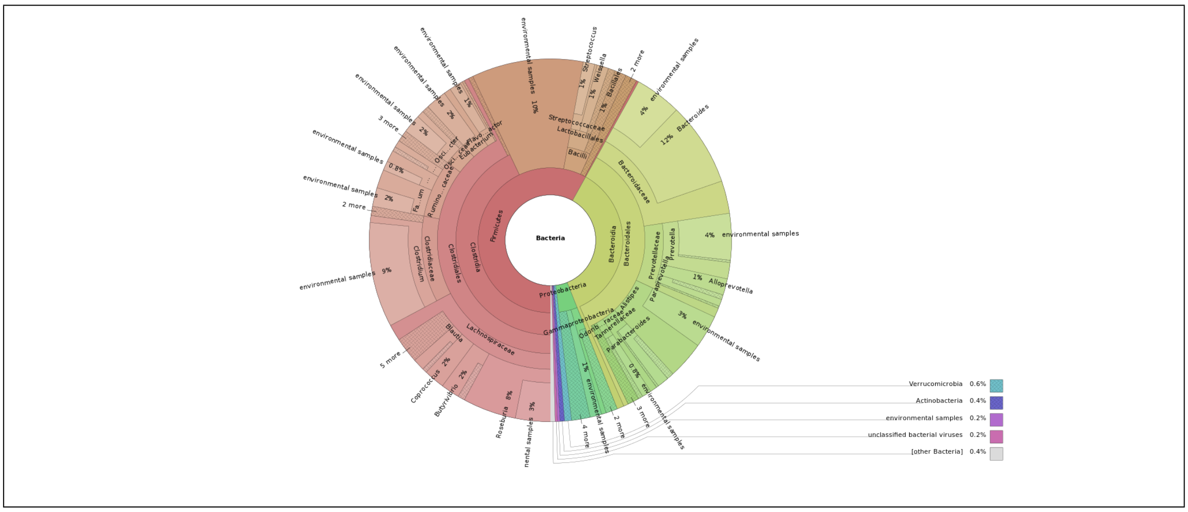


Figure 4.10: The taxonomic profile for M23 proteases that pair with CBMs.

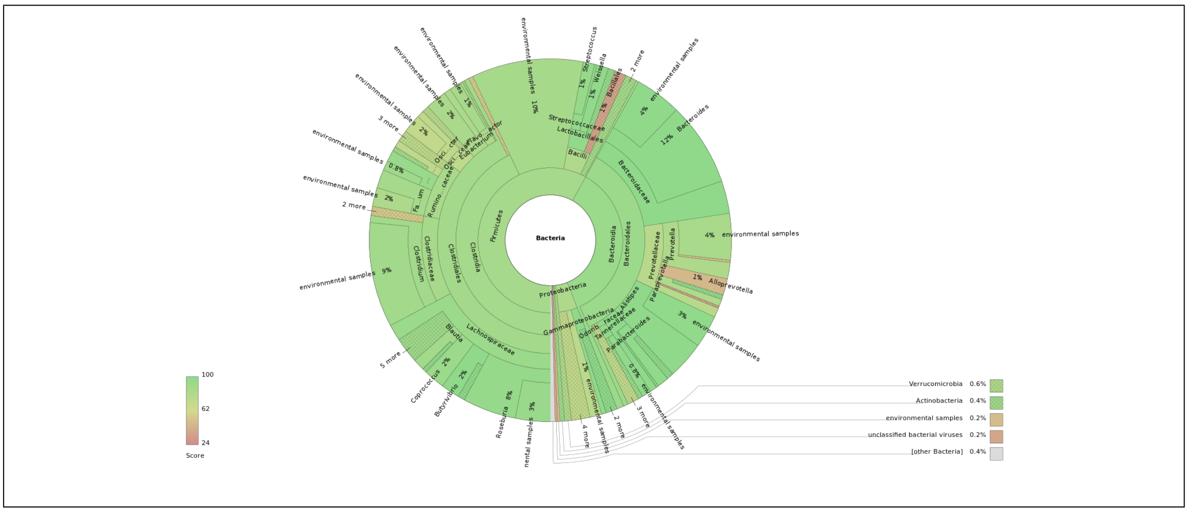


Figure 4.11: The taxonomic profile for M23 proteases that pair with CBMs. The colour here reflects the percentage similarity between the reference gene and the query gene.

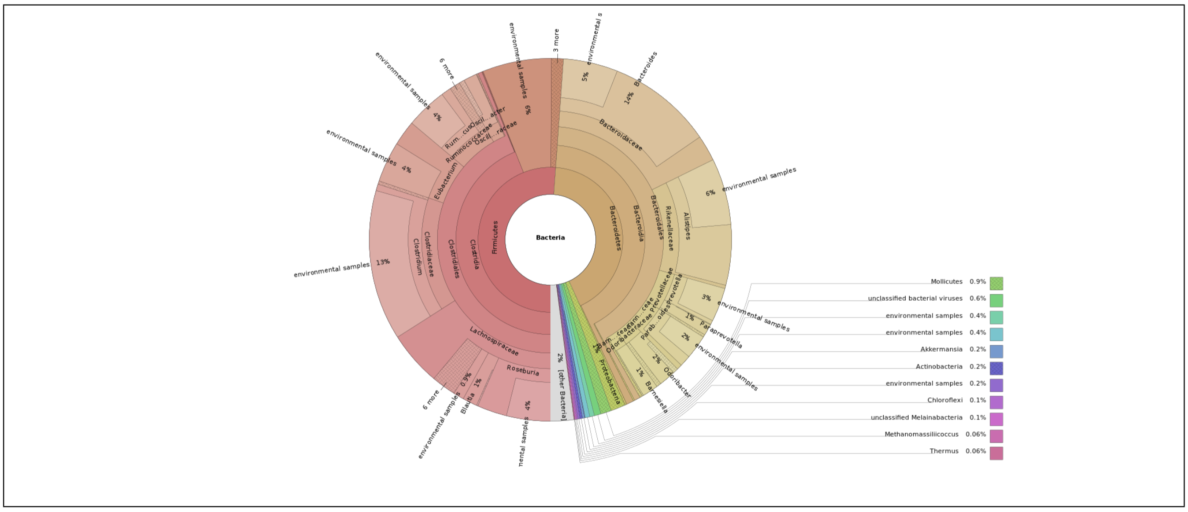


Figure 4.12: The taxonomic profile for C40 proteases that pair with CBMs.

Peptidoglycan hydrolases are a common mechanism for microbial competition and have been shown to contribute in both healthy and dysbiosis states as well as creating predispoisitions by allowing pathogens to be tolerated (Rangan *et al.*, 2016; Hasnain *et al.*, 2012). This result shows how regular peptidoglycans are across the microbiome and where novel instances may be found across the range of the micorbiota and the scope for novel compound discovery in this area from this approach.

4.3.5 Proteases targeting host derived protein and peptide associated glycans

Excluding the high number of cell wall targeting proteases MEROPS families S08 and S09 represent the largest group of proteases. The S08 and S09 families represent generic secreted and biologically active protesaes, respectively, that pair with a broad range of CBM combination. Notabley CBM37, CBM48, CBM44, CBM32 CBM26 are the most common partners, however, the proteases usually pair with these CBMs in isolation indicating a single target. Interestingly, when MEROPS family S08 pairs with CBM32 (a known mucin binding domain) it also combines with CBM51 which shares similarly to the M60-like mucinase discussed earlier.

MEROPS family M60 is shown to almost exclusively pair with CBM32. The family has 152 hits in total, 150 from CBM32 and they all stem from a single CBM pairing in the M60-CBM32 configuration which is encoded in 34 distinct genes. The taxonomic distribution, Figure 4.13, also shows a clear bias towards Bacteroidetes and a likely microhabitat. These 34 distinct genes represents a high chance of novel host mucin targeting proteases, particularly the genes with low similarity to any known homologues in the Prevotella phylum, Figure 4.14. Noting these features of the M60 protease; MEROPS family S33 also shares a similar bias to CBM32 pairing and a taxonomic distribution across Bacteroidetes, Figure 4.15. Therefore, the S33 proteases represent a whole new class of putative proteases targeting host mucins.

Not all host derived protein and peptide associated glycans are structural molecules. MEROPS family S09 has 1124 hits with CBMs, of which 704 are hits with CBM48. MEROPS family S09 is broadly linked with biologically active molecules, including DDP4 a protease that degrades GLP-1, and is regularly found paired with CBM48 which binds glycogen. It would seem sensible to speculate that the S09-CBM48 pair has an impact on host derived peptide signalling in the gut.

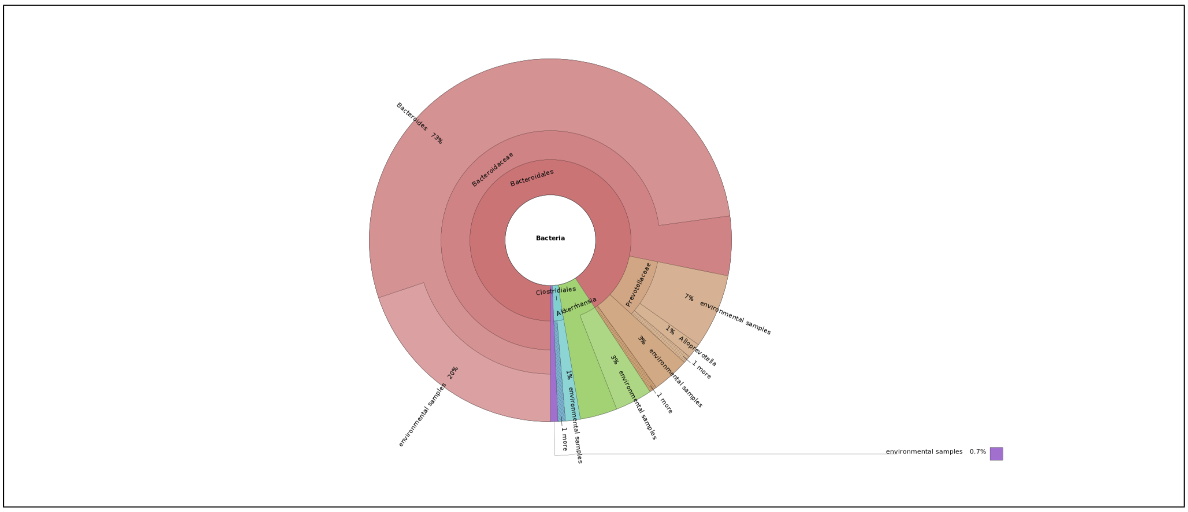


Figure 4.13: The taxonomic profile for M60 proteases that pair with CBMs.

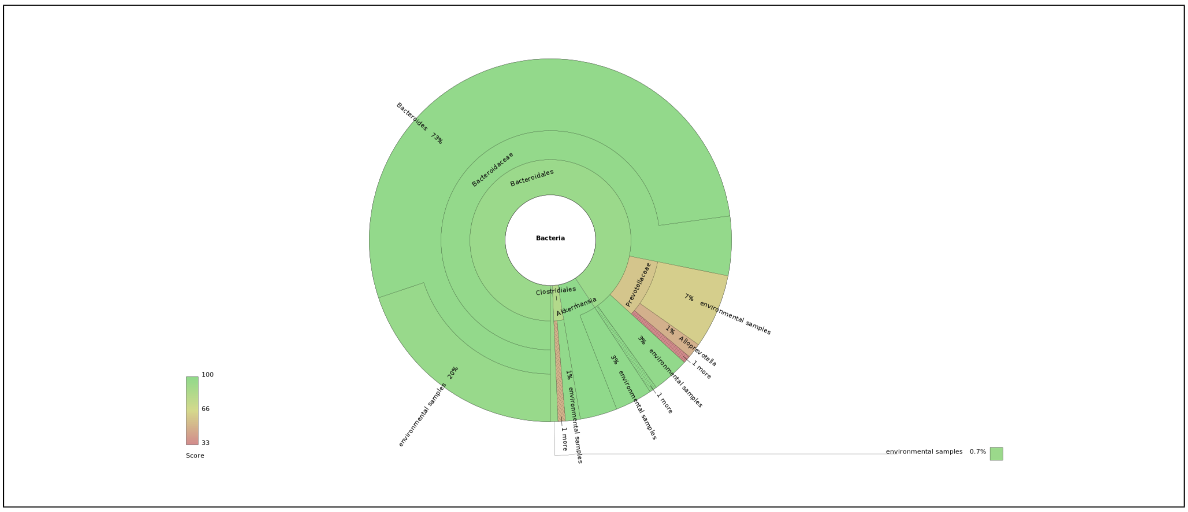


Figure 4.14: The taxonomic profile for M60 proteases that pair with CBMs.The colour represents the percentage similarity to the closest known homologue.

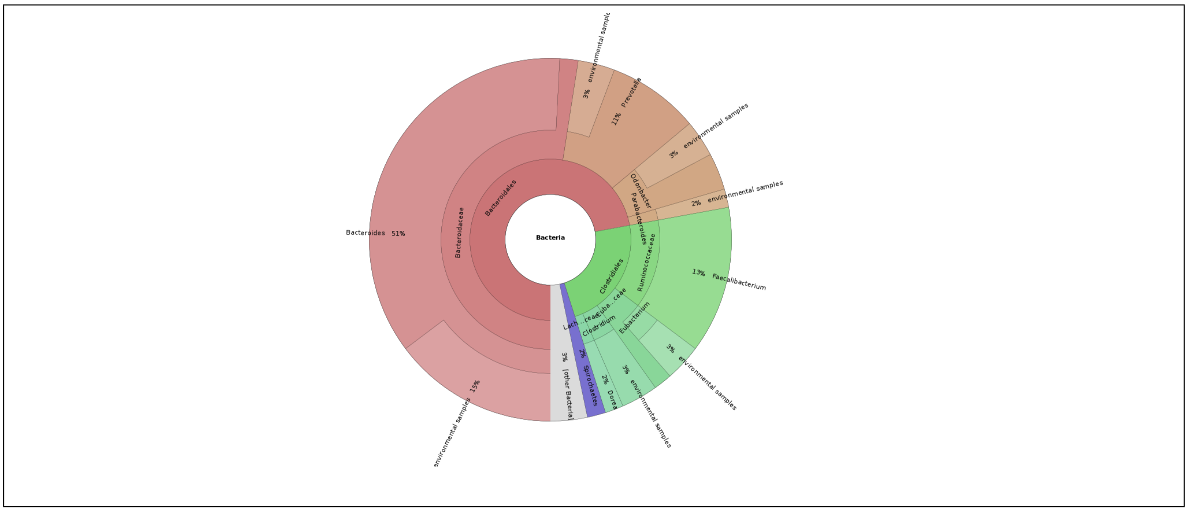


Figure 4.15: The taxonomic profile for S33 proteases that pair with CBMs.

4.3.6 Proteases targeting plant derived protein and peptide associated glycans

MEROPS families M09, C01 and M15 all show a bias towards targeting plant based glycans and all have a clear taxonomic distribution scewed towards Firmicutes and Clostridia, Figure 4.16, Figure 4.17 and Figure 4.18, respectively.

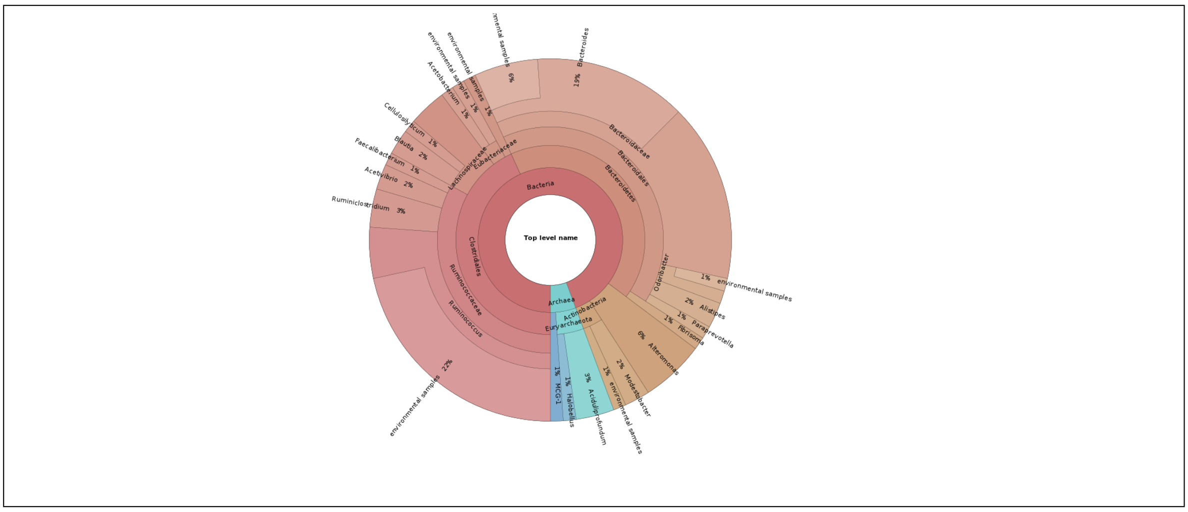


Figure 4.16: The taxonomic profile for M09 proteases that pair with CBMs.

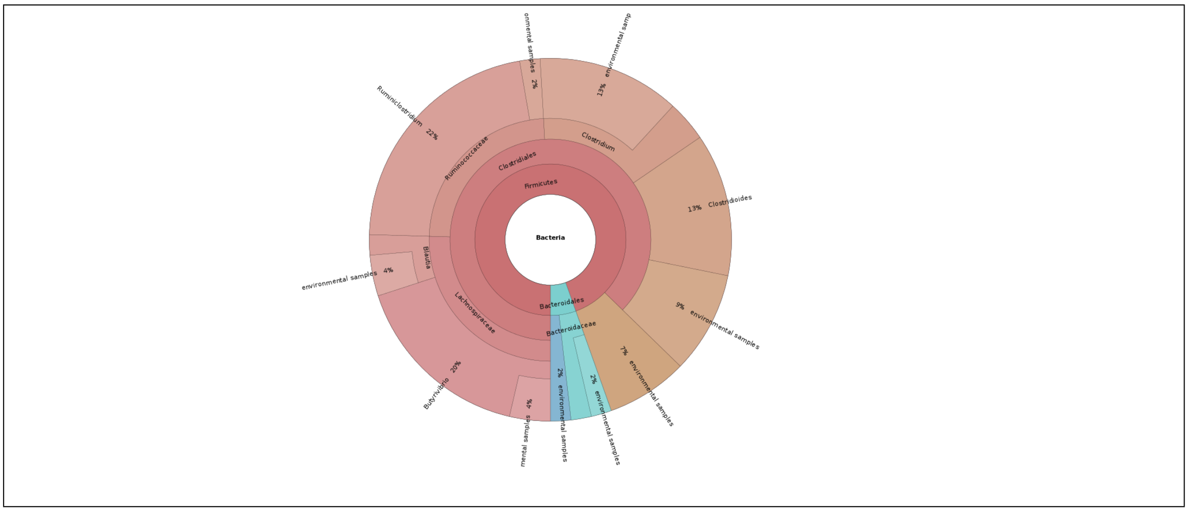


Figure 4.17: The taxonomic profile for C01 proteases that pair with CBMs.

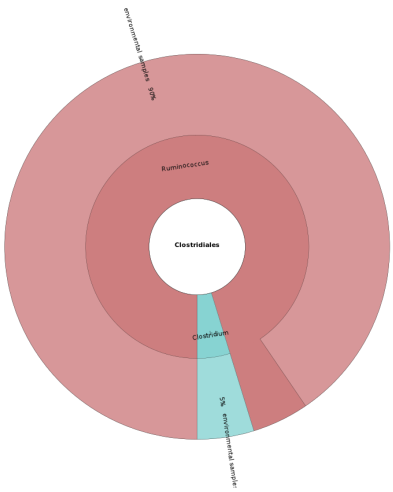


Figure 4.18: The taxonomic profile for M15 proteases that pair with CBMs.

4.4 Discussion

The hypothesis for this study was that there are a range of previously uncharacterised proteases targeting protein and peptide associated glycans in the microbiota that have a broad impact on host-micorbiota interactions. By quantifying the frequency of occurence and diversity of different protease classes that target glycans this study provides context for many isolated studied focussed on single proteases and highlights a limitation of current methods for using metagenomics data to study low level gene occurence. Moreover this study identifies a range of previously uncharacterised protease genes thought to have important roles in host-microbiota interactions.

Existing protease research in the microbiome area has shown gut bacteria can produce extracellular bacterial proteases that degrade host matrix components, such as in cell junctions, and interfere in signalling pathways to facilitate infection (Steck *et al.*, 2011) and that these bacterial proteases have an increased activity in inflammatory bowel disease (IBD) and irritable bowel syndrome (IBS) patients (Carroll *et al.*, 2013). Protease functions in the direct infectious cycle, however, do not account for a large part of the complexity involved in colonisation of the gut particularly in protein and peptide associated glycan metabolism. In this regard this is the first study to investigate the occurrence of protease genes at the microbiome level.

Our global analysis provides a survey of the breadth and frequency of proteases and their functions in metagenomics data. This study identified 611,948 proteases, 15,914 occurrences of CBM-protease pairings, 3,811 distinct CBM-protease genes 1,011 different CBM-protease combinations, and 483 MEROPS family gene configurations.

Glycan metabolism is a central part of the diet-microbiota, microbiota-microbiota and host-microbiota interactions in the gut. As our understanding of the mechanisms underlying microhabitats and colonisation of the gut emerge the links between metagenomics data and dysbiosis associated disease states becomes clearer. Established studies have shown the prevalence of cazymes in the microbiome and their role in the systematic digestion of glycans in the gut. Recent studies have also shown the importance of diverse mechanisms of glycan metabolism for adaptation under specific conditions. Ndeh (2013) demonstrated the first endopeptidase activity from bacterial proteases using a CBM to target host proteoglycans which improves fitness only in the mucosal niche. This study provides a global summary of proteases that may contribute to niche fitness and microbiota-microbiota interactions.

4.4.1 Novel gene discovery

Efforts to study individual genes often speculate on the variance and prevelence of their gene of interest. If the application of the gene is valuable enough large scale screening studies are undertaken that consume extensive man hours and enormous resources. Exploiting metagenomics data for examples of natural diversity is a relatively cheap and complementary alternative that could guide or refine compound screening and provide a preliminary assessment of the natural prevelance and diversity of gene variants given a specific gene configuration.

Specific highlights of this study include M23-CBM50 variants that represent a breadth of potential natural antimicrobial peptides and S33-CBM32 and M60-CBM32 pairings that represent a range of putative mucin targeting genes. Although the scope for further exploration of all aspects of this dataset is vast and domain specific. The main challenge in persuing these application, however, is data accessibility.

4.4.2 Metagenomics for therapeutic intervention design

Current approaches to data mining therapeutic targets in the microbiome use top down statistical testing to detect statistically over represented gene categories. This approach is comparable to comparative genomics and population based testing common in historical drug development. However, metagenomics data describes a complex system with incomplete data so the insights that can be gained from comparative statistics are limited. One of the limitations of gut microbiome data is the faecal sampling bias. Many of the most influential microbes reside next to the mucosa and therefore can be more difficult to sample (Zoetendal *et al.*, 2002). One of the outcomes of this chapter is in the quantification of the inconsistency in hit rate for genes that are known to be prevalent both at low level, as shown in PF13402, and when looking at the entire MEROPS database, which shows up to 10 fold differences in the number of proteases between individuals.

Bottom up analyses in the microbiota will be crucial to designing effective microbiota targeted interventions. Understanding the range of mechanisms for host-microbiota interactions is central to understanding any comprehensive disease model. To design an effective intervention and treatment model it is important to account for the range of species encoding relevant functions. There are currently no resources to effectively map gene occurence data with taxonomic data to facilitate these analyses.

4.4.3 Quantitative annotations from metagenomics context

Our approach to domain based configuration searches has generated new quantitative contextual data about the occurrence frequency for thousands of genes that is complementary to existing functional annotations held annotation databases. This new approach to analysing metagenomics data will enable effective microbiota models to be generated and more comprehensive therapeutic strategies to be designed.

Our ongoing work aims to provide new insights into the potential mechanisms contributing to disturbances in host-microbiota interactions. These analyses will contribute to refining ongoing experimental work dissecting host-microbe interactions at mucosal surfaces and could form the basis of studies into therapies with molecular targets in the microbiota. Furthermore the approach to identify and annotate uncharacterised genes in the microbiome highlights one of the most valuable assets of metagenomics data that has yet to be exploited. A single resource to map the occurrence of metagenomics genes and their neighbors in order to provide context specific information to a multitude of new metagenomics genes. To make this type of resource effectively usable there is a requirement for community standards and an infrastructure for scalable resource handling.

Chapter 5

BioSSIP: A Scalable Semantic Integration Platform for Bioinformatics

5.1 Introduction

Biological data are becoming exponentially larger and more complex as high-throughput sequencing technologies become ubiquitous across life sciences (Fernández-Suárez *et al.*, 2013). This increase in data availability has lead to computational challenges in storing, accessing and integrating bioinformatics resources as well as making secondary analyses reproducible (Cohen-Boulakia *et al.*, 2017). These issues are particularly challenging in microbiome research where data-rich resources are constantly being produced and updated alongside new approaches to data analysis and data mining (Tsilimigras & Fodor, 2016). In this chapter I describe a system to facilitates scalable data handling and a software development life cycle for data integration in microbiome research.

Most highly used bioinformatics resources are provided as federated stores (Zhang *et al.*, 2011). To the user this means the data source is curated by a third party and provided through websites, which produce neat results and negates the requirement for distribution of software and access to high performance computational resources. This approach is useful for enabling user access to a single use application or analysis tool, however, often complementary datasets require integration and adaptation to extract value from multiple data sources through integrative bioinformatics (Zhang *et al.*, 2011; Lapatas *et al.*, 2015). This is when data integration pipelines are developed and data warehousing can be employed by bioinformaticians.

5.1.1 Data Integration pipelines

Generating an integrated dataset can be as simple as matching a string which represents an ID in one file to an identical string in another file. Often though data integration tasks become complicated processes requiring a high level of domain knowledge, understanding of semantic data formats as well as managing high performance computing resources. Moreover the process of developing software to facilitate the integration pipeline can become convoluted and error prone with multiple intermediary files and single use software programs. There are four main stages in most data integration studies; reading data sources, data integration, data storage and data access or distribution.

