THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Systems Biology of the Gut Microbiome in Metabolic Diseases

Fredrik H. Karlsson

CHALMERS

Department of Chemical and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2014

Systems Biology of the Gut Microbiome in Metabolic Diseases
Fredrik Karlsson
ISBN 978-91-7385-965-3

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3646
ISSN 0346-718X

Department of Chemical and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Concept of metagenomics: DNA is extracted and sequenced from a microbial community. Sequences are annotated to metabolic functions.

Back cover photo taken by J-O Yxell

Printed by Chalmers Reproservice
Gothenburg, Sweden 2014

Systems Biology of the Gut Microbiome in Metabolic Diseases

FREDRIK H. KARLSSON

Department of Chemical and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Abstract

The human body is hosting a tremendous number of microbial cells and many of these reside in our gut. The gut microbes perform breakdown of indigestible dietary components and contribute to energy harvesting from food. Furthermore, there is a constant interaction between our microbes and our immune system to fend of pathogens and tolerate commensals. Evidence suggests that the composition of the gut microbiota is altered in obesity and could contribute to development of obesity related metabolic diseases. This thesis presents results that show an association between the gut microbiome, the collective genomes of the microbiota, and symptomatic atherosclerosis. The gut microbiome was also found to be associated with diabetes and a classifying model for diabetic status was developed. A novel method for grouping genes into metagenomic clusters that are likely derived from the same genome is also presented. Bariatric surgery is an effective method for treating obesity and reduces the risk of its co-morbidities. It was also found that the gut metagenome is altered considerably after bariatric surgery.

Two software tools for metagenomic data analysis and hypothesis testing are presented. MEDUSA is software for quality control and annotation of metagenomic sequence reads. MEDUSA was used for the analysis of 782 gut metagenomes and a global human gut microbial gene catalogue was constructed and evaluated. FANTOM is software with a graphical user interphase that provides hypothesis testing in a taxonomical and functional context. To model key metabolic functions of gut microbes, genome-scale metabolic models of three species from the human gut are presented and their interactions are evaluated.

This work contributes to the knowledge of associations between the gut microbiota and metabolic diseases. A number of novel methods for data analysis of gut metagenome data are presented.

Keywords: gut metagenome; atherosclerosis; diabetes, bariatric surgery; gene catalogue; FANTOM; MEDUSA; genome-scale metabolic model; metabolism

List of publications

This thesis is based on the work in the following publications:

I. **Karlsson, F.H.**, Fak, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., Backhed, F., and Nielsen, J. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. Nature Communications *3*, 1245.

II. **Karlsson, F.H.***, Tremaroli, V.*, Nookaew, I., Bergstrom, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Backhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99-103.

III. Tremaroli V.*, **Karlsson F.H.***, Werling, M., Nookaew, I., Olbers, T., Fändriks, L., le Roux C., Nielsen, J., Bäckhed, F., Long-term effects of bariatric surgery on the gut metagenome. (Manuscript)

IV. **Karlsson, F.H.**, Tremaroli, V., Nielsen, J., and Bäckhed, F. (2013). Assessing the Human Gut Microbiota in Metabolic Diseases. Diabetes 62, 3341-3349.

V. Sanli, K.*, **Karlsson, F.H.***, Nookaew, I., and Nielsen, J. (2013). FANTOM: Functional and taxonomic analysis of metagenomes. BMC Bioinformatics *14*, 38.

VI. **Karlsson F.H.**, Nookaew I., Nielsen J., Metagenomic Data Utilization and Analysis and construction of a global gut microbial gene catalogue (Submitted)

VII. **Karlsson, F.H.**, Nookaew, I., Petranovic, D., and Nielsen, J. (2011). Prospects for systems biology and modeling of the gut microbiome. Trends in Biotechnology 29: 251-258.

VIII. Shoaie, S., **Karlsson, F.H**., Mardinoglu, A., Nookaew, I., Bordel, S., and Nielsen, J. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. Scientific Reports *3*, 2532.

* Authors contributed equally

Additional publications not part of this thesis:

IX. **Karlsson, F.H.**, Ussery, D.W., Nielsen, J., and Nookaew, I. (2011). A Closer Look at Bacteroides: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut. Microbial Ecology.

X. ElSemman I., **Karlsson F.H.**, Shoaie, S., Nookaew, I., Soliman, T., and Nielsen, J., (2013). Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction. (Submitted)

Contributions

I. Performed metagenomic data analysis. Drafted and edited the paper.
II. Performed metagenomic data analysis. Participated in drafting and editing of the paper.
III. Performed metagenomic data analysis. Participated in drafting and editing of the paper.
IV. Drafted and edited the paper.
V. Supervised the work in detail and participated in drafting and editing of the paper.
VI. Designed the software and performed metagenomic data analysis. Drafted and edited the paper.
VII. Drafted and edited the paper.
VIII. Reconstructed the model for *Bacteroides thetaiotaomicron*, assisted in data analysis and drafting and editing of the paper.

Additional publications not part of this thesis:

IX. Performed data analysis. Drafted and edited the paper.
X. Supervised the work and participated in drafting and editing of the paper.

Abbreviations

AUC	Area under the curve
BLAST	Basic Local Alignment Search Tool
BMI	Body mass index
bp	Base pairs
COG	Cluster of orthologous groups
CTR	Control (refers to a group of individuals described in Paper III)
CVD	Cardiovascular disease
DNA	Deoxyribonucleic acid
FANTOM	Functional and taxonomic analysis of metagenomes
FISH	Fluorescent in situ hybridization
GEM	Genome-scale metabolic model
GSMM	Genome-scale metabolic model, same as GEM, used in Paper VII
HbA1c	Glycated hemoglobin
HDL	High-density lipoprotein
HMP	Human microbiome project
hsCRP	High sensitivity C-reactive protein
IGT	Impaired glucose tolerance
IQR	Inter-quartile range
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
LCA	Lowest common ancestor
LDL	Low-density lipoprotein
MEDUSA	Metagenomic data utilization and analysis
MGC	Metagenomic cluster
NCBI	National Center for Biotechnology Information
NGT	Normal glucose tolerance
nr	Non-redundant
OBS	Obese (refers to a group of individuals described in Paper III)
PCA	Principal component analysis
PCR	Polymerase chain reaction
PYY	Peptide yy
QIIME	Quantitative Insights Into Microbial Ecology
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RYGB	Roux en-Y gastric bypass
SD	Standard deviation
T2D	Type 2 diabetes
TMA	Trimethylamine
TMAO	Trimethylamine N-oxide
VBG	Vertical banded gastroplasty
WHO	World health organization

List of figures and tables

Table of content

1. Introduction

The world is facing an epidemic increase in obesity with a near doubling in prevalence since 1980. More than 1.4 billion people were overweight and of these 500 million were obese in the year 2008 (WHO, 2013a). The obesity epidemic, which started in the United States and Western Europe, is now widespread across the world to all continents. Obesity is a major risk factor for metabolic diseases such as cardiovascular disease and diabetes. Cardiovascular disease with its manifestation coronary heart disease and stroke is the leading cause of death worldwide with an estimated 17 million deaths or 30% of all deaths in 2008 (WHO, 2013b). Half of individuals with diabetes die of cardiovascular disease and overall the mortality rate is double in diabetic individuals compared to healthy.

On a theoretical level, obesity can be avoided by decreasing energy intake and increasing energy expenditure by exercise, in reality this is a much more complicated issue. Efforts to reduce weight by a person are compensated by biologic responses; morbid obesity is most often not a personal choice but a disease (Friedman, 2004). A range of known and unknown environmental factors, genetic factors, what diet is preferred, how much energy is extracted from diet, energy expenditure in resting and active state play a role in determining the body weight and levels of lipids and glucose in the blood. The microorganisms that live in and on us are an environmental factor that might have a role in the pathogenesis of obesity and its comorbidities cardiovascular disease and diabetes. Recent studies have shown that the gut microbiota and its collective genome, the microbiome, is altered in obesity (Duncan et al., 2008; Furet et al., 2010; Le Chatelier et al., 2013; Ley et al., 2005; Turnbaugh et al., 2009). Furthermore, the gut microbiota is not only associated with obesity but is can also transfer the obese phenotype by gut microbiota transplantation in mice (Turnbaugh et al., 2008; Turnbaugh et al., 2006; Vijay-Kumar et al., 2010) and increase insulin sensitivity in humans (Vrieze et al., 2012).

Given the serious burden obesity and its comorbidities cardiovascular disease and diabetes puts on society, there is a pressing need to find new ways of tackling this problem. Investigating the role of the gut microbiota in metabolic diseases is one important way to address this challenge. The gut microbes can be studied by shotgun sequencing of their collective genomes, the microbiome, at a detailed level to characterize the taxonomic and functional profile of this complex ecosystem. Analysis of metagenomic data and how it can be leveraged has important scientific challenges. With this background, this thesis aims to address these three questions:

How is the human gut metagenome associated with metabolic diseases?

It is known that alterations in the gut microbiota are associated with obesity but early reports have to some extent been inconsistent in the specific correlations. The role of the gut microbiota in obesity related metabolic diseases, cardiovascular disease and diabetes, have been studied using 16S rRNA sequencing and quantitative PCR, respectively (Koren et al., 2011; Larsen et al., 2010). However, 16S rRNA sequencing can reveal differences in the taxonomic makeup of the microbiota but to discern the

functional capacity of the genomes of the microorganisms in the gut, metagenomic sequencing is required. Bariatric surgery is an effective method for weight loss and often quickly cures diabetes. It is therefore interesting to investigate if there are long term changes in the composition of the gut microbiota after bariatric surgery that possibly contribute to weight-loss and improved metabolic status. In this work we investigate if and how the gut microbiome is associated with symptomatic atherosclerosis, diabetes and bariatric surgery by sequencing of the gut metagenome.

Can bioinformatics tools for analyzing metagenomic data be advanced and made more easily accessible?

To analyze gut metagenomic data mentioned above and define its functional and taxonomic composition, a bioinformatics pipeline is needed. There was no tool available that was suitable for the type and amount of data that was generated in this project. Due to the large amount of data, mapping and annotation of the metagenomic sequence reads need to be done on a computational cluster with more processing power than a personal computer. An analysis pipeline requires the use of several different programs with custom scripts for formatting the output of one program to suit the input of another. Some later part of the analysis, when all sequence data has been annotated is possible to do on a personal computer. However, statistical analyses are often done in scripting languages and the access to biological database is not incorporated into the statistical software.

This thesis describes a package, MEDUSA, that can be used on a computational cluster to annotate metagenomic sequence reads and provide a quantitative assessment of taxonomic and gene functional features. FANTOM was further developed to take the output from MEDUSA and perform statistical analysis. FANTOM provides analysis tools for drawing biological conclusions from metagenomic data in a graphical user interphase. FANTOM contains access to KEGG and NCBI taxonomy databases.

Can metabolic modeling be used for studying basic metabolism in the human gut?

Metgenomics provides a parts list of the gut microbiome, a list of potential functions that can be performed by the microbiota. We hyptothesise that metabolic modeling can be used to leverage metagenomic data to draw more detailed conclusions about how the different members of the microbiota interact at the metabolic level. Genome-scale metabolic models (GEMs) contain a mapping of genes, proteins and reactions and could be used for investigating metabolic interactions. Due to the complexity of reconstructing and accurately modeling metabolic fluxes, this needs to initially be done in a simplified system.

This thesis describes the concept of metabolic modeling of the gut microbiota and provides an example of such use. This system can be used for testing hypotheses about metabolism in the gut.

2. Background

The term microbiota is here used to describe a community of microorganisms. The term microbiome was originally coined by Joshua Lederberg as "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space". Lederberg has also been quoted as defining the "microbiome to describe the collective genome of our indigenous microbes (microflora)" (Hooper and Gordon, 2001), and the microbiome is used in this thesis to describe the collective genome of the microbiota and is equivalent to the gut metagenome.

2.1. The human gut microbiota

The microorganisms that live on and inside humans are collectively called the human microbiota. An adult human is composed of about 10^{13} somatic cells whereas the number of microorganisms that live on and in a human is 10^{14} and thus outnumber human somatic cells by a factor of 10. The microorganisms are mainly prokaryotic and some eukaryotic cells that collectively make up the human microbiota (Savage, 1977). Apart from the intestinal microbiota, also the skin, oral, nasal and vaginal microbiota have been studied extensively, especially in the Human Microbiome Project (Huttenhower et al., 2012b). The human microbiota is not restricted to these sites but do also reside in for example the lungs, the blood and atherosclerotic plaques (Amar et al., 2013; Goddard et al., 2012; Koren et al., 2011). The main mass of microorganisms associated with humans resides in our intestinal tract. The weight of the bacteria living in a human intestine is about 1.5 kg and make up about 50% of the fecal matter (Zhao, 2013). This thesis focuses on the gut microbiota and microbiota residing at other anatomical sites are not described in detail. The oral and gastrointestinal microbiota are distinctly different in composition although they are connected via the esophagus and food and microbial cells pass from the oral to the gastrointestinal tract (Huttenhower et al., 2012b).

It is important to note that the main part of the organisms making up the human microbiota is seen as non-pathogenic but rather co-exist in a symbiotic or commensal relationship with the human host.

The density of cells increases along the length of the intestine to reach 10^{11}-10^{12} cells per gram of luminal content in the distal colon (Backhed et al., 2005). The gut microbiota is mainly composed of bacteria from two major phyla, Bacteroidetes and Firmicutes, with less abundant phyla such as Actinobacteria, Proteobacteria and Verrucomicrobia (Ley et al., 2005). Methanogenic archaea from the phyla Euryachaeota are also present, mainly the species *Methanobrevibacter smithii* (Eckburg et al., 2005; Qin et al., 2010). The density and composition of the human gut microbiota is described in Figure 1. The total number of species is variable among humans but a study of 124 individuals estimates over 1000 species in the cohort and each individual carried at least 160 species. The study also found a core of species that were share among all (18 species) or a majority (75 species) of individuals. The abundance of the species in the core is highly variable, up to three orders of magnitude (Qin et al., 2010).

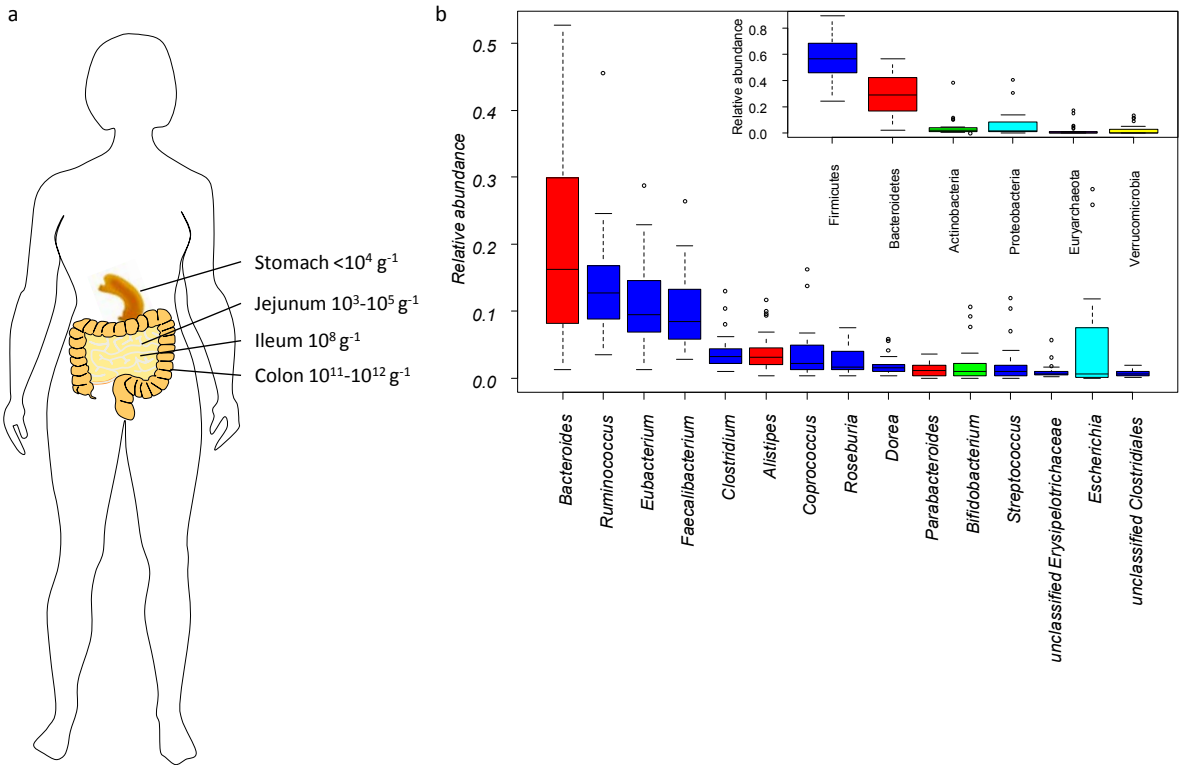

Figure 1 Density and composition of the human gut microbiota. a) Numbers for each section of the gastrointestinal tract represents the number of microorganisms per gram of intestinal content, adapted from (Leser and Molbak, 2009). b) Composition of genera (main) and phyla (inset) in human feces. Data taken from (Karlsson et al., 2012).

Humans are born essentially sterile and acquire microbes during birth in the birth canal and in contact with skin and environmental microbes. The mode of delivery and whether the baby is breast or formula fed are important factors that determines the early colonization (Wall et al., 2009). After birth, an infant is colonized by facultative anaerobes, for example *Escherichia coli* and Streptococcus species, when oxygen levels in the gut decreases, obligate anaerobic species can colonize. A study of infants from three geographical areas, United States, rural Malawi and Venezuela showed how bacterial diversity increases with age and stabilizes after about 3 years of life. Children and their parents have more similar microbiota than children to unrelated individuals and likewise are members of the same household more similar microbially than two individuals from different households (Yatsunenko et al., 2012). This indicates that common environmental exposures are important for shaping the microbiota.

The information about the composition of the gut microbiota is most commonly learnt from fecal samples that can be collected in a non-invasive manner. Commonly when we refer to the gut microbiota, we mean the composition of microorganisms in a fecal sample. Our knowledge about the composition of microorganisms from other parts of the gastrointestinal is limited as sampling of these sections is invasive and difficult to obtain. However, fecal samples reflect well the large and dense composition in the colon where most of the metabolic activity and fermentation occurs. Analysis of samples from ileostomists (individuals who had their colon removed) showed that *Streptococcus*, *Escherichia* and *Clostridium* species were most abundant in the small intestine. A larger

diversity was seen in colon samples compared to small intestinal samples (Zoetendal et al., 2012).

The stability of the gut microbiota over time has not been studied so extensively as variation between individuals. It has been shown in several studies that an individual's microbiota is more similar between two time points than to the microbiota of another individual (Huttenhower et al., 2012b; Rajilic-Stojanovic et al., 2012; Turnbaugh et al., 2009). An extensive study investigating the stability of the gut microbiota of 37 individuals over 5 years found that the gut microbiota was remarkably stable over time and 70% of the strains were remaining after 1 year with few changes occurring the following 4 years (Faith et al., 2013). Strains that were more abundant were also more stable over time. The stability was further manifested in a metagenomic study looking at single nucleotide polymorphisms (SNPs) in the microbiome and found that individual specific strains persist over time (Schloissnig et al., 2013). This indicates that a sample at one time point is representing the composition of the microbiota over time which is important for diagnostic purposes.

2.2. Traditional methods to study the gut microbiota

Culturing of microbes has been used to characterize and quantify microbial taxa of human stool samples. Quantitative culturing is done by spreading serial dilutions of a sample onto selective plates and counting the colonies formed. The taxonomic resolution varies but is typically at genera or above and culturing is only applicable to the live part of the microbiota. Culturing has successfully been used to study the infant gut microbiota where initially a large fraction of the microbes are facultative anaerobes and can be readily cultured (Adlerberth et al., 2007).

Molecular methods have been developed due to the difficulty to culture some microorganisms in the human gut, especially strictly anaerobic species, and to increase the taxonomic resolution. For bacteria and Archaea, which make up the major part of the microorganisms inhabiting the human gut, the 16S rRNA gene has been the main target for analysis since the mid-1980s (Woese, 1987). The 16S ribosomal gene is about 1500 base pairs long and ubiquitous in bacteria and Archaea (Morgan and Huttenhower, 2012). An important feature of the 16S rRNA gene is that it contains conserved regions as well as variable regions in different species which makes it possible for constructing universal primers as well as specific taxonomic identification. The conserved regions make it possible to selectively amplify and characterize only the 16S rRNA genes in a microbial sample using PCR.

Methods that aim to do fingerprint analyses of the 16S rRNA gene content of a microbial sample include temperature gradient gel electrophoresis, denaturing gradient gel electrophoresis and terminal restriction fragment length polymorphism. Gradient gel electrophoresis methods work by separating DNA fragments based on their size and sequence since the latter determines the denaturing condition and single stranded DNA migrates slower than double stranded DNA. Gradient gel electrophoresis methods have a low taxonomic resolution and are most suitable for cheap and quick comparisons for

preliminary purposes. The terminal restriction fragment length polymorphism method involves a PCR amplification of 16S rRNA genes and labeling the terminal fragment. A subsequent restriction with one or more endonucleases cleaves fragments based on sequence which can be separated on a gel. A disadvantage is that also this method does not give direct taxonomic identities to observed fragments. However, fragment lengths can be compared to *in silico* cut fragments from databases of known 16S rRNA sequences (Sjoberg et al., 2013).

