Comparative metagenomic analysis of the human intestinal microbiota

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Manimozhiyan Arumugam
aus
Madurai, Indien

Würzburg 2010

Eingereicht am:

Mitglieder der Promotionskommission:

- Vorsitzender: Prof. Dr. Jörg Schultz
- 1. Gutachter: Dr. habil. Peer Bork
- 2. Gutachter: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums: 31. Mai 2010

Doktorurkunde ausgehändigt am:

Erklärung

I hereby declare that my thesis entitled "Comparative metagenomic analysis of the human intestinal microbiota" is the result of my own work.

I did not receive any help or support from commercial consultants.

All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I verify that this thesis has not been submitted as part of another examination process neither in identical nor similar form.

Manimozhiyan Arumugam

To,

my mom, who, for the love of me,
never questioned my quest for science,
and so taught me how to love,
and
my dad, who, for the love of me,
always questioned my quest for science,
and so taught me how to do science.

காக்கைக்கா காகூகை கூகைக்கா காகாக்கை
கோக்குக்கூ காக்கைக்குக் கொக்கொக்க - கைக்கைக்குக்
காக்கைக்குக் கைக்கைக்கா கா.

-காளமேகப் புலவர்

Crow cannot defeat an owl at night; and an owl cannot defeat a crow during the day. To properly rule a kingdom, the king should wait patiently for the right opportunities and use them, like a crane waiting patiently in the water for fish. Otherwise protection from enemies will be beyond the reach of even powerful kingdoms.

-*Kalamegam*, ca. 15th century.

On the virtue of patience and waiting for the right opportunity.

(Written with only one consonant, *k*.)

जो हुआ वह अच्छा हुआ,	*Whatever happened, was good,*
जो हो रहा है, वह अच्छा हो रहा है ।	*Whatever is happening, is good.*
जो होगा, वह भी अच्छा होगा ।	*Whatever will happen, will also be good.*
तुम्हारा क्या गया, जो तुम रोते हो ?	*What did you lose that you cry for?*
तुम क्या लाए थे, जो तुमने खो दिया ?	*What did you bring that could be lost?*
तुमने क्या पैदा किया, जो नष्ट हो गया ?	*What did you create that could be destroyed?*
तुमने जो लिया, यहीं से लिया ;	*Whatever you took, you took from here;*
जो दिया, यहीं पर दिया ;	*Whatever you gave, you gave here;*
जो आज तुम्हारा है,	*What is yours today,*
कल किसी और का था,	*Was someone else's yesterday,*
कल किसी और का होगा ।	*Will be someone else's tomorrow!*
परिवर्तन ही संसार का नियम है ।	*Change is the law of the universe!*

- from a popular summary of the *Bhagavad Gita*

Acknowledgements

I would like to thank my supervisor Peer Bork for the opportunity to do my PhD in his group. He provided a great setting to work in, an interesting topic to work with, and excellent guidance as well as training to progress steadily in that work. He has taught me, both actively and passively, how to see the big picture, how to look for interesting questions in biology especially in the data at hand, and how to answer them thoroughly. I also thank him for accommodating my "temporal" disabilities and for being accessible almost all the time in spite of his busy schedule.

I thank my Doktorvater Thomas Dandekar, for the opportunity to study at the University of Würzburg as an external student, his guidance and support throughout my PhD. I am grateful for all his efforts to make things very easy for me with only minimal travel to Würzburg, whereas he travelled to be physically present in every TAC meeting in Heidelberg.

I thank my thesis advisory committee members Toby Gibson and Ewan Birney for their guidance, support, constructive criticism and for always being present in the TAC meetings amidst their busy schedules.

I am very thankful to Jeroen Raes who mentored me during the early years of my PhD, taught me most of what I know about metagenomics and was my young Padawan in Badminton☺. Above all, I thank him for being a good friend and the good times we had during conferences around the world.

I thank all the members of the Bork group for their support throughout my PhD and the great atmosphere. Special thanks to the metagenomics team consisting of Jeroen Raes, Takuji Yamada, Daniel Mende and Gabriel Fernandes, who made this thesis possible, and to the new inductee Julien Tap for the helpful microbiological insights into the human gut microbiota.

I thank Ivica Letunic, who sits in the nearest Euclidean space from my desk, for the numerous iTOL hacks (he calls them 'features'), innumerable educational articles, and the involuntary contribution of several Hanutas towards my thesis.

Thanks to Takuji Yamada for takujifying our lives and our graphics, bringing aesthetics into scientific articles including this thesis, and the in-depth knowledge of metabolic pathways that he spreads to all of us.

I thank Alison Waller, Tomas Larsson, Vera van Noort and Daniel Mende for the critical reading of this thesis and constructive comments that significantly improved it. I thank Daniel Mende and Sean Powell for being my linguistic connection to the German society. Special thanks to Daniel for translating the summary of this thesis, and Georg Zeller for translating the curriculum vitae at the end of this thesis. I am indebted to the "dominating" Unreal gang for the pure adrenaline rush during the team-building exercise and the silly fights afterwards – something all of us look forward to every day. Thanks to Peer for keeping this tradition up!