5.1.1.1 Data entry

The process of reading data from a data source is dependent on the format of the input data and typically involves the developer writing a parser to interpret the data. Often data formats are followed loosely and there is a time consuming process of identifying edge cases that need to be captured to avoid errors being propagated through the integration pipeline. Related to this issue reuse and distribution of parsers is uncommon and intermediary file formats are often replicated and relatively hidden on local machines. Therefore the data integration process often becomes untestable and non reproducible. Programming languages tied to data standards for data integration have shown some promise in this area showing that it is possible to automate dataset generation and testing (Warrender & Lord, 2015; Köhler *et al.*, 2006). However, generating datasets with TAWNY OWL is based on single use programs, strict ontological data standards and uncommon programming languages such as Clojure (Lord, 2013).

5.1.1.2 Data integration

Data integration is the process of classifying data and finding a conensus to enable data sources to be merged (Applications, 2007). The semantics and metadata which underlie classification and equivalence assertions in data integration can be defined with varying degrees of formality and specificity. Lightweight metadata approaches, such as eXtensible Markup Language (XML) and restricted vocabularies, are the most common. These formats are flexible but do not usually let you reason over the data. Ontology and Resource Description Framework (RDF) formats are rich and represent a gold standard for data representation. However, ontologies can be time consuming to define and require in-depth domain knowledge that is often not flexible enough for integration projects with time sensitive data sources.

5.1.1.3 Data storage

Data storage and retrieval requirements in bioinformatics applications are usually met through in memory or relational database systems and in some instances through denormalised data warehousing to push performance at scale. However, some bioinformatics projects are pushing the performance limitations of these data storage solutions and there is a requirement in bioinformatics for scalable complex data handling such as graph databases (Have & Jensen, 2013). NoSQL solutions have been developed and deployed elsewhere to enable web technologies but few have bioinformatics support meaning there are high implementation costs to use NoSQL technologies.

If the datasets are small enough, then in-memory data structures are often the easiest and fastest way handle data. However, for larger datasets, persistent storage may be required. Scale, complexity and accessibility of the data and objectives of the mining task are the main concerns in the specification of a database model. Practically, these concerns are measured by the size of the dataset, complexity of the mining process and the latency requirements for the end user. Bespoke in memory data structures are very fast and can work well for complex datasets due to comprehensive accessibility of the database model. However, the size is limited to memory allocation on your machine or distributed cluster. Often in memory formats are used for complex analyses of relatively small datasets and visualisation tasks. ONDEX (Köhler *et al.*, 2006), cytoscape (Smoot *et al.*, 2011) and gephi (Mathieu Bastian *et al.*, 2009) are examples of in memory software systems used for data integration and visualisation. As the hardware limitations of these systems are reached a persistent data storage model is required.

Relational databases work well for continuous data with a predictable data structure. Most biological databases, such as RefSeq, MEROPS and UniProt can be downloaded in SQL format and fit well with the tabular schema required. However, for data integration across multiple data sources the database schema can become complicated. Additionally querying performance in SQL data stores tends to be good for accessing indexes within a small number of tables but if there is a requirement for a high number of 'joins' between data sources to access your data (which would be stored across multiple tables and indexes in the SQL schema) there is often substantial delays in the query time or the data structure has to be denormalised to make the data structure fit the query and requires a high amount of replication and reduced flexibility.

NoSQL databases use non relational data models. NoSQL projects range from flat key-value stores, such as Amazon's Dynamo, to nested document stores, such as MongoDB, and graph based data stores, such as Neo4J. NoSQL data stores typically sacrifice relational features such as referential integrity (asserting that two pointers represent the same value) for vastly increased build and query speeds, and the ability to distribute datasets over multiple machines.

5.1.1.4 Data access and distribution

Data access and distribution are implicitly linked to the data storage. The requirements for data distribution depends on the intended user. The endpoint of an integration pipeline can be a live database or a flat results file of comma separated values, such as in Chapter 4. However, if the data endpoint is part of a evolving system that will be updated over time and replicated by other developers on other projects it is important to support updates and version control. Database dumps are often provided for this type of endpoint or datasets are serialised into flat files for subsequent distribution and reuse. Rarely though is the dataset distributed with the software that was used to generate it therefore limiting the ability to update particular derivative resources. Typically the software is decoupled from the data on private machines and only the data is published for interogation.

In practice software applications for data integration tend to champion one of two approaches. They offer an end to end integration pipeline environment with visualisation and lightweight metadata, such as ONDEX (Köhler *et al.*, 2006) or cytoscape (Smoot *et al.*, 2011). Alternatively, a bespoke integration process with multiple manual steps are built to populate a database (Oh no sequences, 2014; Lord, 2013; Noy *et al.*, 2003). The first approach is useful for small projects although limited in terms of the scale of the data handling and can be contrained for complex system designs. The second approach is useful in terms of flexibility in system design and for scalability of the dataset but limited in terms of reproducibility. Both approaches have advantages and disadvantages but one major factor in deciding which will work for a given project is the performance requirements for how the integrated dataset will be stored, accessed and distributed.

Table 5.1: A table to compare the features of a range of typical applications used data integration tasks.

DB type	file format	Application	Goals	Bio support	Distributed	License
in memory	xml	Gephi	Vis	No	No	GPL
	rdf	ONDEX	Vis / data store	Yes	No	Apache2.0
	rdf	Cytoscape	Vis	Yes	No	Apache2.0
SQL	sql	MySQL	Data store	No	No	Apache2.0
	sql	BioMart	Data store	Yes	No	Apache2.0
	sql	Teradata	Data store	No	Yes	Copyright
NOSQL	JSON	MongoDB	Data store	No	Yes	Apache2.0
	key value	Neo4J	Data store	No	No	AGPL
	key value	Casandra	Data store	No	Yes	GPL

5.1.2 Microbiome use case

The human gut microbiota is a highly complex, diverse and variable microbial ecology (Arumugam *et al.*, 2011; Qin *et al.*, 2010; Caporaso *et al.*, 2011). Advances in high-throughput sequencing technologies have provided new insights into the genetic complement of the gut microbiota and revealed multiple links with diseases that are highly prevalent in modern society (Sekirov *et al.*, 2010). To utilise the vast datasets produced by these technologies to investigate the microbiota a large amount of data complemented with large scale data integration and analysis is crucial (Collison *et al.*, 2012).

The most common tools currently used for microbiome analyses are QIIME (Caporaso *et al.*, 2010b) and MG-RAST (Meyer *et al.*, 2008), for analysis of 16S metagenomic data and whole-genome metagenomic data respectively. These solutions provide primary analyses including quality control, filtering, assembly and annotation. A typical whole-genome metagenomics dataset from the gut microbiota sequenced using an Illumnia Genome Analyser II contains around 3 Giga base-pairs from around 40 million sequence reads (Karlsson *et al.*, 2013). Through pipeline analyses, 70-90% of the reads pass quality control and filtering and the assembled contigs typically predict around 200,000 ORFs of which 60-90% can be annotated with molecular functions and species level taxonomy (Meyer *et al.*, 2008). This results in around 100,000-200,000 predicted genes with secondary (functional or taxonomic) annotations that represent partial coverage of the true genetic complement of the microbial community. Secondary analyses are subsequently used to predict community characteristics by clustering secondary annotations, both functionally and taxonomically, to identify patterns and features across datasets that provide insights into the overall community function.

Secondary analyses are usually carried out on an individual study basis and can be enhanced by reprocessing data from the other studies for comparison. Given the scale, complexity and non-standard nature of microbiome data, the task of secondary analysis and particularly data integration is not simple. MG-RAST (Meyer *et al.*, 2008) and other pipelines offer comparisons between secondary annotations processed within the same pipeline based on common annotation sources and experimental metadata. However, these top-down analyses only produce results comparing the fraction of data that contains secondary annotations from any single annotation source within the constraints of the experimental model. This means integration of datasets with other data sources processed outside of the analysis pipeline, including key dataset produced by major consortium studies, is challenging and often not easily reproducible. Furthermore bottom-up exploration of data is difficult due to data handling challenges associated with handling primary data. For example studying correlations between all predicted gene clusters across experimental datasets instead of only functionally annotated clusters is beyond the capacity of most anlaysis pipelines.

A relevant use case is Chapter 4, a seemingly simple 3 step process to find common proteases targeting protein and peptide associated glycans in metagenomics datasets. This study required 1479 lines of java code, 14 bash scripts, hundreds of Gigabytes of intermediary files and thousands of CPU hours on HPC machines. Moreover, despite my best efforts to publish all intermediary data files and code to a project webpage the data integration process is not easily reproducible and does not contain a intuitive interface to access all of the endpoint data for updating resources.

The hypothesis for this work is that a framework for data integration of metgenomics data would facilitate more reproducible data usage and facilitate the use of metagenomics data for gene discovery and exploration.

5.1.2.1 Datasets

Initiatives such as the European Metagenomics of the Human Intestinal Tract (MetaHIT) project (Ehrlich & Consortium, 2011) and American Human Microbiome Project (HMP) (Nelson *et al.*, 2010) have been the drivers behind a dramatic expansion in the availability of data relating to the microbiota. These consortium driven ventures aimed to address some of the big challenges in the field, such as providing a catalogue of common genes (Qin *et al.*, 2010, 2012; Karlsson *et al.*, 2013) and genomes (Consortium, 2012; Gevers *et al.*, 2012) and characterising a ‘stable’ microbiome (Caporaso *et al.*, 2011) and a ‘core’ microbiome (Arumugam *et al.*, 2011). The corresponding studies have identified or quantified these important features and generated a series of high quality datasets (Qin *et al.*, 2010; Arumugam *et al.*, 2011; Le Chatelier *et al.*, 2013; Karlsson *et al.*, 2013). The datasets provide extensive evidence of metageonmic concepts, such as metagenomic linkage groups (MLGs), as well as providing a large number of high quality datasets that can be used as gold standards. These datasets are often published with GenBank project IDs and processed in-house so intermediate annotation results are not stored in online analysis pipelines and occasionally published in custom formats.

These centralised initiatives are a contrast to many distributed research efforts exploring more subtle experimental hypotheses, both through generation of new data and through identifying new patterns in existing data. Such projects depend on availability of datasets, including secondary analyses that can be related to independent experimental models. Datasets of independent research projects are often submitted to online analysis pipelines and therefore populate IMG/HMP, MG-RAST and other public databases containing intermediate and secondary analysis results.

Table 5.2 summarises the main resources containing secondary analyses of microbiome data including, metagenomic datasets and their metadata from both pooled distributed studies (such as those in MG-RAST) and centralised consortium studies, full genomes isolated from the microbiota by the HMP and distributed through the IMG/HMP resource, curated annotation database resources both for taxonomic and functional annotation as exemplified by GreenGenes, Kegg Orthology (KO) and eggNOG, and metagenomics feature datasets identified from consortium studies as exemplified by metagenomic linkage groups (MLGs). This list is not exhaustive but represents the major data resources and data concepts within the microbiome research field.

Table 5.2: Summary of the main resources for secondary analyses of microbiome data.

type	resource	size (GB)	number of samples	genes per sample	taxa per sample
metagenomes	MetaHIT 2010	3.3Gb	124	1,800,000	4500
	Chinese study	15.2Gb	345	2,200,000	4000
taxonomy	GreeneGenes	2Gb	5,000	1000	1
	MLG	0.4Gb	200	100	1
	HMP	8Gb	4000	2000	1
function	KEGG ontology	0.8Gb	na	na	na
	COG	2.1Gb	na	na	na
	eggNOG	1.7Gb	na	na	na

We consider each of the above significant resources that are currently difficult to consistently integrate in secondary analyses due to the heterogenous distribution of data. In order to utilise this heterogeneous data it is crucial that it is represented in a standardised computationally reasonable format and handled in a scalable manor.

Strategies designed to mine primary annotation and integrated annotations require multiple datasets and annotation sources to be stored in a standardised format and to be accessible within a single datastore. Consequently this requires data storage in the order of hundreds of millions of documents with the highly linked information, which is beyond the capacity of most relational database models but lends itself to graph database technology (Have & Jensen, 2013).

5.1.3 Hypothesis

The hypothesis for this work is that data integration standards will improve access to rich metagenomics datasets and support bottom up approaches to exploiting metagenomics data. This tool aims to support better science and community resources to improve and accelerate progression towards developing models host-microbiota interactions in disease that can be used in the design of therapeutic interventions considering the microbiota.

Here I present BioSSIP, a Scalable Semantic Integration Platform for Bioinformatics. BioSSIP is a software library for bioinformatics data integration that supports standardised data integration pipelines in a modular manor allowing each application to combine flexible parser design with simple semantics and a graph database back end.

5.2 BioSSIP

BioSSIP is a three tier user system, Figure 5.1. The core software API facilitates the building of closely coupled data integration and data mining applications. Each BioSSIP application includes a data integration pipeline that integrates multiple data sources into a single performant data representation which can be used for data mining and exposed as a live database when the application is executed. Secondary users can then download BioSSIP applications and relaunch the integrated dataset locally, or they can reuse the data parsers, data mining strategies and data integration pipelines with new data or extended the pipeline with new modules. Finally client side users can interact with live applications, run live queries on integrated datasets and interogate both the integration pipeline and data mining strategy as well as the resulting data.

1. The developer interface: Uses the core BioSSIP API and provides the ability to develop standardised data integration pipelines built on a performant embedded NOSQL database in an open source version controlled, cross platform application.

2. The server side UI interface: A BioSSIP application that implements the core API. The application can be downloaded, relaunched, extended and rebuilt with an existing dataset or a new dataset without writing any code.

3. The client UI interface: A web interface to a live BioSSIP application. The client side UI provides the ability to query a live dataset that was built in BioSSIP and effectively visualise or interpret the results.

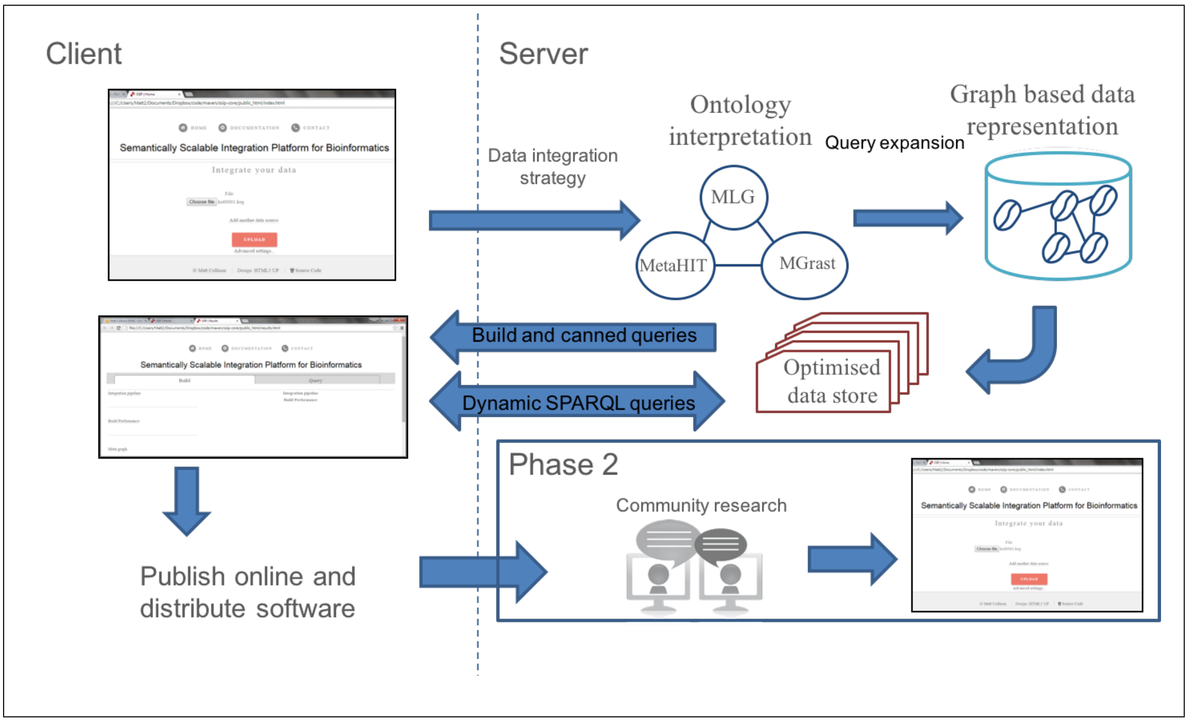


Figure 5.1: Software lifecycle for BioSSIP and the three tier user system.

5.2.1 BioSSIP core: For developers

BioSSIP core is a software API and maven dependency that co-ordinates data handling for bioinformatics data integration and data mining, Figure 5.2. The modular nature of BioSSIP allows the user to develop a standardised and reproducible data integration pipeline in java, the most popular programming language in use at the time of writing (noa, n.d.), with facilities for full version control and cross project dependencies using the maven build tools. BioSSIP natively uses an embedded graph database, Neo4J, for data storage which means the integrated dataset is stored in a scalable and highly performant database that can be distributed and relaunched without manipulating data integrity.

5.2.1.1 Build process - parsers and data integration

BioSSIP is built around the maven build lifecycle. You must first create a project with maven and add the BioSSIP core dependencies. For developers there is a getting started tutorial which guides the user through the process of setting up a project. Once you have a project you will be able to start building parsers and generate an application that will handle the integration process.

Data sources are imported into the BioSSIP platform as `SSIPNode` objects and `SSIPRelationship` objects through parsers that implement the `ParserInterface` and write to a database through a

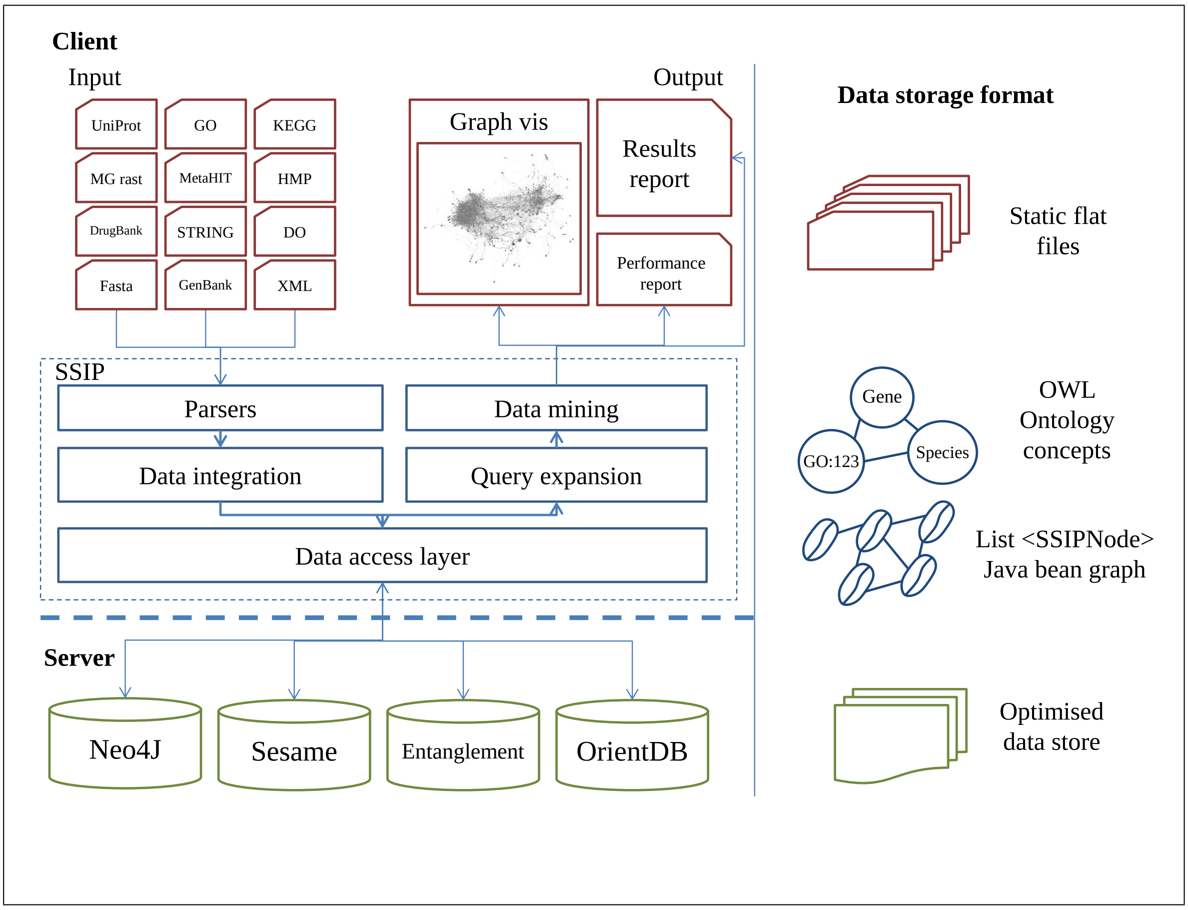


Figure 5.2: **Software architecture for BioSSIP.** Red indicates data stored as static files that make up the input and output of the software. Blue indicates modules of code in the software where data are represented as ontology concepts before being flattened to Java Beans.

data access layer which implements `BackendInterface`, this is a `Neo4J` object by default. Semantic structure is defined by forced `type` and `id` properties on each `SSIPNode`. Metadata is abstracted implicitly from the type properties that are defined in the parser design. There is currently no restrictions on the type property although any new relationships between types are extracted and submitted into a metadata graph that enables the metadata to be stored 'on the fly'.

Uniqueness and equivalence is asserted based on matching the `id` properties in each `SSIPNode`. Multiple IDs can be defined under different `labels` and the labels can be dynamically loaded into the namespace during the data integration process. The merge policy is based on overwriting the existing node so requires a read/write process to retain an existing state. Alternative merge properties can be loaded and discarded from uniqueness checks using the `Neo4J.createIndex(String property)` method. `SSIPNodes` store outgoing relationships in a `Map<SSIPNode,SSIPRelationship>` so within each node it has an empty version of destination nodes without knowing if that entity exists yet. This enables 'hanging' relations. Then at runtime the node is merged with the existing record. Nodes are held in a buffer and committed in batches of 1000 or when the database connection is closed.

5.2.1.2 Database Model and data storage

BioSSIP data is natively stored in an embedded Neo4J database as nodes and edges. Practically there are three levels of abstraction on the data, Figure 5.3. At the machine code level the database is serialised into property hashes and index files. As the database is loaded the identifiers for the indexes contained in each label are loaded into memory to seed queries in the graph and during requests only the information of immediate neighbours are loaded into memory. Subsequent nodes are loaded at request time as the relevant key value stores are engaged through traversals. The Neo4J node and edge concepts are constructed around the key value properties. Neo4J nodes and edges are passed from the BioSSIP client to the database and interpreted during the response. The BioSSIP abstraction level only works on the client side to help structure metadata on the node and edge objects. By structuring ID and metadata concepts during the data build phase it is possible to allows semantics to be layered into the database model without the database model having any knowledge of the semantics. By extracting the type and inferring metadata structure at build time BioSSIP has a live view of the metadata that can be used to interpret the data and structure/query strategies.

As datasets are built in BioSSIP the parsers communicate with the data access layer. The data access layer manages the database connection and stores a buffer of node concepts. This buffer is set to 1000 node concepts by default and is commited when the terminate database connection command is given. Once a database transaction is established the transactionality of the commit is handled within a Transaction container and if the Transaction fails a log is written and the transaction is rolled back as guaranteed by ACID transactions in Neo4J.

5.2.1.3 Query process - data access and distribution

The BioSSIP API provides an interface for writing data mining strategies. This is based on simple rules for traversing the graph to reach an endpoint. The query then represents a formal specification for the data mining strategy.

A major goal in the development of this tool was to enable a development lifecycle for data integration and data mining. The persistent embedded nature of the graph database means the entire database and source code constituting the integration pipeline and any queries can be version controlled, packaged and distributed as a normal maven project. There are multiple advantages to this approach, the optimised data store can be precomputed and distributed, any updated data sources can

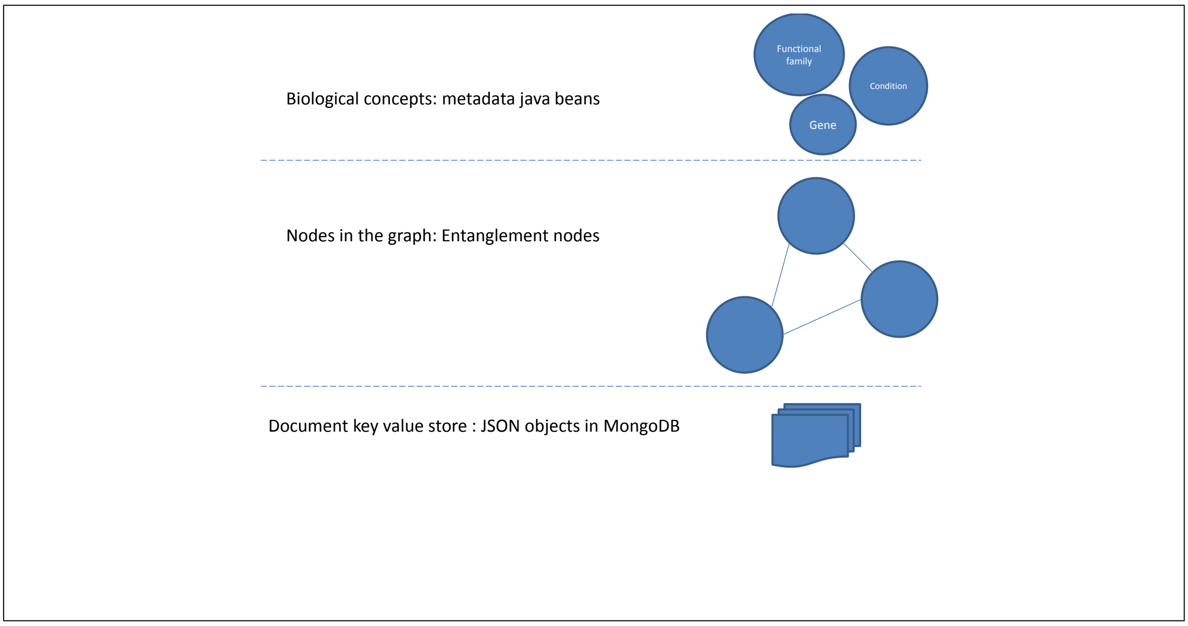


Figure 5.3: **Build times for graphs size with linear and connected graphs.**.

be rebuilt into a new or into the existing database, the exact same data integration pipeline and data mining strategies can be run over multiple versions of the computed database as it changes over time.