Microarrays with probes complementary to 16S rRNA sequences can be used as a high throughput tool to characterize microbial communities. The human intestinal tract chip (HITChip) is designed with 1140 probes targeting the variable region of the 16S rRNA gene. The HITChip provides relative abundance information of probes but is of course limited to the sequences present on the microarray (Rajilic-Stojanovic et al., 2009).

2.2.1. 16s rRNA gene sequencing

Direct sequencing of the 16S rRNA gene is increasingly used as the cost of sequencing is dropping and bioinformatics tools and databases used for analysis are readily available. Initial studies used Sanger sequencing of cloned 16S rRNA genes into *E. coli* and could produce near full length sequences (Eckburg et al., 2005; Ley et al., 2005). Direct sequencing of amplified sequences could be performed with the introduction of the 454 sequencing technology (Andersson et al., 2008; Sogin et al., 2006). This technology has the disadvantage that it can only produce sequence lengths of 100-450 bp but a selection of the hyper variable regions of the 16S rRNA gene can be targeted which proved sufficient for taxonomic identification at a genus or species level. Typically, the hyper-variable regions of the 16S rRNA genes used are the V1, V2, V4 and V6 regions. In the analysis of 16S rRNA genes, near identical sequences are grouped or binned into operational taxonomic units, OTUs, with a similarity of 95%, 97% or 99% because errors can be introduced by sequencing and to group nearly identical species or strains into a common group. OTUs are almost equivalent to the term species but might not be named or characterized previously (Morgan and Huttenhower, 2012). There are large repositories with known sequences for species that have been cultured and isolated from the environment such as GreenGenes (DeSantis et al., 2006), SILVA (Pruesse et al., 2007) and Ribosomal Database Project (Cole et al., 2013) which facilitates easy comparison.

The analysis of 16S rRNA gene sequences can be performed with software packages such as the highly used Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso et al., 2010) and mothur (Schloss et al., 2009) that can run on a laptop or computer cluster and can analyze millions of 16S rRNA gene sequences from microbial communities. These tools are command line scripts that take raw sequences as input and could bin them into OTUs, display phylogenetic trees, calculate diversity and compare the microbial content between groups of samples. Analyses could be performed using annotations of sequences to reference databases as mentioned above or *de novo* for sequences that are not presently in the databases.

16S rRNA sequencing is today widely used as a tool for exploring the content of a microbial sample due to its relative low cost and well developed software analysis tools.

This technology is capable of answering which microorganisms are present and their abundance.

2.3. Metagenomics of the gut microbiota

Shotgun metagenome sequencing of a microbial community's genomic content, the microbiome, can not only describe the taxonomic content but also its functional potential in term of individual gene functions. This is important because reference genomes are lacking for many species in the environment and the human gut. Furthermore, species with similar 16S rRNA gene sequences can have different functional potential e.g. in toxicity and pathogenicity thus making inferences of the functional content of a microbiome from taxonomic markers difficult or incorrect.

2.3.1. DNA extraction and sequencing

A metagenome project starts with sample collection and quick freezing to -80°C as a measure to quench any changes to sample from its original state or introduction of foreign material. Next, DNA is extracted from the sample and the method used is important to recover genetic material from a broad class of cells in the samples and with consistent recovery rates. Mechanical cell lysis by repeated bead beating together with chemical lysis has been shown to yield DNA from a broad range of species from human fecal samples. Overall, 4 different DNA extraction methods using mechanical and enzymatic lysis showed more similar microbial abundance profiles compared to inter-subject variation (Salonen et al., 2010).

Sequencing technology has developed tremendously since the early metagenome projects of the human gut. Prices per base have dropped while the number of bases that could be sequenced per machine has increased several orders of magnitude. Initial studies used Sanger sequencing technology which involves laborious plasmid libraries in *E. coli* cells and subsequent purification and sequencing of individual transformants (Gill et al., 2006; Kurokawa et al., 2007). The number of sequenced bases from studies using Sanger sequencing is in the order of a hundred mega base pairs and around a hundred thousand reads per sample (Arumugam et al., 2011; Gill et al., 2006; Kurokawa et al., 2007). With the introduction of the 454 pyrosequencing technology, the isolated DNA could be sequenced without the cloning step into *E. coli*. A study of lean and obese twins presented a total of 2.1 Gbp of sequences or about half a million reads per sample from the microbiome of 18 individuals using the 454 pyrosequencing technology (Turnbaugh et al., 2009). With the introduction of the Illumina/Solexa technology, the number of reads per sample could be significantly increased. In a study of 124 individuals, Illumina sequencing was used to produce 576.7 Gbp with an average 4.5 Gbp or 62 million reads per sample (Qin et al., 2010). It was shown that even with short read lengths produced by the Illumina technology, 44 and 75 bp at the time, it was possible to assemble sequences into longer contigs which covered previously sequenced human gut metagenomes sequenced with longer reads. The SOLiD technology was recently used to produce 35 bp single reads that could characterize the gene abundance similarly to profiles produced with Illumina sequences (Cotillard et al., 2013). SOLiD reads are produced in color space, not sequence space, meaning that converted to sequence space, they are correct until the first erroneously called color. The use of color space reads means that *de novo* assembly is problematic and studies using SOLiD reads have relied

on using established reference catalogues. A comparison between the 454 pyrosequencing and Illumina technologes showed that derived assemblies overlapped by 90% of the abundance estimated correlated with an R^2 of above 0.9 (Luo et al., 2012).

2.4. Characteristics of the gut metagenome

The first metagenomic study of the human gut microbiome was performed in 2006 on two American individuals by sequencing a total of 78 Mbp. The gene functional content of the metagenome contained enrichment of genes for glycan degradation, amino acid metabolism, xenobiotic metabolism and methanogenesis compared to the human genome (Gill et al., 2006). The gut metagenomes from 13 individuals, including unweaned infants, were compared to other metagenomomes from the environment. Infants had a simpler composition and higher inter-individual variation of the metagenome compared to adults (Kurokawa et al., 2007). A shared core was identified at the gene functional levels rather than at the taxonomic level by sequencing of 18 American obese and lean individuals. Core functions were carbohydrate, glycan and amino acid metabolism whereas cell motility, signaling and membrane transport were identified as variable between the 18 gut metagenomes (Turnbaugh et al., 2009). By deep sequencing of the fecal metagenome from 124 Spanish and Danish individuals, a gene catalogue of 3.3 million genes was assembled. Almost 300,000 genes were found in at least a majority of the individuals and these were identified as a core of common genes. Out of the genes that could be taxonomically annotated almost all belonged to Bacteria and Archaea (Qin et al., 2010). A large American project, The Human Microbiome Project, sequenced the microbiome at different anatomical sites and repeatedly sampled some individuals. The variation between subjects was consistently lower compared to the variation between samples from the same individual taken at different time points both at the taxonomic and functional level (Huttenhower et al., 2012a). The studies described above have been important for describing the diversity and function of the gut microbiome. Several bioinformatics methods for analysis were described that are important for the field and in a few cases these were also distributed as public software.

2.5. Bioinformatics tools for metagenomic data analysis

The bioinformatics tools used for analysis have evolved together with the field and also with the advancement of sequencing technology. Larger datasets and varying read lengths put different requirements on the analysis software e.g. with the output of a Sanger sequencing run, it was possible to BLAST all reads against a database such as NCBI *nr* while the same procedure is impractical with hundreds of millions of short reads delivered by one run on an Illumina sequencing machine. This illustrates the faster development in sequencing technology compared with computational power that has been observed recently.

2.5.1. Taxonomic characterization of metagenomic reads

Obtaining a taxonomic profile of a whole metagenome is commonly one main objective in a bioinformatics analysis of a metagenomic dataset. This is done by classifying each read and thereafter calculateing the relative abundance of a taxonomic unit. Available tools rely on sequenced genomes of microbial species and the available genomes were

traditionally biased towards model organism and pathogenic species. Efforts to fill gaps in the taxonomic tree by the Human Microbiome Project have provided whole genome sequences for 178 strains associated with the human body (Nelson et al., 2010) and the project have delivered more strains after the publication. Taxonomic classification binning methods work by training a classifier algorithm on known reference genomes that is used for characterization of metagenomic reads. PhylophytiaS is a tool that uses the k-mer frequencies in a sequence as input to a support vector machine for taxonomic classification (Patil et al., 2011). In a similar approach, Phymm uses interpolated Markov models trained on reference genomes to classify short metagenomic reads taxonomically (Brady and Salzberg, 2009). These and other binning methods do not rely on alignment and perform reasonably well when there are no sequenced representatives in the reference database but alignment methods work well when there is at least a genome from the same genus known.

Alignment based approaches are common in classifying metagenomic reads and have successfully been used in large scale projects of the human gut microbiota (Huttenhower et al., 2012b; Qin et al., 2010). Parsing a BLAST search of metagenomic reads to a database such as NCBI *nr* can be performed by the software MEGAN and the reads are then annotated to NCBI taxonomies to the lowest common ancestor (Huson et al., 2007). Although a BLAST search to NCBI *nr* is a sensitive method to find the origin of a metagenomic read, it does have a considerable computational cost. A BLAST search against sequenced microbial genomes can have an output in the order of 10 reads per second on a single CPU and a search against larger databases slows down with increasing size of the database. Speeding up the alignment is therefore essential and could be done by either reducing the size of the database or using accelerated alignment algorithms. By identifying clade specific marker genes and including only those in a reference database, the tool Metaphlan provides a speedup compared to alignment to a full database of microbial genomes (Segata et al., 2012). Accelerated alignment tools such as Bowtie2 (Langmead and Salzberg, 2012)and SOAP2 (Li et al., 2009) can perform alignments several orders of magnitude faster than BLAST but with a loss of sensitivity.

Fast methods for the analysis of metagenomes are important and continued development is crucial to keep up with the decreasing costs and increased output from sequencing machines. As an example, to search 500 000 MetaHIT reads against the NCBI *nr* database with BLASTX had a cost of $151 (Angiuoli et al., 2011). Considering that the average sequencing depth of this study was 62 million reads, the cost per sample of performing the above alignment would be over $18 000, many times more than the cost of sequencing.

2.5.2. *De novo* assembly of metagenomic reads

By sequencing a metagenome at sufficient depth compared to its complexity, it is possible to assemble reads into longer contigs. Assemblies of metagenomic data is typically fragmented and complete genomes cannot be expected although near complete genomes have been reconstructed from the cow rumen (Hess et al., 2011).

Software for single genome assembly has successfully been used in assembly of metagenomic data. SOAPdenovo (Li et al., 2010) and velvet (Zerbino and Birney, 2008)

have been used for assembly of metagenomic data from the human gut (Huttenhower et al., 2012a; Qin et al., 2010) and the cow rumen (Hess et al., 2011). Modifications to single genome assemblers have been made to better handle the varying abundance of species in a metagenome, two example being MetaVelvet (Namiki et al., 2012) and Meta-IDBA (Peng et al., 2011). MOCAT is a metagenomic assembly pipeline that can preprocess, assemble by calling SOAPdenovo and revise assembly (Kultima et al., 2012). MetAMOS is an assembly pipeline that has support for 8 different assemblers and facilitates easy comparison on the performance of each of them (Treangen et al., 2013).

2.5.3. Functional annotation and metabolic reconstruction

Metagenomic sequencing, as opposed to most other profiling methods of a microbial community including 16s rRNA sequencing, can be used to study the genetic functional potential and not only the taxonomic profiles. Two main approaches for functional reconstruction and metabolic reconstruction of metagenomes are being used. The first one relies on directly characterizing the function of a sequenced metagenomic read by alignment to a catalogue of known genes with known functions. The second approach makes use of alignment of reads to assembled contigs or genes and infers function of reads by the annotated function of the genes. The former approach avoids the *de novo* assembly step and could potentially detect rare functions which have not been assembled. The latter approach benefits from longer and often complete gene sequences that can be more precisely annotated.

Examples of tools for direct functional annotation of metagenomic reads are the online service MG-RAST (Meyer et al., 2008) and the standalone tools MEGAN (Huson et al., 2007) and HUMaN (Abubucker et al., 2012). Large scale metagenomics projects have annotated the functional potential by first performing *de novo* assembly of reads and inferring abundance by alignment of reads to genes (Qin et al., 2010; Qin et al., 2012). Typically, metagenomic genes are compared to genes in functional databases such as NCBI (www.ncbi.nlm.nih.gov), KEGG (Kanehisa et al., 2004), COG (Tatusov et al., 2003) and the Carbohydrate-Active enZYmes Database (CAZy) (Cantarel et al., 2009). KEGG is a database with complete genomic information for thousands of microbial genomes and detailed annotations of their genes to functions with a special focus on metabolism. Specific functions are grouped into pathways and functional categories in a hierarchical manner in the KEGG database. The detailed annotation of metabolic genes to biochemical reactions and metabolites in KEGG makes this database especially useful for metabolic reconstructions and modeling. The CAZy database stores a detailed description about carbohydrate active enzymes and their genes which is important in the study of the human gut metagenome because undigested polysaccharides make up a major part of the energy and carbon source of the human gut microbiota. Carbohydrate-active enzymes degrade, modify or create glycosidic bonds that make up polysaccharides and these are grouped into hundreds of enzyme families. In summary, the metagenome content and the purpose of the study should guide the use of different tools and databases.

2.5.4. Statistical methods for differential abundance analysis

A taxonomical or functional profile of a metagenome is interesting and useful information on its own for describing a microbial community but often a comparison between communities to describe their commonalities and differences are of interest. The challenge is thus to differentiate true differences between groups as opposed to measurement error and biological noise.

Because of the non-normal distribution of metagenomic data that is often observed, non-parametric tests such as Kruskal-Wallis and Wilcoxon rank-sum tests are used for testing whether samples originate from the same distribution. The Student's t-test assumes that data from both groups is normally distributed which is typically not the case in metagenomics. The assumption in Kruskal-Wallis and Wilcoxon rank-sum tests is that data distribution from the tested groups has the same shape. Large metagenome studies of the human gut comparing lean and obese individuals as well as diabetic to non-diabetic individuals have used Wilcoxon rank-sum test which have provided sufficient power to call differences between groups (Le Chatelier et al., 2013; Qin et al., 2012). In smaller studies, Wilcoxon rank-sum test might not be powerful enough to detect true differences between groups. In such cases parametric tests are available and could be used if the distribution of data allows. Parametric tests are likely more appropriate for testing the differential abundance of functional classes which has lower variance than taxonomic features or individual genes (Turnbaugh et al., 2009). One such approach is to use a Poisson model with the possibility to correct for over-dispersion for statistical comparison of metagenomes (Kristiansson et al., 2009). LefSe is a statistical methods that rely on Wilcoxon rank-sum test with an additional step to estimate the effect size (Segata et al., 2011). METASTAT is another method that uses t-tests with sample permutation for detecting differentially abundant features in a metagenome (White et al., 2009). Using permutation techniques in statistical tests risk being computationally costly, and can be problematic if the number of tests to perform is large. To control the number of false positive results when multiple tests are performed, p-values are typically corrected to control the false discovery rate, e.g. with the method presented by Benjamini and Hochberg (Benjamini and Hochberg, 1995).

2.6. Metabolism by gut microbiota

Along the intestinal tract microorganisms contribute to the degradation and consumption of dietary components. Some dietary components such as polysaccharides can be degraded by microorganisms to a greater extent compared to the capabilities of human enzymes. Polysaccharides that are available for the large intestinal microbiota include resistant starch, non-starch polysaccharides plant fiber, unabsorbed sugar and host derived glycans such as mucins. The total amount of polysaccharides that is available to the microbiota is estimated to be in the order of 10-60 g per day (Rosenberg et al., 2013). Degradation of polysaccharides is a complex process that involves several different enzymes, and is often done stepwise with a variety of enzymes degrading different glycosidic bonds between different sugar monomers. Intermediates of degradation and sugar monomers could thus be made available to other species than primary degraders which could result in extensive cross feeding. *Bacteroides* species are

known to carry a large number of glycoside hydrolases in their genomes, *B. thetaiotaomicron* has 172 glycoside hydrolase genes which makes it well equipped for handling large number of different polysaccharides (Xu et al., 2003).

Degradation of undigested polysaccharides by the gut microbiota results in byproduct excretion which is typically in the form of short chain fatty acids (SCFAs) and gases such as carbon dioxide, hydrogen and methane (Figure 2). The main SCFAs are acetate, propionate and butyrate and their concentrations in feces have been estimated to 50.5±12.6, 13.6±5.2 and 14.1±7.6 mM (Schwiertz et al., 2010). SCFAs serve as an important substrate for human colonocytes and stimulate mucus production and cell proliferation. Butyrate is especially important and constitute 60-70% of the energy used by colonocytes (Topping and Clifton, 2001). Propionate absorbed from the gut lumen could be used for gluconeogenesis by the liver (Wolever et al., 1991) and levels of propionate in the venous blood is very low indicating that most is metabolized by the liver (Wolever et al., 1989). Acetate is detectable in venous blood and rectal infusions result in a fall in serum free fatty acids and a rise in total cholesterol and triglycerides (Wolever et al., 1989). Furthermore, SCFAs act as signaling molecules in the human host and regulate inflammation and host energy balance by signaling through the G-protein coupled receptors 43 (Maslowski et al., 2009) and 41 (Samuel et al., 2008), respectively. Taken together, this clearly shows the important metabolic cross talk and interdependence between the microbiota and the host.

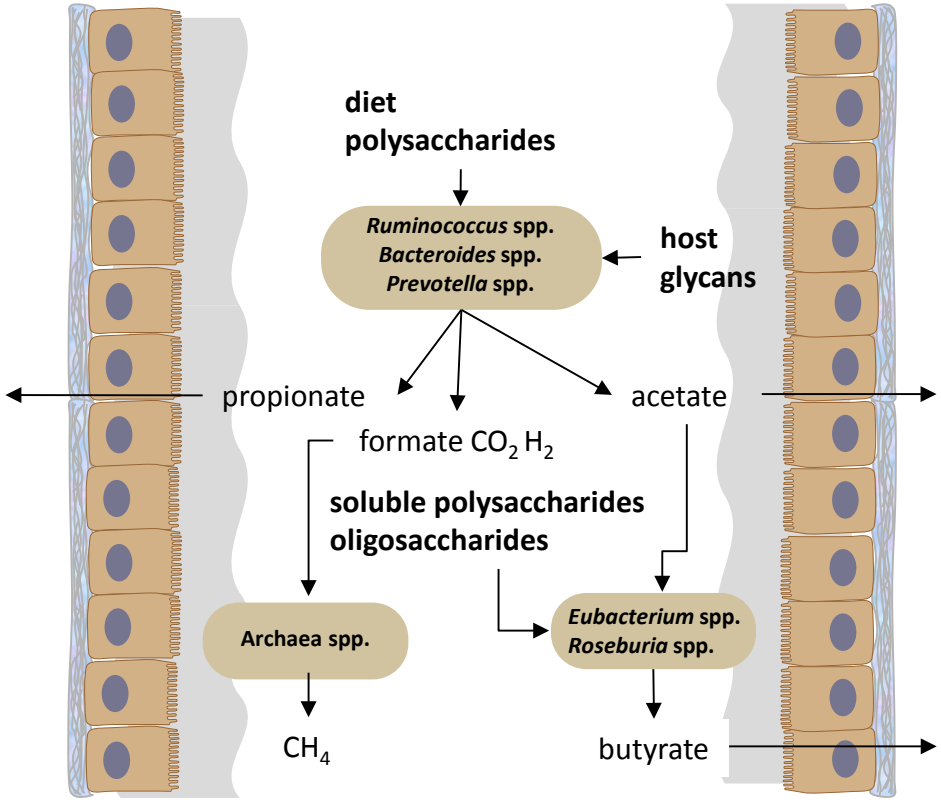

Figure 2 Overview of metabolism by the gut microbiota. The figure shows a very simplified and schematic overview of the main metabolic activity of degrading polysaccharides to short chain fatty acids and other end products.

2.7. The human gut microbiota and metabolic diseases

The gut microbiota and its host interact in a symbiotic relationship but when mutualistic or commensal bacteria are replaced or outcompeted by less favorable or pathogenic species, dysbiosis can occur. A growing amount of literature is showing that metabolic diseases and obesity are associated with changes in the composition of the gut microbiota (Table 1). Initial results were sometimes conflicting and a possible reason could be that the methods used mainly gave coarse taxonomic classifications or the complicated interplay between the diet, gut microbiota and host. Evidence in experimental animals suggests that a disturbed microbiota could cause weight gain and an adiposity associated metabolic profile, an initial study suggests the same causal relationship in humans as in experimental animals. The section will give an overview of known associations between metabolic diseases and the gut microbiota, suggested mechanisms and evidence that support the causal role of the microbiota in disease development.