Thanks to Jean Muller, Chris Creevey and Eoghan Harrington for their help with work related to my thesis. Special thanks to Jean for the U2 concert ☺. Thanks to Yan for keeping the computers running all the time!

Special thanks to Nelly for making our lives much easier by taking care of all the administrative work and letting us concentrate on our work.

Thanks to the EMBL community for the amazing, creative and intellectual atmosphere that provides a great setting for a PhD.

I am very grateful to all the members of the MetaHIT consortium who, through the biannual meetings, taught me so much in the field of human gut microbiology. Special thanks to Dusko Ehrlich and Joel Dore of INRA for the wonderful collaborations, fruitful discussions over manuscripts and the exhilarating time during the MetaHIT meetings. Special thanks also to Eric Pelletier of GenoScope, who distributed the MetaHIT Sanger sequence data and was patient to reply to all my queries regarding the data, and other MetaHIT members of GenoScope. Thanks also to Wang Jun and Qin Junjie for the wonderful collaboration that resulted in a nice publication and a chapter in this thesis.

My thanks to Marc Guell from Luis Serrano's group at CRG for the wonderful collaboration on the *Mycoplasma pneumoniae* work, and Stefan Amlacher from Ed Hurt's group in Heidelberg for the fruitful collaboration on *Chaetomium thermophilum*.

I am thankful to Shantha (Shan) for the delightful typical South Asian conversations and the warm sandwiches that saved me on several occasions.

Special thanks to Guillem who unsuspectingly decided to share an apartment with me. Thanks for feeding me by replenishing the constantly disappearing food supplies, for the great parties we threw at the apartment and the memorable monument at the entrance of the apartment.

I thank my hangout buddies Oriol, Goga, Alvaro, Sara, Maree, Tuba and the rest of the gang for the great time we have spent in Unterestrasse and the rest of Heidelberg. Special thanks to Bennie for the time we spent together, not to mention the amazing Lasagne. I thank the Tango gang – Anna, Florian, Vanessa, Bettina and Bennie, for the great Tango moments we had together.

I thank the talented predocs of the year 2006 for the crazy months of the predoc course, several reunions, ad hoc beer sessions, bowling sessions, ice-skating trips, Wii sessions and the rest of all the fun things we did together.

I thank Usha Middlemas at Masala Heidelberg who helped me translocate myself to India for a couple of hours each Saturday.

I thank Paul Flicek and Melissa Norton for the continuous support before and during my PhD, the Christmas dinners and cooking adventures.

I thank my Badminton gang of Thomas Surrey, Candide Hounsou and Margarete Schnorr, and also the surrogate gang of Damian Brunner, Michael Knop, Adrien Neal, Oliver Wichmann and Jacqueline Dreyer, for refreshing my energy once a week.

My dear friend Fay Christodoulou, the DJ stuck in a scientist's body, was one of the very few who did and could fill these years with fun, laughs, surprises and memorable moments from travelling around Europe and India, crashing random parties in Strasbourg, making deals in Barcelona and the amazingly unbelievably unsuccessful movie nights we ran in Heidelberg. For her and her likes, I thank Greece for the mass production of the Greeks. The rest of the Greek mafia, Alexandra, Kallia, Yannis, Tina and Theo left a significant mark in my life and contributed to me adopting Greece and vice versa.

Amoolya and Chris were my family in Heidelberg and were always there for me helping me pull myself together during tough times. Our dinner & movie sessions, Yoga sessions, Falafel dates and just hanging out together made my stay in Heidelberg very special. Without the two of you, my thesis wouldn't exist, and life wouldn't be what it is!

Mohana Ramaratnam and Kannan Ramaswamy had provided me a home away from home in St. Louis when I was gradually turning into a researcher. The scientific, philosophical and silly discussions we had filled my days in St. Louis with joy, especially the moment their son Tanuj took his first step when the three of us were around him. The rest of the St. Louis gang – Abha, Dilip and Kamala – helped build a little India in St. Louis.

I fondly remember my good friend, the late Arvind Kuppan, with whom I had spent several party nights, an amazing trip to Jamaica and a great New Year's Eve. May his soul rest in peace.

I thank Donald Lee and Alessia Banchieri of St. Louis/Columbia/Monza/Hong Kong for the continuous support during the last several years of my life.

I owe a great deal to the fellow members of Wurstbatch with whom I have shared more than 15 years of my life. They have influenced, inspired and encouraged me all through this time.

Several people have indirectly contributed to my PhD by nurturing my scientific appetite before I even started my PhD. My fellow Wurstbatch members provided the competitive yet friendly atmosphere that got me hooked into science. Dr. Gautam strongly suggested that I pursue a career of science instead of programming my life away. Dr. Stephen Scott, my advisor during the Master's degree at the University of Nebraska, introduced me to the academic environment in higher education and taught me to be curious and inquisitive. Michael Brent provided the perfect setting and training at Washington University that transformed me into a computational biologist from someone holding a degree in biology and computer science. Paul Flicek, Evan Keibler, Jeltje van Baren, Randall Brown and Chaochun Wei influenced that transformation in several ways including through home-brews, Schlafly, burgers, videogames and Christmas dinners.