5.2.2 BioSSIP Application: For server-side user

The endpoint for a BioSSIP data integration project is an executable BioSSIP application. This application is the package containing a fully version controlled copy of the data integration pipeline including a copy of BioSSIP core, project specific data parsers, the integrated dataset as an optimised Neo4J datastore and any data mining strategies formalised in Query classes.

5.2.2.1 Running the application

All BioSSIP applications contain an executable jar built in Java so the executable application can run on any platform where there is a Java runtime environment. When the application is launched it will establish a web server and a website interface is uses to construct the data integration pipeline through a browser, as shown in Figure 5.4. The application has two main pages after the opening data integration pipeline has been registered and submitted; the build page and the query page, Figure 5.5. The build page simply keeps a log of the input data that has been submitted to generate the dataset. The query page enables live queries and canned queries (preprogrammed data mining strategies) to be executed and the results to be downloaded. In the current state the build and query pages are very simple as they rely in metadata from the pipeline configuration.

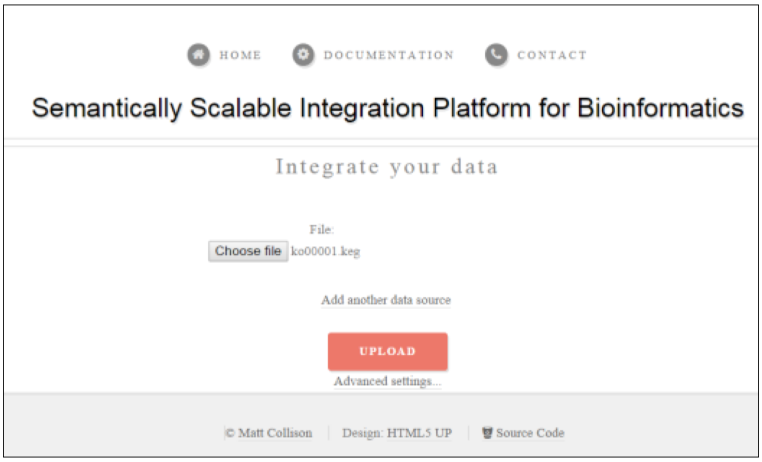


Figure 5.4: The application opening page. The data integration pipeline is built by selecting parsers from the dropdown menu and selecting a datasource to upload or by providing a link to an online resource.

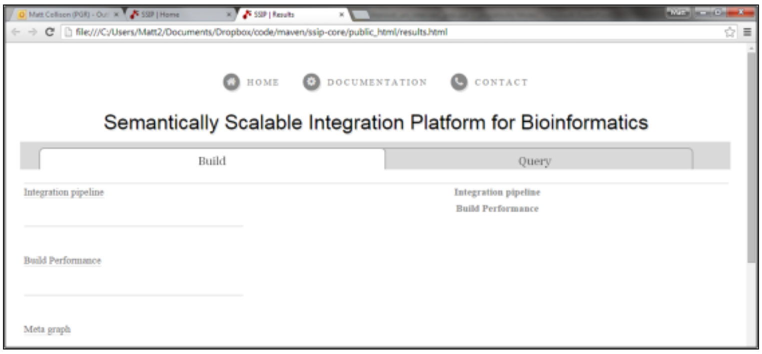


Figure 5.5: Build pages are loaded once the data integration pipeline has been submitted.

5.2.2.2 Data and software distribution and dependencies

BioSSIP is built on maven as a build tool. This platform allows both the core software to be loaded and third party parsers to be shared and loaded easily and effectively. Furthermore the embedded nature of the database means each database can also be distributed and packaged with the software.

5.2.3 BioSSIP users: For client side users

As a client the typical use case is a user that wants to run a new query on an existing dataset or a reviewer that wants to check the development followed rigorous methods but ultimately it's about data visualisation and interpretation.

5.2.3.1 Data visualisation

A simple visualisation was implemented in the browser using sigmajs to get a overview of the returned graph although it is recommended that graph results are explored in Gephi. The Gephi exporter was built to facilitate the export and analysis of results and is called by default on all results returned to the user. An RDF exporter was also developed to enable compatibility with other tools althoug isn't inlcuded in the standard UI.

5.2.4 Branding and licensing

This project was developed under the GNU GPLv3 license. The Neo4J dependency enforces an AGPL license that restricts commercial use of their enterprise software, however, BioSSIP is built on the community edition release that is free for reuse so the entire BioSSIP project has been released under the GPLv3 license.

Branding for BioSSIP was designed as a simple way for people to recognise the platform and the software engineering process it embodies. The research community is a major factor in designing this system so branding and open source licensing was a small but important factors.

Figure 5.6: BioSSIP and SSIP logos.

5.3 Results

5.3.1 KEGG test case

KEGG (Kanehisa & Goto, 2000) was chosen as a suitable test dataset. The data is nicely hierarchical, there is an easily manageable number of nodes that verifying the data structure would be sensible and Kegg is a common annotation source that most developers will be familiar with.

To build the test dataset run the `KeggOrthologyParser`. It is included in the ssip-core dependency at uk.ac.ncl.ssip.parsers, to implement the test case you simply include the ko0001.kegg file in the root directory, create a database object and parser object providing the filepath. On completion you should pass the getAllNodes to GephiExporter and generate a graph of the dataset. You should get a gexf file in you project root directory on completion. Run the file in gephi and should be told there are 10143 nodes and 23043 edges. Once the graph loads you will get a square graph. Select the forced atlas 2 layout and the graph will change shape and start looking like Figure 5.7. Now to check the query function is working, if you construct a query with the properties below and pass the reult to the gephi exporter it should return a grpah like Figure 5.8.

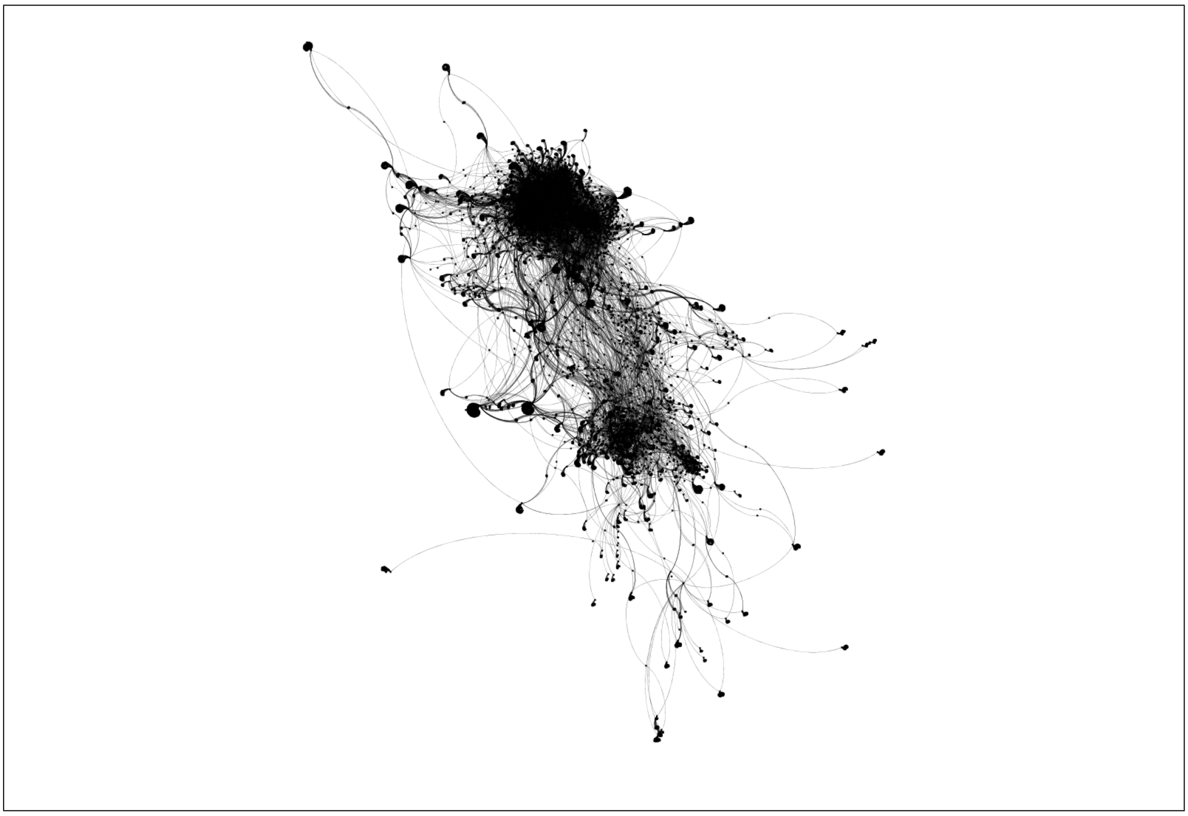


Figure 5.7: A Gephi visualisation of the graph representation of the KEGG orthology in BioSSIP.

5.3.2 Microbiome use case

One of the central resources for microbiome research is the metagenomics gene catalogue. However, without experimental context many of the genes are simply just unannotated genes. This use case maps 20 microbiomes to the gene catalogue from the Qin *et al.* (2010) study. Then maps functional annotations to the catalogue as examples of clustering features. The outcome is an integrated dataset that allows users to discover the metagenomics occurrences of genes in the gene catalogue. The full use case can be seen at the project site.

http://homepages.cs.ncl.ac.uk/matthew.collison/MicroBioSSIP

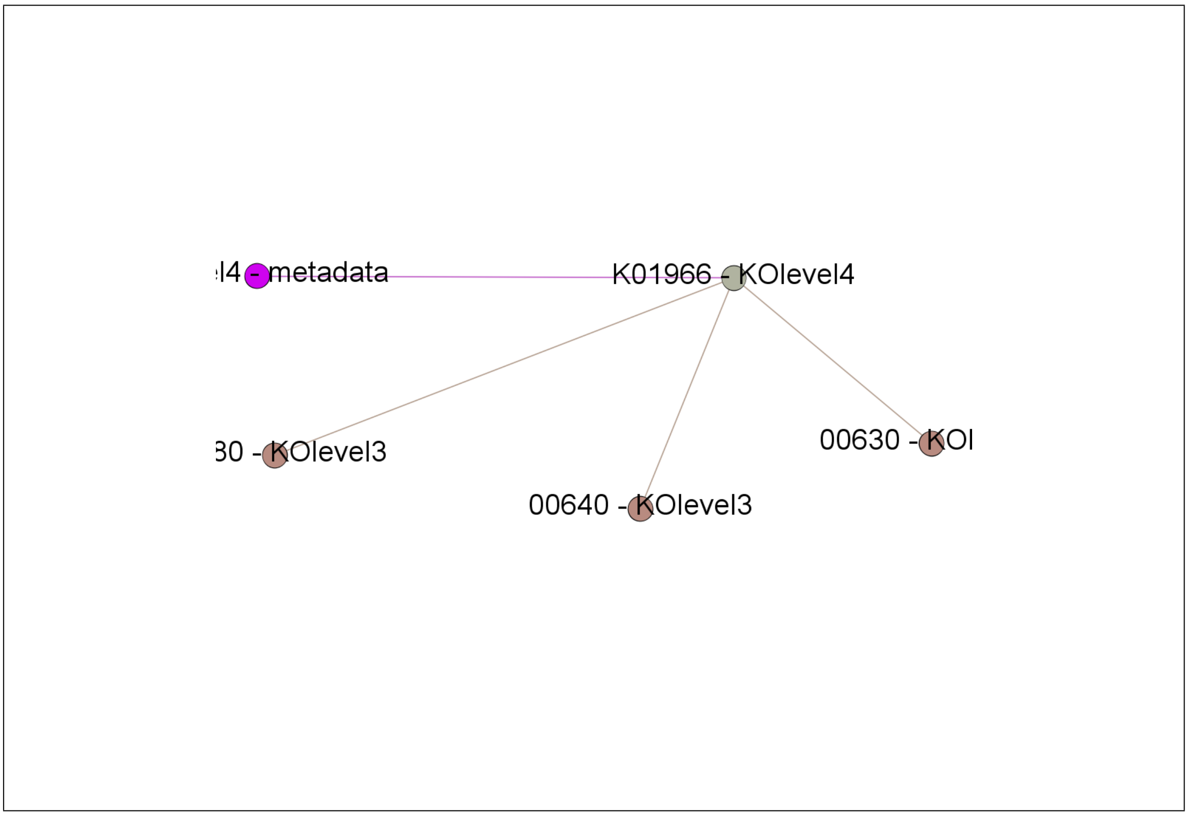


Figure 5.8: A Gephi visualisation of the data visualisation for a simple sub network query.

5.3.3 Third party use case: an integrated dataset for drug repositioning

A BioSSIP application was built by Joseph Mullen at Newcastle University as part of the work contributing to his thesis titled 'Systems Approaches to Drug Repositioning'. The project was also in collaboration with GSK and the Integrative Bioinformatics group in the School of Computing Science. This application was developed concurrently with the BioSSIP framework and makes use of the data integration and pipeline development protocols.

DReNIn, is an integrated dataset developed for drug repositioning. DReNIn uses a high-level drug repositioning ontology to structure data and describes 25 relevant data types and the relationships that exist between these. The integrated dataset includes data from over 20 different sources.

5.4 Discussion

Metagenomics data are inherently large, complex and incomplete datasets. High-throughput technologies have facilitated the emergence of data rich studies and dramatically increased our understanding of how this data can relate to the community level function. This increased data availability has posed challenges to the computational infrastructure for analysing metagenomics data. We have developed a solution to facilitate integrated studies in order to handle the scale and complexity of these data. This approach provides a platform for reproducibility which makes the secondary analyses more accessible to the bioinformaticians. This software is a prototype implementation for a new approach to mircobiome research.

5.4.1 NOSQL for scalable data handling

Graph visualisation tools are common place in integrative bioinformatics. However, graph databases have rarely been used in bioinformatics despite the growing requirement for large scale data handling (Have & Jensen, 2013). This study shows NOSQL graph databases can useful for handling scalability in bioinformatics. Particularly, this study shows one of the first applications of Neo4J and graph databases in bioinformatics.

5.4.2 Data integration developer community

Sharing parsers and derivative datasets could save time, reduce workload and improve standards for the data integration software developer community. Moreover by creating and documenting the data integration and data mining process more clearly it is possible that a new approach to utilising metagenomics data could be adopted by the research community.

5.4.3 Biological insights for drug discovery applications

Host-microbiota interactions are extremely complex and developing effective disease models incorporating the microbiota require a broad understanding of the mechanisms contributing the host-microbiota interactions. Metagenomics data provides a data rich resource, genomics provide context and molecular biology provides mechanisms. Each of these pieces of information has critical information that can inform effective disease models. However, none of them are capable of characterising the microbiota contribution in isolation. Therefore exploring the context of molecular mechanisms in the microbiome is a crucial step in developing understanding of the microbiota, generating effective disease models and designing effective therapeutic strategies. Existing metagenomics analysis tools and annotation pipelines provide a top down view of comparative metagenomics. Here we describe a new approach and a tool to facilitate exploring metagenomics data in a bottom up way to facilitate insight from molecular mechanisms.

Chapter 6

Conclusions, Discussion and Future Work

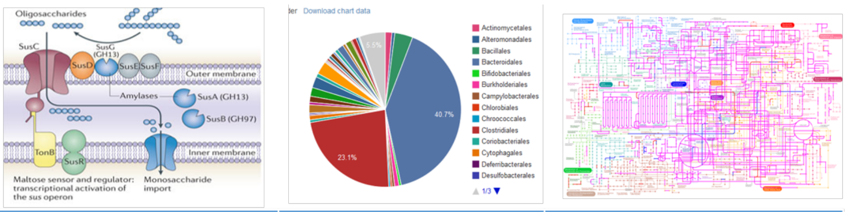
6.1 Project evaluation

The microbiota links human biology with the environment through microbial communities. Microbiome research provides vast opportunities to study and manipulate this relationship to benefit humn health. The aim of this thesis was 'explore the biology of the microbiota through bioinformatics to improve opportunities for new therapeutic strategies'. To summarise, in this thesis I have surveyed the capabilities of existing data mining approaches to identify features of the microbiota associated with disease states, contributed to the development of a novel microbiota targeted therapeutic intervention for type II diabetes, completed a wide scale screening of metagenomics data for previously uncharacterised protease genes proposed to impact host-microbiota and microbiota-microbiota interactions, and developed a data integration tool to facilitate a new bottom up approach to study the microbiome and complete secondary analyses that will be key to developing effective strategies for therapeutic interventions considering the microbiota. The thesis had three objectives:

6.1.1 Evaluate the opportunities for drug discovery from microbiome data analytics

There are three types of experimental methods used to study the microbiome, genomics, 16S metagenomics and whole-genome metagenomics. Each type has a clear limitation in the end points that can be extracted from that data source. Figure 6.1.

This work discusses the limitations of genomics research in that isolated analyses of individual



	Reference Genomes	16S Metagenomics	Whole Genome Metagenomics
Advantages	Characterising cellular mechanisms.	Community trends in taxonomy.	High level gene functions and gene clusters.
Disadvantages	Highly variable and divergent cellular mechanism.	Limited functional insight.	High levels of uncharacterised genes and low coverage.
Drug discovery opportunities	Mostly infectious diseases.	Broad spectrum antibiotics.	Broad antibiotics and small molecules

Figure 6.1: **Experimental methodologies for studying the micorbiome.**

species are fruitless in such as large complex bacterial community. The limitations of a 16S metagenomics data in that only very naieve disease models can be generated. The limitations of whole-genome metagenomics in that the resolution of sampling is not detailed enough to complete reliable analyses of low level genes. Together these studies highlighted thelimitations of existing approaches to analyse each method in isolation and raised the opportunity to exploit the complementary features of each method in integrative analyses that facilitates a new approach to microbiome data analyses.

6.1.2 Evaluate the role of host-microbiota and microbiota-microbiota interactions in health, disease and pharmaceutical interventions

Host-microbiota interactions are complex and understanding the nature of the host-microbiota interface in disease remains a major challenge in development of therapeutic therapies. Developing accuarte testable models for the impact of microbiota in disease is a key step. Chapter 3 contributed to disease understanding by developing a high level disease model and proposing a treatment model for type 2 diabetes. Chapter 4 contributed to disease understanding by revealing thousands of genes linked host-microbiota and microbiota-microbiota interactions that will progress understanding of niche microhabitats key in developing disease models. Chapter 5 contributed to understanding of host-microbiota and microbiota-microbiota interactions by enabling a new approach to microbiome analyses.

6.1.3 Provide computational tools to facilitate microbiome analyses and improve access to microbiome data for drug discovery

The analyses in the early chpaters of this thesis highlight the requirement for reproducible and scientific approaches to data integration. The SSIP framework is an early implementatiopn of a potential solution to these challenges.

6.2 Discussion

Microbiome research has significant challenges in data handling. The increase in data availability has lead to computational challenges in storing, accessing and integrating bioinformatics resources as well as making secondary analyses reproducible. Most commonly used software tools and benchmark datasets, such as MG-RAST and the HMP datasets, are provided as end to end analyses and federated stores to avoid the users needing high performance computational resources or in-depth knowledge of the analysis pipeline. However, this also means most experimental datasets are analysed in isolation and secondary analyses to facilitate screening for novel genes and low level functional analysis required for drug discovery and understanding of molecular mechanisms data integration will be important.

Top-down analyses only produce results comparing the fraction of data that contains secondary annotations from any single annotation source within the constraints of the experimental model. This means integration of datasets with other data sources processed outside of the analysis pipeline, including key dataset produced by major consortium studies, is challenging and often not easily reproducible. Furthermore bottom-up exploration of data is difficult due to data handling challenges associated with handling primary data. For instance studying correlations between all predicted gene clusters across experimental datasets instead of only functionally annotated clusters is beyond the capacity of most anlaysis pipelines.

This thesis documents and quantifies the possibilities and limitations of existing approaches to microbiome analyses in drug development, shows the advantages and scope for bottom up analyses of metagenomics data and finally provides a computational tool to support a new approach to microbiome analyses that could bridge the gap between high level metagenomics studies and low level molecular biology studies that is needed for effective design of therapeutic strategies that considers host-microbiota interactions.

Appendix I - Subject identification

Table 6.1: **Subject identification and assignment to analyses. A tick indicates that subject was included in analyses for the corresponding techniques. An astericks indicates the daat collection for that subject is missing measurements for Insulin, Amylin, Leptin, Ghrelin, GIP, PYY, PPr, active GLP-1 and total GLP-1. Superscript notation means measurements for the names metabolite are missing and @ means multiple others are also missing.**

Subject ID	Condition	Weight	metabolite profiling	16S metagenomics	WG metagenomics
1	L	✔	✔^{LIP}	✔	✔
2	L	✔	✔	✔	✔
3	L	✔	✔^{PPr}	✔	
4	L	✔	✔		
5	L	✔	✔	✔	
6	L	✔	✔*	✔	
7	L	✔	✔	✔	
8	L	✔	✔*	✔	
9	L	✔	✔	✔	
10	L	✔	✔	✔	
11	O	✔	✔		✔
12	O	✔	✔	✔	✔
13	O	✔	✔*	✔	
14	O	✔	✔	✔	
15	O	✔	✔	✔	
16	O	✔	✔	✔	
17	O	✔	✔	✔	
18	O	✔	✔	✔	
19	O	✔	✔	✔	
20	O	✔	✔	✔	
21	A_{500}	✔	✔	✔	
22	A_{500}	✔	✔^{Ghrelin}	✔	
23	A_{500}	✔	✔^{PPr}	✔	
24	A_{500}	✔	✔	✔	
25	A_{500}	✔	✔	✔	
26	A_{500}	✔	✔	✔	

Table 6.1 – continued from previous page

Subject ID	Condition	Weight	metabolite profiling	16S metagenomics	WG metagenomics
27	A_{500}	✔	✔	✔	
28	A_{500}	✔	✔	✔	
29	A_{500}	✔	✔	✔	
30	A_{500}				
31	A_{150}	✔			
32	A_{150}	✔			
33	A_{150}	✔			
34	A_{150}	✔			
35	A_{150}	✔			
36	A_{150}	✔			
37	A_{150}	✔			
38	A_{150}	✔			
39	A_{150}	✔			
40	A_{150}	✔			
41	A_{50}	✔			
42	A_{50}	✔			
43	A_{50}	✔			
44	A_{50}	✔			
45	A_{50}	✔			
46	A_{50}	✔			
47	A_{50}	✔			
48	A_{50}	✔			
49	A_{50}	✔			
50	A_{50}				
51	B_{500}	✔	✔*	✔	
52	B_{500}	✔	✔		
53	B_{500}	✔	✔	✔	
54	B_{500}	✔	✔	✔	
55	B_{500}	✔	✔	✔	
56	B_{500}	✔	✔*@		
57	B_{500}	✔	✔	✔	
58	B_{500}	✔	✔	✔	
59	B_{500}	✔	✔	✔	
60	B_{500}	✔	✔		
61	B_{150}	✔			
62	B_{150}	✔			
63	B_{150}	✔			
64	B_{150}	✔			
65	B_{150}				
66	B_{150}	✔			

Table 6.1 – continued from previous page

Subject ID	Condition	Weight	metabolite profiling	16S metagenomics	WG metagenomics
67	B_{150}	✔			
68	B_{150}	✔			
69	B_{150}	✔			
70	B_{150}	✔			
71	B_{50}	✔	✔^{PPr}		
72	B_{50}	✔	✔		
73	B_{50}	✔	✔	✔	
74	B_{50}	✔	✔	✔	
75	B_{50}	✔	✔	✔	
76	B_{50}	✔	✔	✔	
77	B_{50}	✔	✔	✔	
78	B_{50}	✔	✔		
79	B_{50}	✔	✔	✔	
80	B_{50}	✔	✔		
81	$D_{10\%}$	✔	✔	✔	✔
82	$D_{10\%}$	✔	✔*	✔	
83	$D_{10\%}$	✔	✔	✔	
84	$D_{10\%}$	✔	✔		
85	$D_{10\%}$	✔	✔	✔	
86	$D_{10\%}$	✔	✔	✔	
87	$D_{10\%}$	✔	✔	✔	
88	$D_{10\%}$	✔	✔	✔	
89	$D_{10\%}$	✔	✔	✔	
90	$D_{10\%}$	✔	✔	✔	

Appendix II - Normal probability plots

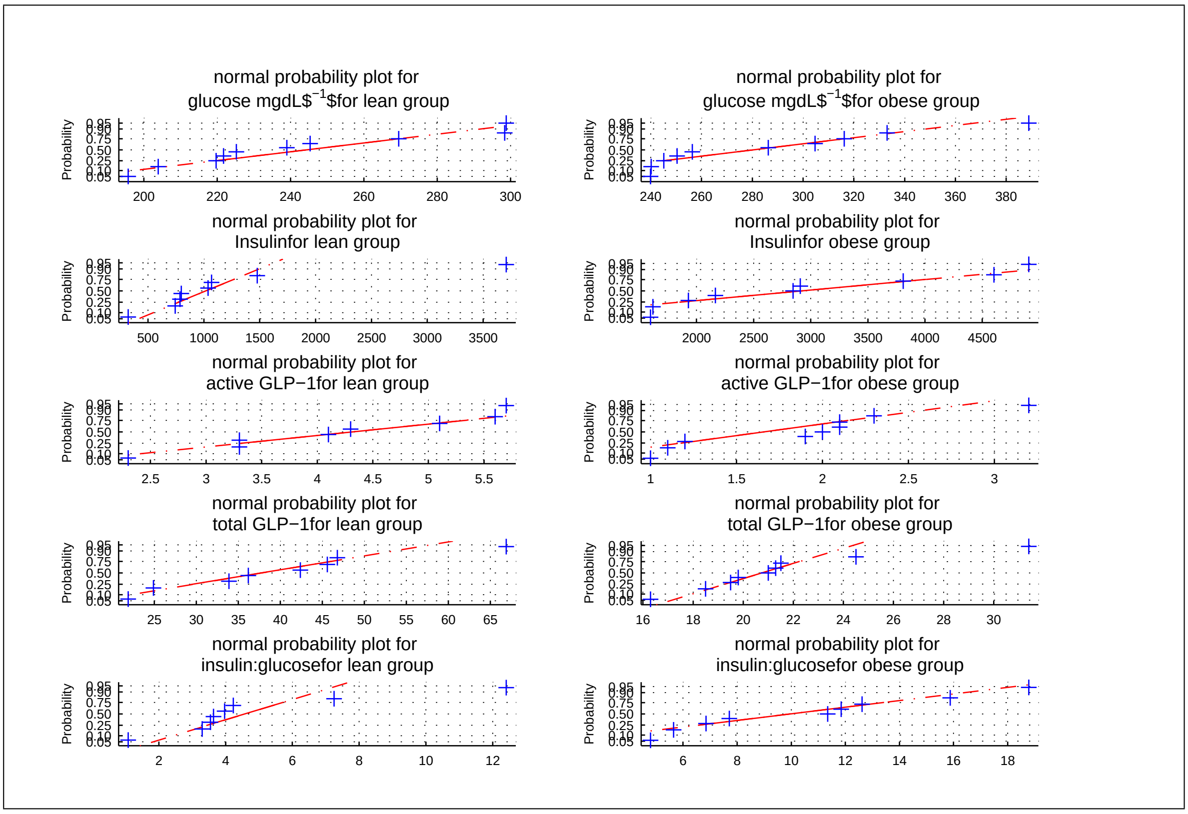


Figure 6.2: **Normal probability plots of each metabolite for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using modelvalidation.m.