Table 1 Studies of associations between gut microbiota and metabolic diseases and weight-loss intervention in humans

Disease/Intervention	Study
Obesity	(Ley et al., 2006) (Kalliomaki et al., 2008) (Duncan et al., 2008) (Turnbaugh et al., 2009) (Zhang et al., 2009) (Schwiertz et al., 2010) (Furet et al., 2010) (Zupancic et al., 2012) (Le Chatelier et al., 2013)
Type 2 diabetes	(Larsen et al., 2010) (Qin et al., 2012) (Karlsson et al., 2013)
Type 1 diabetes	(Brown et al., 2011)
Atherosclerosis/Cardiovascular disease	(Koren et al., 2011) (Karlsson et al., 2012)
Weight-loss interventions	(Ley et al., 2006) (Duncan et al., 2008) (Zhang et al., 2009) (Furet et al., 2010) (Kong et al., 2013) (Graessler et al., 2013) (Cotillard et al., 2013)

2.7.1. Obesity

Development of obesity is due to an excess of energy intake compared to energy expenditure. The energy balance is dependent on several environmental and genetic factors such as diet, exercise and regulation of physiological functions. Inheritability of obesity is 40-70% but even with very large genome wide association studies, the proportion of explained genetic variance of body mass index using 32 validated markers is only 1.45% (Speliotes et al., 2010). This suggests that other inheritable factors are important for the development of obesity. The gut microbiota plays an important role by partly processing the food we eat and regulates the immune system.

Alterations between components of the gut microbiota and obesity have been observed in several studies. The ratio between the two major phyla in the human gut, Bacteroidetes and Firmicutes were found to be associated with obesity with increased level of Firmicutes in the obese flora (Ley et al., 2006). The ratio was restored in individuals following a weight loss program. Later reports could not confirm this altered ratio between the two major phyla and found no difference (Duncan et al., 2008) or an opposite association (Schwiertz et al., 2010). It has also been suggested that the higher levels of SCFAs found in obese subjects are relevant for obesity (Schwiertz et al., 2010). The Bacteroidetes to Firmicutes ratio is a rough measure of the composition in the human

gut, the Firmicutes phyla contains some clearly pathogenic species such as *Clostridium botulinum* and *Listeria monocytogenes* as some that are generally regarded as beneficial to the host such as *Faecalibacterium prausnitzii* and *Eubacterium rectale*. This shows that the broad description of the obese gut microbiota is not enough and that more specific methods are needed. A reduced diversity of the microbiota has been observed in obese Danish and American individuals (Le Chatelier et al., 2013; Turnbaugh et al., 2009).

To elucidate whether alterations is the gut microbiota are causing obesity and underlying mechanisms, intervention studies and work using experimental animals are crucial. Transplantations of gut microbiota of lean and obese mice to germ-free recipients have shown that the obesity phenotype is transferable by the microbiota. Flora from genetically obese (*ob/ob*) and diet induced obese mice has the potential to cause obesity in recipients (Turnbaugh et al., 2008; Turnbaugh et al., 2006). Interestingly, there seems to be a similar causal relationship in humans. Transfer of intestinal microbiota from lean donors to recipients with the metabolic syndrome resulted in improved glucose metabolism and insulin sensitivity together with increased levels of butyrate producing bacteria (Vrieze et al., 2012). Transfer of whole microbial fractions is controversial because the potential risk of transferring pathogenic organisms and isolated cultured fractions of beneficial microbes that has the same improvements to health is desirable.

Several mechanisms for the influence of the gut microbiota on obesity have been proposed. Increased energy harvest by breakdown of otherwise indigestible carbohydrates to short chain fatty acids have been proposed to be contributing to increased energy intake (Turnbaugh et al., 2006). The gut microbiota interplays with the signaling and regulatory network of the host and thereby regulates the energy balance. It has been shown that the gut microbiota promotes monosaccharide absorption and suppresses the fasting-induced adipocyte factor (fiaf) in intestinal tissue. Fiaf is an inhibitor of lipoprotein lipase and increased lipase activity results in increased storage of fat in adipocytes (Backhed et al., 2004). SCFAs play a signaling role by acting on the G-protein coupled receptor 41 (Gpr41) and Gpr41-/- mice are leaner than their wild type littermates but this effect is not evident in germ-free conditions. Gpr41-/- mice have lower expression of PYY, a gut derived hormone acting to slow down gastrointestinal transit. Knockout of Gpr41 results in reduced levels of PYY and increased transit rates resulting in more energy being excreted with feces (Samuel et al., 2008). Given the background above, it is clear that it is not only one single mechanism that could explain how the gut microbiota could increase adiposity and the important species that could play a role are yet not identified.

2.7.2. Type 2 diabetes

Obesity is a major risk factor for Type 2 diabetes (T2D) and the two are closely associated. The associations and mechanisms for the relation between obesity and the gut microbiota are relevant also for T2D but it is also important to investigate the specific associations and mechanisms that might trigger the onset of T2D. T2D is characterized by insulin resistance and sometimes reduced insulin production, resulting in poor cellular uptake of glucose and elevated levels of blood glucose. T2D is the most

common form of diabetes and around 350 million people are presently affected (Danaei et al., 2011).

Using qPCR and sequencing of the V4 region of the 16S rRNA gene to study the gut microbiota in 36 male adults, compositional changes in the gut microbiota was found to be associated with diabetes. *Clostridia* were significantly reduced in diabetic subjects while Betaproteobacteria were enriched. Furthermore, the *Bacteroides-Prevotella* group to *C. coccoides - E. rectale* ratio and *Lactobacillus* correlated positively to plasma glucose after an oral glucose tolerance test (Larsen et al., 2010). A larger study of the gut metagenome in 345 Chinese individuals found that 60 000 genes were associated with T2D and found that butyrate producing bacteria were depleted in T2D individuals (Qin et al., 2012).

T2D is associated with low-grade inflammation, for example increased levels of pro-inflammatory cytokines. The increased levels of cytokines are deleterious for insulin sensitivity. Lipopolysaccharides, a membrane component of Gram-negative bacteria are triggers of inflammation and are elevated in mice on a high fat diet. Feeding of high fat diet resulted in reduced levels of *Bifidobacteria* and *C. coccoides-E. rectale (Cani et al., 2007)*. In a study of mice lacking a receptor for bacterial flagellin, TLR5, alterations in the gut microbiota was observed as well as increased adiposity, low-grade inflammation and insulin resistance (Vijay-Kumar et al., 2010). When the microbiota of mice lacking the TLR5 receptor was transplanted into germ-free mice, recipients had worse glucose metabolism and higher levels of inflammation compared to recipients of a wild-type gut microbiota.

2.7.3. Atherosclerosis and cardiovascular disease

Cardiovascular disease (CVD), with manifestations such as heart attack and stroke, is the most common cause of death representing about 30% of deaths worldwide. Diabetes and obesity are major risk factors for cardiovascular disease. Buildup of plaques in the arterial wall by cholesterol and macrophages could eventually rupture and clog the blood flow downstream resulting in stroke or heart attack. The plaques contain bacterial DNA from the genera *Chryseomonas*, *Veillonella*, and *Streptococcus* which are also present in oral and gut samples (Koren et al., 2011).

Metabolomics studies in humans have identified trimethylamine (TMA) and trimethylamine N-oxide (TMAO) as risk factors for development of CVD. Mechanistic investigations have suggested that microbial metabolism of phosphatidylcholine produces TMA which is absorbed and converted to TMAO in the liver. Supplementation with choline resulted in more plague formation but suppression of the gut microbiota by treatment with antibiotics alleviated the symptoms (Wang et al., 2011). TMA and TMAO were also found to being produced by dietary L-carnitine by the gut microbiota. Vegans and vegetarians were found to be producing less TMA and TMAO from a supplementation of L-carnitine compared to omnivorous individuals (Koeth et al., 2013).

Microbiota metabolism of bile acids by the gut microbiota is of special interest in the context of atherosclerosis. Bile acids are synthesized from cholesterol in the liver, released in the duodenum, serve as detergents that solubilize dietary lipids and are

actively reabsorbed in the distal ileum. Microbial metabolism of bile acids in the small intestine and colon by deconjugation and dehydroxylation produces secondary bile acids. The secondary bile acid lithocolic acid is lost in feces which results in a net loss in the enterohepatic circulation of bile acids, thus draining cholesterol (Ridlon et al., 2006). The role of microbial metabolism of bile acids in atherosclerosis is not clear. Attempts have been made to use probiotic bacteria to lower serum cholesterol with mixed effects in humans (Ooi and Liong, 2010).

2.7.4. Weight-loss interventions

Interventions to reduce weight and improve metabolic profile such as improving insulin sensitivity and lowering cholesterol include dietary interventions, exercise, medication and bariatric surgery. Effects on the microbiota of the above interventions have been shown but it is sometimes difficult to separate the effect of the intervention itself and the weight-loss. The relative abundance of Bacteroidetes was increased in individuals who were assigned a fat restricted or carbohydrate restricted weight loss diet followed over 1 year (Ley et al., 2005). The total sequencing depth of the study was just over 18 000 16S rRNA sequences which is low by today's standards. In a study of 23 individuals undergoing a weight loss regimen did not find any differences in the abundance of *Bacteroides* using quantitative fluorescent *in situ* hybridization (FISH) to study the composition of fecal microbiota (Duncan et al., 2008). From these two studies, conclusions are not coherent possibly because diets, study design and methods for assessing the microbiota differed. In a study of 49 individuals using deep metagenomic sequencing to assess the microbiota, it was found that dietary intervention by an energy restricted high protein diet resulted in higher diversity in individuals who initially had a low diversity (Cotillard et al., 2013).

Bariatric surgery is an efficient method to reduce weight in severely obese individuals and reduce the risk of diabetes and cardiovascular disease (Sjostrom et al., 2004; Sjostrom et al., 2007). Several different bariatric surgical procedures exist that restrict the size of the stomach or gastric bypass by rerouting the stomach and the small intestine. The microbiota is altered after bariatric surgery by increased levels of Proteobacteria, as was shown in a study comparing lean, obese and post-gastric-bypass surgery individuals (Zhang et al., 2009). The increase in Proteobacteria and in particular *E. coli* was observed in a study following changes in the microbiota before and after gastric bypass surgery. Lactic acid bacteria such as *Lactobacillus* decreased after surgery (Furet et al., 2010). Yet another more recent study also found an expansion of Proteobacteria after gastric bypass and a decrease in *Lactobacillus*, *Dorea* and *Bifidobacterium*, overall the diversity in the microbiota increased (Kong et al., 2013). Overall, the changes in microbiota after gastric bypass is more clear and coherent between studies compared to diet interventions. This is likely due to the fact that bariatric surgery is a more drastic intervention compared to a diet intervention.

2.8. Systems biology and metabolic modeling

The term systems biology relates to a field that uses mathematical models and networks to study complex biological systems containing several interacting components. Systems

biology has the power to identify emergent properties from simple interacting components that a reductionist approach might not reveal. Data may be generated in a high-throughput way e.g. genomics, transcriptomics, proteomics or metabolomics. By use of networks and mathematical models that considers the interactions between the components, a better understanding can be obtained compared to looking at the components in isolation. An interesting result from systems biology analysis could be emergent properties that arise from simple interactions of several components and could not be seen or predicted by analysis of individual components.

Metabolism plays a particular important role in the interaction between the human host and its gut microbiota. As have been described in previous chapters, metabolism of dietary components by the gut microbiota can be extensive and microbially produced metabolites are readily found in human blood (Li et al., 2008). Metabolism is responsible for providing the building block for microbial biomass and the free energy needed to maintain life. Genome scale metabolic models (GEMs) are collections of metabolic genes and their stoichiometric reactions of an organism and constitute a powerful tool for addressing metabolic questions. The first organism to be reconstructed was *Haemophilus influenzae* in the year 2000 (Schilling and Palsson, 2000). Since then, a large number of GEMs have been reconstructed for model organisms, medically and industrially relevant species.

Reconstruction of a GEM for an organism starts with collection of gene-protein-reaction associations and is typically based on experimental or genomic inferences. Several bioinformatics tools are available that automatize many steps of the reconstruction (Agren et al., 2013; Henry et al., 2010). Biochemical reactions are defined in a matrix S with the stoichiometric coefficients, rows correspond to metabolites and columns correspond to reactions. Genomic and biochemical reaction databases such as KEGG (Kanehisa et al., 2004) are very useful for automatic reconstruction (Figure 3). A number of manual steps are necessary, such as definition of biomass components, gap filling and fitting of parameters for growth rate.

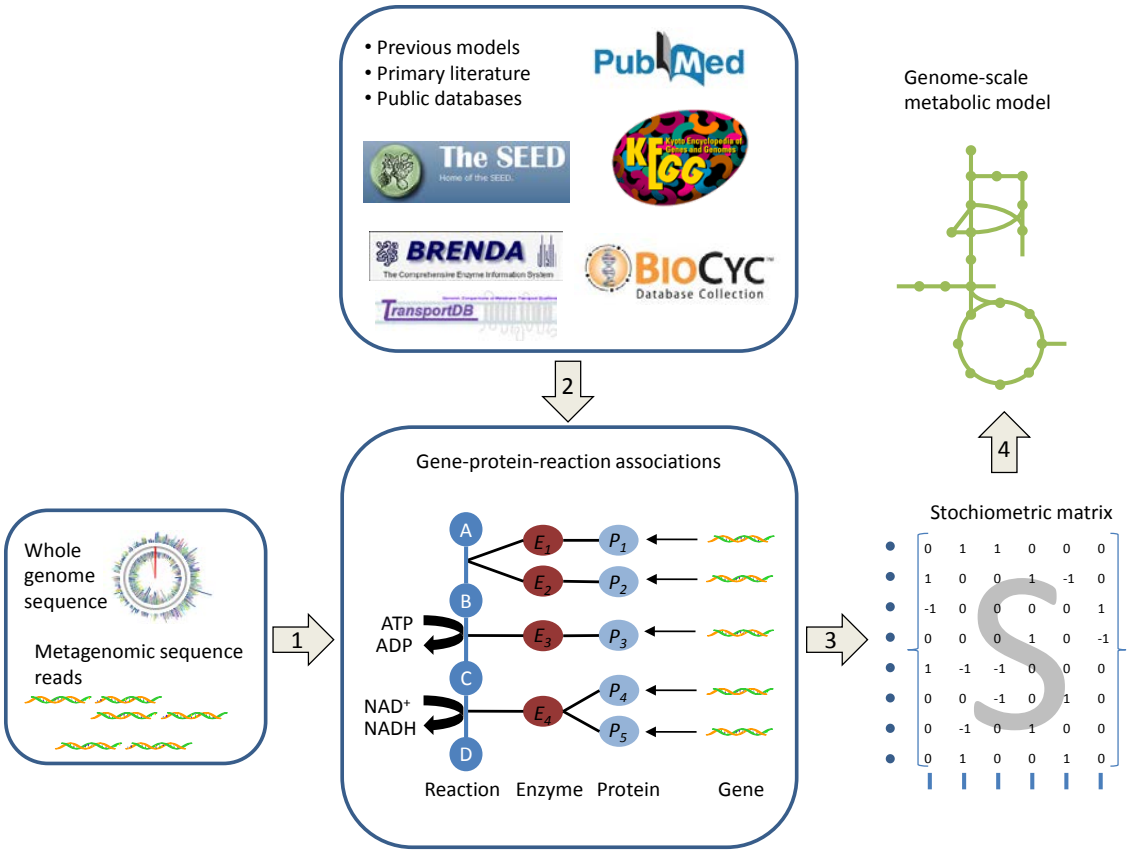

Figure 3 Reconstruction of a genome-scale metabolic model. 1) Reconstruction starts with a genome sequence of the organism. 2) Information from the literature and public databases about the biochemical conversion performed by genes of the organism is collected. 3) Reaction converting metabolites are assembled into a stoichiometric matrix. 4) Gene-protein-reaction associations together with the exact stoichiometric description of reactions and metabolites are combined into what is called a genome scale metabolic model.

The reactions connect metabolites into a metabolic network and constitute a framework for mapping high-throughput data onto. One important example is to use the link between metabolites, reactions and genes in a concept called reporter features. Transcriptomic data can be mapped onto the metabolic network and reveal reporter metabolites around which there are extensive transcriptional changes (Oliveira et al., 2008; Patil and Nielsen, 2005). However, GEMs are not merely gene-metabolite mappings, they are detailed collections of biochemical reactions that have undergone manual curation and gap filling to constitute a functional metabolic network with complete pathways from substrate to biomass components. Reactions are checked for mass/charge balance, thermodynamic feasibility and gaps, dead ends and blocked reactions are resolved. Biomass composition is determined by experimental measurements or literature and a reaction for biomass formation is added. Flux balance analysis can be used to simulate fluxes of an organism operating at steady state that fulfill maximization of an objective function under given constraints. This can be formulated mathematically:

$$\max \quad \boldsymbol{c^T \cdot v}$$

Subject to:

$$\boldsymbol{S \cdot v = 0}$$

$$\textit{lower bounds} \leq \boldsymbol{v} \leq \textit{upper bounds}$$

Where c is a vector with a coefficient for each reaction that specifies a linear combination of fluxes to be maximized and v is a vector with the rate of each reaction, S is the stoichiometric matrix that defines the metabolic network. The fluxes are constrained by bounds that limit the solution space. The origin of the constraints could be thermodynamics, compartmentalization, diffusion, enzyme capacity or experimental observations and constitute limitations for the system.

3. Results and discussion

This section summarizes the publications that are the basis of this thesis. The results can be sectioned into three parts. In the first part (section 3.1), the results from three metagenomic studies are presented where the association between the gut metagenome and metabolic diseases and bariatric surgery are studied. The second part (section 3.2) presents two bioinformatics tools for metagenomic data analysis that have been developed alongside with the data analysis of gut metagenomes. The third part presents (section 3.3) a systems biology and modeling approach to study the gut microbiota. Metabolic models of three important species in the human gut are described and validated.

3.1. Association of the human gut metagenome with metabolic diseases

In this section, results from three studies (**Paper I-III**) of the gut metagenome association with symptomatic atherosclerosis, diabetes and bariatric surgery are presented separately. **Paper IV** is a review of recent results in the field of gut metagenome and metabolic diseases. Common lessons from the three metagenome studies and recent results from the literature are compared in section 3.1.4.

3.1.1. Paper I: Associations between the gut metagenome and symptomatic atherosclerosis

Cardiovascular disease, with its manifestations myocardial infarction and stroke, is caused by accumulation of cholesterol and macrophages to the arterial wall that eventually ruptures and restricts the blood flow to the heart and brain, respectively. The gut microbiota has been implicated as an environmental factor that modulates host lipid metabolism (Backhed et al., 2004; Backhed et al., 2005; Cani et al., 2007; Ley et al., 2006). The gut microbiota can be a source of inflammatory molecules such as lipopolysaccharides and peptidoglycan that can contribute to metabolic disease (Cani et al., 2007; Erridge et al., 2007; Schertzer et al., 2011). To address the question whether the gut metagenome is associated with cardiovascular disease, we sequenced the gut metagenome of patients (n=12) who had manifestation of emboli to the brain or retinal artery with severely stenotic plaques in the carotid artery. As a control group (n=13), gender and age matched controls without large and potentially vulnerable plaques in the carotid artery were recruited.

DNA from fecal samples was extracted by a previously published method (Salonen et al., 2010). The isolated metagenomic DNA was sequenced using the Illumina HiSeq2000 instrument and 100 bp paired end reads were generated. On average, 12.5 ±4.7 (SD) million reads were generated per sample. Low quality and contaminant reads were removed.

Taxonomic characterization was done by aligning the metagenomic reads to a catalogue of 2382 sequenced prokaryotic genomes. Most reads aligning to the genomes were bacterial (98±4% (s.d.)) and the dominating phyla were Bacteroidetes and Firmicutes (see Figure 1 for an overview of the main phyla and genera). The genus *Collinsella* was

enriched in patients whereas *Eubacterium* and *Roseburia* and three species of Bacteroides were enriched in control subjects (adj. P<0.05, Wilcoxon rank-sum test; Figure 4). *Eubacterium* and *Roseburia* species are known butyrate producers and acetate utilizers (Duncan et al., 2002; Mahowald et al., 2009) and the importance of butyrate producing bacteria and their decrease in inflammatory bowel disease have been reviewed by Lois and Flint (Louis et al., 2010).

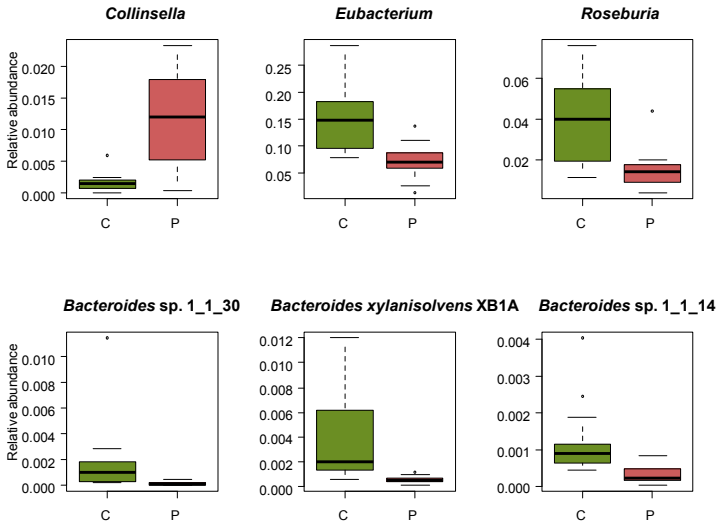

Figure 4 Microbial composition associated with symptomatic atherosclerosis. Abundance of bacterial genera (top) and species (bottom) that differ between patients (P) and controls (C), Adj. P <0.05 for all comparisons. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

Several *Clostridiales* genera correlated negatively with the inflammatory marker high-sensitivity C-reactive protein (hsCRP) (Figure 5). At the species level, *Clostridium* sp. SS2/1 and SSC/2 negatively correlated (Spearman's correlation, adj. P<0.05) with hsCRP and these are both characterized as butyrate producing bacteria.