Finally, I would like to thank my family for all the support, care, love and encouragement they have given me throughout my life.

There are so many others who have influenced my life and career in so many different ways. Although I have not listed every single one of you here, know that I am very grateful to each one of you.

Contents

List of Figures

List of Tables

Abbreviations

BMI	Body mass index
bp	Basepairs
CD	Crohn's disease
COG	Clusters of orthologous groups
DNA	Deoxyribonucleic acid
Gb	Gigabases
GFF	General feature format
IBD	Inflammatory bowel disease
Kb	Kilobases
KEGG	Kyoto Encyclopedia of Genes and Genomes
Mb	Megabases
MetaHIT	Metagenomics of the Human Intestinal Tract
NOG	Non-supervised orthologous group
OG	Orthologous group
ORF	Open reading frame
PCA	Principal component analysis
RDP	Ribosomal Database Project
RNA	Ribonucleic acid
SD	Standard deviation
SMASH	Simple metagenomic analysis shell
UC	Ulcerative Colitis

Summary

The human gut is home for thousands of microbes that are important for human life. As most of these cannot be cultivated, metagenomics is an important means to understand this important community. To perform comparative metagenomic analysis of the human gut microbiome, I have developed SMASH (Simple metagenomic analysis shell), a computational pipeline. SMASH can also be used to assemble and analyze single genomes, and has been successfully applied to the bacterium *Mycoplasma pneumoniae* and the fungus *Chaetomium thermophilum*. In the context of the MetaHIT (Metagenomics of the human intestinal tract) consortium our group is participating in, I used SMASH to validate the assembly and to estimate the assembly error rate of 576.7 Gb metagenome sequence obtained using Illumina Solexa technology from fecal DNA of 124 European individuals. I also estimated the completeness of the gene catalogue containing 3.3 million open reading frames obtained from these metagenomes. Finally, I used SMASH to analyze human gut metagenomes of 39 individuals from 6 countries encompassing a wide range of host properties such as age, body mass index and disease states. We find that the variation in the gut microbiome is not continuous but stratified into enterotypes. Enterotypes are complex host-microbial symbiotic states that are not explained by host properties, nutritional habits or possible technical biases. The concept of enterotypes might have far reaching implications, for example, to explain different responses to diet or drug intake. We also find several functional markers in the human gut microbiome that correlate with a number of host properties such as body mass index, highlighting the need for functional analysis and raising hopes for the application of microbial markers as diagnostic or even prognostic tools for microbiota-associated human disorders.

Zusammenfassung

Der menschliche Darm beheimatet tausende Mikroben, die für das menschliche Leben wichtig sind. Da die meisten dieser Mikroben nicht kultivierbar sind, ist „Metagenomics“ ein wichtiges Werkzeug zum Verständnis dieser wichtigen mikrobiellen Gemeinschaft. Um vergleichende Metagenomanalysen durchführen zu können, habe ich das Computerprogramm SMASH (Simple metagenomic analysis shell) entwickelt. SMASH kann auch zur Assemblierung und Analyse von Einzelgenomen benutzt werden und wurde erfolgreich auch das Bakterium *Mycoplasma pneumoniae* und den Pilz *Chaetomium thermophilum* angewandt. Im Zusammenhang mit der Beteiligung unserer Arbeitsgruppe am MetaHIT (Metagenomics of the human intestinal tract) Konsortium, habe ich SMASH benutzt um die Assemblierung zu validieren und die Fehlerrate der Assemblierung von 576.7 Gb Metagenomsequenzen, die mit der Illumina Solexa Technologie aus der fäkalen DNS von 124 europäischen Personen gewonnen wurde, zu bestimmen. Des Weiteren habe ich die Vollständigkeit des Genkatalogs dieser Metagenome, der 3.3 Millionen offene Leserahmen enthält, geschätzt. Zuletzt habe ich SMASH benutzt um die Darmmetagenome von 39 Personen aus 6 Ländern zu analysieren. Hauptergebnis dieser Analyse war, dass die Variation der Darmmikrobiota nicht kontinuierlich ist. Anstatt dessen fanden wir so genannte Enterotypen. Enterotypen sind komplexe Zustände der Symbiose zwischen Wirt und Mikroben, die sich nicht durch Wirteigenschaften, wie Alter, Body-Mass-Index, Erkrankungen und Ernährungseigenschaften oder ein mögliches technisches Bias erklären lassen. Das Konzept der Enterotypen könnte weitgehende Folgen haben. Diese könnten zum Beispiel die unterschiedlichen Reaktionen auf Diäten oder Medikamenteneinahmen erklären. Weiterhin konnten wir eine Anzahl an Markern im menschlichen Darmmikrobiome finden, die mit unterschiedlichen Wirtseigenschaften wie dem Body-Mass-Index korrelieren. Dies hebt die Wichtigkeit dieser Analysemethode hervor und erweckt Hoffnungen auf Anwendung mikrobieller Marker als diagnostisches oder sogar prognostisches Werkzeug für menschliche Erkrankungen in denen das Mikrobiom eine Rolle spielt.