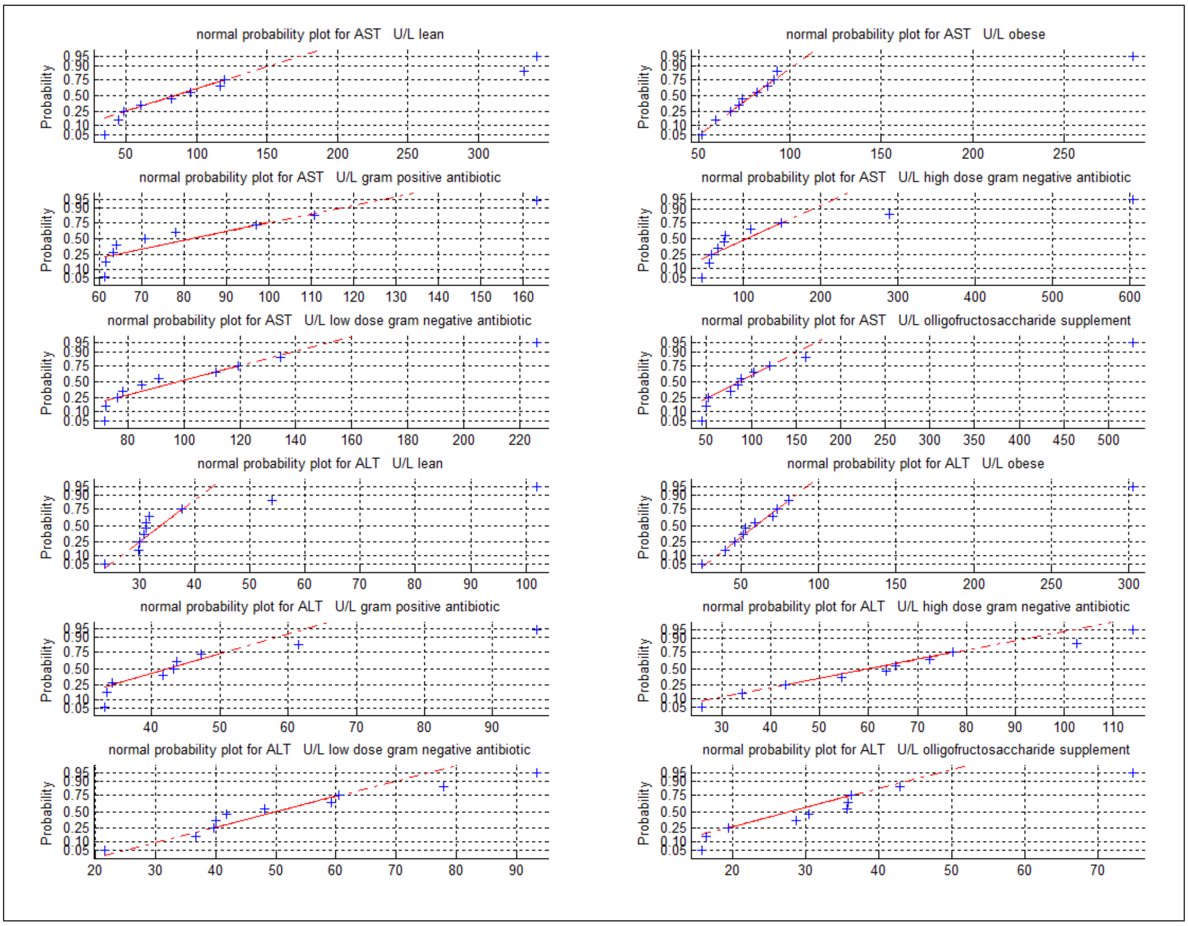


Figure 6.3: **Normal probability plots of each liver function marker for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

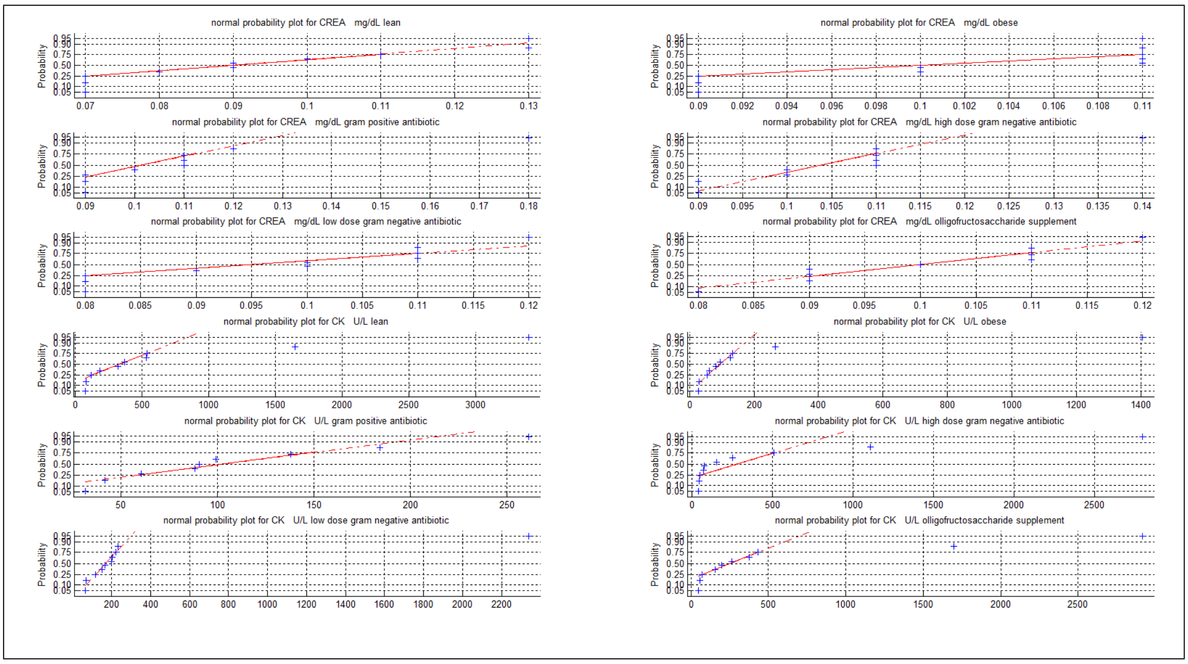


Figure 6.4: **Normal probability plots of each kidney function marker for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

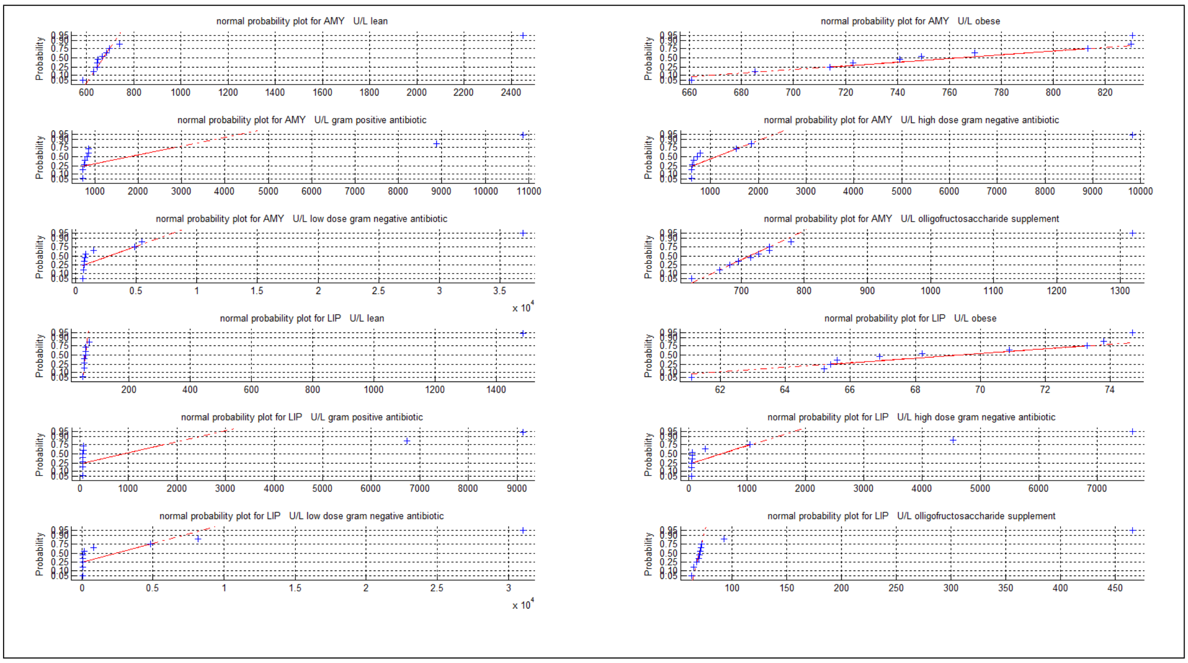


Figure 6.5: **Normal probability plots of each pancreatic function marker for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

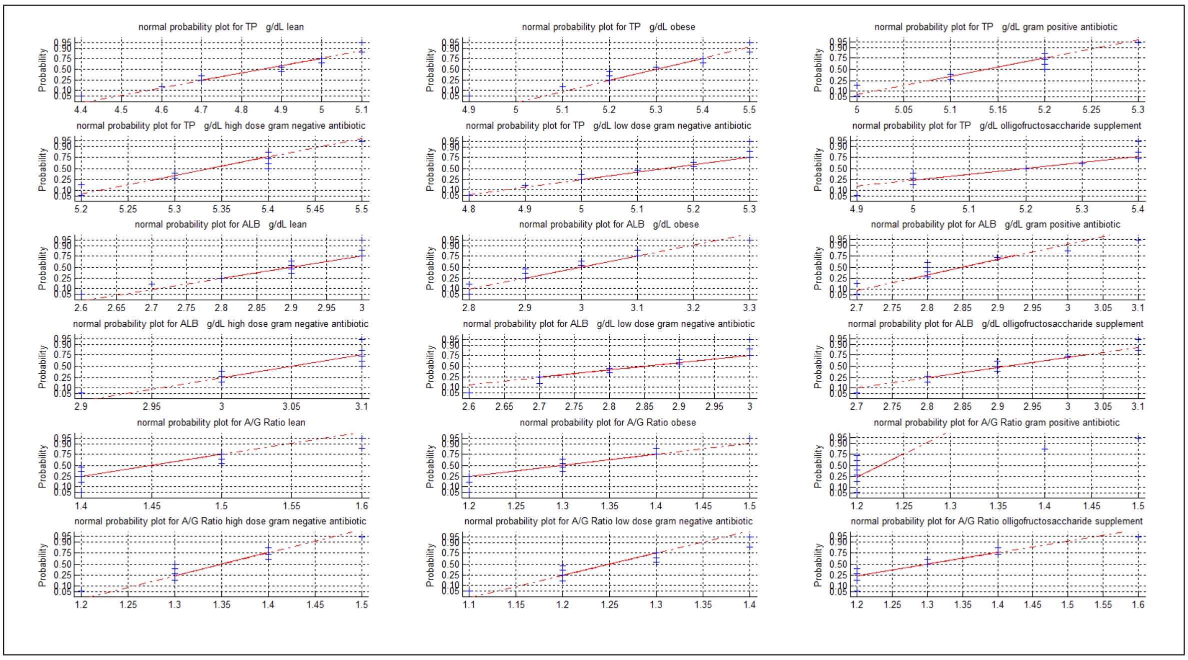


Figure 6.6: **Normal probability plots of each plasma protein marker for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

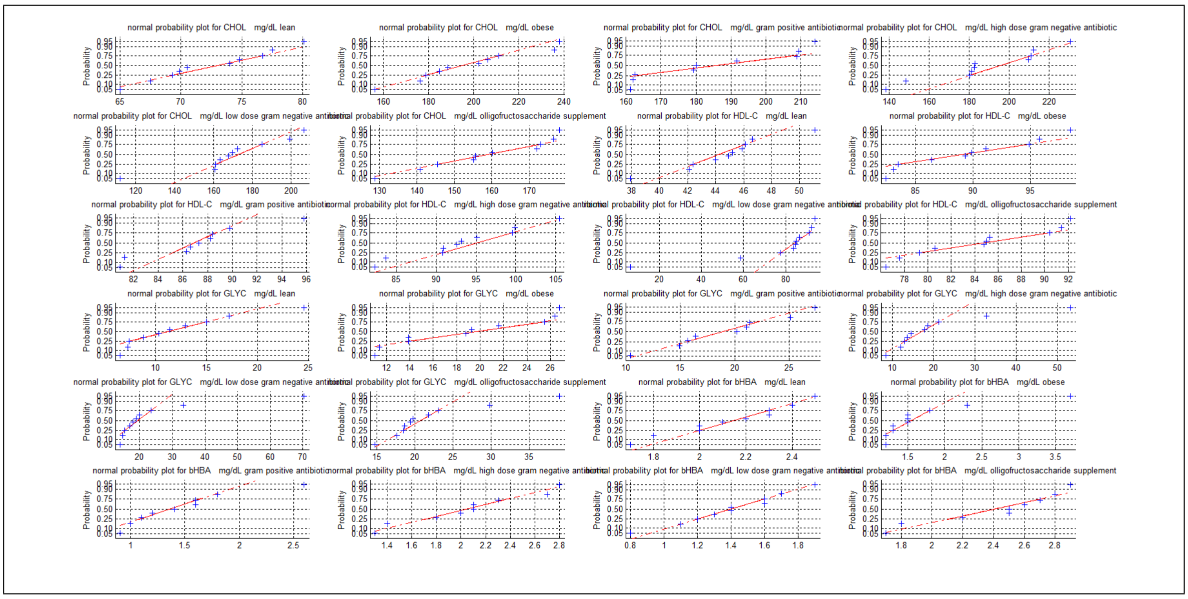


Figure 6.7: **Normal probability plots of each energy substrate for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

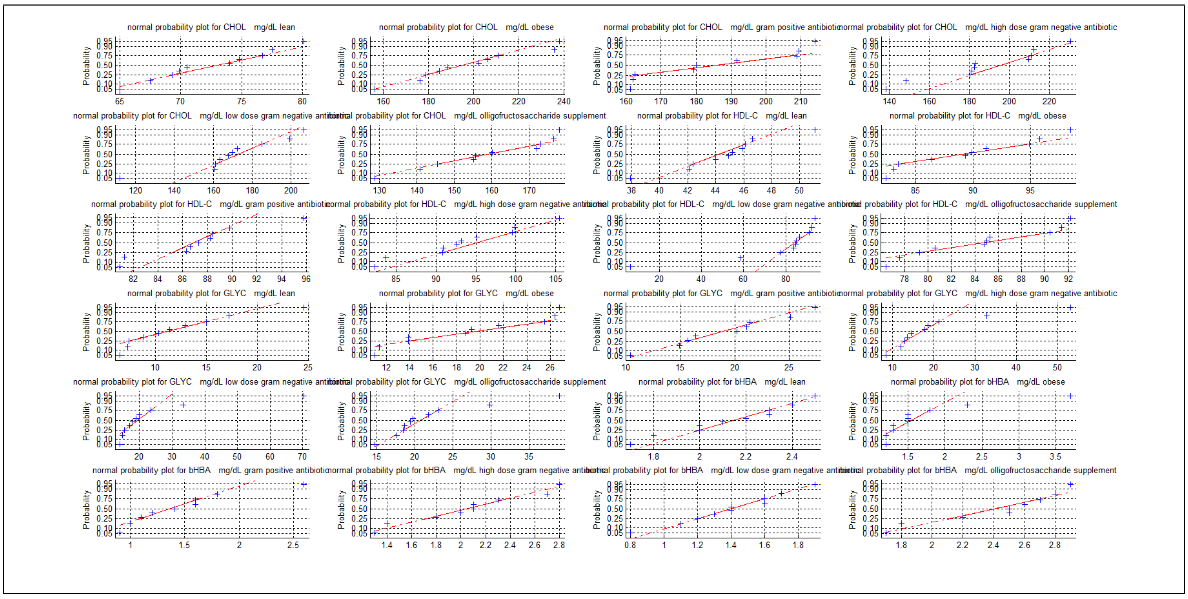


Figure 6.8: **Normal probability plots of each energy substrate for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

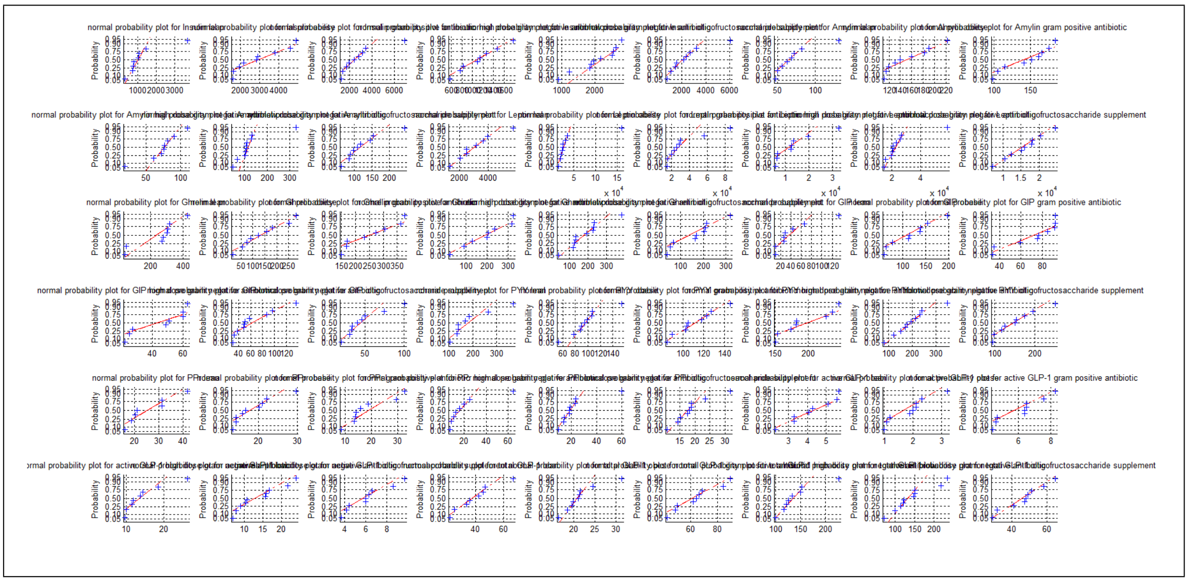


Figure 6.9: **Normal probability plots of each endocrine signalling molecule for each group.** Simulated data following a predicted normal distribution with real mean and standard deviation is present in the background in red. If the real data is normally distributed it will be linear and denser in the middle section, following the red simulation. This graph and the statistical values were calculcated and generated using treatmentphysiology.m.

Appendix III - Complete gene configuration table

Table 6.2: **Gene configuration information for all CBM-protease pairs.**

MEROPS family	**Gene configuration**	**No of hits**	**No of distinct genes**
A01	A01-CBM50	10	6
A01	A01-CBM2	1	1
A01	A01-CBM32	1	1
A01	A01-CBM54	1	1
A11	A11-CBM48	1	1
A22	A22-CBM67	1	1
C01	C01-CBM37	25	9
C01	C01-CBM37-CBM37-CBM37-CBM37	8	3
C01	C01-CBM37-CBM37	7	4
C01	C01-CBM46	6	4
C01	C01-CBM32	3	3
C01	C01-CBM58	2	1
C01	C01-CBM56	2	1
C01	C01-CBM32-CBM46	1	1
C01	C01-CBM37-CBM37-CBM37	1	1
C02	C02-CBM32	12	4
C10	C10-CBM61	3	2
C10	C10-CBM37-CBM37	2	1
C10	C10-CBM6	1	1
C11	C11-CBM56	9	5
C12	C12-CBM26	1	1
C13	C13-CBM32-CBM32	1	1
C14	C14-CBM37	14	8
C14	C14-CBM37-CBM37	1	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
C14	C14-CBM32	1	1
C14	C14-CBM37-CBM37-CBM37	1	1
C15	C15-CBM35	1	1
C16	C16-CBM67	1	1
C19	C19-CBM32	1	1
C25	C25-CBM35	2	2
C26	C26-CBM65	1	1
C26	C26-CBM32	1	1
C39	C39-CBM13	2	1
C40	C40-CBM50	953	298
C40	C40-CBM50-CBM50	312	95
C40	C40-CBM44	137	32
C40	C40-CBM50-CBM50-CBM50	131	43
C40	C40-CBM50-CBM50-CBM50-CBM50-CBM50	11	5
C40	C40-CBM40	10	1
C40	C40-CBM50-CBM50-CBM50-CBM50	9	4
C40	C40-CBM32	8	3
C40	C40-CBM13	7	4
C40	C40-CBM44-CBM44	7	2
C40	C40-CBM56	5	3
C40	C40-CBM54	4	4
C40	C40-CBM44-CBM40	3	1
C40	C40-CBM37	3	2
C40	C40-CBM35	2	2
C40	C40-CBM9	2	1
C40	C40-CBM61	2	2
C40	C40-CBM37-CBM37	2	2
C40	C40-CBM4	1	1
C40	C40-CBM26-CBM26	1	1
C40	C40-CBM6	1	1
C40	C40-CBM37-CBM37-CBM37	1	1
C40	C40-CBM12	1	1
C40	C40-CBM44-CBM37	1	1
C40	C40-CBM50-CBM50-CBM50-CBM50-CBM50-CBM50	1	1
C40	C40-CBM66	1	1
C44	C44-CBM67	9	1
C44	C44-CBM32	1	1
C44	C44-CBM48-CBM68	1	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
C45	C45-CBM32	15	3
C50	C50-CBM50	1	1
C51	C51-CBM13	8	6
C51	C51-CBM37	4	4
C51	C51-CBM13-CBM13	1	1
C56	C56-CBM34	1	1
C59	C59-CBM67	1	1
C60	C60-CBM20	1	1
C60	C60-CBM62	1	1
C69	C69-CBM67	4	1
C82	C82-CBM35	26	2
C82	C82-CBM37-CBM37	22	3
C82	C82-CBM37	17	7
C82	C82-CBM37-CBM37-CBM37	9	2
C82	C82-CBM50	6	5
C82	C82-CBM61	5	2
C82	C82-CBM69	4	2
C82	C82-CBM32	3	1
C82	C82-CBM37-CBM32	2	1
C82	C82-CBM13-CBM13	1	1
C82	C82-CBM6	1	1
C97	C97-CBM50	1	1
I32	I32-CBM4	1	1
I39	I39-CBM42	21	2
I39	I39-CBM22	9	6
I39	I39-CBM34	6	3
I39	I39-CBM16	3	2
I39	I39-CBM23	2	2
I39	I39-CBM56	2	1
I39	I39-CBM51	2	2
I39	I39-CBM39	1	1
I39	I39-CBM13	1	1
I39	I39-CBM68	1	1
I39	I39-CBM21	1	1
I39	I39-CBM32	1	1
I39	I39-CBM4	1	1
I43	I43-CBM37	297	120
I43	I43-CBM37-CBM37	182	89