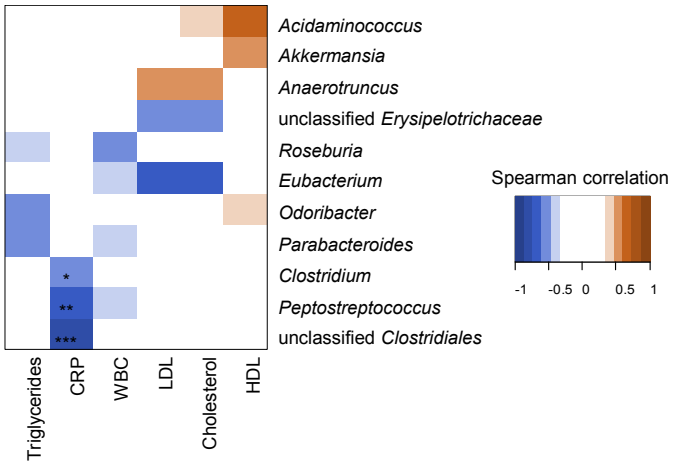

Figure 5 Genera correlating with clinical biomarkers. Spearman's correlation was calculated between abundance of genera and clinical biomarkers and the strength of correlation is indicated by color. *Adj. P<0.05, **adj. P<0.01 and ***adj. P<0.001

It has been suggested that the human gut microbiota variation can be stratified into three enterotypes and that the variation is not continuous (Arumugam et al., 2011). Using the same methods as presented in the publication, we could also identify enterotypes in our cohort (Figure 6). Recently, there has been a debate whether there are distinct enterotypes and if these are stratified or continuous (Koren et al., 2013). Furthermore, the cutoff that should be used for identification of discrete clusters has also been debated and whether the clusters are universal to all regions or restricted certain geographies (Yatsunenko et al., 2012). In this cohort, we find enterotypes driven by *Bacteroides*, *Prevotella* and a third enterotype where the driver genus is less clearly defined but in which *Ruminococcus* is enriched. Here we found that patients were overrepresented in the *Ruminococcus* enterotype and controls were overrepresented in the *Bacteroides* enterotype.

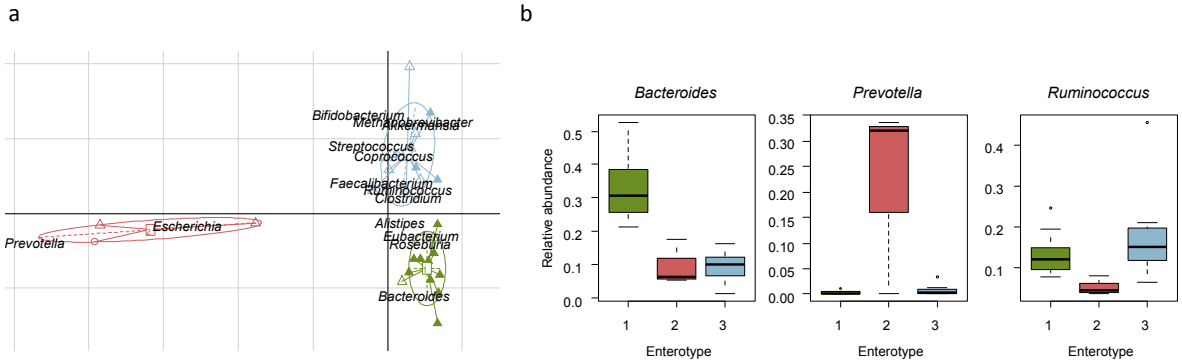

Figure 6 Enterotypes of the gut microbiota.a) Based on the abundance of genera in the cohort using the clustering method presented in (Arumugam et al., 2011), three enterotypes could be identified. Controls and patients are denoted by filled triangles and empty triangles, respectively and two subjects not included in the comparison are represented by empty circles. b) Abundance of three genera suggested being drivers of the enterotypes. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

We performed *de novo* assembly of the sequence data, first for each individual sample separately and subsequently for a pool of all the non-assembled data from the individual samples to create one global gene catalog of our cohort. Genes were predicted from the contig set and these genes were functionally annotated to KEGG. Reads were aligned to the contigs and their position was recorded to estimate gene abundance. When a comparison was made of gene abundance between patients and controls, a total of 225 KOs were differentially abundant (adj. P<0.05, Wilcoxon rank-sum test). By using the reporter feature algorithm we could identify KEGG pathways associated with symptomatic atherosclerosis and the highest scoring was the peptidoglycan biosynthesis pathway. Peptidoglycan is known to activate the immune system through the nucleotide oligomerization domain proteins and activation has been linked to metabolic disease (Schertzer et al., 2011) and inflammation is known to contribute to atherosclerotic disease (Hansson, 2005). Furthermore, we found metabolic genes that had negative correlation with inflammation; the highest scoring association being butyrate-acetoacetate CoA-transferase (K01036) with hsCRP (Spearman's r=0.73, adj. P=0.04). This finding is in agreement with the taxonomic analysis above which also identified known butyrate producers negatively correlated with hsCRP. Butyrate has been identified as a negative regulator of inflammation through G-protein coupled receptor 43 (Maslowski et al., 2009). The most significantly enriched function in controls was the phytoene dehydrogenase (K10027), involved in the metabolism of lipid-soluble antioxidants such as the carotenoids lycopene and ß-carotene (Figure 7). We evaluated whether controls also had increased levels of carotenoids, and found increased levels of ß-carotene (P=0.05, Student's t-test), but not lycopene, in serum of healthy controls compared with patients.

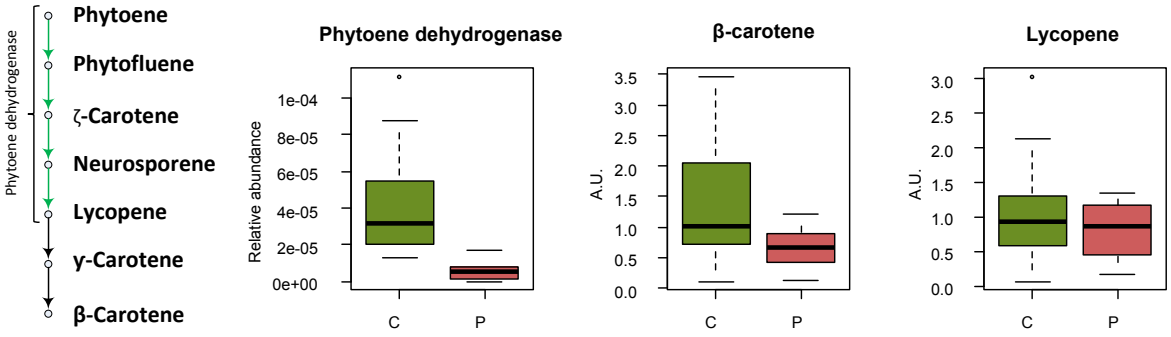

Figure 7 Phytoene dehydrogenase genes are enriched in the metagenome of healthy controls. ß-carotene (P=0.05 Student's t-test) is enriched in the serum of healthy controls but not lycopene. Boxes denote the interquartile range (IQR) between the first and third quartiles and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

High levels of ß-carotene and lycopene are associated with a reduced risk of cardiovascular disease (Kardinaal et al., 1993; Kohlmeier et al., 1997) but supplementation of these compounds have not proven to be protective (Hennekens et al., 1996; Kritchevsky, 1999). On the other hand, a study of over 500 individuals failed to observe an association between lycopene intake and plasma lycopene levels (Bermudez et al., 2005) indicating that other mechanisms might be more important in determining

plasma levels than oral intake of lycopene. Furthermore, bacterial species from the human gut have been shown to produce carotenoids (Khaneja et al., 2010; Perez-Fons et al., 2011). These findings represent an important step towards elucidating the role of carotenoids in atherosclerotic by highlighting the potential role that the microbiota could play in producing carotenoids.

3.1.2. Paper II: Gut metagenome in women with normal, impaired and diabetic glucose control

Diabetes is a disease where blood level glucose is elevated and in the case of T2D, this is a result of insulin resistance and in some cases failure of the pancreas to produce insulin. Studies have shown that genetic markers are poor predictor of future diabetes whereas environmental factors and socio-economic factors have greater influence (Noble et al., 2011). In this study, the fecal metagenome of 145 European women with T2D (n=53), impaired glucose tolerance (IGT; n=49) or normal glucose tolerance (NGT; n=43) glucose control were studied. From these samples, 453 Gbp of sequence was generated or 3.1±1.8 Gbp per sample by the Illumina Hiseq 2000 instrument. Data was analyzed by comparing to a set of reference genomes and also *do novo* assembled and functional analysis. A set of 2382 reference genomes were collected from NCBI and HMP (http://www.hmpdacc.org) and metagenomic reads were aligned to these genomes using Bowtie (Langmead et al., 2009). The relative abundance of each genome was calculated and increases in four *Lactobacillus* species while decrease in five *Clostridium* species was observed in T2D subjects compared to NGT (adjusted $P<0.05$, Wilcoxon rank-sum test). The abundance of *Lactobacillus* species correlated positively with fasting glucose and HbA1c in the total cohort (adjusted $P<0.05$, Spearman correlation). The abundance of *Clostridium* species correlated negatively with fasting glucose, HbA1c, insulin, C-peptide and plasma triglycerides and positively with adiponectin and HDL (adjusted $P<0.05$, Spearman correlation).

To fully make use of the metagenomic data, *de novo* assembly was performed on individual gut metagenomes and then a global assembly on unassembled reads. Genes were predicted on the assembly and a non-redundant gene catalogue was constructed which was then merged with the MetaHIT gene catalogue (Qin et al., 2010). Reads were aligned to the combined gene catalogue to acquire individual quantitative measures of the gut metagenomes. Microbial genes come in sets of genomes and genes from the same genome should follow the same abundance pattern across individual samples. Using this assumption, we clustered genes based on their profile across samples. We considered only genes present in at least 10 subjects and calculated the correlation coefficient between genes across subjects (Figure 8).

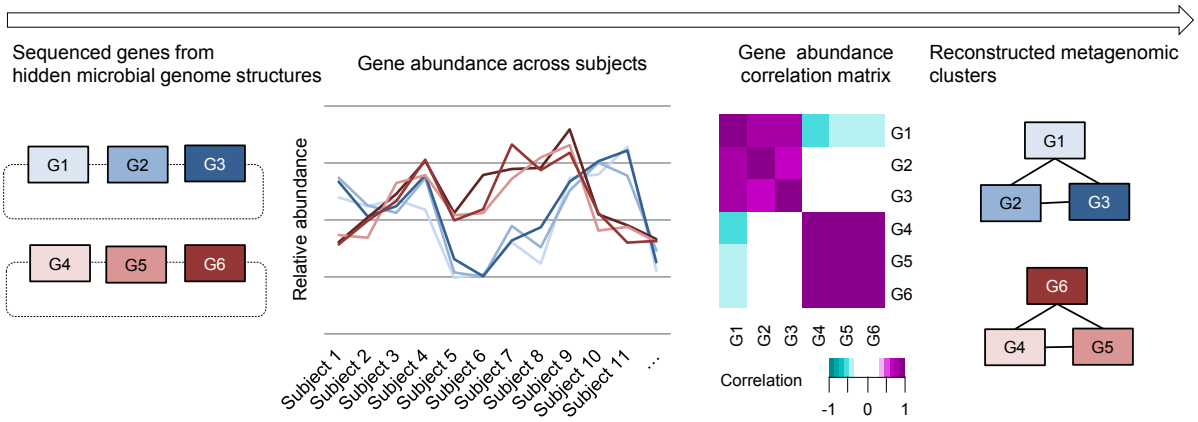

Figure 8 Reconstruction of metagenomic clusters. Genes in the same genome should have a similar abundance across samples and were clustered based on their co-occurrence.

Highly correlated sets were defined as metagenomic clusters (MGCs) and the 800 largest clusters contained at least 104 genes or 550,188 genes in total (Figure 9). The phylogenetic origin of the MGCs was determined by alignment of genes against the NCBI *nr* with blastp and determination of the lowest common ancestor (LCA) by requiring that 50% of the genes had a best hit to the same phylogenetic group. Results show that 36% of the MGCs could be determined at the species level (Figure 9). MGCs with a LCA at the order level were mainly from the Clostridiales (98%) and these are known to be taxonomically difficult to define (Arumugam et al., 2011).

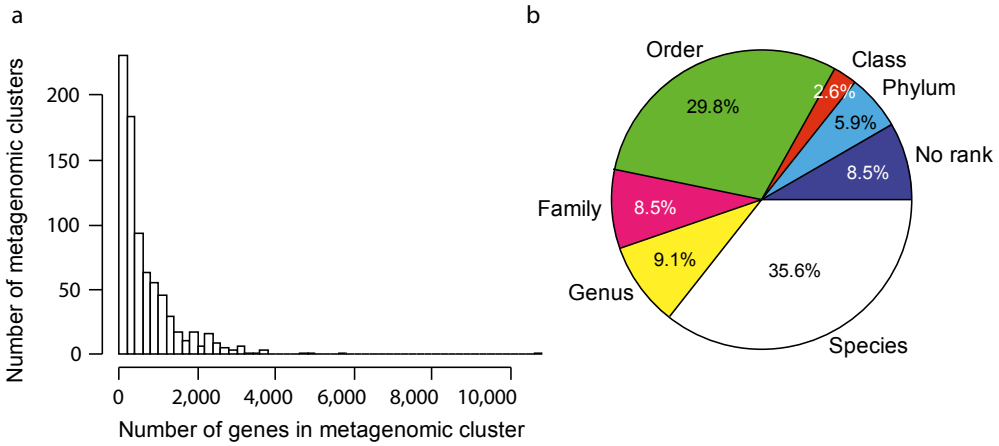

Figure 9 Characterization of the 800 largest metagenomic clusters (MGCs) (>104 genes). a) Histogram of the number of genes in each MGC. b) Taxonomic classification detail of MGCs.

The abundance of a MGC was calculated by summing the abundance of the member genes. The 800 largest MGCs were compared between the NGT and T2D group and 26 were found to be differentially abundant between the two groups (adjusted P <0.05, Wilcoxon rank-sum test). The MGCs enriched in T2D subjects and identified at the species level were *Lactobacillus gasseri, Clostridium clostridioforme* and *Streptococcus*

mutans. *L. gasseri* correlated positively with fasting glucose and HbA1c, whereas *C. clostridioforme* correlated positively with triglyceride and C-peptide levels. The MGCs enriched in the NGT group were from *Roseburia*, *Clostridium*, two *Eubacterium eligens*, several Clostridiales, Coriobacteriaceae and one *Bacteroides intestinalis.*

Using gut metagenomic data as an identifier for diabetes status was tested by training a random forest model to classify NGT and T2D individuals based on their species and MGC abundance profiles. The performance was evaluated with tenfold cross-validation and the performance assessed using a receiver operating characteristic (ROC) analysis. The discriminatory power of the models were calculated as the area under the ROC curve (AUC), T2D was identified more accurately using MGCs (highest AUC=0.83) than with microbial species (highest AUC=0.71) (Figure 10). The most important predictors in the models were very similar to the differentially abundant species and MGCs and are shown in Figure 10, the most important predictor in both models was *L. gasseri.*

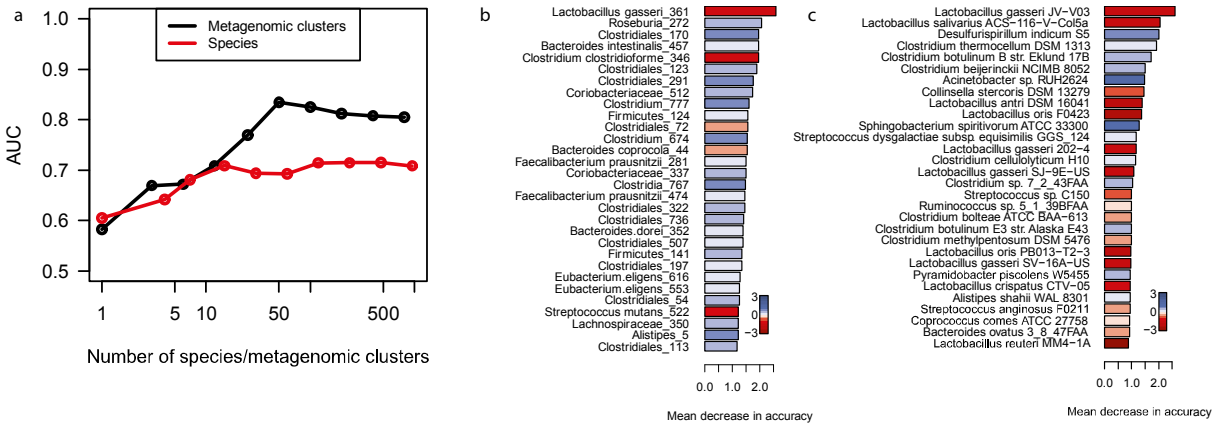

Figure 10 Classification of T2D by species and MGC abundance. a) Performance of random forest models on discriminating NGT and T2D subjects assessed by area under the receiver-operating characteristic curve (AUC). Performance was evaluated with an increasing number of species/MGCs. b) The 30 most important MGCs discriminating between NGT and T2D. c) The 30 most important species discriminating between NGT and T2D. The bar lengths in b and c indicate the importance of the MGC/species, and colors represent enrichment in NGT (blue shades) or T2D (red shades).

We then used the model for discriminating diabetic status on the IGT subjects (n=49) and 34 of these were classified as T2D, 10 as NGT and 5 were uncertain based on their gut metagenome (Figure 11). Plasma levels of triglycerides and C-peptide were significantly higher in the subgroup classified as T2D compared to NGT. This suggests that the model identifies subjects in the IGT group with altered metabolism associated with risk factors for diabetes.

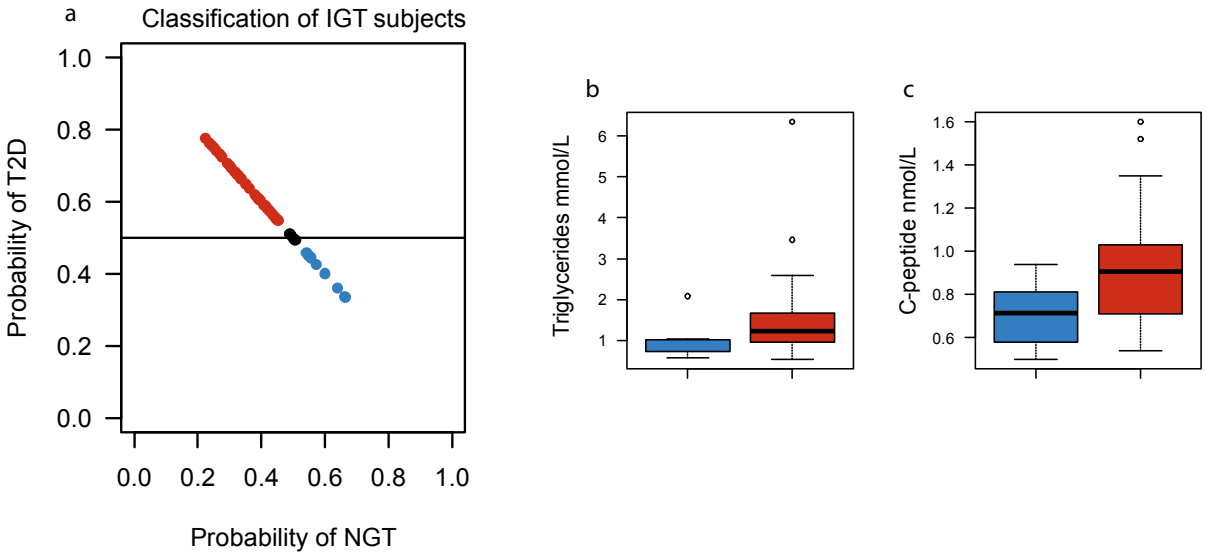

Figure 11 Stratification of IGT women using MGCs. a) Prediction results of classification of IGT subjects as NGT (blue) and T2D (red). b) IGT women predicted to be T2D had higher triglyceride levels (P=0.019 Wilcoxon rank-sum test). c) IGT women predicted to be T2D had higher C-peptide levels (P=0.030 Wilcoxon-rank sum test).

Analysis of the functional composition of the gut metagenomes was performed by annotating the genes to the KEGG database. KEGG ortholog abundance was calculated and compared across groups. Pathway annotations of KEGG orthologs and results from the differential abundance analysis were used in the reporter algorithm (Oliveira et al., 2008) to identify reporter pathways. The pathways that were highest scored for enrichment in T2D metagenomes were starch and glucose metabolism, fructose and mannose metabolism and ABC transporters. The pathways that were highest scored for enrichment in NGT metagenomes were flagellar assembly and riboflavin metabolism. This suggests that the gut metagenome of T2D individuals is enriched in genes for simple sugar degradation while the metagenome of NGT individuals is enriched in vitamin production. Similar results were also observed in the Chinese T2D cohort (Qin et al., 2012).