Chapter 1

Introduction

"Life would not long remain possible in the absence of microbes."

Louis Pasteur, ca. 1880

1.1 Human gut microbiome

The adult human body consists of 10^{13} human cells on average. However it harbors ten times as many microbial cells – symbionts, commensals and pathogens, collectively called the human microbiota[1]. Nobel laureate Joshua Lederberg coined the term microbiome to address them – "to signify the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space"[2]. He also states, quite accurately, that they "have been all but ignored as determinants of health and disease" – although the human microbiome has been studied for several years now, its implications to and associations with human diseases have been ignored altogether until recently. Although scientists have been realizing the influence of these microbes in human health, development and immunity more and more recently, they have barely understood how they exert this influence. One important reason for this lack of knowledge is that most of these microbes have never been cultivated since they require specific environments to survive and it is hard to reproduce such environments in the laboratory.

A vast majority of the human microbiota resides in the human intestinal tract[1], reaching approximately 1.5kg of biomass coming from 500 to 1000 species[3]. The microbial community of the gut has a pronounced impact on human life[4]. Gut microbiome breaks down indigestible large polysaccharides including resistant starches, cellulose, hemicelluloses, pectins and gums and ferments them to produce short chain fatty acids such as acetate, propionate and butyrate, all of which have important functions in human physiology[4]. It degrades oxalate, found in a variety of food and drinks such as tea, coffee, chocolate, fruits and vegetables, which plays an important role in the formation of calcium oxalate kidney stones[5]. It synthesizes certain vitamins including vitamins B and K[6]. It continuously stimulates the maturation of cells secreting immunoglobulin IgA, which serve as the first line of defense against foreign antigens at the mucosal membranes[7]. Postnatal colonization of the gut by the microbes and subsequent microbial contact is also thought to play an important role in

preventing allergic diseases[3,7]. The non-pathogenic resident microbes serve as a protective barrier by adhering to the mucosal lining and preventing attachment and entry of entero-invasive bacteria as well as by competing for nutrients and consuming all resources thereby preventing intrusion by pathogenic bacteria[4]. Thus the gut microbiota provides us metabolic traits that we do not possess, maintains an active immune system in the mucosal interfaces with external environment and serves as a protective barrier against potential pathogens.

1.2 Characterizing the human gut microbiome

Various studies have characterized the human gut microbiome using the 16S ribosomal RNA gene, a universally present microbial marker gene with highly conserved regions, which was used to establish prokaryotic phylogeny[8]. Together, these studies have revealed the species diversity of the gut microbiota within and between individuals[9-12] and have lead to a general consensus about the phylum level composition that more than 90% of the gut microbes belong to two phyla – Bacteroidetes and Firmicutes[9,11,13-14]. The balance between these two phyla has been controversially associated with obesity in human and mice[15-17]. In mouse models, the relative abundance of these two phyla has a direct impact on the energy harvested from diet and the microbiomes of obese mice have a higher harvest capacity[18]. Changes in the composition of the gut microbiota have also been associated with other non-infectious diseases such as inflammatory bowel diseases (IBD)[19-20], diabetes and autism[21]. Although these 16S rRNA gene based studies have generated a wealth of knowledge on species present in the gut and have suggested correlations between the microbiome and some human diseases, there is still a dearth of knowledge concerning the metabolic potential of these species and the molecular mechanisms of host-microbial interactions underpinning the aforementioned correlations. For example, the 16S rRNA gene only reveals the species it belongs to, and the metabolic potential must then be derived from the gene repertoire of the species, usually from the genome sequence. However, only a fraction of the estimated 500-1000 species living in the gut have been sequenced completely – 35 species in the STRING database[22] are gut-associated (Supplementary Figure C-2), and as of May 2009, GenBank contains complete genome sequences of a mere 38 gut-associated species[23]. A vast majority of the remaining gut species cannot be cultivated in a laboratory by standard procedures, which is a

requirement for genome sequencing. Therefore culture-free characterization of the phylogenetic and functional (gene) repertoire of the microbiome, e.g. using metagenomic approaches, is required to understand the host-microbial crosstalk in the human gut environment.