MEROPS family	Gene configuration	No of hits	No of distinct genes
I43	I43-CBM37-CBM37-CBM37	67	35
I43	I43-CBM37-CBM37-CBM37-CBM37	37	11
I43	I43-CBM26	10	1
I43	I43-CBM37-CBM37-CBM37-CBM37-CBM37	9	6
I43	I43-CBM16	9	5
I43	I43-CBM32	9	8
I43	I43-CBM4-CBM37-CBM37-CBM37-CBM37	7	1
I43	I43-CBM37-CBM32-CBM37-CBM37	5	1
I43	I43-CBM4	4	2
I43	I43-CBM23	4	1
I43	I43-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37	3	2
I43	I43-CBM37-CBM37-CBM37-CBM16-CBM37-CBM37	3	1
I43	I43-CBM32-CBM37-CBM37	3	1
I43	I43-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37	3	1
I43	I43-CBM16-CBM37-CBM37-CBM37-CBM37-CBM37	3	2
I43	I43-CBM61	2	2
I43	I43-CBM46-CBM23-CBM46-CBM37-CBM37-CBM46-CBM37	1	1
I43	I43-CBM46-CBM46-CBM37-CBM46-CBM23-CBM37-CBM37	1	1
I43	I43-CBM6-CBM37	1	1
I43	I43-CBM37-CBM37-CBM16	1	1
I43	I43-CBM37-CBM37-CBM37-CBM37-CBM16	1	1
I43	I43-CBM37-CBM46-CBM23	1	1
I43	I43-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37-CBM37	1	1
I43	I43-CBM37-CBM37-CBM16-CBM37-CBM37-CBM37	1	1
I43	I43-CBM37-CBM16-CBM37-CBM37-CBM37-CBM37	1	1
I43	I43-CBM23-CBM37-CBM46	1	1
I43	I43-CBM37-CBM46-CBM23-CBM46-CBM37-CBM37-CBM46	1	1
I43	I43-CBM13-CBM13-CBM37	1	1
I43	I43-CBM37-CBM16	1	1
I43	I43-CBM46-CBM46-CBM37-CBM46-CBM23-CBM37	1	1
I43	I43-CBM46-CBM37-CBM46-CBM23-CBM46-CBM37	1	1
I43	I43-CBM20	1	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
I43	I43-CBM46-CBM23-CBM37-CBM46-CBM46-CBM37-CBM37	1	1
I43	I43-CBM32-CBM37	1	1
I43	I43-CBM46	1	1
I43	I43-CBM23-CBM46-CBM37	1	1
I43	I43-CBM22	1	1
I43	I43-CBM44	1	1
I63	I63-CBM13	8	3
I63	I63-CBM56	2	2
I63	I63-CBM50	2	1
I63	I63-CBM13-CBM13	1	1
I63	I63-CBM2-CBM2	1	1
I63	I63-CBM6	1	1
I63	I63-CBM51	1	1
I63	I63-CBM32	1	1
M01	M01-CBM32	1	1
M02	M02-CBM26	1	1
M03	M03-CBM11	2	2
M03	M03-CBM47	1	1
M03	M03-CBM32	1	1
M03	M03-CBM35	1	1
M04	M04-CBM35	26	4
M04	M04-CBM4	3	2
M04	M04-CBM51	1	1
M04	M04-CBM41-CBM25-CBM41-CBM48	1	1
M04	M04-CBM8	1	1
M06	M06-CBM44	29	8
M06	M06-CBM37	18	8
M06	M06-CBM32	7	3
M06	M06-CBM16	4	4
M06	M06-CBM38	3	2
M06	M06-CBM46	2	1
M06	M06-CBM44-CBM44	2	2
M06	M06-CBM11	1	1
M06	M06-CBM61	1	1
M07	M07-CBM46	1	1
M09	M09-CBM44	79	27
M09	M09-CBM32-CBM44	5	3

MEROPS family	Gene configuration	No of hits	No of distinct genes
M09	M09-CBM44-CBM37	2	2
M09	M09-CBM37-CBM44	1	1
M09	M09-CBM62-CBM62	1	1
M10	M10-CBM20-CBM20	32	3
M10	M10-CBM20	12	2
M10	M10-CBM32-CBM62	6	1
M10	M10-CBM13	1	1
M12	M12-CBM50	2	1
M12	M12-CBM58	1	1
M13	M13-CBM37	9	1
M13	M13-CBM67	2	2
M13	M13-CBM20-CBM20	2	2
M14	M14-CBM50	119	30
M14	M14-CBM37	24	8
M14	M14-CBM13	7	5
M14	M14-CBM40	2	2
M14	M14-CBM13-CBM13-CBM13-CBM13-CBM13	1	1
M14	M14-CBM13-CBM13-CBM13-CBM13	1	1
M14	M14-CBM4-CBM32	1	1
M14	M14-CBM32	1	1
M14	M14-CBM13-CBM13	1	1
M14	M14-CBM37-CBM37	1	1
M15	M15-CBM37-CBM37-CBM37-CBM37	14	1
M15	M15-CBM37	3	1
M15	M15-CBM37-CBM37-CBM37	2	1
M15	M15-CBM37-CBM37	1	1
M15	M15-CBM35	1	1
M16	M16-CBM6	2	1
M16	M16-CBM61	1	1
M16	M16-CBM4-CBM4	1	1
M17	M17-CBM9	1	1
M18	M18-CBM5	4	2
M18	M18-CBM12	1	1
M18	M18-CBM48	1	1
M19	M19-CBM34	14	1
M19	M19-CBM10	1	1
M20	M20-CBM48	6	2
M20	M20-CBM66	6	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
M20	M20-CBM22	3	1
M20	M20-CBM31	3	1
M20	M20-CBM67	2	2
M20	M20-CBM20-CBM20	1	1
M20	M20-CBM35	1	1
M20	M20-CBM40	1	1
M20	M20-CBM32	1	1
M22	M22-CBM62-CBM62	2	1
M22	M22-CBM4	2	1
M22	M22-CBM14	1	1
M22	M22-CBM37	1	1
M22	M22-CBM62	1	1
M23	M23-CBM50	4078	1211
M23	M23-CBM50-CBM50	1060	335
M23	M23-CBM50-CBM50-CBM50	413	113
M23	M23-CBM50-CBM50-CBM50-CBM50	76	32
M23	M23-CBM50-CBM50-CBM50-CBM50-CBM50	31	14
M23	M23-CBM37-CBM50	9	5
M23	M23-CBM13	5	4
M23	M23-CBM50-CBM37	4	3
M23	M23-CBM13-CBM13	3	3
M23	M23-CBM63-CBM50	3	1
M23	M23-CBM61-CBM50	2	1
M23	M23-CBM66	2	2
M23	M23-CBM61-CBM61-CBM50-CBM61	2	2
M23	M23-CBM50-CBM61-CBM61-CBM61	2	2
M23	M23-CBM50-CBM61	1	1
M23	M23-CBM6-CBM50	1	1
M23	M23-CBM61	1	1
M23	M23-CBM4	1	1
M23	M23-CBM50-CBM28	1	1
M23	M23-CBM37-CBM13	1	1
M23	M23-CBM11	1	1
M23	M23-CBM50-CBM50-CBM50-CBM50-CBM50-CBM50	1	1
M23	M23-CBM16	1	1
M23	M23-CBM37	1	1
M23	M23-CBM56	1	1
M26	M26-CBM37	7	3

MEROPS family	Gene configuration	No of hits	No of distinct genes
M26	M26-CBM56	5	3
M26	M26-CBM62	2	2
M26	M26-CBM61	2	2
M26	M26-CBM32	2	2
M26	M26-CBM46	2	2
M26	M26-CBM2	1	1
M26	M26-CBM62-CBM32	1	1
M26	M26-CBM20	1	1
M26	M26-CBM35	1	1
M26	M26-CBM32-CBM32-CBM32	1	1
M28	M28-CBM35	1	1
M28	M28-CBM16	1	1
M38	M38-CBM8	3	1
M38	M38-CBM32	3	2
M38	M38-CBM56	1	1
M41	M41-CBM22	8	1
M41	M41-CBM35	1	1
M41	M41-CBM64	1	1
M42	M42-CBM50	1	1
M42	M42-CBM62	1	1
M43	M43-CBM40	404	144
M43	M43-CBM66	55	20
M43	M43-CBM32	45	7
M43	M43-CBM40-CBM13	15	3
M43	M43-CBM40-CBM40	7	5
M43	M43-CBM13-CBM40	5	2
M43	M43-CBM40-CBM32	4	2
M43	M43-CBM66-CBM40	3	2
M43	M43-CBM32-CBM40	3	1
M43	M43-CBM56-CBM40	2	1
M43	M43-CBM36	2	1
M43	M43-CBM32-CBM32	1	1
M43	M43-CBM51-CBM66	1	1
M43	M43-CBM32-CBM6	1	1
M43	M43-CBM44	1	1
M43	M43-CBM57	1	1
M43	M43-CBM13	1	1
M43	M43-CBM40-CBM51	1	1

MEROPS family	**Gene configuration**	**No of hits**	**No of distinct genes**
M43	M43-CBM40-CBM56	1	1
M43	M43-CBM35-CBM40	1	1
M43	M43-CBM9	1	1
M43	M43-CBM44-CBM44	1	1
M50	M50-CBM9	21	1
M50	M50-CBM66	7	4
M50	M50-CBM23	3	1
M56	M56-CBM70	1	1
M60	M60-CBM32	148	34
M60	M60-CBM37-CBM32	1	1
M60	M60-CBM13	1	1
M60	M60-CBM32-CBM51	1	1
M61	M61-CBM51	1	1
M67	M67-CBM50	114	27
M81	M81-CBM51	1	1
M82	M82-CBM27	1	1
N06	N06-CBM2	1	1
N10	N10-CBM16	2	2
N10	N10-CBM61	1	1
P01	P01-CBM25	2	2
P01	P01-CBM35	1	1
S01	S01-CBM32	10	2
S01	S01-CBM37	9	5
S01	S01-CBM51	7	5
S01	S01-CBM67	2	2
S01	S01-CBM32-CBM32	2	1
S01	S01-CBM40	1	1
S01	S01-CBM46	1	1
S01	S01-CBM13	1	1
S01	S01-CBM26	1	1
S01	S01-CBM3	1	1
S01	S01-CBM2	1	1
S01	S01-CBM19	1	1
S06	S06-CBM32	3	2
S08	S08-CBM44	104	36
S08	S08-CBM26	73	16
S08	S08-CBM37	29	10
S08	S08-CBM40	27	13

MEROPS family	**Gene configuration**	**No of hits**	**No of distinct genes**
S08	S08-CBM32	26	11
S08	S08-CBM26-CBM26	26	9
S08	S08-CBM44-CBM44	24	9
S08	S08-CBM11	23	1
S08	S08-CBM26-CBM48	15	2
S08	S08-CBM37-CBM37	8	3
S08	S08-CBM13	8	6
S08	S08-CBM35	7	4
S08	S08-CBM22	6	5
S08	S08-CBM40-CBM13	6	1
S08	S08-CBM62	5	1
S08	S08-CBM2-CBM2	5	3
S08	S08-CBM66	4	2
S08	S08-CBM13-CBM13	4	4
S08	S08-CBM2	4	4
S08	S08-CBM46-CBM32	4	2
S08	S08-CBM26-CBM26-CBM48	4	1
S08	S08-CBM32-CBM32	3	2
S08	S08-CBM46	3	3
S08	S08-CBM44-CBM40	3	1
S08	S08-CBM67	2	2
S08	S08-CBM66-CBM66-CBM67	2	1
S08	S08-CBM6	2	2
S08	S08-CBM56	2	2
S08	S08-CBM13-CBM13-CBM13	2	2
S08	S08-CBM23	1	1
S08	S08-CBM61	1	1
S08	S08-CBM48-CBM26	1	1
S08	S08-CBM32-CBM32-CBM32	1	1
S08	S08-CBM66-CBM6-CBM66	1	1
S08	S08-CBM51-CBM32-CBM32	1	1
S08	S08-CBM32-CBM46	1	1
S08	S08-CBM13-CBM32-CBM6-CBM32-CBM32	1	1
S08	S08-CBM40-CBM40	1	1
S08	S08-CBM44-CBM32	1	1
S08	S08-CBM35-CBM35	1	1
S08	S08-CBM26-CBM48-CBM26	1	1
S08	S08-CBM3	1	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
S08	S08-CBM32-CBM32-CBM6-CBM32	1	1
S08	S08-CBM32-CBM32-CBM51	1	1
S08	S08-CBM4	1	1
S08	S08-CBM22-CBM6-CBM22	1	1
S08	S08-CBM22-CBM22	1	1
S08	S08-CBM57	1	1
S08	S08-CBM58	1	1
S08	S08-CBM32-CBM32-CBM13-CBM6-CBM32	1	1
S08	S08-CBM30-CBM4	1	1
S08	S08-CBM44-CBM22-CBM16	1	1
S08	S08-CBM32-CBM32-CBM6-CBM13-CBM32	1	1
S08	S08-CBM48	1	1
S09	S09-CBM48	611	110
S09	S09-CBM44	58	17
S09	S09-CBM37	51	30
S09	S09-CBM37-CBM37-CBM37	38	4
S09	S09-CBM20	32	6
S09	S09-CBM48-CBM6	30	3
S09	S09-CBM48-CBM48	27	3
S09	S09-CBM37-CBM44	17	2
S09	S09-CBM37-CBM37	10	7
S09	S09-CBM32	10	6
S09	S09-CBM6-CBM48	9	2
S09	S09-CBM26	9	4
S09	S09-CBM62-CBM62	8	1
S09	S09-CBM22	4	4
S09	S09-CBM4	4	4
S09	S09-CBM44-CBM44	3	3
S09	S09-CBM34	3	1
S09	S09-CBM51	3	1
S09	S09-CBM66	2	1
S09	S09-CBM51-CBM51	2	1
S09	S09-CBM29	2	1
S09	S09-CBM6	2	1
S09	S09-CBM67	1	1
S09	S09-CBM23	1	1
S09	S09-CBM62	1	1
S09	S09-CBM40	1	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
S09	S09-CBM64	1	1
S09	S09-CBM26-CBM37	1	1
S09	S09-CBM37-CBM23	1	1
S09	S09-CBM13	1	1
S10	S10-CBM67	1	1
S10	S10-CBM11	1	1
S11	S11-CBM37	6	4
S11	S11-CBM32	4	2
S11	S11-CBM22	3	2
S11	S11-CBM67	3	1
S12	S12-CBM32	16	2
S12	S12-CBM9	1	1
S13	S13-CBM13-CBM13	1	1
S13	S13-CBM48	1	1
S14	S14-CBM32	3	1
S14	S14-CBM64	1	1
S14	S14-CBM61	1	1
S14	S14-CBM67	1	1
S16	S16-CBM20	2	1
S16	S16-CBM32	1	1
S24	S24-CBM48	1	1
S26	S26-CBM12	2	2
S26	S26-CBM30	1	1
S26	S26-CBM19	1	1
S26	S26-CBM48	1	1
S33	S33-CBM32	38	9
S33	S33-CBM20	7	1
S33	S33-CBM51	4	1
S33	S33-CBM67	3	1
S33	S33-CBM4	2	1
S33	S33-CBM13	1	1
S33	S33-CBM12	1	1
S33	S33-CBM66	1	1
S33	S33-CBM44	1	1
S33	S33-CBM16	1	1
S33	S33-CBM48	1	1
S33	S33-CBM3	1	1
S41	S41-CBM4	2	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
S45	S45-CBM48	1	1
S49	S49-CBM20	2	1
S54	S54-CBM32	1	1
S55	S55-CBM16	12	2
S66	S66-CBM5	1	1
T01	T01-CBM9	2	1
T03	T03-CBM6	2	1
T03	T03-CBM67	2	1
T03	T03-CBM62	1	1
T03	T03-CBM37-CBM32	1	1
T03	T03-CBM48	1	1
T03	T03-CBM48-CBM48	1	1
T05	T05-CBM70	1	1
U32	U32-CBM39	5	1
U32	U32-CBM62-CBM32	1	1
U32	U32-CBM16	1	1
U62	U62-CBM36	2	1
U69	U69-CBM56	60	37
U69	U69-CBM32	47	15
U69	U69-CBM37	19	7
U69	U69-CBM4-CBM4-CBM4-CBM4	11	1
U69	U69-CBM4	11	9
U69	U69-CBM35	7	7
U69	U69-CBM9-CBM9	6	1
U69	U69-CBM22	5	4
U69	U69-CBM67	5	2
U69	U69-CBM9	4	1
U69	U69-CBM22-CBM9-CBM9	4	1
U69	U69-CBM61	3	1
U69	U69-CBM2	3	2
U69	U69-CBM6	3	2
U69	U69-CBM51	3	1
U69	U69-CBM4-CBM4-CBM4	3	1
U69	U69-CBM26	2	2
U69	U69-CBM66	2	2
U69	U69-CBM26-CBM26-CBM26	2	2
U69	U69-CBM26-CBM56	2	1
U69	U69-CBM9-CBM22-CBM9	2	1

MEROPS family	Gene configuration	No of hits	No of distinct genes
U69	U69-CBM2-CBM2	1	1
U69	U69-CBM56-CBM26	1	1
U69	U69-CBM32-CBM32-CBM32	1	1
U69	U69-CBM13-CBM13	1	1
U69	U69-CBM59	1	1
U69	U69-CBM16	1	1
U69	U69-CBM56-CBM56	1	1
U69	U69-CBM13	1	1
U69	U69-CBM26-CBM26	1	1

Bibliography

TIOBE Index | TIOBE - The Software Quality Company.

Abraham, Clara, & Medzhitov, Ruslan. 2011. Interactions between the host innate immune system and microbes in inflammatory bowel disease. *Gastroenterology*, **140**(6), 1729–1737.

Alivisatos, A. P., Blaser, M. J., Brodie, E. L., Chun, M., Dangl, J. L., Donohue, T. J., Dorrestein, P. C., Gilbert, J. A., Green, J. L., Jansson, J. K., Knight, R., Maxon, M. E., McFall-Ngai, M. J., Miller, J. F., Pollard, K. S., Ruby, E. G., Taha, S. A., & Consortium, Unified Microbiome Initiative. 2015. A unified initiative to harness Earth's microbiomes. *Science*, **350**(6260), 507–508.

Allen, Adrian, & Pearson, Jeffrey P. 2000. The Gastrointestinal Adherent Mucous Gel Barrier. *Pages 57–64 of: Glycoprotein Methods and Protocols*. Methods in Molecular Biology™. Humana Press. DOI: 10.1385/1-59259-048-9:057.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Amann, R I, Ludwig, W, & Schleifer, K H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, **59**(1), 143–169.

American Diabetes Association. 2012. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, **36**(Supplement_1), S67–S74.

Aminov, Rustam I. 2010. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Frontiers in Microbiology*, **1**(Dec.).

Angelakis, Emmanouil, Million, Matthieu, Kankoe, Sallah, Lagier, Jean-Christophe, Armougom, Fabrice, Giorgi, Roch, & Raoult, Didier. 2014. Abnormal Weight Gain and Gut Microbiota Modifications Are Side Effects of Long-Term Doxycycline and Hydroxychloroquine Treatment. *Antimicrobial Agents and Chemotherapy*, **58**(6), 3342–3347.

Apache, Software Foundation. *Maven – Welcome to Apache Maven*.

Applications, National Research Council (US) Committee on Metagenomics: Challenges and Functional. 2007. *Data Management and Bioinformatics Challenges of Metagenomics*. National Academies Press (US).

Arslanoglu, Sertac, Moro, Guido E, & Boehm, Gunther. 2007. Early supplementation of prebiotic oligosaccharides protects formula-fed infants against infections during the first 6 months of life. *The Journal of Nutrition*, **137**(11), 2420–2424.

Arumugam, Manimozhiyan, Raes, Jeroen, Pelletier, Eric, Le Paslier, Denis, Yamada, Takuji, Mende, Daniel R, Fernandes, Gabriel R, Tap, Julien, Bruls, Thomas, Batto, Jean-Michel, Bertalan, Marcelo, Borruel, Natalia, Casellas, Francesc, Fernandez, Leyden, Gautier, Laurent, Hansen, Torben, Hattori, Masahira, Hayashi, Tetsuya, Kleerebezem, Michiel, Kurokawa, Ken, Leclerc, Marion, Levenez, Florence, Manichanh, Chaysavanh, Nielsen, H Bjørn, Nielsen, Trine, Pons, Nicolas, Poulain, Julie, Qin, Junjie, Sicheritz-Ponten, Thomas, Tims, Sebastian, Torrents, David, Ugarte, Edgardo, Zoetendal, Erwin G, Wang, Jun, Guarner, Francisco, Pedersen, Oluf, de Vos, Willem M, Brunak, Søren, Doré, Joel, Antolín, María, Artiguenave, François, Blottiere, Hervé M, Almeida, Mathieu, Brechot, Christian, Cara, Carlos, Chervaux, Christian, Cultrone, Antonella, Delorme, Christine, Denariaz, Gérard, Dervyn, Rozenn, Foerstner, Konrad U, Friss, Carsten, van de Guchte, Maarten, Guedon, Eric, Haimet, Florence, Huber, Wolfgang, van Hylckama-Vlieg, Johan, Jamet, Alexandre, Juste, Catherine, Kaci, Ghalia, Knol, Jan, Lakhdari, Omar, Layec, Severine, Le Roux, Karine, Maguin, Emmanuelle, Mérieux, Alexandre, Melo Minardi, Raquel, M'rini, Christine, Muller, Jean, Oozeer, Raish, Parkhill, Julian, Renault, Pierre, Rescigno, Maria, Sanchez, Nicolas, Sunagawa, Shinichi, Torrejon, Antonio, Turner, Keith, Vandemeulebrouck, Gaetana, Varela, Encarna, Winogradsky, Yohanan, Zeller, Georg, Weissenbach, Jean, Ehrlich, S Dusko, & Bork, Peer. 2011. Enterotypes of the human gut microbiome. *Nature*, **473**(7346), 174–180.

Bailey, Jennifer R, Probert, Christopher S J, & Cogan, Tristan A. 2011. Identification and characterisation of an iron-responsive candidate probiotic. *PloS One*, **6**(10), e26507.

Baker, G C, Smith, J J, & Cowan, D A. 2003. Review and re-analysis of domain-specific 16S primers. *Journal of microbiological methods*, **55**(3), 541–555.

Bell, Michael J., Collison, Matthew, & Lord, Phillip. 2013. Can Inferred Provenance and Its Visualisation Be Used to Detect Erroneous Annotation? A Case Study Using UniProtKB. *PLoS ONE*, **8**(10), e75541.

Bengtsson-Palme, Johan, Angelin, Martin, Huss, Mikael, Kjellqvist, Sanela, Kristiansson, Erik, Palmgren, Helena, Larsson, D. G. Joakim, & Johansson, Anders. 2015. The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. *Antimicrobial Agents and Chemotherapy*, **59**(10), 6551–6560.

Biagi, Elena, Nylund, Lotta, Candela, Marco, Ostan, Rita, Bucci, Laura, Pini, Elisa, Nikkïla, Janne, Monti, Daniela, Satokari, Reetta, Franceschi, Claudio, Brigidi, Patrizia, & De Vos, Willem. 2010. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PloS One*, **5**(5), e10667.

Bland, Jeffrey. 2016. Intestinal Microbiome, Akkermansia muciniphila, and Medical Nutrition Therapy. *Integrative Medicine: A Clinician's Journal*, **15**(5), 14–16.

Blaser, Martin. 2011. Antibiotic overuse: Stop the killing of beneficial bacteria. *Nature*, **476**(7361), 393–394.

Blaxter, Mark, Mann, Jenna, Chapman, Tom, Thomas, Fran, Whitton, Claire, Floyd, Robin, & Abebe, Eyualem. 2005. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1462), 1935–1943.

Boorom, Kenneth Fiske. 2007. Is this recently characterized gastrointestinal pathogen responsible for rising rates of inflammatory bowel disease (IBD) and IBD associated autism in Europe and the United States in the 1990s? *Medical Hypotheses*, **69**(3), 652–659.

Brown, James R, Magid-Slav, Michal, Sanseau, Philippe, & Rajpal, Deepak K. 2011. Computational biology approaches for selecting host-pathogen drug targets. *Drug Discovery Today*, **16**(5-6), 229–236.

Brown, Lisa, Wolf, Julie M., Prados-Rosales, Rafael, & Casadevall, Arturo. 2015. Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi. *Nature Reviews Microbiology*, **13**(10), 620–630.

Brännmark, Cecilia, Nyman, Elin, Fagerholm, Siri, Bergenholm, Linnéa, Ekstrand, Eva-Maria, Cedersund, Gunnar, & Strålfors, Peter. 2013. Insulin Signaling in Type 2 Diabetes EXPERIMENTAL AND MODELING ANALYSES REVEAL MECHANISMS OF INSULIN RESISTANCE IN HUMAN ADIPOCYTES. *Journal of Biological Chemistry*, **288**(14), 9867–9880.

Buist, Girbe, Steen, Anton, Kok, Jan, & Kuipers, Oscar P. 2008. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, **68**(4), 838–847.

Burcelin, Rémy, Serino, Matteo, Chabo, Chantal, Blasco-Baque, Vincent, & Amar, Jacques. 2011. Gut microbiota and diabetes: from pathogenesis to therapeutic perspective. *Acta Diabetologica*, Oct.

Cani, Patrice D, & Delzenne, Nathalie M. 2011. The gut microbiome as therapeutic target. *Pharmacology & Therapeutics*, **130**(2), 202–212.

Cani, Patrice D., Amar, Jacques, Iglesias, Miguel Angel, Poggi, Marjorie, Knauf, Claude, Bastelica, Delphine, Neyrinck, Audrey M., Fava, Francesca, Tuohy, Kieran M., Chabo, Chantal, Waget, Aurélie, Delmée, Evelyne, Cousin, Béatrice, Sulpice, Thierry, Chamontin, Bernard, Ferrières, Jean, Tanti, Jean-François, Gibson, Glenn R., Casteilla, Louis, Delzenne, Nathalie M., Alessi, Marie Christine, & Burcelin, Rémy. 2007. Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes*, **56**(7), 1761–1772.

Cani, Patrice D., Bibiloni, Rodrigo, Knauf, Claude, Waget, Aurélie, Neyrinck, Audrey M., Delzenne, Nathalie M., & Burcelin, Rémy. 2008. Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. *Diabetes*, **57**(6), 1470–1481.

Cani, Patrice D., Possemiers, Sam, Wiele, Tom Van de, Guiot, Yves, Everard, Amandine, Rottier, Olivier, Geurts, Lucie, Naslain, Damien, Neyrinck, Audrey M., Lambert, Didier M., Muccioli, Giulio G., & Delzenne, Nathalie M. 2009. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. *Gut*, Feb.

Cani, Patrice D., Osto, Melania, Geurts, Lucie, & Everard, Amandine. 2012. Involvement of gut microbiota in the development of low-grade inflammation and type 2 diabetes associated with obesity. *Gut Microbes*, **3**(4), 279–288.

Cantarel, Brandi L., Coutinho, Pedro M., Rancurel, Corinne, Bernard, Thomas, Lombard, Vincent, & Henrissat, Bernard. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, **37**(Database issue), D233–D238.

Caporaso, J. Gregory, Bittinger, Kyle, Bushman, Frederic D., DeSantis, Todd Z., Andersen, Gary L., & Knight, Rob. 2010a. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**(2), 266–267.