The above results were compared to a study of the gut metagenome in Chinese T2D and NGT individuals. In both studies, *Clostridium clostridioforme* MGCs were increased whereas *Roseburia* was decreased in T2D metagenomes. The Chinese T2D subjects had increased levels of *Akkermansia muciniphila, Clostridium ramosum* and depleted in *Roseburia intestinalis, Faecalibacterium prausnitzii, Eubacterium* and *Erysipelotrichaceae* which agrees with the published results (Qin et al., 2012). By using the procedure described above to train and evaluate a predictive model for diabetic status, an AUC of 0.82 was obtained for prediction within the Chinese population. However, cross comparison between populations yielded low AUC values. The most important predictor species and MGCs differed and also their relative abundance in the two cohorts.

3.1.3. Paper III: Long-term effects of bariatric surgery on the gut metagenome

Bariatric surgery is the most effective treatment for severe obesity and its comorbidities while dietary interventions have limited efficacy for treating severe obesity. Several types of bariatric surgery exists and Roux-en-Y gastric bypass (RYGB) is the most widely used and have been shown to result in greater weight loss compared to Vertical banded gastroplasty (VBG), the weight-loss averages are 25% and 16% after a follow-up of 10 years, respectively (Sjostrom et al., 2007). RYGB is also associated with improved response to a meal with higher levels of satiety hormones glucagon-like peptide 1 and peptide YY (Werling et al., 2013). The mechanisms that mediate weight loss after bariatric surgery are not fully understood. Reduced food intake, gastric emptying, bile acid metabolism and gut hormones have been suggested to contribute to weight loss. Alteration in the composition of the gut microbiota is a plausible contributor that needs further investigation.

The gut microbiota is altered 3 months to about a year after RYGB surgery with increased levels of Proteobacteria, especially *E. coli* (Furet et al., 2010; Kong et al., 2013; Zhang et al., 2009). The gut microbiota was studied with quantitative PCR or pyrosequencing of the 16S rRNA gene which gives information about the taxonomic composition. To gain understanding of also the functional composition, metagenomic sequencing is required. A study of the gut metagenome 3 months after bariatric surgery in 6 individuals showed an increase in genes assigned to the phosphotransferase system. In the study performed here, long term effects, with a follow up of more than 9 years, of bariatric surgery on the gut microbiota are investigated.

The aim of this study was to sequence the gut metagenomes in patients who have undergone RYGB (n=7) and VBG (n=7) and compare these to those of severely obese (OBS, n=7, BMI=44.9±4.7 (SD)) and overweight or obese (CTR, n=9, BMI=31.9±2.7 (SD)) individuals. CTR individuals were included from the previous study described in **Paper II**, to control for differences in BMI and age but individuals in this group had a normal glucose control. Gut microbial DNA from fecal samples was sequenced with the Illumina Hiseq2000 instrument and in total 63 Gbp of paired-end reads was generated from the 21 new samples from this study.

The taxonomic composition of the gut microbiota was determined by alignment of metagenomic reads to a catalogue of 2,382 reference genomes obtained from the NCBI and HMP databases. Genus abundance was determined and is presented for each sample in Figure 12. Both *Escherichia* and *Klebsiella* are enriched in the RYGB compared to the OBS and CTR groups (Wilcoxon rank-sum test, adj. $P<0.05$). The same trend of enrichment of *Escherichia* is seen in the VBG group but does not reach statistical significance. At the species level, several *E. coli* species were enriched in the RYGB group and a few Firmicutes species were decreased from *Clostridia* and *Gemella* genera. VBG subjects show a similar trend of increasing Proteobacterial species but comparisons does reach statistical significance, if the adj. $P<0.1$ cutoff is used and then many of the same species are both enriched in RYGB and VBG.

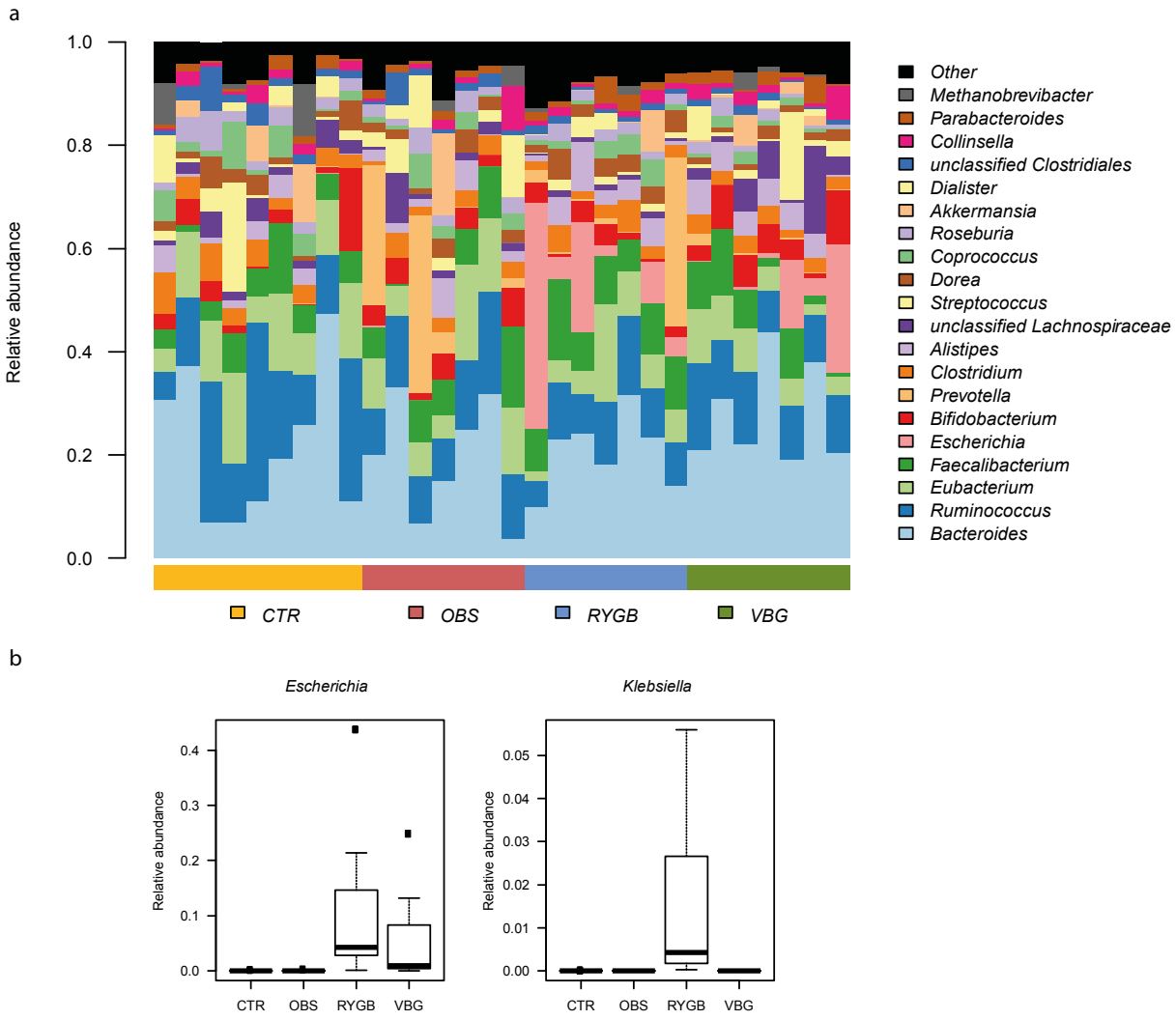

Figure 12 Genus abundance profiles of bariatric surgery patients and controls. a) Genus abundance profiles of the 30 subjects ordered by surgery group. The 20 most abundant genera are shown and less abundant genera are grouped into others. b) *Escherichia* and *Klebsiella* are enriched in RYGB groups compared to OBS and CTR.

The metagenomic reads were aligned to the MetaHIT gut microbial gene catalogue (Qin et al., 2010) using Bowtie2 (Langmead and Salzberg, 2012) and reads aligning to genes were counted. To calculate the abundance of the gene functions KEGG KOs, the provided annotations to the gene catalogue was used and the gene counts for each KO were summed. To assess differential KO abundance between groups, the R package edgeR (Robinson et al., 2010), was used. The catalogue was further annotated in more detail in a bile acid metabolizing enzymes as explained in detail in **Paper III.**

To evaluate the diversity harbored in the gut metagenomes, a maximum of 11 million reads were sampled from the aligned reads and genes with aligning reads were counted. Results show that RYGB have a higher gene count compared to VBG individuals (Student's t-test p=0.036 and Wilcoxon rank-sum test p=0.026) but other comparisons

were non-significant (Figure 13). An increased diversity assessed by the number of genera has been seen previously 3 and 6 months after RYGB (Kong et al., 2013).

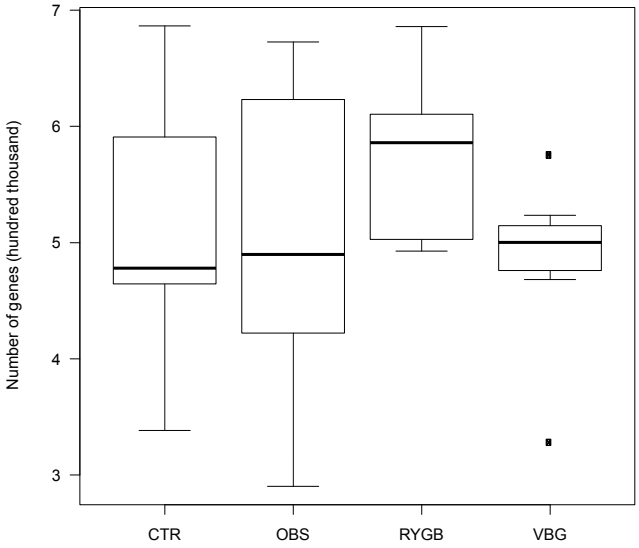

Figure 13 Gene richness in the gut metagenomes. Gene richness is higher in RYGB compared to VBG.

The gene functional differences between the groups were large, 932 and 1117 KOs were enriched in the RYGB compared to the OBS and CTR groups, respectively. In a similar way, 684 and 846 KOs were found to be enriched in the VBG group compared to the OBS and CTR groups respectively. Few changes were seen in the comparison of OBS and CTR or RYGB and VBG indicating that the two operated groups were affected similarly and that the two control groups were comparable. To summarize the overall changes in functional pathways, the reporter feature algorithm was employed as implemented in the R package Piano (Varemo et al., 2013). Fatty acid metabolism and two component systems were identified as reporter pathways enriched in the RYGB and VBG groups compared to control groups. Furthermore, also genes in the phosphotransferase system pathway were found to be enriched which has been observed also previously in a comparison of the gut metagenome 3 months after RYGB surgery (Graessler et al., 2013).

Bile acid metabolism is altered in obese compared to lean individuals and this is important because bile acids are closely related to cholesterol metabolism and act as detergents for uptake of dietary fat. Bile acid levels were measured in the serum of RYGB, VBG and OBS subjects after a standardized meal. RYGB subjects had higher postprandial levels of total bile acids as well as glyco- and tauro-conjugated bile acids in serum while VBG had intermediate and OBS low levels of total and conjugated bile acids. The observation that bile acid metabolism was different between the groups made it interesting to investigate the abundance of genes for bile acid metabolism in the gut metagenomes. Levels of genes in the specific pathway for 7α-dehydroxylation, BaiB, BaiCD, BaiE, BaiF, BaiG, BaiH and BaiI, had a trend for enrichment in the RYGB group but did not reach statistical significance (Adj. P 0.25-0.51).

The causal role of the gut microbial contribution to weight-loss after bariatric surgery was investigated by transplantation of whole microbiota into germ-free mice. Mice receiving either RYGB or VBG microbiota gained less body fat compared with recipients of OBS microbiota. The fat gain in RYGB recipients was also significantly lower compared to VBG recipients.

In summary, the gut metagenome after bariatric surgery is altered to a large extent both taxonomically and functionally. The alterations are likely due to altered intestinal growth conditions such as pH, bile acid levels and nutrients related to host malabsorption. Fecal transplantation from human donors to germ-free mice here suggests that the altered gut metagenome contributes to the improved metabolism and weight reduction. Similar results have been shown by transplantation of gut flora RYGB operated mice to germ-free recipients, RYGB flora results in less body fat compared to flora from sham-operated mice (Liou et al., 2013). These results indicate that shifts in the microbiota contribute to reduced weight after bariatric surgery.

3.1.4. Common lessons from the gut microbiome in metabolic diseases

A common finding in the diabetes and symptomatic atherosclerosis patients was that they both have reduced levels of butyrate producing bacteria, *Roseburia*, *Eubacterium* and other *Clostridiales* species. Known butyrate producing species and genes for butyrate production was also negatively correlated with the inflammatory marker hsCRP in **Paper I**. Butyrate producing bacteria have also been associated with a number of healthy states compared to diseased. An example is that butyrate producing bacteria such as *Roseburia* species were found to be reduced in diabetic patients in China (Qin et al., 2012), obese individuals with poor metabolic and inflammatory profiles (Le Chatelier et al., 2013) and similarly a reduction in inflammatory bowel disease (Sokol et al., 2007; Sokol et al., 2008). There is a possibility that butyrate producers are a proxy for an unknown factor that is associated with a healthy microbiota. However, much evidence suggests that butyrate itself is beneficial to the host. Specifically butyrate has been shown to induce differentiation of T-regulatory cells, anti-inflammatory immune suppressing cells, and ameliorate the development of colitis (Arpaia et al., 2013; Furusawa et al., 2013). Butyrate is also known for providing energy and carbon to cells lining the intestine and supports the intestinal barrier function. Taken together, butyrate seems to be an important component of a healthy host-microbiota interaction but it cannot be ruled out that also other factors of bacteria characterized as butyrate producers is providing beneficial elements. To increase the levels of butyrate producing bacteria, several different approaches could be taken. Transplantation of gut microbiota to recipient with metabolic syndrome resulted in increased insulin sensitivity and higher levels of butyrate producing bacteria in the colon (Vrieze et al., 2012). Other strategies to increase the levels of butyrate producing bacteria have been reviewed by Louis and Flint (Louis and Flint, 2009). Diet rich in resistant starch and a colonic pH of around 5.5 instead of 6.5 have been shown to promote butyrate production (Louis and Flint, 2009).

There is less consensus into what direction the microbiota takes in a metabolic diseased state. In **Paper I**, *Collinsella* was enriched in patients with symptomatic atherosclerosis,

whereas in **Paper II** *Lactobacillus* and *Streptococcus* were enriched. A common finding in several studies of obesity and diabetes is the enrichment of *Clostridium clostridioforme* and related species in diabetic or obese subjects. *C. clostridioforme* has traditionally been used to group three related species, *Clostridium bolteae*, *C. clostridioforme* and *Clostridium hathewayi,* all who are infectious species with some resistance to antibiotics (Finegold et al., 2005). In **Paper II**, *C. clostridioforme* was found to be enriched in diabetic Swedish patients and patients from China. In the original publication *C. bolteae and C. hathewayi* were specifically reported to be enriched in Chinese diabetic patients (Qin et al., 2012). A study of obese and lean individuals associated *C. clostridioforme* and *C. bolteae* with a low diversity microbiome and overall adiposity, insulin resistance and dyslipidaemia (Le Chatelier et al., 2013). An experiment using germ-free mice as recipient of lean and obese human gut microbiota showed that lean donor microbiota conferred less weight gain and adiposity in recipient mice. *C. clostridioforme* and *C. hathewayi* were identified as sources from the obese microbiota and the former could colonize lean recipients if co-housed with obese while the abundance of the latter correlated with levels of branched chain amino acid (Ridaura et al., 2013). Taken together, these findings suggest that *C. clostridioforme* and related species should be further investigated to discern if it plays a role in promoting adiposity and insulin resistance.

Bariatric surgery results in extensive alterations of the gut microbiota, mainly an expansion of Gammaproteobacteria and *Escherichia*, from <1% to ~20%, have been observed shortly after RYGB (Kong et al., 2013; Zhang et al., 2009) and also in a longer term as reported in **Paper III.** Experiments using germ-free mice as recipients of diet-induced-obese mice who have undergone either RYGB or sham operations showed that RYGB flora recipients had less adiposity compared to recipients of sham flora (Liou et al., 2013). *Escherichia* was enriched after RYGB also in mice. The expansion of *Escherichia* is not generally considered to be associated with health compared to a metabolic diseased state. The mechanism by which the microbiota can confer lower adiposity in mice is therefore interesting and important to investigate. Furthermore, the microbial changes could also confer health complications besides its suggested contribution to reduced weight.

Paper IV discussed the current literature about the associations between the gut microbiota, methods to study it as well as the use of germ-free mice as tools for discerning causal relationship between the host physiology and the microbiota.

3.2. Bioinformatic tools for metagenomic data analysis

Analysis of a metagenomic data set for discerning associations between metagenomic components and disease state and other clinical parameters involves several steps. The process can be divided into two main steps: (i) quality control, data filtering and annotation of sequence reads (ii) linking data clinical data to quantitative metagenomic features. Step (i) typically requires knowledge about a Linux operating system and knowledge about how to execute a range of bioinformatics programs and parallel computation. Step (ii) is less computationally demanding but could still demand programming and computational knowledge. To make these two tasks easier and lower

the hurdles for more researchers to analyze metagenomic data, two tools for are presented in **Paper V** and **Paper VI**.

3.2.1. Paper V: FANTOM, an easy to use tool for metagenomic data analysis

Analysis of metagenomic data has traditionally been a computationally intensive work that required knowledge in one or more programming languages. Many bioinformatics tools are designed for the Linux operating system and thus require knowledge in Linux to be executed. Furthermore, a bioinformatics analysis pipeline typically involves several different programs and databases and the output format of one program needs to be parsed to fit the input of the next one in the pipeline. If the requirements of computational skills were lowered, the number of people in the scientific community that could contribute to analyzing and interpreting metagenomic data would increase. With the above in mind, FANTOM, Functional ANnotation and Taxonomic analysis Of Metagenomes, was developed. FANTOM is a standalone tool that runs on Windows, OSX and Linux with a graphical user interphase to analyze quantitative metagenomic data in a functional and taxonomic content (Figure 14). The abundance data is easily integrated with hierarchical databases such as NCBI taxonomy and KEGG.

FANTOM was implemented in Python and make use of some core scientific packages such as numpy. The graphical user interphase was implemented with wxPython. Installers are provided for the platforms, Windows, OSX and Linux which makes it easy to install and execute.

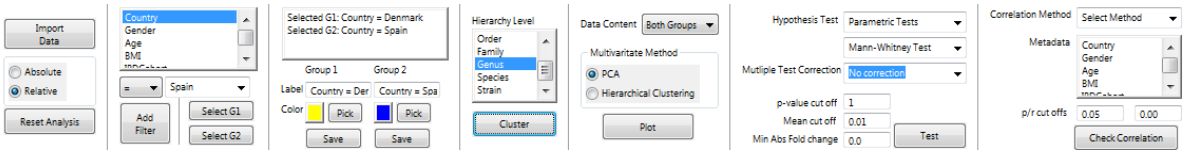

Figure 14 Screenshot from the command panel of FANTOM.

FANTOM needs two input files, an abundance file of metagenomic features, taxonomic or functional, and a sample metadata file. The abundance file should be a tab delimited file with identifiers such as NCBI Taxonomy IDs or KEGG KOs and the abundance information in the form of read counts. The metadata file should be a tab delimited file with the same identifiers as the samples in the abundance file and numerical or categorical variables. In the data import step, the user selects the type of data that should be imported and which database to use. FANTOM makes use of the hierarchical structure in databases such as NCBI Taxonomy and KEGG pathways. The abundance of a higher node in the hierarchy (e.g . Genera or pathways) is calculated by summing the abundance of all member nodes (e.g. species or KEGG KOs). An example is shown in Figure 15 where the input data is species abundance and these are summed to phyla and displayed in an area plot.