1.3 Metagenomics of the human gut microbiome

Metagenomics (or environmental genomics) allows culture-free characterization of natural or host-associated microbial communities by capturing a genomic snapshot of the microbial environment. This powerful tool helps us to observe and analyze microbes in their natural habitat, including the ones that are hard to culture by standard methods. It enables us to understand the structure and dynamics of the microbial community of an environment, the environmental factors and pressures that are shaping this community, and the response from the community to these forces. Early metagenomic studies characterized the microbial communities of a specific environment, e.g., by reconstructing genomes of dominant species in an acidic mine environment[24], and identifying 148 previously unidentified bacterial phylotypes as well as 1.2 million previously unknown genes in seawater samples[25]. Later studies compared different environments and identified habitat specific fingerprints of gene content in terrestrial and marine microbial communities[26] and metabolic footprints that co-varied with combination of environmental variables[27]. Comparative metagenomics enables us to understand the different ways in which (1) same or closely related microbes respond to different factors and forces in different environments, and (2) different microbes respond to same forces from one environment.

The first metagenomic analysis of the human gut microbiome of two healthy American individuals found a significant enrichment of genes involved in the metabolism of glycans, amino acids and xenobiotics; methanogenesis; and biosynthesis of vitamins and isoprenoids[28], confirming the significant role played by the microbiome. A comparative metagenomic analysis of 13 Japanese individuals and two American individuals (mentioned earlier) identified 237 known gene families that were commonly enriched in adult gut microbiome and revealed a high compositional complexity of *Bacteroides*, a dominant genus in human gut from the phylum Bacteroidetes[29]. Neither of these studies identified correlations between host properties and the gut

microbiome. Another comparative metagenomic analysis of six twin pairs and their mothers assessed the impact of genotype and shared early environmental exposure on the gut microbiome and found a comparable degree of co-variation between monozygotic and dizygotic twin pairs[30]. It also revealed a core microbiome at the gene level and hypothesized that deviations from that core were associated with different physiological states of host adiposity (obese vs. lean). Taking these approaches further, the role of gut microbiota in diseases (which they are thought to be associated with) can be better investigated by comparing the microbial dynamics in the gut environment of healthy individuals and patients. Systematic studies carefully designed to elucidate these dynamics will lead to an improved understanding of the human-microbial interactions and no doubt will improve human health and well-being.

The Metagenomics of the Human Intestinal Tract (MetaHIT) consortium is a Europe-wide collaborative effort that aims at understanding the interactions between the human gut microbiota and their hosts, particularly in the context of two diseases – obesity and inflammatory bowel diseases (IBD). The consortium has generated high quality gut metagenome sequences using Sanger sequencing technology from fecal samples of eight individuals (healthy as well as with obesity and IBD conditions) as a resource to understand these interactions. It also generates a set of reference genes and genomes of the intestinal microbes and tools for identifying correlations between gut microbiota and the two diseases. P. Bork's group at EMBL is the designated Data Analysis and Coordination Center (DACC) for this project and we perform the bioinformatic analysis of the samples. We wanted to analyze the metagenomic sequence data at many levels including genomic, phylogenetic, metabolic and functional levels, and perform comparative analysis of multiple samples to elucidate the correlations between gut microbiota and host properties.

1.4 Need for a comparative metagenomic analysis tool

Comparative analysis of the human gut microbiome requires that the samples that are being compared be treated in the same manner. Different metagenomic studies mentioned in Sections 1.2 and 1.3 have obtained their results and insights using different methods – methods that are consistent in their own rights. However, due to the differences in treatments extended to the samples and the data, these results are

not directly comparable across studies (and hence datasets) to draw comparative conclusions[31-32]. Furthermore, most of these methods are not available to the general scientific community to enable them to either reproduce the results on published datasets, or analyze novel datasets. Some of these methods have still not been adapted to handle data generated by next generation sequencing technologies.

A few publicly available tools allow users to analyze their own metagenomic datasets. MEGAN was the first stand-alone computer program allowing laptop analysis of large metagenomic datasets[33]. Metagenomic DNA read sequences are first compared (outside of MEGAN) against a database of known sequences, such as the NCBI nrdb (non-redundant database). MEGAN then uses the results of this comparison to explore the taxonomical content of a single metagenomic dataset. A later version of MEGAN[34] allowed comparative analysis of the taxonomical content of multiple datasets as well as minimal functional analysis using COGs[35]. The metagenomics RAST server[36] is a web-based tool that produces metabolic and phylogenetic profiles of metagenomic datasets and also allows comparison of these profiles from multiple datasets, but it requires the sequence data to be transferred to the server for the analysis.

None of these tools assembles shorter sequence reads into longer contiguous sequences (contigs). The amount of information gathered from a genomic element correlates with the length of that element[37]. Although short functional signatures, whole domains and single domain genes can be identified by reads that are between 100bp and 1000bp, longer contigs reveal multidomain genes and operons[37], and even enable function prediction of uncharacterized genes by gene neighborhood information[31]. Although very little sequence assembly is possible for a highly complex environment like the soil, a significant fraction of reads from the human gut metagenome assemble into contigs. Thus assembling reads into longer contigs will significantly improve the functional characterization of the gut metagenome.

None of the available tools provides quantitative phylogenetic characterization of metagenomes, e.g., by accounting for variations in genome size and the copy number of the 16S rRNA gene. They do not allow for comparisons of samples using these profiles with statistical insight from bootstrap analysis.