Caporaso, J Gregory, Kuczynski, Justin, Stombaugh, Jesse, Bittinger, Kyle, Bushman, Frederic D, Costello, Elizabeth K, Fierer, Noah, Peña, Antonio Gonzalez, Goodrich, Julia K, Gordon, Jeffrey I, Huttley, Gavin A, Kelley, Scott T, Knights, Dan, Koenig, Jeremy E, Ley, Ruth E, Lozupone, Catherine A, McDonald, Daniel, Muegge, Brian D, Pirrung, Meg, Reeder, Jens, Sevinsky, Joel R, Turnbaugh, Peter J, Walters, William A, Widmann, Jeremy, Yatsunenko, Tanya, Zaneveld, Jesse,

& Knight, Rob. 2010b. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**(5), 335–336.

Caporaso, J Gregory, Lauber, Christian L, Costello, Elizabeth K, Berg-Lyons, Donna, Gonzalez, Antonio, Stombaugh, Jesse, Knights, Dan, Gajer, Pawel, Ravel, Jacques, Fierer, Noah, Gordon, Jeffrey I, & Knight, Rob. 2011. Moving pictures of the human microbiome. *Genome Biology*, **12**(5), R50.

Carroll, Ian M., Ringel-Kulka, Tamar, Ferrier, Laurent, Wu, Michael C., Siddle, Jennica P., Bueno, Lionel, & Ringel, Yehuda. 2013. Fecal Protease Activity Is Associated with Compositional Alterations in the Intestinal Microbiota. *PLOS ONE*, **8**(10), e78017.

Chaudet, Marcia M., & Rose, David R. 2016. Suggested alternative starch utilization system from the human gut bacterium Bacteroides thetaiotaomicron. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire*, **94**(3), 241–246.

Cho, Yoon Shin, Chen, Chien-Hsiun, Hu, Cheng, Long, Jirong, Ong, Rick Twee Hee, Sim, Xueling, Takeuchi, Fumihiko, Wu, Ying, Go, Min Jin, Yamauchi, Toshimasa, Chang, Yi-Cheng, Kwak, Soo Heon, Ma, Ronald C W, Yamamoto, Ken, Adair, Linda S, Aung, Tin, Cai, Qiuyin, Chang, Li-Ching, Chen, Yuan-Tsong, Gao, Yutang, Hu, Frank B, Kim, Hyung-Lae, Kim, Sangsoo, Kim, Young Jin, Lee, Jeannette Jen-Mai, Lee, Nanette R, Li, Yun, Liu, Jian Jun, Lu, Wei, Nakamura, Jiro, Nakashima, Eitaro, Ng, Daniel Peng-Keat, Tay, Wan Ting, Tsai, Fuu-Jen, Wong, Tien Yin, Yokota, Mitsuhiro, Zheng, Wei, Zhang, Rong, Wang, Congrong, So, Wing Yee, Ohnaka, Keizo, Ikegami, Hiroshi, Hara, Kazuo, Cho, Young Min, Cho, Nam H, Chang, Tien-Jyun, Bao, Yuqian, Hedman, Åsa K, Morris, Andrew P, McCarthy, Mark I, DIAGRAM Consortium, MuTHER Consortium, Takayanagi, Ryoichi, Park, Kyong Soo, Jia, Weiping, Chuang, Lee-Ming, Chan, Juliana C N, Maeda, Shiro, Kadowaki, Takashi, Lee, Jong-Young, Wu, Jer-Yuarn, Teo, Yik Ying, Tai, E Shyong, Shu, Xiao Ou, Mohlke, Karen L, Kato, Norihiro, Han, Bok-Ghee, & Seielstad, Mark. 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature genetics*, **44**(1), 67–72.

Choices, N. H. S. 2012 (Jan.). *Diabetes, type 2 - Complications - NHS Choices*. If diabetes is not treated, it can lead to a number of different health problems. High glucose levels can damage blood vessels, nerves and organs.

Cockell, Simon J, Weile, Jochen, Lord, Phillip, Wipat, Claire, Andriychenko, Dmytro, Pocock, Matthew, Wilkinson, Darren, Young, Malcolm, & Wipat, Anil. 2010. An integrated dataset for in silico drug discovery. *Journal of Integrative Bioinformatics*, **7**(3).

Cohen-Boulakia, Sarah, Belhajjame, Khalid, Collin, Olivier, Chopard, Jérôme, Froidevaux, Christine, Gaignard, Alban, Hinsen, Konrad, Larmande, Pierre, Bras, Yvan Le, Lemoine, Frédéric, Mareuil, Fabien, Ménager, Hervé, Pradal, Christophe, & Blanchet, Christophe. 2017. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, **75**(Supplement C), 284–298.

Collison, Matthew, Hirt, Robert P, Wipat, Anil, Nakjang, Sirintra, Sanseau, Philippe, & Brown, James R. 2012. Data mining the human gut microbiota for therapeutic targets. *Briefings in Bioinformatics*, Mar.

Consortium, The Human Microbiome Project. 2012. A framework for human microbiome research. *Nature*, **486**(7402), 215–221.

Costello, Elizabeth K., Lauber, Christian L., Hamady, Micah, Fierer, Noah, Gordon, Jeffrey I., & Knight, Rob. 2009. Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science*, **326**(5960), 1694 –1697.

Costello, Elizabeth K, Gordon, Jeffrey I, Secor, Stephen M, & Knight, Rob. 2010. Postprandial remodeling of the gut microbiota in Burmese pythons. *The ISME Journal*, **4**(11), 1375–1385.

Crawford, Peter A, Crowley, Jan R, Sambandam, Nandakumar, Muegge, Brian D, Costello, Elizabeth K, Hamady, Micah, Knight, Rob, & Gordon, Jeffrey I. 2009. Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(27), 11276–11281.

Danielsson, Dan, Teigen, Per Kristen, & Moi, Harald. 2011. The genital econiche: focus on microbiota and bacterial vaginosis. *Annals of the New York Academy of Sciences*, **1230**(Aug.), 48–58.

Dao, Maria Carlota, Everard, Amandine, Aron-Wisnewsky, Judith, Sokolovska, Nataliya, Prifti, Edi, Verger, Eric O., Kayser, Brandon D., Levenez, Florence, Chilloux, Julien, Hoyles, Lesley, Consortium, MICRO-Obes, Dumas, Marc-Emmanuel, Rizkalla, Salwa W., Doré, Joel, Cani, Patrice D., & Clément, Karine. 2015. Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut*, May, gutjnl–2014–308778.

de Castro, Alinne P., Fernandes, Gabriel da R., & Franco, Octávio L. 2014. Insights into novel antimicrobial compounds and antibiotic resistance genes from soil metagenomes. *Frontiers in Microbiology*, **5**(Sept.).

De Filippo, Carlotta, Cavalieri, Duccio, Di Paola, Monica, Ramazzotti, Matteo, Poullet, Jean Baptiste, Massart, Sebastien, Collini, Silvia, Pieraccini, Giuseppe, & Lionetti, Paolo. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(33), 14691–14696.

den Besten, Gijs, van Eunen, Karen, Groen, Albert K., Venema, Koen, Reijngoud, Dirk-Jan, & Bakker, Barbara M. 2013. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research*, **54**(9), 2325–2340.

Dethlefsen, Les, Huse, Sue, Sogin, Mitchell L, & Relman, David A. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology*, **6**(11), e280.

Diabetes UK. 2014 (Jan.). *Diabetes complications - Diabetes UK*.

Dominguez-Bello, Maria G, Costello, Elizabeth K, Contreras, Monica, Magris, Magda, Hidalgo, Glida, Fierer, Noah, & Knight, Rob. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(26), 11971–11975.

Donaldson, Gregory P., Lee, S. Melanie, & Mazmanian, Sarkis K. 2016. Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, **14**(1), 20–32.

Drucker, Daniel J. 2003. Glucagon-like peptides: regulators of cell proliferation, differentiation, and apoptosis. *Molecular endocrinology (Baltimore, Md.)*, **17**(2), 161–171.

Dubilier, Nicole, McFall-Ngai, Margaret, & Zhao, Liping. 2015. Microbiology: Create a global microbiome effort. *Nature News*, **526**(7575), 631.

Dumas, Marc-Emmanuel, Barton, Richard H, Toye, Ayo, Cloarec, Olivier, Blancher, Christine, Rothwell, Alice, Fearnside, Jane, Tatoud, Roger, Blanc, Véronique, Lindon, John C, Mitchell, Steve C, Holmes, Elaine, McCarthy, Mark I, Scott, James, Gauguier, Dominique, & Nicholson, Jeremy K.

2006. Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(33), 12511–12516.

Eckburg, Paul B, Bik, Elisabeth M, Bernstein, Charles N, Purdom, Elizabeth, Dethlefsen, Les, Sargent, Michael, Gill, Steven R, Nelson, Karen E, & Relman, David A. 2005. Diversity of the human intestinal microbial flora. *Science (New York, N.Y.)*, **308**(5728), 1635–1638.

Eddy, Sean R. 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, **7**(10), e1002195.

Edgar, Robert C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), 2460–2461.

Ehrlich, S Dusko, & Consortium, The Metahit. 2011. MetaHIT : The European Union Project on Metagenomics of the Human Intestinal Tract. *October*, 307–316.

Elliott, David E, & Weinstock, Joel V. 2009. Helminthic therapy: using worms to treat immune-mediated disease. *Advances in Experimental Medicine and Biology*, **666**, 157–166.

Ellrott, Kyle, Jaroszewski, Lukasz, Li, Weizhong, Wooley, John C, & Godzik, Adam. 2010. Expansion of the protein repertoire in newly explored environments: human gut microbiome specific protein families. *PLoS Computational Biology*, **6**(6), e1000798.

Everard, Amandine, Belzer, Clara, Geurts, Lucie, Ouwerkerk, Janneke P., Druart, Céline, Bindels, Laure B., Guiot, Yves, Derrien, Muriel, Muccioli, Giulio G., Delzenne, Nathalie M., Vos, Willem M. de, & Cani, Patrice D. 2013. Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proceedings of the National Academy of Sciences*, May.

Falkow, S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. *Reviews of Infectious Diseases*, **10 Suppl 2**(Aug.), S274–276.

Federsel, Hans-Jürgen. 2010. How to conduct research in the pharmaceutical industry? Facing the dilemma: small, autonomous teams versus large, integrated centers. *Expert Opinion on Drug Discovery*, **5**(9), 813–818.

Feria-Gervasio, David, Denis, Sylvain, Alric, Monique, & Brugère, Jean-François. 2011. In vitro maintenance of a human proximal colon microbiota using the continuous fermentation system P-ECSIM. *Applied Microbiology and Biotechnology*, **91**(5), 1425–1433.

Fernández-Suárez, Xosé M, Rigden, Daniel J, & Galperin, Michael Y. 2013. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic acids research*, Dec.

Fierer, Noah, Lauber, Christian L, Zhou, Nick, McDonald, Daniel, Costello, Elizabeth K, & Knight, Rob. 2010. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(14), 6477–6481.

Frank, Daniel N, Robertson, Charles E, Hamm, Christina M, Kpadeh, Zegbeh, Zhang, Tianyi, Chen, Hongyan, Zhu, Wei, Sartor, R Balfour, Boedeker, Edgar C, Harpaz, Noam, Pace, Norman R, & Li, Ellen. 2011. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflammatory Bowel Diseases*, **17**(1), 179–184.

Frese, Steven A, Benson, Andrew K, Tannock, Gerald W, Loach, Diane M, Kim, Jaehyoung, Zhang, Min, Oh, Phaik Lyn, Heng, Nicholas C K, Patil, Prabhu B, Juge, Nathalie, Mackenzie, Donald A, Pearson, Bruce M, Lapidus, Alla, Dalin, Eileen, Tice, Hope, Goltsman, Eugene, Land, Miriam, Hauser, Loren, Ivanova, Natalia, Kyrpides, Nikos C, & Walter, Jens. 2011. The evolution of host specialization in the vertebrate gut symbiont Lactobacillus reuteri. *PLoS Genetics*, **7**(2), e1001314.

Gao, Hui, & Dahlman-Wright, Karin. 2011. From DNA binding to metabolic control: integration of -omics data reveals drug targets for prostate cancer. *The EMBO Journal*, **30**(13), 2516–2517.

Garrett, Wendy S, Gallini, Carey A, Yatsunenko, Tanya, Michaud, Monia, DuBois, Andrea, Delaney, Mary L, Punit, Shivesh, Karlsson, Maria, Bry, Lynn, Glickman, Jonathan N, Gordon, Jeffrey I, Onderdonk, Andrew B, & Glimcher, Laurie H. 2010. Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host & Microbe*, **8**(3), 292–300.

Garrido, Daniel, Kim, Jae Han, German, J. Bruce, Raybould, Helen E., & Mills, David A. 2011. Oligosaccharide binding proteins from Bifidobacterium longum subsp. infantis reveal a preference for host glycans. *PloS One*, **6**(3), e17315.

Gevers, Dirk, Knight, Rob, Petrosino, Joseph F., Huang, Katherine, McGuire, Amy L., Birren, Bruce W., Nelson, Karen E., White, Owen, Methé, Barbara A., & Huttenhower, Curtis. 2012. The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. *PLoS Biol*, **10**(8), e1001377.

Godzik, Adam. 2011. Metagenomics and the protein universe. *Current Opinion in Structural Biology*, **21**(3), 398–403.

Government of Canada, Canadian Institutes of Health Research. 2009 (Aug.). *Canadian Microbiome Initiative - CIHR*.

Grice, Elizabeth A, & Segre, Julia A. 2011. The skin microbiome. *Nature Reviews. Microbiology*, **9**(4), 244–253.

Hamady, Micah, & Knight, Rob. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, **19**(7), 1141–1152.

Hanisch, F. G. 2001. O-glycosylation of the mucin type. *Biological Chemistry*, **382**(2), 143–149.

Hartman, Amber L, Lough, Denver M, Barupal, Dinesh K, Fiehn, Oliver, Fishbein, Thomas, Zasloff, Michael, & Eisen, Jonathan A. 2009. Human gut microbiome adopts an alternative state following small bowel transplantation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(40), 17187–17192.

Hasnain, Sumaira Z., McGuckin, Michael A., Grencis, Richard K., & Thornton, David J. 2012. Serine protease(s) secreted by the nematode Trichuris muris degrade the mucus barrier. *PLoS neglected tropical diseases*, **6**(10), e1856.

Have, Christian Theil, & Jensen, Lars Juhl. 2013. Are graph databases ready for bioinformatics? *Bioinformatics*, Oct., btt549.

Hayes, M. V., & Orr, D. C. 1983. Mode of action of ceftazidime: affinity for the penicillin-binding proteins of Escherichia coli K12, Pseudomonas aeruginosa and Staphylococcus aureus. *The Journal of Antimicrobial Chemotherapy*, **12**(2), 119–126.

Hehemann, Jan-Hendrik, Correc, Gaëlle, Barbeyron, Tristan, Helbert, William, Czjzek, Mirjam, & Michel, Gurvan. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, **464**(7290), 908–912.

Hickson, Mary, D'Souza, Aloysius L, Muthu, Nirmala, Rogers, Thomas R, Want, Susan, Rajkumar, Chakravarthi, & Bulpitt, Christopher J. 2007. Use of probiotic Lactobacillus preparation to prevent diarrhoea associated with antibiotics: randomised double blind placebo controlled trial. *BMJ*, **335**(7610), 80.

Hildebrand, Andrea, Remmert, Michael, Biegert, Andreas, & Söding, Johannes. 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77 Suppl 9**, 128–132.

Hirt, Robert P., de Miguel, Natalia, Nakjang, Sirintra, Dessi, Daniele, Liu, Yuk-Chien, Diaz, Nicia, Rappelli, Paola, Acosta-Serrano, Alvaro, Fiori, Pier-Luigi, & Mottram, Jeremy C. 2011. Trichomonas vaginalis pathobiology new insights from the genome sequence. *Advances in Parasitology*, **77**, 87–140.

Holmes, Elaine, Wilson, Ian D, & Nicholson, Jeremy K. 2008. Metabolic phenotyping in health and disease. *Cell*, **134**(5), 714–717.

Hooper, Lora V, & Macpherson, Andrew J. 2010. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature Reviews. Immunology*, **10**(3), 159–169.

Hopkins, Andrew L, & Groom, Colin R. 2002. The druggable genome. *Nature Reviews. Drug Discovery*, **1**(9), 727–730.

Hu, Frank B. 2011. Globalization of Diabetes The role of diet, lifestyle, and genes. *Diabetes Care*, **34**(6), 1249–1257.

Hugenholtz, P, Goebel, B M, & Pace, N R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, **180**(18), 4765–4774.

Huson, Daniel H, Auch, Alexander F, Qi, Ji, & Schuster, Stephan C. 2007. MEGAN analysis of metagenomic data. *Genome Research*, **17**(3), 377–386.

Hviid, Anders, Svanström, Henrik, & Frisch, Morten. 2011. Antibiotic use and inflammatory bowel diseases in childhood. *Gut*, **60**(1), 49–54.

Hyatt, Doug, LoCascio, Philip F., Hauser, Loren J., & Uberbacher, Edward C. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28**(17), 2223–2230.

International Diabetes Federation. 2013. *IDF Diabetes Atlas*. 6th edition edn. Brussels, Belgium.: International Diabetes Federation.

Jansson, Janet, Knight, Rob, & Gilbert, Jack. *Earth Microbiome Project*.

Kaddurah-Daouk, Rima, Baillie, Rebecca A, Zhu, Hongjie, Zeng, Zhao-Bang, Wiest, Michelle M, Nguyen, Uyen Thao, Wojnoonski, Katie, Watkins, Steven M, Trupp, Miles, & Krauss, Ronald M. 2011. Enteric microbiome metabolites correlate with response to simvastatin treatment. *PloS One*, **6**(10), e25482.

Kalev, Ivan, & Habeck, Michael. 2011. HHfrag: HMM-based fragment detection using HHpred. *Bioinformatics (Oxford, England)*, **27**(22), 3110–3116.

Kanehisa, M, & Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kaoutari, Abdessamad El, Armougom, Fabrice, Gordon, Jeffrey I., Raoult, Didier, & Henrissat, Bernard. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, **11**(7), 497–504.

Karlsson, Fredrik H., Tremaroli, Valentina, Nookaew, Intawat, Bergström, Göran, Behre, Carl Johan, Fagerberg, Björn, Nielsen, Jens, & Bäckhed, Fredrik. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, **498**(7452), 99–103.

Kau, Andrew L, Ahern, Philip P, Griffin, Nicholas W, Goodman, Andrew L, & Gordon, Jeffrey I. 2011. Human nutrition, the gut microbiome and the immune system. *Nature*, **474**(7351), 327–336.

Kellermayer, Richard, Dowd, Scot E, Harris, R Alan, Balasa, Alfred, Schaible, Tiffany D, Wolcott, Randy D, Tatevian, Nina, Szigeti, Reka, Li, Zhijie, Versalovic, James, & Smith, C Wayne. 2011. Colonic mucosal DNA methylation, immune response, and microbiome patterns in Toll-like receptor 2-knockout mice. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, **25**(5), 1449–1460.

Kembel, Steven W, Eisen, Jonathan A, Pollard, Katherine S, & Green, Jessica L. 2011. The phylogenetic diversity of metagenomes. *PloS One*, **6**(8), e23214.

Kiley, James P. 2011. Advancing respiratory research. *Chest*, **140**(2), 497–501.

Klindworth, Anna, Pruesse, Elmar, Schweer, Timmy, Peplies, Jörg, Quast, Christian, Horn, Matthias, & Glöckner, Frank Oliver. 2012. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, Aug., gks808.

Knight, Paul, Campbell, Barry J, & Rhodes, Jonathan M. 2008. Host-bacteria interaction in inflammatory bowel disease. *British Medical Bulletin*, **88**(1), 95–113.

Kootte, R S, Vrieze, A, Holleman, F, Dallinga-Thie, G M, Zoetendal, E G, de Vos, W M, Groen, A K, Hoekstra, J B L, Stroes, E S, & Nieuwdorp, M. 2011. The therapeutic potential of manipulating gut microbiota in obesity and type 2 diabetes mellitus. *Diabetes, Obesity & Metabolism*, Aug.

Koropatkin, Nicole M., Cameron, Elizabeth A., & Martens, Eric C. 2012. How glycan metabolism shapes the human gut microbiota. *Nature reviews. Microbiology*, **10**(5), 323–335.

Kurokawa, Ken, Itoh, Takehiko, Kuwahara, Tomomi, Oshima, Kenshiro, Toh, Hidehiro, Toyoda, Atsushi, Takami, Hideto, Morita, Hidetoshi, Sharma, Vineet K, Srivastava, Tulika P, Taylor, Todd D, Noguchi, Hideki, Mori, Hiroshi, Ogura, Yoshitoshi, Ehrlich, Dusko S, Itoh, Kikuji, Takagi, Toshihisa, Sakaki, Yoshiyuki, Hayashi, Tetsuya, & Hattori, Masahira. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, **14**(4), 169–181.

Köhler, Jacob, Baumbach, Jan, Taubert, Jan, Specht, Michael, Skusa, Andre, Rüegg, Alexander, Rawlings, Chris, Verrier, Paul, & Philippi, Stephan. 2006. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**(11), 1383–1390.

Lander, E S, Linton, L M, Birren, B, Nusbaum, C, Zody, M C, Baldwin, J, Devon, K, Dewar, K, Doyle, M, FitzHugh, W, Funke, R, Gage, D, Harris, K, Heaford, A, Howland, J, Kann, L, Lehoczky, J, LeVine, R, McEwan, P, McKernan, K, Meldrim, J, Mesirov, J P, Miranda, C, Morris, W, Naylor, J, Raymond, C, Rosetti, M, Santos, R, Sheridan, A, Sougnez, C, Stange-Thomann, N, Stojanovic, N, Subramanian, A, Wyman, D, Rogers, J, Sulston, J, Ainscough, R, Beck, S, Bentley, D, Burton, J, Clee, C, Carter, N, Coulson, A, Deadman, R, Deloukas, P, Dunham, A, Dunham, I, Durbin, R, French, L, Grafham, D, Gregory, S, Hubbard, T, Humphray, S, Hunt, A, Jones, M, Lloyd, C, McMurray, A, Matthews, L, Mercer, S, Milne, S, Mullikin, J C, Mungall, A, Plumb, R, Ross, M, Shownkeen, R, Sims, S, Waterston, R H, Wilson, R K, Hillier, L W, McPherson, J D, Marra, M A, Mardis, E R, Fulton, L A, Chinwalla, A T, Pepin, K H, Gish, W R, Chissoe, S L, Wendl, M C, Delehaunty, K D, Miner, T L, Delehaunty, A, Kramer, J B, Cook, L L, Fulton, R S, Johnson, D L, Minx, P J, Clifton, S W, Hawkins, T, Branscomb, E, Predki, P, Richardson, P, Wenning, S, Slezak, T, Doggett, N, Cheng, J F, Olsen, A, Lucas, S, Elkin, C, Uberbacher, E, Frazier, M, Gibbs, R A, Muzny, D M, Scherer, S E, Bouck, J B, Sodergren, E J, Worley, K C, Rives, C M, Gorrell, J H, Metzker, M L, Naylor, S L, Kucherlapati, R S, Nelson, D L, Weinstock, G M, Sakaki, Y, Fujiyama, A, Hattori, M, Yada, T, Toyoda, A, Itoh, T, Kawagoe, C, Watanabe, H, Totoki, Y, Taylor, T, Weissenbach, J, Heilig, R, Saurin, W, Artiguenave, F, Brottier, P, Bruls, T, Pelletier, E, Robert, C, Wincker, P, Smith, D R, Doucette-Stamm, L, Rubenfield, M, Weinstock, K, Lee, H M, Dubois, J, Rosenthal, A, Platzer, M, Nyakatura, G, Taudien, S, Rump, A, Yang, H, Yu, J, Wang, J, Huang, G, Gu, J, Hood, L, Rowen, L, Madan, A, Qin, S, Davis, R W, Federspiel, N A, Abola, A P, Proctor, M J, Myers, R M, Schmutz, J, Dickson, M, Grimwood, J, Cox, D R, Olson, M V, Kaul, R, Raymond, C, Shimizu, N, Kawasaki, K, Minoshima, S, Evans, G A, Athanasiou, M, Schultz, R, Roe, B A, Chen, F, Pan, H, Ramser, J, Lehrach, H, Reinhardt, R, McCombie, W R, de la Bastide, M, Dedhia, N, Blöcker, H, Hornischer, K, Nordsiek, G, Agarwala, R, Aravind, L, Bailey, J A, Bateman, A, Batzoglou, S, Birney, E, Bork, P, Brown, D G, Burge, C B, Cerutti, L, Chen, H C, Church, D, Clamp, M, Copley, R R, Doerks, T, Eddy, S R, Eichler, E E, Furey, T S, Galagan, J, Gilbert, J G, Harmon, C, Hayashizaki, Y, Haussler, D, Hermjakob, H, Hokamp, K, Jang, W, Johnson, L S, Jones, T A, Kasif, S, Kaspryzk, A, Kennedy, S, Kent, W J, Kitts, P, Koonin, E V, Korf, I, Kulp, D, Lancet, D, Lowe, T M, McLysaght, A, Mikkelsen, T, Moran, J V, Mulder, N, Pollara, V J, Ponting, C P, Schuler, G, Schultz, J, Slater, G, Smit, A F, Stupka, E, Szustakowski, J, Thierry-Mieg, D, Thierry-Mieg, J, Wagner, L, Wallis, J, Wheeler, R, Williams, A, Wolf, Y I, Wolfe, K H, Yang, S P, Yeh, R F, Collins, F, Guyer, M S, Peterson, J, Felsenfeld, A, Wetterstrand, K A, Patrinos, A, Morgan, M J, de Jong, P, Catanese, J J, Osoegawa, K, Shizuya, H, Choi, S, Chen, Y J, Szustakowki, J, & International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lander, Eric S. 2011. Initial impact of the sequencing of the human genome. *Nature*, **470**(7333), 187–197.