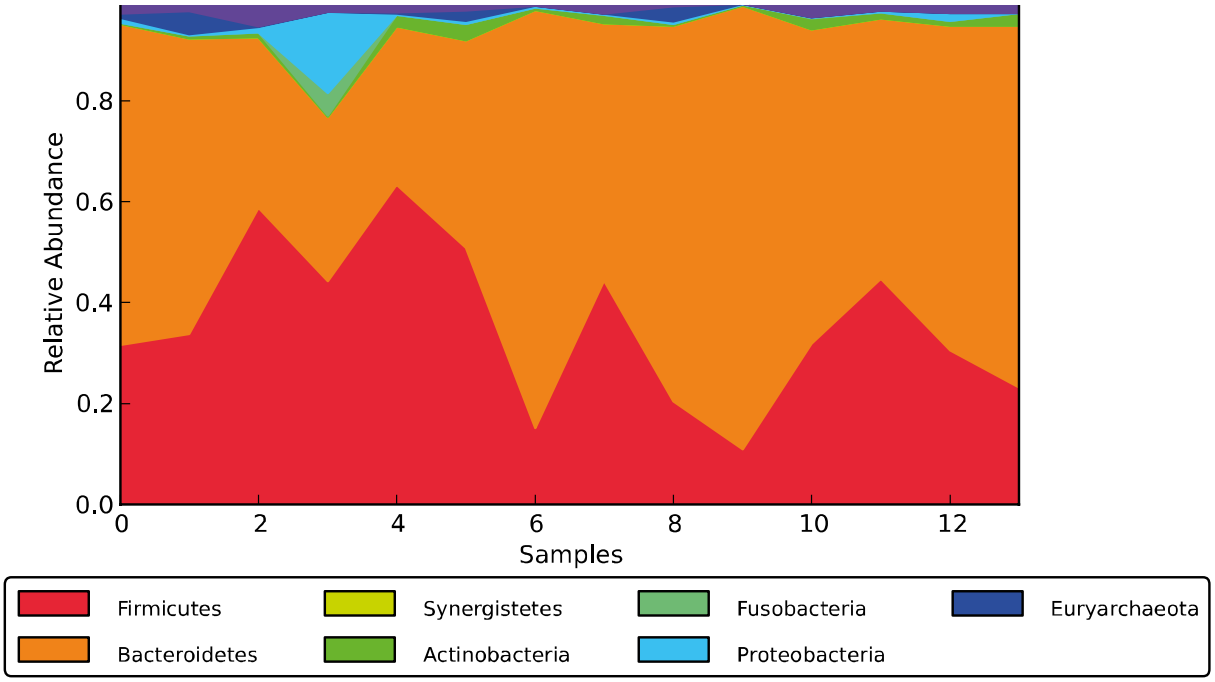

Figure 15 Area plot of phyla abundance in 13 gut metagenomes

Statistical hypotheses tests can be performed between groups of samples defined in the metadata information about samples. Both parametric and non-parametric tests were included in the FANTOM and non-parametric tests are encouraged because metagenomic data typically do not follow common distributions. The software does also contain the possibility to correct obtained p-values for multiple testing with e.g. the Bonferroni or Benjamini-Hochberg methods. Data at varying hierarchical levels can be visualized in several different ways by area, bar and box plots. In summary, FANTOM is an easy to use downstream tool for metagenome data analysis with integration to databases for biological interpretation.

3.2.2. Paper VI: MEDUSA and construction of a global gut microbial gene catalogue

Modern DNA sequencing machines can produce hundred millions of sequence reads from a metagenome and annotation and characterization of such a dataset can be a daunting task. The analysis involves several steps, often performed sequentially, and includes data quality control, filtering contaminant reads (e.g. human) and comparison to a reference catalogue. The reference catalogue can be a set of sequenced genomes or a non-redundant gene catalogue of genes assembled from metagenomes. The data size often requires that these tasks are performed on a computational cluster with parallel execution. To address the challenge of quantitative characterization of metagenomes, MEDUSA was developed to perform quality control, filtering and counting alignments to up to two databases in one computational stream. Furthermore, it includes scripts for handling downstream tasks and annotation to taxonomic and functional databases (Figure 16). MEDUSA was implemented in Python with the use of the package numpy and the standalone tools fastx (http://hannonlab.cshl.edu/fastx_toolkit/), bowtie2 (Langmead and Salzberg, 2012) and GEM (Marco-Sola et al., 2012).

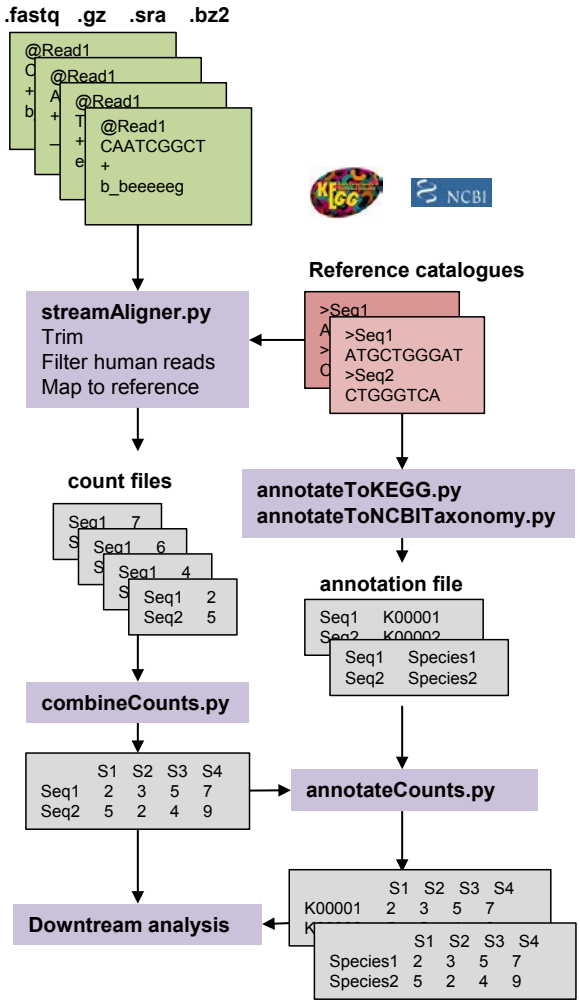

Figure 16 Overview of the MEDUSA pipeline. Fastq sequence files are input data and can be compressed in various ways. MEDUSA aligns reads to a reference database and counts aligning reads. Count files can be merged into a count table and annotated to taxonomic and functional (KEGG KO) levels.

MEDUSA was tested on 4 large metagenomic datasets including subjects from 3 different continents (Huttenhower et al., 2012a; Karlsson et al., 2013; Qin et al., 2010; Qin et al., 2012). A global gut microbial gene catalogue was constructed from the 4 studies by starting with assembled contigs, predicting genes on contigs with Metagenemark (Zhu et al., 2010). Predicted genes were clustered with Usearch (Edgar, 2010) using the criteria 95% sequence identity and 90% coverage of the shorter sequence. Genes from each study were first clustered separately and later merged in a global gene catalogue containing 11 million genes. Each study showed a substantial number of unique genes while the number of genes found shared between all studies was 488,482 (Figure 17). Importantly, the shared genes were abundant and attracted 38 ± 8% of the reads when mapped onto the gene catalogue and a similarly large part of reads mapped on to study-unique genes (36 ± 4%). Considering the smaller number of shared genes compared to study-unique genes, if normalized to the number of genes in each category, shared genes are highly abundant.

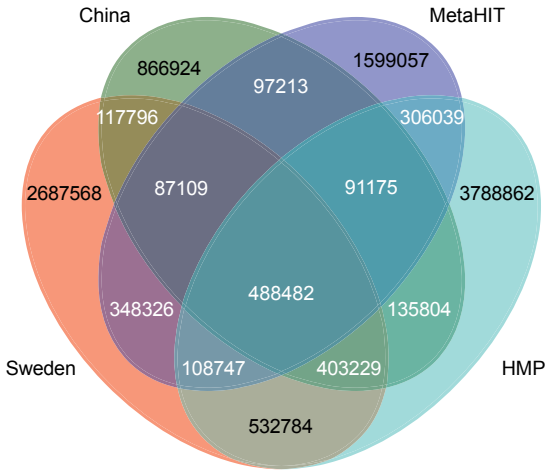

Figure 17 Venn diagram of gene distribution in the 4 studies included. A core of almost half a million genes was shared between all studies. The largest number of unique genes was found in the HMP study which was also the one with most samples and deepest sequencing.

Reads from each of the 782 samples were subjected to quality control; human reads were filtered and mapped onto a catalogue of 1747 prokaryotic species genomes and the constructed gut microbial gene catalogue. Almost 98% of the reads passed the quality cutoff and of these, 75% were aligned to the gene catalogue while 39% were aligned to the species genome catalogue. The taxonomic profiles of the metagenomes were determined by analyzing the reads aligning to the species genome catalogue. The most abundant genus was *Bacteroides* but the abundance varied greatly within the samples (Figure 18). The abundance of *Bacteroides* was higher in HMP and Chinese samples compared to European samples. In two different reports *Bacteroides* abundance have been associated with a diet rich in animal protein and fat (David et al., 2013; Wu et al., 2011). The abundance of genera from the Firmicutes varied across study populations and in general the Swedish and to some extent the Metahit population had more *Faecalibacterium*, *Eubacterium*, *Clostridium* and *Dorea.* The European populations also had a more diverse species richness assessed by the Shannon diversity index compared to HMP and Chinese samples.

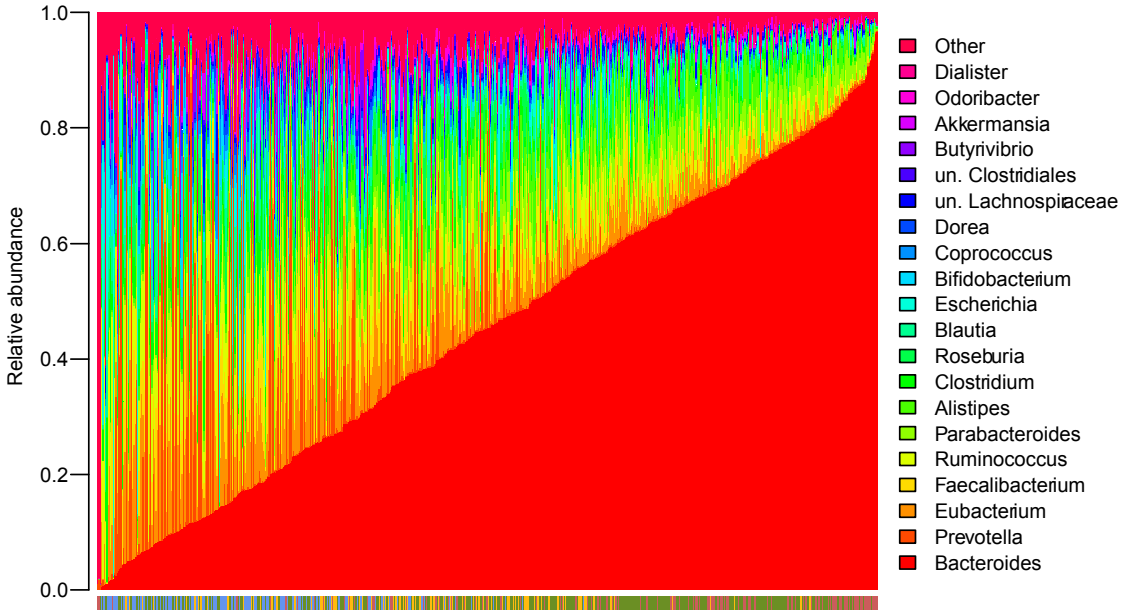

Figure 18 Genus abundance in the 782 samples. Samples are ordered by increasing abundance of *Bacteroides* represents an increasing gradient. Color bar at the bottom shows the study origin of the sample, Light blue; Sweden, yellow; MetaHIT, red; HMP, green; China.

The existence of a core gut microbiome at the taxonomic level have been debated and arguments for a core microbiome only at the functional gene level rather than at the taxonomic species level (Turnbaugh et al., 2009) or that there is a core of species (Qin et al., 2010; Tap et al., 2009) have been presented. The answer to this question is depending on how the core is defined as this is to some extent arbitrary. In this work, a core species is defined as being a species that has a relative abundance above 10^{-4} and being present in more than 50% of the studied individuals. The number of species present in at least 50% of the individuals is 116 and 71 are also present in 90% of the individuals (Figure 19). This suggests that there is a core also at the organismal or taxonomic level in this set of individuals from three different continents.

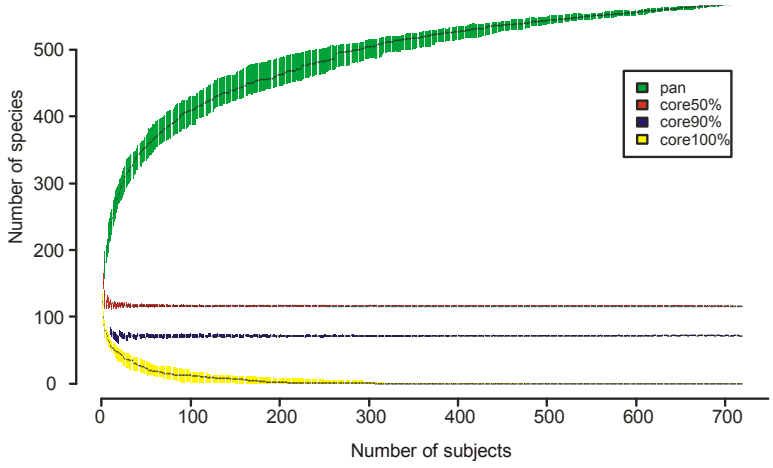

Figure 19 Pan and core species. The core percentage means that the species was present in at least that fraction of the studied subjects.

Most core species belong to the genus *Bacteroides* with 20 representatives present in at least 50 % of the subjects (Table 2). Furthermore, genera in the Firmicutes such as *Clostridium*, *Ruminococcus Eubacterium* and *Faecalibacterium* are all clearly present in the core as well as some species from the Actinobacteria in the *Bifidobacterium* genus.

Table 2 Origin and number of core species present in at least 50% of the subjects

Genus name	Number of species	Genus name	Number of species
Bacteroides	20	unclassified butyrate producing bacteria	1
Clostridium	13	*Butyrivibrio*	1
Prevotella	10	*Capnocytophaga*	1
Ruminococcus	9	*Clostridiales*	1
Eubacterium	8	*Collinsella*	1
Lachnospiraceae	6	*Eggerthella*	1
Faecalibacterium	4	*Holdemania*	1
Alistipes	3	*Marvinbryantia*	1
Bifidobacterium	3	*Megasphaera*	1
Coprococcus	3	*Odoribacter*	1
Roseburia	3	*Oribacterium*	1
Blautia	2	*Paraprevotella*	1
Coprobacillus	2	*Parasutterella*	1
Dorea	2	*Phascolarctobacterium*	1
Erysipelotrichaceae	2	*Pseudoflavonifractor*	1
Escherichia	2	*Ruminococcaceae*	1
Parabacteroides	2	*Streptococcus*	1
Anaerostipes	1	*Subdoligranulum*	1
Anaerotruncus	1	*Tannerella*	1
Bilophila	1		

The richness of the gut microbiota was assessed by counting genes after normalization to 11 million reads. A gene was counted as present if at least two reads mapped on to it. When the richness was compared for subjects from the 4 studies, it again appears that European samples have higher richness compared to Chinese and HMP samples (Figure 20). Low richness of the microbiota has been reported to be associated with a number of diseases such as inflammatory bowel disease (Manichanh et al., 2006), inflammation in elderly (Claesson et al., 2012) and obesity (Le Chatelier et al., 2013; Turnbaugh et al., 2009). Furthermore, large differences in diversity are also seen between populations and lower diversity has been observed in subjects from America compared to Amerindians from Venezuela and Malawians (Yatsunenko et al., 2012). Although there are differences in diversity between individuals, there is a common core that is present in at least 50% of the individuals. The size of this core is 287,921 genes which indicated that a large portion of the genes carried by an individual is shared. This size of genes can be roughly compared to the species core of 116 species which carries 348,000 genes assuming that each species has 3000 genes on average. Interestingly, over 10 million genes are shared

by at least 2 individuals which indicates that even rare genes are shared by some individuals.

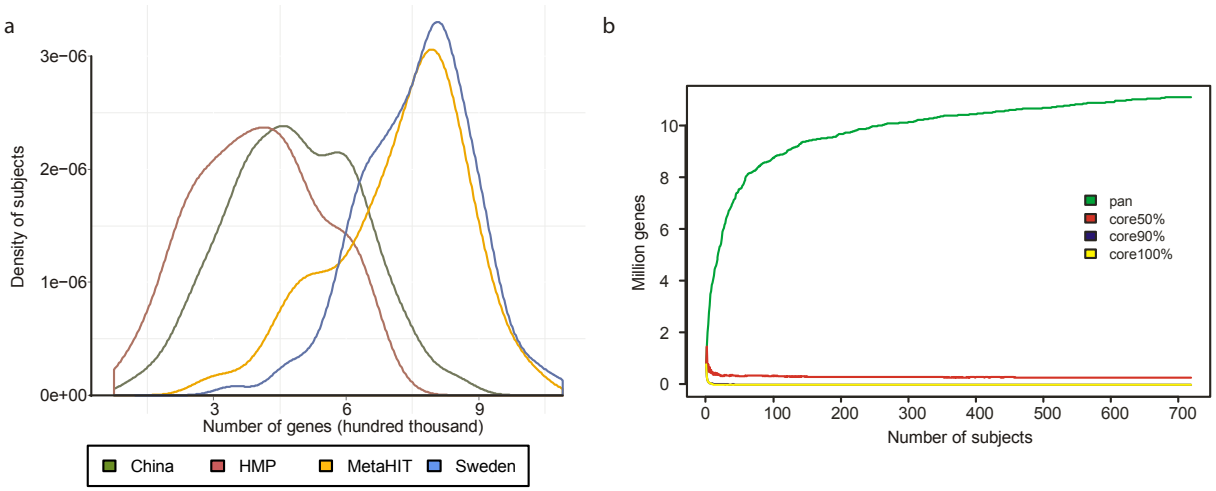

Figure 20 Gene richness and pan and core genes. a) Gene richness in samples grouped by study. The number of genes were counted using 11 million reads from each sample. Swedish and MetaHIT samples have a higher gene richness compared to American and Chinese. b) The number of pan and core genes in the 4 studies are shown as a function of the number of subjects. The number of core genes present in at least 50% of the population is 287,921.

3.3. Systems biology and metabolic modeling applied to the gut microbiota

To increase the information gained from data generated in metagenomic studies, systems biology approaches and metabolic modeling can be applied. Metagenomics provides a part list of the components present and associated with different disease states or diets. However, to link metagenomic information with likely metabolic consequences of a particular state and how the parts interact, a modeling approach is needed. A model in this context takes an input and calculates an output under a number of assumptions. The type of model used in this work is genome-scale metabolic models. The motivation use these models and their historic use in relation to human health are outlined in **Paper VII** and an example of their application to the gut microbiota is presented in **Paper VIII**.

3.3.1. Paper VII: Genome-scale metabolic models for human health and the gut microbiota.

Genome-scale metabolic models have served as a very useful tool for studying the interactions between human metabolism and human related microbes. In **Paper VII**, the field of GEM modeling of human related microbes was reviewed and no reconstruction of microbial species from the gut microbiota could be found. Furthermore, the previous literature of modeling more than one species was very limited. With this background, ideas for a framework for modeling the human gut microbiota were presented (Figure 21). Importantly, GEMs can be used to generate hypotheses that could be tested in an

experimental setting. An example could be to evaluate the production of butyrate from a series of different consortia of microbes. These hypotheses could be quickly tested *in silico* and the most promising could then be tested experimentally.

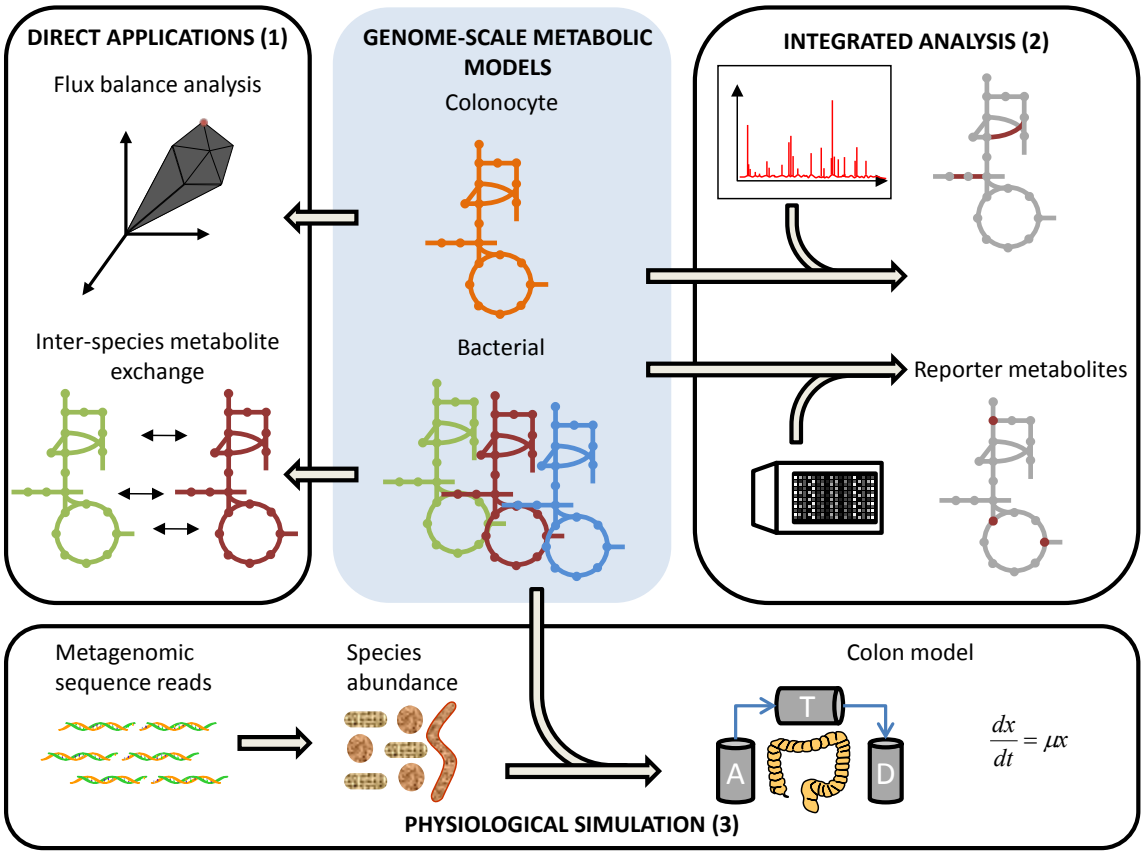

Figure 21 Framework for modeling the gut microbiota using GEMs. 1) GEMs can be directly used for simulations using FBA to predict metabolic fluxes under different growth conditions and interspecies metabolite exchanges. 2) GEMs constitute excellent scaffolds for mapping transcriptomic and metabolomics data and inferring a metabolic context. 3) Metagenomics data of species abundance can be used in physiological simulations of the intestinal tract to infer metabolic fluxes.