Therefore I developed SMASH (Simple Metagenomic Analysis Shell) – the first comparative metagenomic analysis tool that provides efficient metagenome assembly, quantitative phylogenetic/functional characterization and clustering of samples with bootstrap analysis.

1.5 Thesis outline

One of the main objectives of my PhD was to provide the scientific community with a suite of bioinformatic tools to assemble, annotate and analyze metagenomic sequences. Chapter 3 presents the work involved in creating this pipeline (SMASH), the design principles behind it and its features. SMASH provides a complete suite of tools for assembling raw reads to contigs, predicting genes on contigs, functional annotation of the genes, functional characterization of the metagenome through this annotation, accurate taxonomic assignment of reads using sequence similarity to known sequences or using marker genes such as 16S rRNA, phylogenetic characterization of the dataset using this assignment, comparing multiple datasets using phylogenetic and functional characteristics and clustering the datasets.

The rest of the thesis presents studies that used SMASH to analyze genomes and metagenomes. Although designed for metagenomic analysis, SMASH can also be easily applied to assemble and analyze single genomes. During the early design stage of SMASH, we were involved in a study of the transcriptome of *Mycoplasma pneumoniae*. To generate an accurate genomic tiling array of *M. pneumoniae*, we resequenced the genome and estimated the number of genomic changes (protein coding and non-coding) it accumulated over 10 years of laboratory culturing process, by comparing to the annotated reference sequence[38-39]. This involved distinguishing real genomic changes from changes due to sequencing artifacts, which is an important concern in metagenomics as well. I used SMASH to validate the genome assembly and to analyze the genome sequence variation. This exercise laid the groundwork for designing the strand-specific tiling array used in deducing the transcriptome complexity of *M. pneumoniae*[40] presented in Chapter 4. It also resulted in an efficient sequence assembly component of SMASH. I have also successfully applied SMASH to analyze the genome of a recently evolved thermophilic fungus *Chaetomium thermophilum*[41].

In the context of the MetaHIT consortium, we used the Illumina Genome Analyzer (GA) technology to carry out deep sequencing of total DNA from fecal samples of 124 European adults (of Nordic and Mediterranean origins). We generated 576.7 Gb of sequence, the largest metagenomic dataset to-date and almost 200 times more than any previous study on gut microbiome, assembled them into contigs and predicted 3.3 million unique open reading frames. I used SMASH to validate the assembly of these metagenomes, to estimate the assembly error rates and to estimate the completeness of the gene catalogue by comparing to 89 frequent gut microbial genomes. This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut and indicates that many bacterial species are shared by different individuals. The results of this study are presented in Chapter 5. Our results also show that short-read metagenomic sequencing can be used for global characterization of the genetic potential of ecologically complex environments.

By the time the MetaHIT consortium reached the goal of generating high quality human gut metagenome sequences from eight individuals, there were published metagenomes from 13 Japanese[29] and four American[28,30] individuals. Meanwhile, members of the MetaHIT consortium generated metagenome sequences from eight French individuals for an obesity study (MicroObes) and six Italian individuals for studying the microbiota of the elderly (MicroAge), which they contributed for what transpired to be the comparative metagenomic analysis of the human gut microbiomes of 39 individuals from 6 countries. Using SMASH to analyze these datasets, I found that the variation in the gut microbiota is stratified and not continuous. I identified several robust clusters of gut metagenomes (hereafter referred to as enterotypes) that are not explained by host properties, nutritional habits or various possible technical biases. We also found several functional markers in the human gut microbiome that correlate with a number of host properties such as body mass index (BMI), raising hopes for the application of microbial markers as diagnostic or even prognostic tools for microbiota-associated human disorders. The results of this study are presented in Chapter 6. This study also introduces several novel conceptual and methodological advances (details of which are presented in Section 2.2) that should be useful for numerous other environmental microbial studies to come in the near future.

1.6 Further reading

For a more detailed introduction to metagenomics and human gut microbiome beyond what is presented here, readers are referred to reviews on metagenomics and environmental sequencing[42-44], insightful reviews and commentaries on the microbial ecology of the human gut and the impact of microbes on human health[1,3,6,45-47] and practical treatments on how to use metagenomics as a tool to understand the microbial community at different levels[32,37,48-49].

Chapter 2

Methods

2.1 Establishing the gut microbial reference gene set

2.1.1 Human fecal sample collection

Danish individuals were from the Inter-99 cohort[50], varying in phenotypes according to BMI and status towards obesity/diabetes, whereas Spanish individuals were either healthy controls or patients with chronic inflammatory bowel diseases (Crohn's disease or ulcerative colitis) in clinical remission. Patients and healthy controls were asked to provide a frozen stool sample. Fresh stool samples were obtained at home, and samples were immediately frozen by storing them in their home freezer. Frozen samples were delivered to the Hospital using insulating polystyrene foam containers, and then they were stored at −80 °C until analysis.