Landy, J, Al-Hassi, H O, McLaughlin, S D, Walker, A W, Ciclitira, P J, Nicholls, R J, Clark, S K, & Hart, A L. 2011. Review article: faecal transplantation therapy for gastrointestinal disease. *Alimentary Pharmacology & Therapeutics*, **34**(4), 409–415.

Langdon, Amy, Crook, Nathan, & Dantas, Gautam. 2016. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Medicine*, **8**(Apr.).

Lapatas, Vasileios, Stefanidakis, Michalis, Jimenez, Rafael C., Via, Allegra, & Schneider, Maria Victoria. 2015. Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, **22**(Sept.), 9.

Larsen, Nadja, Vogensen, Finn K., van den Berg, Frans W. J., Nielsen, Dennis Sandris, Andreasen, Anne Sofie, Pedersen, Bente K., Al-Soud, Waleed Abu, Sørensen, Søren J., Hansen, Lars H., & Jakobsen, Mogens. 2010. Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. *PLoS ONE*, **5**(2), e9085.

Larsen, Peter E, Collart, Frank R, Field, Dawn, Meyer, Folker, Keegan, Kevin P, Henry, Christopher S, McGrath, John, Quinn, John, & Gilbert, Jack A. 2011. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microbial Informatics and Experimentation*, **1**, 4.

Lazar, Mitchell A. 2005. How obesity causes diabetes: not a tall tale. *Science (New York, N.Y.)*, **307**(5708), 373–375.

Le Chatelier, Emmanuelle, Nielsen, Trine, Qin, Junjie, Prifti, Edi, Hildebrand, Falk, Falony, Gwen, Almeida, Mathieu, Arumugam, Manimozhiyan, Batto, Jean-Michel, Kennedy, Sean, Leonard, Pierre, Li, Junhua, Burgdorf, Kristoffer, Grarup, Niels, Jørgensen, Torben, Brandslund, Ivan, Nielsen, Henrik Bjørn, Juncker, Agnieszka S., Bertalan, Marcelo, Levenez, Florence, Pons, Nicolas, Rasmussen, Simon, Sunagawa, Shinichi, Tap, Julien, Tims, Sebastian, Zoetendal, Erwin G., Brunak, Søren, Clément, Karine, Doré, Joël, Kleerebezem, Michiel, Kristiansen, Karsten, Renault, Pierre, Sicheritz-Ponten, Thomas, de Vos, Willem M., Zucker, Jean-Daniel, Raes, Jeroen, Hansen, Torben, MetaHIT Consortium, Bork, Peer, Wang, Jun, Ehrlich, S. Dusko, Pedersen, Oluf, & MetaHIT consortium additional Members. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature*, **500**(7464), 541–546.

Lee, Insuk, Date, Shailesh V, Adai, Alex T, & Marcotte, Edward M. 2004. A probabilistic functional network of yeast genes. *Science (New York, N.Y.)*, **306**(5701), 1555–1558.

Lee, Ju-Hoon, Karamychev, V N, Kozyavkin, S A, Mills, D, Pavlov, A R, Pavlova, N V, Polouchine, N N, Richardson, P M, Shakhova, V V, Slesarev, A I, Weimer, B, & O'Sullivan, D J. 2008. Comparative genomic analysis of the gut bacterium Bifidobacterium longum reveals loci susceptible to deletion during pure culture growth. *BMC Genomics*, **9**, 247.

Ley, Ruth E, Bäckhed, Fredrik, Turnbaugh, Peter, Lozupone, Catherine A, Knight, Robin D, & Gordon, Jeffrey I. 2005. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(31), 11070–11075.

Ley, Ruth E., Peterson, Daniel A., & Gordon, Jeffrey I. 2006a. Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell*, **124**(4), 837–848.

Ley, Ruth E, Turnbaugh, Peter J, Klein, Samuel, & Gordon, Jeffrey I. 2006b. Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**(7122), 1022–1023.

Li, Ruiqiang, Zhu, Hongmei, Ruan, Jue, Qian, Wubin, Fang, Xiaodong, Shi, Zhongbin, Li, Yingrui, Li, Shengting, Shan, Gao, Kristiansen, Karsten, Li, Songgang, Yang, Huanming, Wang, Jian, & Wang, Jun. 2009. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, Dec.

Lingner, Thomas, Asshauer, Kathrin Petra, Schreiber, Fabian, & Meinicke, Peter. 2011. CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research*, **39**(Web Server issue), W518–523.

Liu, Bo, & Pop, Mihai. 2011. MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. *BMC Proceedings*, **5 Suppl 2**, S9.

Lombard, Vincent, Ramulu, Hemalatha Golaconda, Drula, Elodie, Coutinho, Pedro M., & Henrissat, Bernard. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, **42**(D1), D490–D495.

Lord, Phillip. 2013. The Semantic Web takes Wing: Programming Ontologies with Tawny-OWL. *arXiv:1303.0213 [cs]*, Mar. arXiv: 1303.0213.

Lozupone, Catherine, Hamady, Micah, & Knight, Rob. 2006. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.

Lozupone, Catherine, Lladser, Manuel E, Knights, Dan, Stombaugh, Jesse, & Knight, Rob. 2011. UniFrac: an effective distance metric for microbial community comparison. *The ISME journal*, **5**(2), 169–172.

Lyte, Mark. 2011. Probiotics function mechanistically as delivery vehicles for neuroactive compounds: Microbial endocrinology in the design and use of probiotics. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **33**(8), 574–581.

Macfarlane, G. T., Steed, H., & Macfarlane, S. 2008. Bacterial metabolism and health-related effects of galacto-oligosaccharides and other prebiotics. *Journal of Applied Microbiology*, **104**(2), 305–344.

Marchesi, Julian R, Dutilh, Bas E, Hall, Neil, Peters, Wilbert H M, Roelofs, Rian, Boleij, Annemarie, & Tjalsma, Harold. 2011. Towards the human colorectal cancer microbiome. *PloS One*, **6**(5), e20447.

Markowitz, Victor M., Chen, I-Min A., Palaniappan, Krishna, Chu, Ken, Szeto, Ernest, Grechkin, Yuri, Ratner, Anna, Jacob, Biju, Huang, Jinghua, Williams, Peter, Huntemann, Marcel, Anderson, Iain, Mavromatis, Konstantinos, Ivanova, Natalia N., & Kyrpides, Nikos C. 2012. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, **40**(Database issue), D115–D122.

Martens, Eric C, Chiang, Herbert C, & Gordon, Jeffrey I. 2008. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host & Microbe*, **4**(5), 447–457.

Mathieu Bastian, Sebastien Heymann, & Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks.

McLoughlin, Rachel M, & Mills, Kingston H G. 2011. Influence of gastrointestinal commensal bacteria on the immune responses that mediate allergy and asthma. *The Journal of Allergy and Clinical Immunology*, **127**(5), 1097–1107; quiz 1108–1109.

McNeil, N. I. 1984. The contribution of the large intestine to energy supplies in man. *The American Journal of Clinical Nutrition*, **39**(2), 338–342.

Medici, F, Hawa, M, Ianari, A, Pyke, D A, & Leslie, R D. 1999. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis. *Diabetologia*, **42**(2), 146–150.

Medini, Duccio, Donati, Claudio, Tettelin, Hervé, Masignani, Vega, & Rappuoli, Rino. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development*, **15**(6), 589–594.

Mesnage, Stéphane, Dellarole, Mariano, Baxter, Nicola J., Rouget, Jean-Baptiste, Dimitrov, Jordan D., Wang, Ning, Fujimoto, Yukari, Hounslow, Andrea M., Lacroix-Desmazes, Sébastien, Fukase, Koichi, Foster, Simon J., & Williamson, Michael P. 2014. Molecular basis for bacterial peptidoglycan recognition by LysM domains. *Nature Communications*, **5**(June).

Meyer, F, Paarmann, D, D'Souza, M, Olson, R, Glass, E M, Kubal, M, Paczian, T, Rodriguez, A, Stevens, R, Wilke, A, Wilkening, J, & Edwards, R A. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Million, M, Lagier, J-C, Yahav, D, & Paul, M. 2013. Gut bacterial microbiota and obesity. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, **19**(4), 305–313.

molecular, Stratec. 2012. *PSP® Spin Stool DNA Plus Kit*.

Moore, Jason H, Lari, Richard Cowper Sal, Hill, Douglas, Hibberd, Patricia L, & Madan, Juliette C. 2011. Human microbiome visualization using 3d technology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 154–164.

Morris, Andrew P, Voight, Benjamin F, Teslovich, Tanya M, Ferreira, Teresa, Segre, Ayellet V, Steinthorsdottir, Valgerdur, Strawbridge, Rona J, Khan, Hassan, Grallert, Harald, Mahajan, Anubha, Prokopenko, Inga, Kang, Hyun Min, Dina, Christian, Esko, Tonu, Fraser, Ross M, Kanoni, Stavroula, Kumar, Ashish, Lagou, Vasiliki, Langenberg, Claudia, Luan, Jian'an, Lindgren, Cecilia M, Muller-Nurasyid, Martina, Pechlivanis, Sonali, Rayner, N William, Scott, Laura J, Wiltshire, Steven, Yengo, Loic, Kinnunen, Leena, Rossin, Elizabeth J, Raychaudhuri, Soumya, Johnson, Andrew D, Dimas, Antigone S, Loos, Ruth J F, Vedantam, Sailaja, Chen, Han, Florez, Jose C, Fox, Caroline, Liu, Ching-Ti, Rybin, Denis, Couper, David J, Kao, Wen Hong L, Li, Man, Cornelis, Marilyn C, Kraft, Peter, Sun, Qi, van Dam, Rob M, Stringham, Heather M, Chines, Peter S, Fischer, Krista, Fontanillas, Pierre, Holmen, Oddgeir L, Hunt, Sarah E, Jackson, Anne U, Kong, Augustine, Lawrence, Robert, Meyer, Julia, Perry, John RB, Platou, Carl GP, Potter, Simon, Rehnberg, Emil, Robertson, Neil, Sivapalaratnam, Suthesh, Stancakova, Alena, Stirrups, Kathleen, Thorleifsson, Gudmar, Tikkanen, Emmi, Wood, Andrew R, Almgren, Peter, Atalay, Mustafa, Benediktsson, Rafn, Bonnycastle, Lori L, Burtt, Noel, Carey, Jason, Charpentier, Guillaume, Crenshaw, Andrew T, Doney, Alex S F, Dorkhan, Mozhgan, Edkins, Sarah, Emilsson, Valur, Eury, Elodie, Forsen, Tom, Gertow, Karl, Gigante, Bruna, Grant, George B, Groves, Christopher J, Guiducci, Candace, Herder, Christian, Hreidarsson, Astradur B, Hui, Jennie, James, Alan, Jonsson, Anna, Rathmann, Wolfgang, Klopp, Norman, Kravic, Jasmina, Krjutskov, Kaarel, Langford, Cordelia, Leander, Karin, Lindholm, Eero, Lobbens, Stephane, Mannisto, Satu, Mirza, Ghazala, Muhleisen, Thomas W, Musk, Bill, Parkin, Melissa, Rallidis, Loukianos, Saramies, Jouko, Sennblad, Bengt, Shah, Sonia, Sigur?sson, Gunnar, Silveira, Angela, Steinbach, Gerald, Thorand, Barbara, Trakalo, Joseph, Veglia, Fabrizio, Wennauer, Roman, Winckler, Wendy, Zabaneh, Delilah, Campbell, Harry, van Duijn, Cornelia, Uitterlinden89, Andre G, Hofman, Albert, Sijbrands, Eric, Abecasis, Goncalo R, Owen, Katharine R, Zeggini, Eleftheria, Trip, Mieke D, Forouhi, Nita G, Syvanen, Ann-Christine, Eriksson, Johan G, Peltonen, Leena, Nothen, Markus M, Balkau, Beverley, Palmer, Colin N A, Lyssenko, Valeriya, Tuomi, Tiinamaija, Isomaa, Bo, Hunter, David J, Qi, Lu, Shuldiner, Alan R, Roden, Michael, Barroso, Ines, Wilsgaard, Tom, Beilby, John, Hovingh, Kees, Price, Jackie F, Wilson, James F, Rauramaa, Rainer, Lakka, Timo A, Lind, Lars, Dedoussis, George, Nj?lstad, Inger, Pedersen, Nancy L, Khaw, Kay-Tee, Wareham, Nicholas J, Keinanen-Kiukaanniemi, Sirkka M, Saaristo, Timo E, Korpi-Hyovalti, Eeva, Saltevo, Juha, Laakso, Markku, Kuusisto, Johanna, Metspalu, Andres, Collins, Francis S, Mohlke, Karen L, Bergman, Richard N, Tuomilehto, Jaakko, Boehm, Bernhard O, Gieger, Christian, Hveem, Kristian, Cauchi, Stephane, Froguel, Philippe, Baldassarre, Damiano, Tremoli, Elena, Humphries, Steve E, Saleheen, Danish, Danesh, John, Ingelsson, Erik, Ripatti, Samuli, Salomaa, Veikko, Erbel, Raimund, Jockel, Karl-Heinz, Moebus, Susanne, Peters, Annette, Illig, Thomas, de Faire, Ulf, Hamsten, Anders, Morris, Andrew D, Donnelly, Peter J, Frayling, Timothy M, Hattersley, Andrew T, Boerwinkle, Eric, Melander, Olle, Kathiresan, Sekar, Nilsson, Peter M, Deloukas, Panos, Thorsteinsdottir, Unnur, Groop, Leif C, Stefansson, Kari, Hu, Frank, Pankow, James S, Dupuis, Josee, Meigs, James B, Altshuler, David, Boehnke, Michael, & McCarthy, Mark I. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, **44**(9), 981–990.

Muccioli, Giulio G., Naslain, Damien, Bäckhed, Fredrik, Reigstad, Christopher S., Lambert, Didier M., Delzenne, Nathalie M., & Cani, Patrice D. 2010. The endocannabinoid system links gut microbiota to adipogenesis. *Molecular Systems Biology*, **6**(July), 392.

Mueller, Noel T., Bakacs, Elizabeth, Combellick, Joan, Grigoryan, Zoya, & Dominguez-Bello, Maria G. 2015. The infant microbiome development: mom matters. *Trends in molecular medicine*, **21**(2), 109–117.

Muyzer, G, de Waal, E C, & Uitterlinden, A G. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and environmental microbiology*, **59**(3), 695–700.

Nakjang, Sirintra. 2011. *Newcastle University eTheses: Comparative genomics for studying the proteomes of mucosal microorganisms*. Ph.D. thesis.

Nakjang, Sirintra, Ndeh, Didier A., Wipat, Anil, Bolam, David N., & Hirt, Robert P. 2012. A Novel Extracellular Metallopeptidase Domain Shared by Animal Host-Associated Mutualistic and Pathogenic Microbes. *PLoS ONE*, **7**(1).

Nation Institute of Health, NIH. *Human Microbiome Project - Program Initiatives*.

Navarro-Garcia, Fernando, Gutierrez-Jimenez, Javier, Garcia-Tovar, Carlos, Castro, Luis A., Salazar-Gonzalez, Hector, & Cordova, Vanessa. 2010. Pic, an Autotransporter Protein Secreted by Different Pathogens in the Enterobacteriaceae Family, Is a Potent Mucus Secretagogue. *Infection and Immunity*, **78**(10), 4101–4109.

Nayfach, Stephen, & Pollard, Katherine S. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology*, **16**(1).

Ndeh, Didier A. 2013. *Newcastle University eTheses: Novel glycan-targeted extracellular proteases from divergent mucosal microbes*. Ph.D. thesis.

Nelson, Karen E, Weinstock, George M, Highlander, Sarah K, Worley, Kim C, Creasy, Heather Huot, Wortman, Jennifer Russo, Rusch, Douglas B, Mitreva, Makedonka, Sodergren, Erica, Chinwalla, Asif T, Feldgarden, Michael, Gevers, Dirk, Haas, Brian J, Madupu, Ramana, Ward, Doyle V, Birren, Bruce W, Gibbs, Richard A, Methe, Barbara, Petrosino, Joseph F, Strausberg, Robert L, Sutton, Granger G, White, Owen R, Wilson, Richard K, Durkin, Scott, Giglio, Michelle Gwinn, Gujja, Sharvari, Howarth, Clint, Kodira, Chinnappa D, Kyrpides, Nikos, Mehta, Teena, Muzny, Donna M, Pearson, Matthew, Pepin, Kymberlie, Pati, Amrita, Qin, Xiang, Yandava, Chandri, Zeng, Qiandong, Zhang, Lan, Berlin, Aaron M, Chen, Lei, Hepburn, Theresa A, Johnson, Justin, McCorrison, Jamison, Miller, Jason, Minx, Pat, Nusbaum, Chad, Russ, Carsten, Sykes, Sean M, Tomlinson, Chad M, Young, Sarah, Warren, Wesley C, Badger, Jonathan, Crabtree, Jonathan, Markowitz, Victor M, Orvis, Joshua, Cree, Andrew, Ferriera, Steve, Fulton, Lucinda L, Fulton, Robert S, Gillis, Marcus, Hemphill, Lisa D, Joshi, Vandita, Kovar, Christie, Torralba, Manolito, Wetterstrand, Kris A, Abouellleil, Amr, Wollam, Aye M, Buhay, Christian J, Ding, Yan, Dugan, Shannon, FitzGerald, Michael G, Holder, Mike, Hostetler, Jessica, Clifton, Sandra W, Allen-Vercoe, Emma, Earl, Ashlee M, Farmer, Candace N, Liolios, Konstantinos, Surette, Michael G, Xu, Qiang, Pohl, Craig, Wilczek-Boney, Katarzyna, & Zhu, Dianhui. 2010. A catalog of reference genomes from the human microbiome. *Science (New York, N.Y.)*, **328**(5981), 994–999.

Newman, B, Selby, J V, King, M C, Slemenda, C, Fabsitz, R, & Friedman, G D. 1987. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*, **30**(10), 763–768.

Nicholson, Jeremy K, Wilson, Ian D, & Lindon, John C. 2011. Pharmacometabonomics as an effector for personalized medicine. *Pharmacogenomics*, **12**(1), 103–111.

Noy, Natalya F., Crubézy, Monica, Fergerson, Ray W., Knublauch, Holger, Tu, Samson W., Vendetti, Jennifer, & Musen, Mark A. 2003. Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment. *AMIA Annual Symposium Proceedings*, **2003**, 953.

Oh no sequences. 2014. *- bio4j*.

O'Hara, Ann M, & Shanahan, Fergus. 2006. The gut flora as a forgotten organ. *EMBO Reports*, **7**(7), 688–693.

Ohnuma, Takayuki, Onaga, Shoko, Murata, Katsuyoshi, Taira, Toki, & Katoh, Etsuko. 2008. LysM Domains from Pteris ryukyuensis Chitinase-A A STABILITY STUDY AND CHARACTERIZATION OF THE CHITIN-BINDING SITE. *Journal of Biological Chemistry*, **283**(8), 5178–5187.

Okuda, H, Ogura, K, Kato, A, Takubo, H, & Watabe, T. 1998. A possible mechanism of eighteen patient deaths caused by interactions of sorivudine, a new antiviral drug, with oral 5-fluorouracil prodrugs. *The Journal of Pharmacology and Experimental Therapeutics*, **287**(2), 791–799.

Ondov, Brian D, Bergman, Nicholas H, & Phillippy, Adam M. 2011. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Oracle, Corp. *Java Software | Java Technologies | Oracle UK*.

Oracle, Corp. *Welcome to NetBeans*.

O'Sullivan, Orla, O'Callaghan, John, Sangrador-Vegas, Amaia, McAuliffe, Olivia, Slattery, Lydia, Kaleta, Pawel, Callanan, Michael, Fitzgerald, Gerald F, Ross, R Paul, & Beresford, Tom. 2009. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC Microbiology*, **9**, 50.

Panda, Suchita, El khader, Ismail, Casellas, Francesc, López Vivancos, Josefa, García Cors, Montserrat, Santiago, Alba, Cuenca, Silvia, Guarner, Francisco, & Manichanh, Chaysavanh. 2014. Short-term effect of antibiotics on human gut microbiota. *PloS One*, **9**(4), e95476.

Payne, David J, Gwynn, Michael N, Holmes, David J, & Pompliano, David L. 2007. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nature Reviews. Drug Discovery*, **6**(1), 29–40.

Pearson, Jeffrey P., & Brownlee, Iain A. 2010. The Interaction of Large Bowel Microflora with the Colonic Mucus Barrier. *International Journal of Inflammation*, **2010**(Oct.).

Penders, John, Stobberingh, Ellen E., Savelkoul, Paul H. M., & Wolffs, Petra F. G. 2013. The human microbiome as a reservoir of antimicrobial resistance. *Frontiers in Microbiology*, **4**(Apr.).

Peterson, Jane, Garges, Susan, Giovanni, Maria, McInnes, Pamela, Wang, Lu, Schloss, Jeffery A, Bonazzi, Vivien, McEwen, Jean E, Wetterstrand, Kris A, Deal, Carolyn, Baker, Carl C, Di Francesco, Valentina, Howcroft, T Kevin, Karp, Robert W, Lunsford, R Dwayne, Wellington, Christopher R, Belachew, Tsegahiwot, Wright, Michael, Giblin, Christina, David, Hagit, Mills, Melody, Salomon, Rachelle, Mullins, Christopher, Akolkar, Beena, Begg, Lisa, Davis, Cindy, Grandison, Lindsey, Humble, Michael, Khalsa, Jag, Little, A Roger, Peavy, Hannah, Pontzer, Carol, Portnoy, Matthew, Sayre, Michael H, Starke-Reed, Pamela, Zakhari, Samir, Read, Jennifer, Watson, Bracie, & Guyer, Mark. 2009. The NIH Human Microbiome Project. *Genome Research*, **19**(12), 2317–2323.

Pickup, John C. 2004. Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes. *Diabetes care*, **27**(3), 813–823.

Plewczynski, Dariusz, & Rychlewski, Leszek. 2009. Meta-basic estimates the size of druggable human genome. *Journal of Molecular Modeling*, **15**(6), 695–699.

Pokusaeva, Karina, Fitzgerald, Gerald F., & van Sinderen, Douwe. 2011. Carbohydrate metabolism in Bifidobacteria. *Genes & Nutrition*, **6**(3), 285–306.

Prakash, Tulika, & Taylor, Todd D. 2012. Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics*, **13**(6), 711–727.

Preidis, Geoffrey A, & Versalovic, James. 2009. Targeting the human microbiome with antibiotics, probiotics, and prebiotics: gastroenterology enters the metagenomics era. *Gastroenterology*, **136**(6), 2015–2031.

Proal, Amy D, Albert, Paul J, Blaney, Greg P, Lindseth, Inge A, Benediktsson, Chris, & Marshall, Trevor G. 2011. Immunostimulation in the era of the metagenome. *Cellular & Molecular Immunology*, **8**(3), 213–225.

Pudlo, Nicholas A., Urs, Karthik, Kumar, Supriya Suresh, German, J. Bruce, Mills, David A., & Martens, Eric C. 2015a. Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *mBio*, **6**(6), e01282–15.

Pudlo, Nicholas A., Urs, Karthik, Kumar, Supriya Suresh, German, J. Bruce, Mills, David A., & Martens, Eric C. 2015b. Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. *mBio*, **6**(6), e01282–01215.

Pylro, Victor Satler, Roesch, Luiz Fernando Wurdig, Ortega, José Miguel, do Amaral, Alexandre Morais, Tótola, Marcos Rogério, Hirsch, Penny Ruth, Rosado, Alexandre Soares, Góes-Neto, Aristóteles, da Costa da Silva, Artur Luiz, Rosa, Carlos Augusto, Morais, Daniel Kumazawa, Andreote, Fernando Dini, Duarte, Gabriela Frois, de Melo, Itamar Soares, Seldin, Lucy, Lambais, Márcio Rodrigues, Hungria, Mariangela, Peixoto, Raquel Silva, Kruger, Ricardo Henrique, Tsai, Siu Mui, Azevedo, Vasco, & Brazilian Microbiome Project Organization Committee. 2014. Brazilian Microbiome Project: revealing the unexplored microbial diversity–challenges and prospects. *Microbial Ecology*, **67**(2), 237–241.

Qin, Junjie, Li, Ruiqiang, Raes, Jeroen, Arumugam, Manimozhiyan, Burgdorf, Kristoffer Solvsten, Manichanh, Chaysavanh, Nielsen, Trine, Pons, Nicolas, Levenez, Florence, Yamada, Takuji, Mende, Daniel R., Li, Junhua, Xu, Junming, Li, Shaochuan, Li, Dongfang, Cao, Jianjun, Wang, Bo, Liang, Huiqing, Zheng, Huisong, Xie, Yinlong, Tap, Julien, Lepage, Patricia, Bertalan, Marcelo, Batto, Jean-Michel, Hansen, Torben, Le Paslier, Denis, Linneberg, Allan, Nielsen, H. Bjorn, Pelletier, Eric, Renault, Pierre, Sicheritz-Ponten, Thomas, Turner, Keith, Zhu, Hongmei, Yu, Chang, Li, Shengting, Jian, Min, Zhou, Yan, Li, Yingrui, Zhang, Xiuqing, Li, Songgang, Qin, Nan, Yang, Huanming, Wang, Jian, Brunak, Soren, Dore, Joel, Guarner, Francisco, Kristiansen, Karsten, Pedersen, Oluf, Parkhill, Julian, Weissenbach, Jean, Bork, Peer, Ehrlich, S. Dusko, & Wang, Jun. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**(7285), 59–65.