Initial modeling studies using GEMs must be validated with experimental data. **Paper VII** identified a set of three species, *Bacteroides thetaiotaomicron, Eubacterium rectale* and *Methanobrevibacter smithii,* as possible species to model in a gut environment with valuable data for validation in the literature.

3.3.2. Paper VIII: Metabolic modeling of three bacteria in the gut.

Metabolism of SCFAs is particularly important in the gut microbiota, both as a necessary byproduct of microbial fermentation and as an important substrate for colonocytes and other host cells. In **Paper VIII**, three important species in the human gut were reconstructed, *Bacteroides thetaiotaomicron* (iBth1201), *Eubacterium rectale* (iEre400) and *Methanobrevibacter smithii* (iMsi385). *B. thetaiotaomicron* is a well-studied representative from the Bacteroidetes phyla and one of the first gut bacteria to be sequenced (Xu et al., 2003). It contains a large repertoire of polysaccharide degrading enzymes and its main byproduct during fermentation is acetate and propionate (Mahowald et al., 2009). *E. rectale* is a member of the Firmicutes phyla and a known

butyrate producer while *M. smithii* is and archaea that plays a special and important role in the gut by disposing hydrogen and formate and produces methane.

The RAVEN toolbox was used for reconstruction of the three species (Agren et al., 2013). Manually constructed GEMs of well characterized species (Feist et al., 2007; Heinemann et al., 2005; Satish Kumar et al., 2011) were used as templates for reconstruction, manual curation was performed and reactions were also added from KEGG (Kanehisa et al., 2004). Each GEM was validated individually with available experimental data and metabolic task such as amino acid production and substrate utilization were evaluated with the RAVEN toolbox.

The GEMs were evaluated in and *in vivo* setting and compared with well-characterized germ-free mice colonized with combinations of the studied microbes. In this context, the metabolic fluxes of substrate uptake and byproduct secretion are predicted from the abundance information provided in the experiments. Metabolic fluxes are then compared to experimental observations. An important assumption in these simulations is that each microbe uses a minimal amount of substrate for a given production of its biomass which is equivalent to maximizing the biomass production for a given amount of substrate.

B. thetaiotaomicron simulations were compared to experimental results in mono-colonized mice and are shown in Figure 22. Acetate and propionate production could be compared and were in agreement with experimental results (Mahowald et al., 2009). *E. rectale* simulations were also compared to results in mono-colonized mice and produced butyrate, CO_2 and H_2. Again, results were in agreement with experimental observations (Mahowald et al., 2009). *M smithii* was also mono-colonized in germ free mice and simulations were compared to experimental values (data shown in **Paper VIII**). The main produced byproduct is methane gas for which the production in mice is experimentally difficult to measure.

The most interesting aspect of this type of modeling is to simulate how two or more species interact with each other. In order to simulate the interaction, the two models were merged into a common stoichiometric matrix and the extracellular compartment was merged to a common compartment. *B. thetaiotaomicron* and *E. rectale* were simulated together in the mouse gut and results compared to experimental values. The main exchange between the two species is acetate that is produced by *B. thetaiotaomicron* and taken up by *E. rectale.* Also here, the biomass production rate was set to its experimental value and the substrate uptake rate was minimized. The predicted metabolite concentrations and the deviations to the experimental values for acetate, propionate and butyrate were 0.7 (14%), 1.2 (40%) and 0.07 (30%) mmol/g Cecal content, respectively (Figure 22). The errors in the prediction values compared to experimental data are comparable to the measurement error in the experiments. Values for the biomass content are used in the simulations and do also contain error to which the error in SCFAs are directly proportional to. The predictions presented here are thus as good as the accuracy and quality of the input data allows.

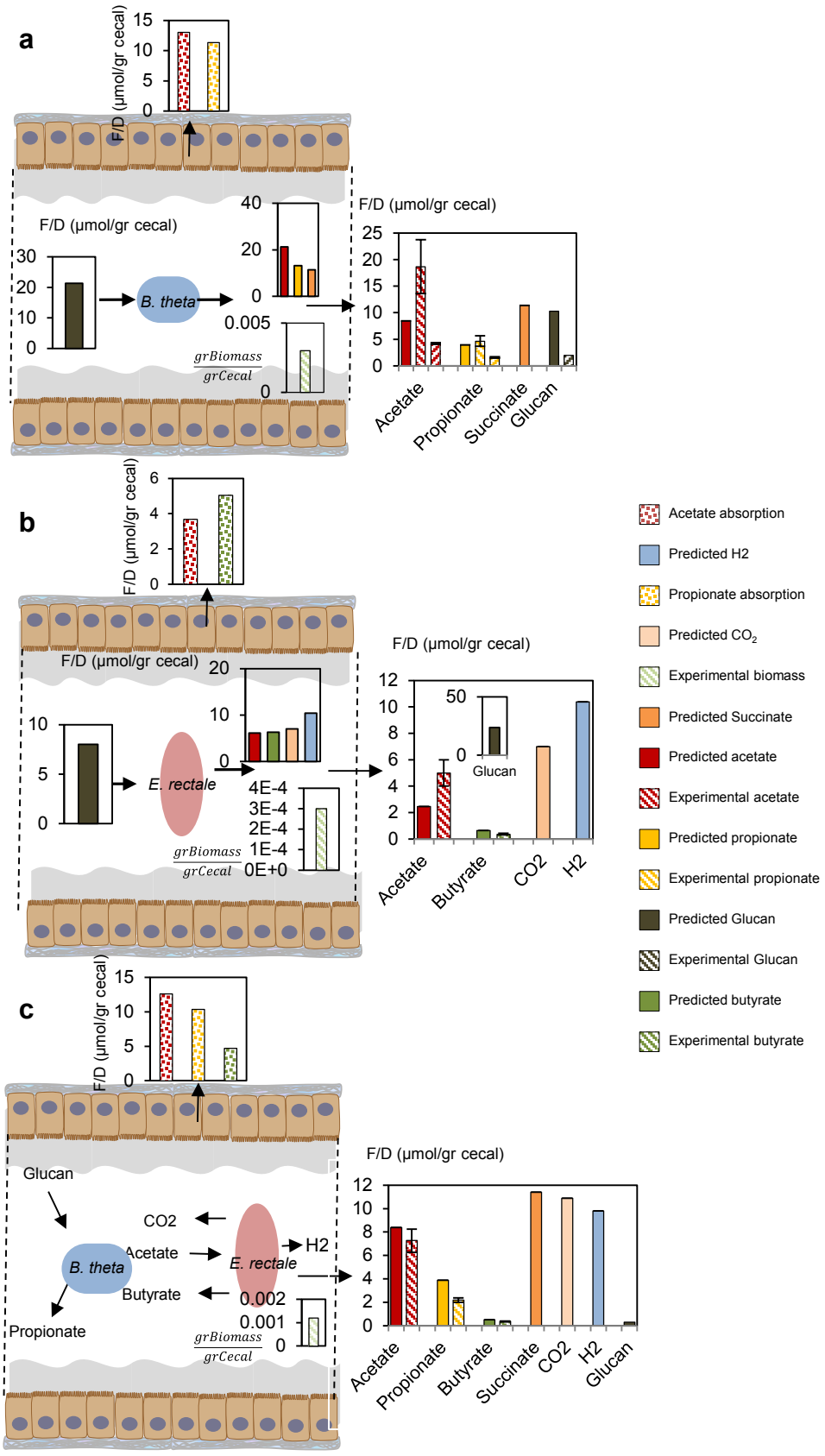

Figure 22 Simulation of mono and co-colonizations in germ-free mice. (a) Monocolonization of *Bacteroides thetaiotaomicron* (b) Mono-colonization of *Eubacterium rectale (c)* Co-colonization of *Bacteroides thetaiotaomicron* and *Eubacterium rectale.*

GEMs are excellent tool for integrative analysis and interpretation of transcriptomic data in a metabolic context. In this case study, transcriptional profiles of *E. rectale* and *B. thetaiotaomicron* in a co-colonization were compared to mono-colonization of the two. By using the reporter subnetworks algorithm (Patil and Nielsen, 2005), a number of metabolic subnetworks were identified that represent a concerted metabolic response upon co-colonization (Figure 23). *E. rectale* responds to co-colonization with *B. thetaiotaomicron* by down-regulation of most of its polysaccharide utilizing enzymes and up-regulating several genes for amino acid biosynthesis and cell wall production. A number of amino acid transporters were also up-regulated indicating that *E. rectale* shifts substrate utilization strategy from carbohydrates to amino acids. *B. thetaiotaomicron* responds to the presence of *E. rectale* by up-regulating the expression of poly- and mono-saccharide utilization genes. *B. thetaiotaomicron* broadens the variety of polysaccharides used, also by increased expression of enzymes for degradation of host glycans.

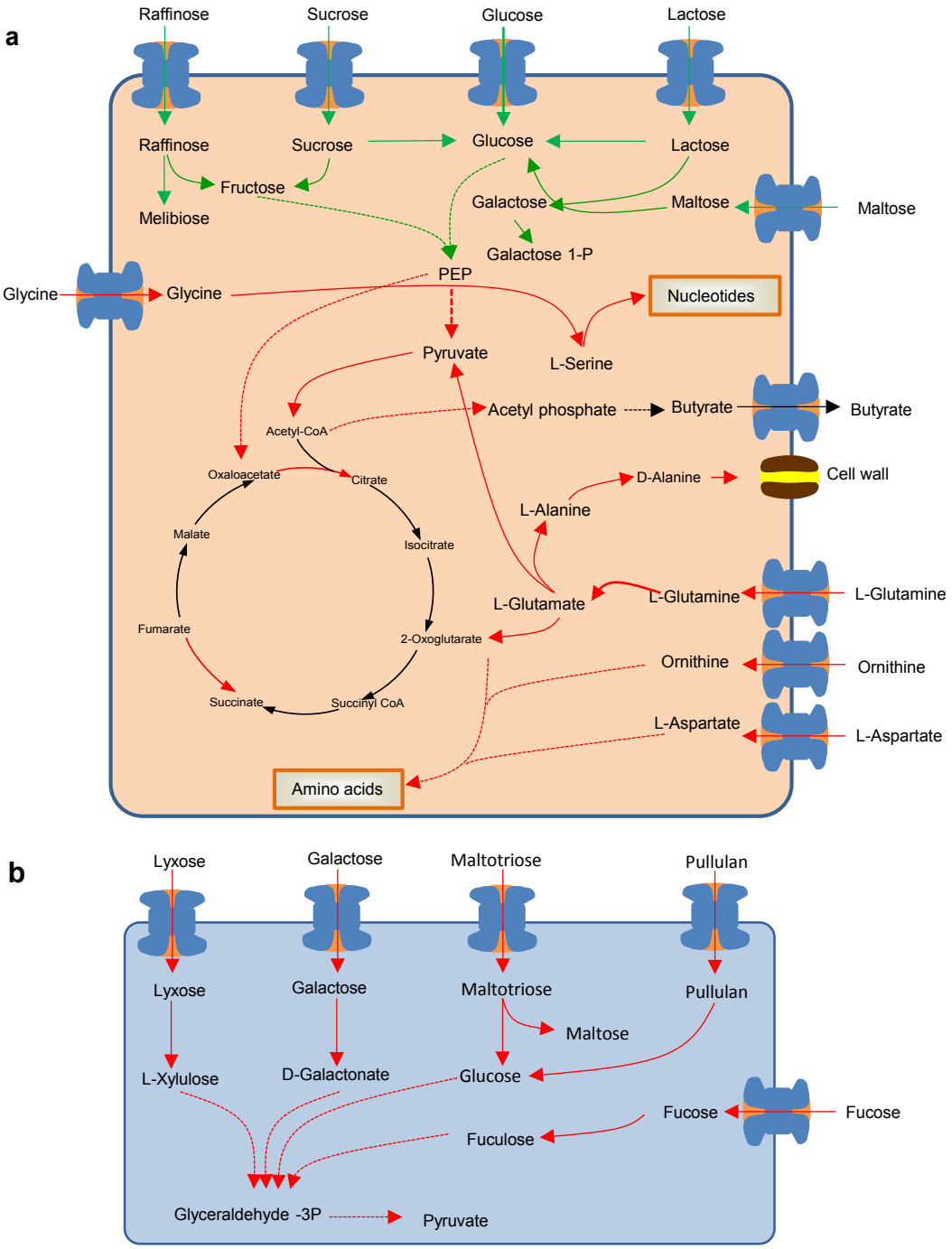

Figure 23 Reporter subnetworks for the transcriptional response of co-colonization. (a) Transcriptional changes in *E. rectale* when co-colonized with *B. thetaiotaomicron* are mainly in down regulation of saccharide metabolizing enzymes and upregulation of amino acid metabolism. (b) Transcriptional changes in *B. thetaiotaomicron* when co-colonized with *E. rectale* are mainly up-regulation of polysaccharide utilization. Red are up-regulated while green are down-regulated reactions.

4. Conclusions and future perspectives

This thesis aims to address the three questions that were introduced in the introduction. The three questions were:

- **How is the human gut metagenome associated with metabolic diseases?**
- **Can bioinformatics tools for analyzing metagenomic data be advanced and made more easily accessible?**
- **Can metabolic modeling be used for studying basic metabolism in the human gut?**

In **Paper I,** we observed that the gut metagenome was altered in patients that have had an atherosclerotic event. Control subjects had higher levels of *Roseburia* and *Eubacterium*, both genera are known to contain species that are butyrate producers. Furthermore, the key gene for butyrate production was strongly negatively correlated with the inflammatory marker hsCRP. Control subjects also had higher plasma levels of the antioxidant ß-carotene and genes for the synthesis of this compound in their metagenomes. Patients had higher levels of genes for the synthesis of peptidoglycan and higher abundance of species from the *Collinsella* genus. Taken together, we have found clear differences in the gut metagenome of patients with symptomatic atherosclerosis compared to controls.

In **Paper II,** we studied a larger cohort of subjects with diabetic, impaired and normal glucose control. With the higher number of samples in this study, it is possible to group genes into MGCs based on their co-occurrence across individuals. This made it possible to construct taxonomic markers based on groups of genes likely from the same species of which some might not be sequenced yet. By comparing species and MGCs abundance between T2D and NGT individuals, again species which are known butyrate producers such as those from *Roseburia* had higher abundance in controls. T2D individuals had higher levels of *Lactobacillus* and *C. clostridioforme.* In **Paper II** we also demonstrated that the gut metagenome can be used for accurately classifying diabetic status which is a step towards identifying individuals at risk of developing diabetes. Indeed, the model identified IGT individuals that also had higher triglyceride and C-peptide levels which are risk factors for metabolic disease.

Taken together, **Paper I** and **Paper II** demonstrate that metabolic diseases are associated with deviations from a healthy gut microbiota. The healthy state appears to be one rich in butyrate producing bacteria as has been seen also in obesity and inflammatory bowel disease. However, a disease associated microbiota might have deviated in several different directions. The comparison of gut metagenomes from 4 different studies with individuals from 3 continents in **Paper VI** highlights large differences between populations. Different diets are likely to play an important role in shaping the microbiota and could be the reason for regional variation. **Paper III** demonstrated large long term effects of bariatric surgery on the gut microbiota with an expansion of *Escherichia* and altered abundance of many gene functions. VBG subjects show similar deviations from the obese state as do RYGB although to a lesser extent. **Paper IV** reviews the current knowledge about the gut microbiota and metabolic diseases.

Two software tools for metagenomic data analysis and interpretation were developed. **Paper VI** describes MEDUSA that can be used for efficient pre-processing and annotation of metagenomic reads. Tools for handling the data tables and annotation to KEGG and NCBI taxonomy are provided. MEDUSA was tested by analyzing 782 human gut metagenomes and a global human gut microbial gene catalogue is presented. Downstream processing of quantitative metagenomic features together with clinical data can be analyzed in the program FANTOM described in **Paper V.** FANTOM has a graphical user interphase that should be easy to use for researchers who are not familiar with computer programming. FANTOM also has tools for analyzing data at different taxonomic ranks and functional pathways.

Paper VII outlines a systems biology approach of modeling the metabolism of the gut microbiota and reviewed the use of genome scale metabolic models in human health and disease. In **Paper VIII,** the modeling approach is implemented. Genome-scale metabolic models of three different species in the human gut, *B. thetaiotaomicron, E. rectale* and *M. smithii*, with diverse metabolic functions were reconstructed. The models were validated individually and then merged to interact with each other. Several metabolites were cross-feeding between the 3 models and results were compared to that of mice colonized with the 3 bacteria with good agreement. The models are excellent tools for interpreting transcriptomic data in metabolic context and such use demonstrated how *B. thetaiotaomicron* and *E. rectale* adapts to the presence of each other in the mouse gut.

4.1. Future perspectives

Several different associations between the gut microbiota and obesity have been proposed, sometimes these were conflicting. Work on larger cohorts using deep sequencing will likely clear some of the present uncertainties. It is clear that broad taxonomic characterization is not enough but that detailed characterization of the microbiota is needed to discern relevant associations between human health and the gut microbiota.

In order for findings of the association between the gut microbiota and metabolic diseases to be clinically relevant and used, a number of things need to be established. The associations between the gut microbiota and metabolic diseases have so far mainly been performed on case-control studies. This is of course an important first step to establish if there is a correlation but does not provide evidence about causal relationships. Prospective studies where samples are collected some time before a disease develops are useful for developing biomarkers and risk factors in the gut microbiota for predicting disease development. The final goal is intervention studies, possibly using findings from case-control and prospective studies, to alter a disease related microbiota to a healthy state. An interventional study using fecal transplantation has been used to improve glucose metabolism in patients with metabolic syndrome (Vrieze et al., 2012). However, transfer of whole microbiota carries the risk of also transferring pathogens, some of which are unknown or undetectable with present methods, to recipients. Crucial species delivered as a probiotic cocktail that are tested to be safe and evaluated in a randomized controlled trial is a better alternative. Prebiotics, non-digestible food ingredients that stimulate growth of beneficial bacteria, could be an alternative to promote a healthier microbiota. I believe that the field is moving in this

direction and that there is enough evidence that merits further investigation into the contribution of the gut microbiota to metabolic diseases and possible interventions.

Several tool for advanced bioinformatics analyses of metagenomic data have been released which is important to make metagenomic data analyses available to a wider range of scientist. Compared to RNAseq and genome-wide association studies, metagenomic studies do not use standardized bioinformatics and data analysis methods to the same extent. This is partly due to that the metagenomic field is young and our "other genome" has not been studied to the same extent as the human genome. Additionally, the "other genome" is considerably less conserved between individuals compared to our human genomes and is thus a moving target. For this reason, the gene catalogues will likely have to be updated also in the future.

Genome-scale metabolic models are well suited for modeling of microbial metabolism in the human gut. Modeling approaches are promising for enriching metagenomic data by predicting metabolic outcomes but there are some hurdles to overcome. Accurate reconstruction of a metabolic model is time consuming and requires manual curation of a computer generated draft model and this process needs to be accelerated if all main species in the human gut should be reconstructed. Modeling could be used for aiding the design of probiotics and prebiotics. Considering the complexity of degradation of dietary fibers into sugar monomers and the interspecies feeding that goes on in this process, it is not trivial at all to predict the outcome. Detailed metabolic modeling of fiber decomposition by key species is likely very useful in understanding experimental results and generating new hypotheses.

It has been very exciting to work in the rapidly developing field of the gut metagenome and I expect that it will continue in the same pace in the years to come.

Acknowledgements

First of all I would like to thank my supervisor Jens Nielsen for his support and belief in my capabilities. This project was bold and ambitious from the start and your courage for exploring the unknown has been very inspiring. I have learnt a lot from you for which I am very grateful.

Intawat Nookaew, you have been a great help in getting things to work, coming with great ideas, inspiration and helping out when we have faced difficult bioinformatics problems. Dina Petranovic was early on a great help for general questions about bacteria and you taught me how to write a paper for which I am very grateful. You also guided me in teaching undergraduate students which was a great experience and something that forced me away from the computer and into the lab at times.

Fredrik Bäckhed has been tremendously important for this work. Your support and knowledge has been very valuable. Meeting you and your group for discussions biweekly for the last 3 years has been great and from those meetings I have learnt immensely about microbiology and medicine.

Many people have contributed to this work and have been very important for me. Saeed Shoaie, you have been a great discussion partner and collaborator in the Systems and Synthetic Biology group at Chalmers and a long time office mate. Kemal Sanli, Adil Mardinoglu, Sergio Bordel and Ibrahim El-Semman were greatly contributing to this work. Some people from your group have been great for scientific discussions; Leif Väremo, Tobias Österlund, Wanwipa Vongsangnak, Natapol Pornputtapong, Shaq Hosseini. Some parts of this work were carried out in the lab and for this I would like to thank Suwanee Jansa-Ard, Marie Nordqvist, Malin Nordvall, Ximena Rozo Sevilla and Emma Ribbenhed for helping out and answering my questions when I was lost. Thanks go to Erica Dahlin and Martina Butorac for practical and administrational help.