2.1.2 DNA extraction

A frozen aliquot (200 mg) of each fecal sample was suspended in 250 μl of guanidine thiocyanate, 0.1 M Tris (pH 7.5) and 40 μl of 10% N-lauroyl sarcosine. Then, DNA extraction was conducted as previously described[19]. The DNA concentration and its molecular size were estimated by nanodrop (Thermo Scientific) and agarose gel electrophoresis.

2.1.3 DNA library construction and sequencing

DNA library preparation followed the manufacturer's instruction (Illumina). We used the same workflow as described elsewhere to perform cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturization and hybridization of the sequencing primers. The base-calling pipeline (version IlluminaPipeline-0.3) was used to process the raw fluorescent images and call sequences. We constructed one library (clone insert size 200 bp) for each of the first 15 samples, and two libraries with different clone insert sizes (135 bp and 400 bp) for each of the remaining 109 samples for validation of experimental reproducibility.

To estimate the optimal return between the generation of novel sequence and sequencing depth, we aligned the Illumina GA reads from samples MH0006 and MH0012 onto 468,335 Sanger reads totalling to 311.7 Mb generated from the same two samples (156.9 and 154.7 Mb, respectively, Supplementary Table 2), using the Short Oligonucleotide Alignment Program (SOAP)[51] and a match requirement of 95%

sequence identity. With about 4 Gb of Illumina sequence, 94% and 89% of the Sanger reads (for MH0006 and MH0012, respectively) were covered. Further extensive sequencing, to 12.6 and 16.6 Gb for MH0006 and MH0012, respectively, brought only a moderate increase of coverage to about 95% (Supplementary Figure B-1). More than 90% of the Sanger reads were covered by the Illumina sequences to a very high and uniform level (Supplementary Figure B-2), indicating that there is little or no bias in the Illumina GA sequence. As expected, a large proportion of Illumina sequences (57% and 74% for M0006 and M0012, respectively) was novel and could not be mapped onto the Sanger reads. This fraction was similar at the 4 and 12–16 Gb sequencing levels, confirming that most of the novelty was captured already at 4 Gb. We generated 35.4–97.6 million reads for the remaining 122 samples, with an average of 62.5 million reads. Sequencing read length of the first batch of 15 samples was 44 bp and the second batch was 75 bp.

2.1.4 Public data used

The sequenced bacteria genomes (totally 806 genomes) deposited in GenBank were downloaded from NCBI database (`http://www.ncbi.nlm.nih.gov`) on 10 January 2009. The known human gut bacteria genome sequences were downloaded from HMP database (`http://www.hmpdacc-resources.org/cgi-bin/hmp_catalog/main.cgi`), GenBank (67 genomes), Washington University in St Louis (85 genomes, version April 2009, `http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/`), and sequenced by the MetaHIT project (17 genomes, version September 2009, `http://www.sanger.ac.uk/pathogens/metahit/`). The other gut metagenome data used in this project include: (1) human gut metagenomic data sequenced from US individuals[30], which was downloaded from NCBI with the accession SRA002775; (2) human gut metagenomic data from Japanese individuals[29], which was downloaded from P. Bork's group at EMBL (`http://www.bork.embl.de`). The integrated NR database we constructed in this study included GenBank NR database (version April 2009) and all genes from the known human gut bacteria genomes.

2.1.5 Illumina GA short reads *de novo* assembly

High-quality short reads of each DNA sample were assembled by the SOAPdenovo assembler[52]. In brief, we first filtered the low abundant sequences from the assembly

according to 17-mer frequencies. The 17-mers with depth less than 5 were screened in front of assembly, for these low-frequency sequences were very unlikely to be assembled, whereas removing them would significantly reduce memory requirement and make assembly feasible in an ordinary supercomputer (512 GB memory in our institute). Then the sequences were processed one by one and the de Bruijn graph data format was used to store the overlap information among the sequences. The overlap paths supported by a single read were unreliable and removed. Short low-depth tips and bubbles that were caused by sequencing errors or genetic variations between microbial strains were trimmed and merged, respectively. Read paths were used to solve the tiny repeats. Finally, we broke the connections at repeat boundaries, and outputted the continuous sequences with unambiguous connections as contigs. The metagenomic special model was chosen, and parameters '−K 21' and '−K 23' were used for 44 bp and 75 bp reads, respectively, to indicate the minimal sequence overlap required.

After *de novo* assembly for each sample independently, we merged all the unassembled reads together and performed assembly for them, as to maximize the usage of data and assemble the microbial genomes that have low frequency in each read set, but have sufficient sequence depth for assembly by putting the data of all samples together.