Qin, Junjie, Li, Yingrui, Cai, Zhiming, Li, Shenghui, Zhu, Jianfeng, Zhang, Fan, Liang, Suisha, Zhang, Wenwei, Guan, Yuanlin, Shen, Dongqian, Peng, Yangqing, Zhang, Dongya, Jie, Zhuye, Wu, Wenxian, Qin, Youwen, Xue, Wenbin, Li, Junhua, Han, Lingchuan, Lu, Donghui, Wu, Peixian, Dai, Yali, Sun, Xiaojuan, Li, Zesong, Tang, Aifa, Zhong, Shilong, Li, Xiaoping, Chen, Weineng, Xu, Ran, Wang, Mingbang, Feng, Qiang, Gong, Meihua, Yu, Jing, Zhang, Yanyan, Zhang, Ming, Hansen, Torben, Sanchez, Gaston, Raes, Jeroen, Falony, Gwen, Okuda, Shujiro, Almeida, Mathieu, LeChatelier, Emmanuelle, Renault, Pierre, Pons, Nicolas, Batto, Jean-Michel, Zhang, Zhaoxi, Chen, Hua, Yang, Ruifu, Zheng, Weimou, Li, Songgang, Yang, Huanming, Wang, Jian, Ehrlich, S. Dusko, Nielsen, Rasmus, Pedersen, Oluf, Kristiansen, Karsten, & Wang, Jun. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**(7418), 55–60.

Quigley, Eamonn Mm. 2011. Gut microbiota and the role of probiotics in therapy. *Current Opinion in Pharmacology*, Oct.

Rangan, Kavita J., Pedicord, Virginia A., Wang, Yen-Chih, Kim, Byungchul, Lu, Yun, Shaham, Shai, Mucida, Daniel, & Hang, Howard C. 2016. A secreted bacterial peptidoglycan hydrolase enhances tolerance to enteric pathogens. *Science*, **353**(6306), 1434–1437.

Rawlings, Neil D., Morton, Fraser R., Kok, Chai Yin, Kong, Jun, & Barrett, Alan J. 2008. MEROPS: the peptidase database. *Nucleic Acids Research*, **36**(Database issue), D320–D325.

Rawlings, Neil D., Barrett, Alan J., & Finn, Robert. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, **44**(D1), D343–350.

Reeder, Jens, & Knight, Rob. 2010. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature methods*, **7**(9), 668–669.

Research, Center for Drug Evaluation and. 2014 (Oct.). *New Drugs at FDA: CDER's New Molecular Entities and New Therapeutic Biological Products of 2014*. New Drugs at FDA: CDER's New Molecular Entities of 2013.

Reue, Karen, & Donkor, Jimmy. 2007. Genetic Factors in Type 2 Diabetes All in the (Lipin) Family. *Diabetes*, **56**(12), 2842–2843.

Rey, Federico E., Gonzalez, Mark D., Cheng, Jiye, Wu, Meng, Ahern, Philip P., & Gordon, Jeffrey I. 2013. Metabolic niche of a prominent sulfate-reducing human gut bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(33), 13582–13587.

Ridaura, Vanessa K., Faith, Jeremiah J., Rey, Federico E., Cheng, Jiye, Duncan, Alexis E., Kau, Andrew L., Griffin, Nicholas W., Lombard, Vincent, Henrissat, Bernard, Bain, James R., Muehlbauer, Michael J., Ilkayeva, Olga, Semenkovich, Clay F., Funai, Katsuhiko, Hayashi, David K., Lyle, Barbara J., Martini, Margaret C., Ursell, Luke K., Clemente, Jose C., Treuren, William Van, Walters, William A., Knight, Rob, Newgard, Christopher B., Heath, Andrew C., & Gordon, Jeffrey I. 2013. Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice. *Science*, **341**(6150), 1241214.

Roopchand, Diana E., Carmody, Rachel N., Kuhn, Peter, Moskal, Kristin, Rojas-Silva, Patricio, Turnbaugh, Peter J., & Raskin, Ilya. 2015. Dietary Polyphenols Promote Growth of the Gut Bacterium Akkermansia muciniphila and Attenuate High-Fat Diet–Induced Metabolic Syndrome. *Diabetes*, **64**(8), 2847–2858.

Round, June L, & Mazmanian, Sarkis K. 2009. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology*, **9**(5), 313–323.

Sahin, Mehmet, Cam, Hakan, Olgar, Seref, Tunc, Sevket Ercan, Arslan, Cagatay, Uz, Efkan, & Yilmaz, H Ramazan. 2006. Protective role of erdosteine on vancomycin-induced oxidative stress in rat liver. *Molecular and cellular biochemistry*, **291**(1-2), 155–160.

Saleh, Maya, & Trinchieri, Giorgio. 2011. Innate immune mechanisms of colitis and colitis-associated colorectal cancer. *Nature Reviews. Immunology*, **11**(1), 9–20.

Salyers, A A, West, S E, Vercellotti, J R, & Wilkins, T D. 1977. Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Applied and Environmental Microbiology*, **34**(5), 529–533.

Saulnier, Delphine M, Riehle, Kevin, Mistretta, Toni-Ann, Diaz, Maria-Alejandra, Mandal, Debasmita, Raza, Sabeen, Weidler, Erica M, Qin, Xiang, Coarfa, Cristian, Milosavljevic, Aleksandar, Petrosino, Joseph F, Highlander, Sarah, Gibbs, Richard, Lynch, Susan V, Shulman, Robert J, & Versalovic, James. 2011. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology*, **141**(5), 1782–1791.

Savage, D C. 1977. Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, **31**, 107–133.

Scheen, A J. 2003. Pathophysiology of type 2 diabetes. *Acta clinica Belgica*, **58**(6), 335–341.

Schmidt, T M, DeLong, E F, & Pace, N R. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, **173**(14), 4371–4378.

Scott, Laura J., Mohlke, Karen L., Bonnycastle, Lori L., Willer, Cristen J., Li, Yun, Duren, William L., Erdos, Michael R., Stringham, Heather M., Chines, Peter S., Jackson, Anne U., Prokunina-Olsson, Ludmila, Ding, Chia-Jen, Swift, Amy J., Narisu, Narisu, Hu, Tianle, Pruim, Randall, Xiao, Rui, Li, Xiao-Yi, Conneely, Karen N., Riebow, Nancy L., Sprau, Andrew G., Tong, Maurine, White, Peggy P., Hetrick, Kurt N., Barnhart, Michael W., Bark, Craig W., Goldstein, Janet L., Watkins, Lee, Xiang, Fang, Saramies, Jouko, Buchanan, Thomas A., Watanabe, Richard M., Valle, Timo T., Kinnunen, Leena, Abecasis, Goncalo R., Pugh, Elizabeth W., Doheny, Kimberly F., Bergman, Richard N., Tuomilehto, Jaakko, Collins, Francis S., & Boehnke, Michael. 2007. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science (New York, N.Y.)*, **316**(5829), 1341–1345.

Segata, Nicola, Izard, Jacques, Waldron, Levi, Gevers, Dirk, Miropolsky, Larisa, Garrett, Wendy S, & Huttenhower, Curtis. 2011. Metagenomic biomarker discovery and explanation. *Genome Biology*, **12**(6), R60.

Sekirov, Inna, Russell, Shannon L., Antunes, L. Caetano M., & Finlay, B. Brett. 2010. Gut Microbiota in Health and Disease. *Physiological Reviews*, **90**(3), 859 –904.

Shanahan, Fergus. 2010. Probiotics in perspective. *Gastroenterology*, **139**(6), 1808–1812.

Sharpton, Thomas J, Riesenfeld, Samantha J, Kembel, Steven W, Ladau, Joshua, O'Dwyer, James P, Green, Jessica L, Eisen, Jonathan A, & Pollard, Katherine S. 2011. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology*, **7**(1), e1001061.

Shaw, Souradet Y, Blanchard, James F, & Bernstein, Charles N. 2011. Association Between the Use of Antibiotics and New Diagnoses of Crohn's Disease and Ulcerative Colitis. *The American Journal of Gastroenterology*, Sept.

Shen, Jia, Goyal, Abhinav, & Sperling, Laurence. 2012. The Emerging Epidemic of Obesity, Diabetes, and the Metabolic Syndrome in China. *Cardiology Research and Practice*, **2012**.

Shenderov, B A. 2011. Probiotic (symbiotic) bacterial languages. *Anaerobe*, May.

Smoot, Michael E, Ono, Keiichiro, Ruscheinski, Johannes, Wang, Peng-Liang, & Ideker, Trey. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, **27**(3), 431–432.

Smyth, Simon, & Heron, Andrew. 2006. Diabetes and obesity: the twin epidemics. *Nature Medicine*, **12**(1), 75–80.

Sousa, Tiago, Paterson, Ronnie, Moore, Vanessa, Carlsson, Anders, Abrahamsson, Bertil, & Basit, Abdul W. 2008. The gastrointestinal microbiota as a site for the biotransformation of drugs. *International Journal of Pharmaceutics*, **363**(1-2), 1–25.

Spor, Aymé, Koren, Omry, & Ley, Ruth. 2011. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat Rev Micro*, **9**(4), 279–290.

Steck, Natalie, Mueller, Kerstin, Schemann, Michael, & Haller, Dirk. 2011. Bacterial proteases in IBD and IBS. *Gut*, Sept.

Stumvoll, Michael, Goldstein, Barry J, & van Haeften, Timon W. 2005. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, **365**(9467), 1333–1346.

Sun, Y., Wipat, Anil, Pocock, M., Lee, P. A, Watson, P., Flanagan, K., & Worthington, J. T. 2005. A grid-based system for microbial genome comparison and analysis. *Pages 977– 984 Vol. 2 of: IEEE International Symposium on Cluster Computing and the Grid, 2005. CCGrid 2005*, vol. 2. IEEE.

Surwit, R S, Kuhn, C M, Cochrane, C, McCubbin, J A, & Feinglos, M N. 1988a. Diet-induced type II diabetes in C57BL/6J mice. *Diabetes*, **37**(9), 1163–1167.

Surwit, Richard S., Kuhn, Cynthia M., Cochrane, Christina, McCubbin, James A., & Feinglos, Mark N. 1988b. Diet-Induced Type II Diabetes in C57BL/6J Mice. *Diabetes*, **37**(9), 1163–1167.

Tailford, Louise E., Crost, Emmanuelle H., Kavanaugh, Devon, & Juge, Nathalie. 2015. Mucin glycan foraging in the human gut microbiome. *Frontiers in Genetics*, **6**(Mar.).

Taylor, Roy. 2013. Type 2 Diabetes Etiology and reversibility. *Diabetes Care*, **36**(4), 1047–1055.

Temple, Max J., Cuskin, Fiona, Baslé, Arnaud, Hickey, Niall, Speciale, Gaetano, Williams, Spencer J., Gilbert, Harry J., & Lowe, Elisabeth C. 2017. A Bacteroidetes locus dedicated to fungal 1,6-beta-glucan degradation: unique substrate conformation drives specificity of the key endo-1,6-beta-glucanase. *Journal of Biological Chemistry*, May, jbc.M117.787606.

Terrapon, Nicolas, Lombard, Vincent, Gilbert, Harry J., & Henrissat, Bernard. 2015. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics (Oxford, England)*, **31**(5), 647–655.

The MathWorks, Inc. *MATLAB 8.0 and Statistics Toolbox 8.1*.

The MathWorks, Inc. *MATLAB and Statistics Toolbox Release 2012b*.

Thuny, Franck, Richet, Hervé, Casalta, Jean-Paul, Angelakis, Emmanouil, Habib, Gilbert, & Raoult, Didier. 2010. Vancomycin treatment of infective endocarditis is linked with recently acquired obesity. *PloS One*, **5**(2), e9074.

Tilg, Herbert, & Kaser, Arthur. 2011. Gut microbiome, obesity, and metabolic dysfunction. *Journal of Clinical Investigation*, **121**(6), 2126–2132.

Tsilimigras, Matthew C. B., & Fodor, Anthony A. 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, **26**(5), 330–335.

Turnbaugh, Peter J, Ley, Ruth E, Mahowald, Michael A, Magrini, Vincent, Mardis, Elaine R, & Gordon, Jeffrey I. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**(7122), 1027–1031.

Turnbaugh, Peter J, Hamady, Micah, Yatsunenko, Tanya, Cantarel, Brandi L, Duncan, Alexis, Ley, Ruth E, Sogin, Mitchell L, Jones, William J, Roe, Bruce A, Affourtit, Jason P, Egholm, Michael, Henrissat, Bernard, Heath, Andrew C, Knight, Rob, & Gordon, Jeffrey I. 2009. A core gut microbiome in obese and lean twins. *Nature*, **457**(7228), 480–484.

Valencia, Alfonso. 2005. Automatic annotation of protein function. *Current Opinion in Structural Biology*, **15**(3), 267–274.

Venter, J C, Adams, M D, Myers, E W, Li, P W, Mural, R J, Sutton, G G, Smith, H O, Yandell, M, Evans, C A, Holt, R A, Gocayne, J D, Amanatides, P, Ballew, R M, Huson, D H, Wortman, J R, Zhang, Q, Kodira, C D, Zheng, X H, Chen, L, Skupski, M, Subramanian, G, Thomas, P D, Zhang, J, Gabor Miklos, G L, Nelson, C, Broder, S, Clark, A G, Nadeau, J, McKusick, V A, Zinder, N, Levine, A J, Roberts, R J, Simon, M, Slayman, C, Hunkapiller, M, Bolanos, R, Delcher, A, Dew, I, Fasulo, D, Flanigan, M, Florea, L, Halpern, A, Hannenhalli, S, Kravitz, S, Levy, S, Mobarry, C, Reinert, K, Remington, K, Abu-Threideh, J, Beasley, E, Biddick, K, Bonazzi, V, Brandon, R, Cargill, M, Chandramouliswaran, I, Charlab, R, Chaturvedi, K, Deng, Z, Di Francesco, V, Dunn, P, Eilbeck, K, Evangelista, C, Gabrielian, A E, Gan, W, Ge, W, Gong, F, Gu, Z, Guan, P, Heiman, T J, Higgins, M E, Ji, R R, Ke, Z, Ketchum, K A, Lai, Z, Lei, Y, Li, Z, Li, J, Liang, Y, Lin, X, Lu, F, Merkulov, G V, Milshina, N, Moore, H M, Naik, A K, Narayan, V A, Neelam, B, Nusskern, D, Rusch, D B, Salzberg, S, Shao, W, Shue, B, Sun, J, Wang, Z, Wang, A, Wang, X, Wang, J, Wei, M, Wides, R, Xiao, C, Yan, C, Yao, A, Ye, J, Zhan, M, Zhang, W, Zhang, H, Zhao, Q, Zheng, L, Zhong, F, Zhong, W, Zhu, S, Zhao, S, Gilbert, D, Baumhueter, S, Spier, G, Carter, C, Cravchik, A, Woodage, T, Ali, F, An, H, Awe, A, Baldwin, D, Baden, H, Barnstead, M, Barrow, I, Beeson, K, Busam, D, Carver, A, Center, A, Cheng, M L, Curry, L, Danaher, S, Davenport, L, Desilets, R, Dietz, S, Dodson, K, Doup, L, Ferriera, S, Garg, N, Gluecksmann, A, Hart, B, Haynes, J, Haynes, C, Heiner, C, Hladun, S, Hostin, D, Houck, J, Howland, T, Ibegwam, C, Johnson, J, Kalush, F, Kline, L, Koduru, S, Love, A, Mann, F, May, D, McCawley, S, McIntosh, T, McMullen, I, Moy, M, Moy, L, Murphy, B, Nelson, K, Pfannkoch, C, Pratts, E, Puri, V, Qureshi, H, Reardon, M, Rodriguez, R, Rogers, Y H, Romblad, D, Ruhfel, B, Scott, R, Sitter, C, Smallwood, M, Stewart, E, Strong, R, Suh, E, Thomas, R, Tint, N N, Tse, S, Vech, C, Wang, G, Wetter, J, Williams, S, Williams, M, Windsor, S, Winn-Deen, E, Wolfe, K, Zaveri, J, Zaveri, K, Abril, J F, Guigó, R, Campbell, M J, Sjolander, K V, Karlak, B, Kejariwal, A, Mi, H, Lazareva, B, Hatton, T, Narechania, A, Diemer, K, Muruganujan, A, Guo, N, Sato, S, Bafna, V, Istrail, S, Lippert, R, Schwartz, R, Walenz, B, Yooseph, S, Allen, D, Basu, A, Baxendale, J, Blick, L, Caminha, M, Carnes-Stine, J, Caulk, P, Chiang, Y H, Coyne, M, Dahlke, C, Mays, A, Dombroski, M, Donnelly, M, Ely, D, Esparham, S, Fosler, C, Gire, H, Glanowski, S, Glasser, K, Glodek, A, Gorokhov, M, Graham, K, Gropman, B, Harris, M, Heil, J, Henderson, S, Hoover, J, Jennings, D, Jordan, C, Jordan, J, Kasha, J, Kagan, L, Kraft, C, Levitsky, A, Lewis, M, Liu, X, Lopez, J, Ma, D, Majoros, W, McDaniel, J, Murphy, S, Newman, M, Nguyen, T, Nguyen, N, Nodell, M, Pan, S, Peck, J, Peterson, M, Rowe, W, Sanders, R, Scott, J, Simpson, M, Smith, T, Sprague, A, Stockwell, T, Turner, R, Venter, E, Wang, M, Wen, M, Wu, D, Wu, M, Xia, A, Zandieh, A, & Zhu, X. 2001. The sequence of the human genome. *Science (New York, N.Y.)*, **291**(5507), 1304–1351.

Viney, Mark. 2014. The failure of genomics in biology. *Trends in Parasitology*, **30**(7), 319–321.

Visweswaran, Ganesh Ram R., Leenhouts, Kees, van Roosmalen, Maarten, Kok, Jan, & Buist, Girbe. 2014. Exploiting the peptidoglycan-binding motif, LysM, for medical and industrial applications. *Applied Microbiology and Biotechnology*, **98**(10), 4331–4345.

Wallace, Bret D, Wang, Hongwei, Lane, Kimberly T, Scott, John E, Orans, Jillian, Koo, Ja Seol, Venkatesh, Madhukumar, Jobin, Christian, Yeh, Li-An, Mani, Sridhar, & Redinbo, Matthew R. 2010. Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science (New York, N.Y.)*, **330**(6005), 831–835.

Walter, Jens, & Ley, Ruth. 2011. The human gut microbiome: ecology and recent evolutionary changes. *Annual Review of Microbiology*, **65**, 411–429.

Wang, Chao-Yung, & Liao, James K. 2012. A Mouse Model of Diet-Induced Obesity and Insulin Resistance. *Methods in molecular biology (Clifton, N.J.)*, **821**, 421–433.

Wang, Qiong, Garrity, George M., Tiedje, James M., & Cole, James R. 2007. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, **73**(16), 5261–5267.

Ward, M M. 1988. Factors predictive of acute renal failure in rhabdomyolysis. *Archives of internal medicine*, **148**(7), 1553–1557.

Warrender, Jennifer D., & Lord, Phillip. 2015. How, What and Why to test an ontology. *arXiv:1505.04112 [cs]*, May. arXiv: 1505.04112.

Weile, Jochen, Pocock, Matthew, Cockell, Simon J, Lord, Phillip, Dewar, James M, Holstein, Eva-Maria, Wilkinson, Darren, Lydall, David, Hallinan, Jennifer, & Wipat, Anil. 2011. Customizable views on semantically integrated networks for systems biology. *Bioinformatics (Oxford, England)*, **27**(9), 1299–1306.

Weinstock, Joel V, & Elliott, David E. 2009. Helminths and the IBD hygiene hypothesis. *Inflammatory Bowel Diseases*, **15**(1), 128–133.

Weisburg, W G, Barns, S M, Pelletier, D A, & Lane, D J. 1991. 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, **173**(2), 697–703.

Wells, Jerry M, Rossi, Oriana, Meijerink, Marjolein, & van Baarlen, Peter. 2011. Epithelial crosstalk at the microbiota-mucosal interface. *Proceedings of the National Academy of Sciences of the United States of America*, **108 Suppl 1**(Mar.), 4607–4614.

White, James Robert, Nagarajan, Niranjan, & Pop, Mihai. 2009. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol*, **5**(4), e1000352.

Whitman, W B, Coleman, D C, & Wiebe, W J. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(12), 6578–6583.

Wikoff, William R, Anfora, Andrew T, Liu, Jun, Schultz, Peter G, Lesley, Scott A, Peters, Eric C, & Siuzdak, Gary. 2009. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(10), 3698–3703.

Wild, Sarah, Roglic, Gojka, Green, Anders, Sicree, Richard, & King, Hilary. 2004. Global Prevalence of Diabetes Estimates for the year 2000 and projections for 2030. *Diabetes Care*, **27**(5), 1047–1053.

Willing, Benjamin P, Gill, Navkiran, & Finlay, B Brett. 2010. The role of the immune system in regulating the microbiota. *Gut Microbes*, **1**(4), 213–223.

Wilson, Ian D. 2009. Drugs, bugs, and personalized medicine: pharmacometabonomics enters the ring. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(34), 14187–14188.

Winzell, M. S., & Ahren, B. 2004. The High-Fat Diet-Fed Mouse: A Model for Studying Mechanisms and Treatment of Impaired Glucose Tolerance and Type 2 Diabetes. *Diabetes*, **53**(suppl_3), S215–S219.

World Health Organization. 2013 (Oct.). *WHO | Diabetes*.

Wu, Gary D, Lewis, James D, Hoffmann, Christian, Chen, Ying-Yu, Knight, Rob, Bittinger, Kyle, Hwang, Jennifer, Chen, Jun, Berkowsky, Ronald, Nessel, Lisa, Li, Hongzhe, & Bushman, Frederic D. 2010. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiology*, **10**, 206.

Wu, Gary D, Chen, Jun, Hoffmann, Christian, Bittinger, Kyle, Chen, Ying-Yu, Keilbaugh, Sue A, Bewtra, Meenakshi, Knights, Dan, Walters, William A, Knight, Rob, Sinha, Rohini, Gilroy, Erin, Gupta, Kernika, Baldassano, Robert, Nessel, Lisa, Li, Hongzhe, Bushman, Frederic D, & Lewis, James D. 2011. Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)*, **334**(6052), 105–108.

Xu, Jian, Mahowald, Michael A, Ley, Ruth E, Lozupone, Catherine A, Hamady, Micah, Martens, Eric C, Henrissat, Bernard, Coutinho, Pedro M, Minx, Patrick, Latreille, Philippe, Cordum, Holland, Van Brunt, Andrew, Kim, Kyung, Fulton, Robert S, Fulton, Lucinda A, Clifton, Sandra W, Wilson, Richard K, Knight, Robin D, & Gordon, Jeffrey I. 2007. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, **5**(7), e156.

Yin, Yanbin, Mao, Xizeng, Yang, Jincai, Chen, Xin, Mao, Fenglou, & Xu, Ying. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, **40**(Web Server issue), W445–451.

Zaneveld, Jesse, Turnbaugh, Peter J, Lozupone, Catherine, Ley, Ruth E, Hamady, Micah, Gordon, Jeffrey I, & Knight, Rob. 2008. Host-bacterial coevolution and the search for new drug targets. *Current Opinion in Chemical Biology*, **12**(1), 109–114.

Zeggini, Eleftheria, Scott, Laura J., Saxena, Richa, Voight, Benjamin F., Marchini, Jonathan L, Hu, Tainle, de Bakker, Paul IW, Abecasis, Goncalo R, Almgren, Peter, Andersen, Gitte, Ardlie, Kristin, Bostrom, Kristina Bengtsson, Bergman, Richard N, Bonnycastle, Lori L, Borch-Johnsen, Knut, Burtt, Noel P, Chen, Hong, Chines, Peter S, Daly, Mark J, Deodhar, Parimal, Ding, Charles, Doney, Alex S F, Duren, William L, Elliott, Katherine S, Erdos, Michael R, Frayling, Timothy M, Freathy, Rachel M, Gianniny, Lauren, Grallert, Harald, Grarup, Niels, Groves, Christopher J, Guiducci, Candace, Hansen, Torben, Herder, Christian, Hitman, Graham A, Hughes, Thomas E, Isomaa, Bo, Jackson, Anne U, J?rgensen, Torben, Kong, Augustine, Kubalanza, Kari, Kuruvilla, Finny G, Kuusisto, Johanna, Langenberg, Claudia, Lango, Hana, Lauritzen, Torsten, Li, Yun, Lindgren, Cecilia M, Lyssenko, Valeriya, Marvelle, Amanda F, Meisinger, Christa, Midthjell, Kristian, Mohlke, Karen L, Morken, Mario A, Morris, Andrew D, Narisu, Narisu, Nilsson, Peter, Owen, Katharine R, Palmer, Colin NA, Payne, Felicity, Perry, John RB, Pettersen, Elin, Platou, Carl, Prokopenko, Inga, Qi, Lu, Qin, Li, Rayner, Nigel W, Rees, Matthew, Roix, Jeffrey J, Sandbaek, Anelli, Shields, Beverley, Sjogren, Marketa, Steinthorsdottir, Valgerdur, Stringham, Heather M, Swift, Amy J, Thorleifsson, Gudmar, Thorsteinsdottir, Unnur, Timpson, Nicholas J, Tuomi, Tiinamaija, Tuomilehto, Jaakko, Walker, Mark, Watanabe, Richard M, Weedon, Michael N, Willer, Cristen J, Illig, Thomas, Hveem, Kristian, Hu, Frank B, Laakso, Markku, Stefansson, Kari, Pedersen, Oluf, Wareham, Nicholas J, Barroso, Ines, Hattersley, Andrew T, Collins, Francis S, Groop, Leif, McCarthy, Mark I, Boehnke, Michael, & Altshuler, David. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, **40**(5), 638–645.

Zhang, Zhang, Bajic, Vladimir B., Yu, Jun, Cheung, Kei-Hoi, & Townsend, Jeffrey P. 2011. Data Integration in Bioinformatics: Current Efforts and Challenges.

Zhang, Zhigang, Geng, Jiawei, Tang, Xiaodan, Fan, Hong, Xu, Jinchao, Wen, Xiujun, Ma, Zhanshan (Sam), & Shi, Peng. 2014. Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal*, **8**(4), 881–893.

Zhao, Dong, Zhao, Fan, Li, Yan, & Zheng, Zheng. 2012. Projected and Observed Diabetes Epidemics in China and Beyond. *Current Cardiology Reports*, **14**(1), 106–111.

Zoetendal, Erwin G., von Wright, Atte, Vilpponen-Salmela, Terttu, Ben-Amor, Kaouther, Akkermans, Antoon D. L., & de Vos, Willem M. 2002. Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Applied and Environmental Microbiology*, **68**(7), 3401–3407.