Many people from Fredrik Bäckhed's lab and other labs at Sahlgrenska University Hospital have contributed to very important parts of this project. I would especially like to thank Valentina Tremaroli for your very important contribution to this work, we made a great team. Special thanks go to Frida Fåk, Petia Kovatcheva-Datchary and Felix Sommer. Also I would like to thank Björn Fagerberg, Göran Bergström and Carl Johan Behre for clinical input to this work.

The social experience in the Systems and Synthetic Biology group at Chalmers has been great and the people there are great friends. Rasmus, Tobias, Leif, Christoph, Verena, Rahul, Francesco, Shaq and many more have made the working environment a great place to be full of laughs.

Early on in my life I loved experimenting with everything and I am very grateful to my parents for your encouragement and support. Your unconditional support and belief in me has been very important.

Ida and Valdemar, I love you more than I can describe in words. I am very grateful for your support. To share my life with you is fantastic, every day!

References

Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol *8*, e1002358.

Adlerberth, I., Strachan, D.P., Matricardi, P.M., Ahrne, S., Orfei, L., Aberg, N., Perkin, M.R., Tripodi, S., Hesselmar, B., Saalman, R., *et al.* (2007). Gut microbiota and development of atopic eczema in 3 European birth cohorts. J Allergy Clin Immunol *120*, 343-350.

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS Comput Biol *9*, e1002980.

Amar, J., Lange, C., Payros, G., Garret, C., Chabo, C., Lantieri, O., Courtney, M., Marre, M., Charles, M.A., Balkau, B., *et al.* (2013). Blood microbiota dysbiosis is associated with the onset of cardiovascular events in a large general population: the D.E.S.I.R. study. PLoS One *8*, e54461.

Andersson, A.F., Lindberg, M., Jakobsson, H., Backhed, F., Nyren, P., and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One *3*, e2836.

Angiuoli, S.V., White, J.R., Matalka, M., White, O., and Fricke, W.F. (2011). Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. PLoS One *6*, e26624.

Arpaia, N., Campbell, C., Fan, X., Dikiy, S., van der Veeken, J., deRoos, P., Liu, H., Cross, J.R., Pfeffer, K., Coffer, P.J., *et al.* (2013). Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. Nature *504*, 451-455.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M., *et al.* (2011). Enterotypes of the human gut microbiome. Nature *473*, 174-180.

Backhed, F., Ding, H., Wang, T., Hooper, L.V., Koh, G.Y., Nagy, A., Semenkovich, C.F., and Gordon, J.I. (2004). The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A *101*, 15718-15723.

Backhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., and Gordon, J.I. (2005). Host-bacterial mutualism in the human intestine. Science *307*, 1915-1920.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

Bermudez, O.I., Ribaya-Mercado, J.D., Talegawkar, S.A., and Tucker, K.L. (2005). Hispanic and non-Hispanic white elders from Massachusetts have different patterns of carotenoid intake and plasma concentrations. J Nutr *135*, 1496-1502.

Brady, A., and Salzberg, S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods *6*, 673-676.

Brown, C.T., Davis-Richardson, A.G., Giongo, A., Gano, K.A., Crabb, D.B., Mukherjee, N., Casella, G., Drew, J.C., Ilonen, J., Knip, M., *et al.* (2011). Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. PLoS One *6*, e25792.

Cani, P.D., Amar, J., Iglesias, M.A., Poggi, M., Knauf, C., Bastelica, D., Neyrinck, A.M., Fava, F., Tuohy, K.M., Chabo, C., *et al.* (2007). Metabolic endotoxemia initiates obesity and insulin resistance. Diabetes *56*, 1761-1772.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res *37*, D233-238.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. Nat Methods *7*, 335-336.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res.

Cotillard, A., Kennedy, S.P., Kong, L.C., Prifti, E., Pons, N., Le Chatelier, E., Almeida, M., Quinquis, B., Levenez, F., Galleron, N., *et al.* (2013). Dietary intervention impact on gut microbial gene richness. Nature *500*, 585-588.

Danaei, G., Finucane, M.M., Lu, Y., Singh, G.M., Cowan, M.J., Paciorek, C.J., Lin, J.K., Farzadfar, F., Khang, Y.H., Stevens, G.A., *et al.* (2011). National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet *378*, 31-40.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., *et al.* (2013). Diet rapidly and reproducibly alters the human gut microbiome. Nature.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol *72*, 5069-5072.

Duncan, S.H., Hold, G.L., Barcenilla, A., Stewart, C.S., and Flint, H.J. (2002). Roseburia intestinalis sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. Int J Syst Evol Microbiol *52*, 1615-1620.

Duncan, S.H., Lobley, G.E., Holtrop, G., Ince, J., Johnstone, A.M., Louis, P., and Flint, H.J. (2008). Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes (Lond) *32*, 1720-1724.

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

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

Erridge, C., Attina, T., Spickett, C.M., and Webb, D.J. (2007). A high-fat meal induces low-grade endotoxemia: evidence of a novel mechanism of postprandial inflammation. Am J Clin Nutr *86*, 1286-1292.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., *et al.* (2013). The long-term stability of the human gut microbiota. Science *341*, 1237439.

Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.O. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol *3*, 121.

Finegold, S.M., Song, Y., Liu, C., Hecht, D.W., Summanen, P., Kononen, E., and Allen, S.D. (2005). Clostridium clostridioforme: a mixture of three clinically important species. Eur J Clin Microbiol Infect Dis *24*, 319-324.

Friedman, J.M. (2004). Modern science versus the stigma of obesity. Nat Med *10*, 563-569.

Furet, J.P., Kong, L.C., Tap, J., Poitou, C., Basdevant, A., Bouillot, J.L., Mariat, D., Corthier, G., Dore, J., Henegar, C., *et al.* (2010). Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. Diabetes *59*, 3049-3057.

Furusawa, Y., Obata, Y., Fukuda, S., Endo, T.A., Nakato, G., Takahashi, D., Nakanishi, Y., Uetake, C., Kato, K., Kato, T., *et al.* (2013). Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature *504*, 446-450.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355-1359.

Goddard, A.F., Staudinger, B.J., Dowd, S.E., Joshi-Datar, A., Wolcott, R.D., Aitken, M.L., Fligner, C.L., and Singh, P.K. (2012). Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. Proc Natl Acad Sci U S A *109*, 13769-13774.

Graessler, J., Qin, Y., Zhong, H., Zhang, J., Licinio, J., Wong, M.L., Xu, A., Chavakis, T., Bornstein, A.B., Ehrhart-Bornstein, M., *et al.* (2013). Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters. Pharmacogenomics J *13*, 514-522.

Hansson, G.K. (2005). Inflammation, atherosclerosis, and coronary artery disease. N Engl J Med *352*, 1685-1695.

Heinemann, M., Kummel, A., Ruinatscha, R., and Panke, S. (2005). In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. Biotechnol Bioeng *92*, 850-864.

Hennekens, C.H., Buring, J.E., Manson, J.E., Stampfer, M., Rosner, B., Cook, N.R., Belanger, C., LaMotte, F., Gaziano, J.M., Ridker, P.M.*, et al.* (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. N Engl J Med *334*, 1145-1149.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol *28*, 977-982.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T.*, et al.* (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science *331*, 463-467.

Hooper, L.V., and Gordon, J.I. (2001). Commensal host-bacterial relationships in the gut. Science *292*, 1115-1118.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res *17*, 377-386.

Huttenhower, Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, Heather H. Creasy, Ashlee M. Earl, Michael G. FitzGerald, Robert S. Fulton*, et al.* (2012a). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207-214.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S.*, et al.* (2012b). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207-214.

Kalliomaki, M., Collado, M.C., Salminen, S., and Isolauri, E. (2008). Early differences in fecal microbiota composition in children may predict overweight. Am J Clin Nutr *87*, 534-538.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res *32*, D277-280.

Kardinaal, A.F., Kok, F.J., Ringstad, J., Gomez-Aracena, J., Mazaev, V.P., Kohlmeier, L., Martin, B.C., Aro, A., Kark, J.D., Delgado-Rodriguez, M.*, et al.* (1993). Antioxidants in adipose tissue and risk of myocardial infarction: the EURAMIC Study. Lancet *342*, 1379-1384.

Karlsson, F.H., Fak, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., Backhed, F., and Nielsen, J. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun *3*, 1245.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Backhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99-103.

Khaneja, R., Perez-Fons, L., Fakhry, S., Baccigalupi, L., Steiger, S., To, E., Sandmann, G., Dong, T.C., Ricca, E., Fraser, P.D.*, et al.* (2010). Carotenoids found in Bacillus. Journal of Applied Microbiology *108*, 1889-1902.

Koeth, R.A., Wang, Z., Levison, B.S., Buffa, J.A., Org, E., Sheehy, B.T., Britt, E.B., Fu, X., Wu, Y., Li, L.*, et al.* (2013). Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med.

Kohlmeier, L., Kark, J.D., Gomez-Gracia, E., Martin, B.C., Steck, S.E., Kardinaal, A.F., Ringstad, J., Thamm, M., Masaev, V., Riemersma, R.*, et al.* (1997). Lycopene and myocardial infarction risk in the EURAMIC Study. Am J Epidemiol *146*, 618-626.

Kong, L.C., Tap, J., Aron-Wisnewsky, J., Pelloux, V., Basdevant, A., Bouillot, J.L., Zucker, J.D., Dore, J., and Clement, K. (2013). Gut microbiota after gastric bypass in human obesity: increased richness and associations of bacterial genera with adipose tissue genes. Am J Clin Nutr *98*, 16-24.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., and Ley, R.E. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol *9*, e1002863.

Koren, O., Spor, A., Felin, J., Fak, F., Stombaugh, J., Tremaroli, V., Behre, C.J., Knight, R., Fagerberg, B., Ley, R.E.*, et al.* (2011). Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc Natl Acad Sci U S A *108 Suppl 1*, 4592-4598.

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics *25*, 2737-2738.
Kritchevsky, S.B. (1999). beta-Carotene, carotenoids and the prevention of coronary heart disease. J Nutr *129*, 5-8.
Kultima, J.R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D.R., Arumugam, M., Pan, Q., Liu, B., Qin, J.*, et al.* (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One *7*, e47656.
Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V.K., Srivastava, T.P.*, et al.* (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res *14*, 169-181.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.
Larsen, N., Vogensen, F.K., van den Berg, F.W., Nielsen, D.S., Andreasen, A.S., Pedersen, B.K., Al-Soud, W.A., Sorensen, S.J., Hansen, L.H., and Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS One *5*, e9085.
Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S.*, et al.* (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541-546.
Leser, T.D., and Molbak, L. (2009). Better living through microbial action: the benefits of the mammalian gastrointestinal microbiota on the host. Environ Microbiol *11*, 2194-2206.
Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A *102*, 11070-11075.
Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022-1023.
Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen, J., Pang, X., Wei, H.*, et al.* (2008). Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci U S A *105*, 2117-2122.
Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.
Li, Y., Hu, Y., Bolund, L., and Wang, J. (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics *4*, 271-277.
Liou, A.P., Paziuk, M., Luevano, J.M., Jr., Machineni, S., Turnbaugh, P.J., and Kaplan, L.M. (2013). Conserved shifts in the gut microbiota due to gastric bypass reduce host weight and adiposity. Sci Transl Med *5*, 178ra141.
Louis, P., and Flint, H.J. (2009). Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. FEMS Microbiol Lett *294*, 1-8.
Louis, P., Young, P., Holtrop, G., and Flint, H.J. (2010). Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. Environ Microbiol *12*, 304-314.
Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K.T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One *7*, e30087.
Mahowald, M.A., Rey, F.E., Seedorf, H., Turnbaugh, P.J., Fulton, R.S., Wollam, A., Shah, N., Wang, C., Magrini, V., Wilson, R.K.*, et al.* (2009). Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. Proc Natl Acad Sci U S A *106*, 5859-5864.
Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P.*, et al.* (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut *55*, 205-211.
Marco-Sola, S., Sammeth, M., Guigo, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods *9*, 1185-1188.
Maslowski, K.M., Vieira, A.T., Ng, A., Kranich, J., Sierro, F., Yu, D., Schilter, H.C., Rolph, M.S., Mackay, F., Artis, D.*, et al.* (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. Nature *461*, 1282-1286.
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A.*, et al.* (2008). The metagenomics RAST server - a public resource for

the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics *9*, 386.
Morgan, X.C., and Huttenhower, C. (2012). Chapter 12: Human microbiome analysis. PLoS Comput Biol *8*, e1002808.
Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res *40*, e155.
Nelson, K.E., Weinstock, G.M., Highlander, S.K., Worley, K.C., Creasy, H.H., Wortman, J.R., Rusch, D.B., Mitreva, M., Sodergren, E., Chinwalla, A.T.*, et al.* (2010). A catalog of reference genomes from the human microbiome. Science *328*, 994-999.
Noble, D., Mathur, R., Dent, T., Meads, C., and Greenhalgh, T. (2011). Risk models and scores for type 2 diabetes: systematic review. BMJ *343*, d7163.
Oliveira, A.P., Patil, K.R., and Nielsen, J. (2008). Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst Biol *2*, 17.
Ooi, L.G., and Liong, M.T. (2010). Cholesterol-lowering effects of probiotics and prebiotics: a review of in vivo and in vitro findings. Int J Mol Sci *11*, 2499-2522.
Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., and McHardy, A.C. (2011). Taxonomic metagenome sequence assignment with structured output models. Nat Methods *8*, 191-192.
Patil, K.R., and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A *102*, 2685-2689.
Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics *27*, i94-101.
Perez-Fons, L., Steiger, S., Khaneja, R., Bramley, P.M., Cutting, S.M., Sandmann, G., and Fraser, P.D. (2011). Identification and the developmental formation of carotenoid pigments in the yellow/orange Bacillus spore-formers. Biochim Biophys Acta *1811*, 177-185.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res *35*, 7188-7196.
Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T.*, et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59-65.
Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D.*, et al.* (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55-60.
Rajilic-Stojanovic, M., Heilig, H.G., Molenaar, D., Kajander, K., Surakka, A., Smidt, H., and de Vos, W.M. (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol *11*, 1736-1751.
Rajilic-Stojanovic, M., Heilig, H.G., Tims, S., Zoetendal, E.G., and de Vos, W.M. (2012). Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol.
Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R.*, et al.* (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science *341*, 1241214.
Ridlon, J.M., Kang, D.J., and Hylemon, P.B. (2006). Bile salt biotransformations by human intestinal bacteria. J Lipid Res *47*, 241-259.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
Rosenberg, E., DeLong, E.F., and SpringerLink (2013). The prokaryotes: Human microbiology (New York, Springer Reference).
Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen, R.A., Palva, A., and de Vos, W.M. (2010). Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. J Microbiol Methods *81*, 127-134.
Samuel, B.S., Shaito, A., Motoike, T., Rey, F.E., Backhed, F., Manchester, J.K., Hammer, R.E., Williams, S.C., Crowley, J., Yanagisawa, M.*, et al.* (2008). Effects of the gut microbiota on host

adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. Proc Natl Acad Sci U S A *105*, 16767-16772.
Satish Kumar, V., Ferry, J.G., and Maranas, C.D. (2011). Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. BMC Syst Biol *5*, 28.
Savage, D.C. (1977). Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol *31*, 107-133.
Schertzer, J.D., Tamrakar, A.K., Magalhaes, J.G., Pereira, S., Bilan, P.J., Fullerton, M.D., Liu, Z., Steinberg, G.R., Giacca, A., Philpott, D.J., *et al.* (2011). NOD1 Activators Link Innate Immunity to Insulin Resistance. Diabetes *60*, 2206-2215.
Schilling, C.H., and Palsson, B.O. (2000). Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. J Theor Biol *203*, 249-283.
Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., *et al.* (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45-50.
Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol *75*, 7537-7541.
Schwiertz, A., Taras, D., Schafer, K., Beijer, S., Bos, N.A., Donus, C., and Hardt, P.D. (2010). Microbiota and SCFA in lean and overweight healthy subjects. Obesity (Silver Spring) *18*, 190-195.
Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol *12*, R60.
Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods *9*, 811-814.
Sjoberg, F., Nowrouzian, F., Rangel, I., Hannoun, C., Moore, E., Adlerberth, I., and Wold, A.E. (2013). Comparison between terminal-restriction fragment length polymorphism (T-RFLP) and quantitative culture for analysis of infants' gut microbiota. J Microbiol Methods *94*, 37-46.
Sjostrom, L., Lindroos, A.K., Peltonen, M., Torgerson, J., Bouchard, C., Carlsson, B., Dahlgren, S., Larsson, B., Narbro, K., Sjostrom, C.D., *et al.* (2004). Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. N Engl J Med *351*, 2683-2693.
Sjostrom, L., Narbro, K., Sjostrom, C.D., Karason, K., Larsson, B., Wedel, H., Lystig, T., Sullivan, M., Bouchard, C., Carlsson, B., *et al.* (2007). Effects of bariatric surgery on mortality in Swedish obese subjects. N Engl J Med *357*, 741-752.
Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A *103*, 12115-12120.
Sokol, H., Lepage, P., Seksik, P., Dore, J., and Marteau, P. (2007). Molecular comparison of dominant microbiota associated with injured versus healthy mucosa in ulcerative colitis. Gut *56*, 152-154.
Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L.G., Gratadoux, J.J., Blugeon, S., Bridonneau, C., Furet, J.P., Corthier, G., *et al.* (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci U S A *105*, 16731-16736.
Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Magi, R., *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet *42*, 937-948.
Tap, J., Mondot, S., Levenez, F., Pelletier, E., Caron, C., Furet, J.P., Ugarte, E., Munoz-Tamayo, R., Paslier, D.L., Nalin, R., *et al.* (2009). Towards the human intestinal microbiota phylogenetic core. Environ Microbiol *11*, 2574-2584.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., *et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.
Topping, D.L., and Clifton, P.M. (2001). Short-Chain Fatty Acids and Human Colonic Function: Roles of Resistant Starch and Nonstarch Polysaccharides. Physiol Rev *81*, 1031-1064.

Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A.E., Phillippy, A.M., and Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol *14*, R2.

Turnbaugh, P.J., Backhed, F., Fulton, L., and Gordon, J.I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe *3*, 213-223.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P.*, et al.* (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480-484.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature *444*, 1027-1031.

Wall, R., Ross, R.P., Ryan, C.A., Hussey, S., Murphy, B., Fitzgerald, G.F., and Stanton, C. (2009). Role of Gut Microbiota in Early Infant Development. Clinical Medicine Insights: Pediatrics *3*, 45-54.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.M.*, et al.* (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature *472*, 57-63.

Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res *41*, 4378-4391.

Werling, M., Fandriks, L., Bjorklund, P., Maleckas, A., Brandberg, J., Lonroth, H., le Roux, C.W., and Olbers, T. (2013). Long-term results of a randomized clinical trial comparing Roux-en-Y gastric bypass with vertical banded gastroplasty. Br J Surg *100*, 222-230.

White, J.R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol *5*, e1000352.

WHO (2013a). Fact sheet N°311.

WHO (2013b). Fact sheet N°317.

Vijay-Kumar, M., Aitken, J.D., Carvalho, F.A., Cullender, T.C., Mwangi, S., Srinivasan, S., Sitaraman, S.V., Knight, R., Ley, R.E., and Gewirtz, A.T. (2010). Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science *328*, 228-231.

Woese, C.R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.

Wolever, T., Spadafora, P., and Eshuis, H. (1991). Interaction between colonic acetate and propionate in humans. Am J Clin Nutr *53*, 681-687.

Wolever, T.M., Brighenti, F., Royall, D., Jenkins, A.L., and Jenkins, D.J. (1989). Effect of rectal infusion of short chain fatty acids in human subjects. Am J Gastroenterol *84*, 1027-1033.

Vrieze, A., Van Nood, E., Holleman, F., Salojarvi, J., Kootte, R.S., Bartelsman, J.F., Dallinga-Thie, G.M., Ackermans, M.T., Serlie, M.J., Oozeer, R.*, et al.* (2012). Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. Gastroenterology *143*, 913-916 e917.

Wu, G.D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.Y., Keilbaugh, S.A., Bewtra, M., Knights, D., Walters, W.A., Knight, R.*, et al.* (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science.

Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V., and Gordon, J.I. (2003). A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. Science *299*, 2074-2076.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P.*, et al.* (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222-227.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821-829.

Zhang, H., DiBaise, J.K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., Parameswaran, P., Crowell, M.D., Wing, R., Rittmann, B.E.*, et al.* (2009). Human gut microbiota in obesity and after gastric bypass. Proc Natl Acad Sci U S A *106*, 2365-2370.

Zhao, L. (2013). The gut microbiota and obesity: from correlation to causality. Nat Rev Microbiol *11*, 639-647.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res *38*, e132.

Zoetendal, E.G., Raes, J., van den Bogert, B., Arumugam, M., Booijink, C.C., Troost, F.J., Bork, P., Wels, M., de Vos, W.M., and Kleerebezem, M. (2012). The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. ISME J *6*, 1415-1426.

Zupancic, M.L., Cantarel, B.L., Liu, Z., Drabek, E.F., Ryan, K.A., Cirimotich, S., Jones, C., Knight, R., Walters, W.A., Knights, D.*, et al.* (2012). Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. PLoS One *7*, e43052.