2.1.6 Validating Illumina contigs using Sanger reads

We used BLASTN (WU-BLAST 2.0) to map Sanger reads from samples MH0006 and MH0012 (156.9 Mb and 154.7 Mb, respectively) to Illumina contigs (single best hit longer than 75 bp and over 95% identity) from the same samples. Each alignment was scanned for breakage of collinearity where both sequences have at least 50 bases left unaligned at one end of the alignment. Each such breakage was considered an assembly error in the Illumina contig at the location where collinearity breaks. Errors within 30 bp from each other were merged. An error was discarded if there exists a Sanger read that agrees with the contig structure for 60 bp on both sides of the error. For comparison, we repeated this on a Newbler2 assembly of 454 Titanium reads from MH0006 (550 Mb reads). Supplementary Figure B-5 shows the number of errors per Mb of assembled Illumina/454 contigs. We estimate 14.12 errors per Mb of contigs for the Illumina assembly, which is comparable to that of the 454 assembly (20.73 per Mb). 98.7% of Illumina contigs that map at least one Sanger read were collinear over

99.55% of the mapped regions, which is comparable to 97.86% of such 454 contigs being collinear over 99.48% of the mapped regions.

2.1.7 Evaluation of human gut microbiome coverage

The Illumina GA reads were aligned against the assembled contigs and known bacteria genomes using SOAP[51] by allowing at most two mismatches in the first 35-bp region and 90% identity over the read sequence. The Roche/454 and Sanger sequencing reads were aligned against the same reference using BLASTN with 1×10^{-8}, over 100 bp alignment length and minimal 90% identity cutoff. Two mismatches were allowed and identity was set 95% over the read sequence when aligned to the GA reads of MH0006 and MH0012 to Sanger reads from the same samples by SOAP.

2.1.8 Gene prediction and construction of the non-redundant gene set

We use MetaGene[53] – which uses di-codon frequencies estimated by the GC content of a given sequence, and predicts a whole range of ORFs based on the anonymous genomic sequences – to find ORFs from the contigs of each of the 124 samples as well as the contigs from the merged assembly.

The predicted ORFs were then aligned to each other using BLAT[54]. A pair of genes with greater than 95% identity and aligned length covered over 90% of the shorter gene was grouped together. The groups sharing genes were then merged, and the longest ORF in each merged group was used to represent the group, and the other members of the group were taken as redundancy. Therefore, we organized the non-redundant gene set from all the predicted genes by excluding the redundancy. Finally, the ORFs with length less than 100 bp were filtered. We translated the ORFs into protein sequences using the NCBI Genetic Code 11.

2.1.9 Identification of genes

To make a balance between identifying low-abundance genes and reducing the error-rate of identification, we explored the impact of the threshold set for read coverage required to identify a gene in individual microbiomes. The number of genes decreased about twice when the number of reads required for identification was increased from 2

to 6, and changed slowly thereafter (Supplementary Figure B-6a). Nevertheless, to include the rare genes into the analysis, we selected the threshold of 2 reads.

2.1.10 Gene taxonomic assignment

Taxonomic assignment of predicted genes was carried out using BLASTP alignment against the integrated NR database. BLASTP alignment hits with *e*-values larger than 1×10^{-5} were filtered, and for each gene the significant matches which were defined by *e*-values $\leq 10 \times$ *e*-value of the top hit were retained to distinguish taxonomic groups. Then we determined the taxonomical level of each gene by the lowest common ancestor (LCA)-based algorithm that was implemented in MEGAN[33]. The LCA-based algorithm assigns genes to taxa in the way that the taxonomical level of the assigned taxon reflects the level of conservation of the gene. For example, if a gene was conserved in many species, it was assigned to the LCA rather than to a species.

2.1.11 Gene functional classification

We used BLASTP to search the protein sequences of the predicted genes in the eggNOG database[55] and KEGG database[56] with *e*-value $\leq 1 \times 10^{-5}$. The genes were annotated as the function of the NOGs or KEGG homologues with lowest *e*-value. The eggNOG database is an integration of the COG and KOG databases. The genes annotated by COG were classified into the 25 COG categories, and genes that were annotated by KEGG were assigned into KEGG pathways.

2.1.12 Determination of minimal gut bacterial genome

The number of non-redundant genes assigned to the eggNOG clusters was normalized by gene length and cluster copy number (Supplementary Figure B-8). The clusters were ranked by normalized gene number and the range that included the clusters encoding essential *Bacillus subtilis* genes was determined, computing the proportion of these clusters among the successive groups of 100 clusters. Analysis of the range gene clusters involved, besides iPath projections, use of KEGG and manual verification of the completeness of the pathways and protein machineries they encode.

2.1.13 Determination of total functional complement and minimal metagenome

We computed the total and shared number of orthologous groups and/or gene families present in random combinations of n individuals (with $n = 2$ to 124, 100 replicates per bin). This analysis was performed on three groups of gene clusters: (1) known eggNOG orthologous groups (that is, those with functional annotation, excluding those in which the terms [Uu]ncharacteri[sz]ed, [Uu]nknown, [Pp]redicted or[Pp]utative occurred); (2) all eggNOG orthologous groups; (3) all orthologous groups plus gene families constructed from remaining genes not assigned to the two above categories. Families were clustered from all-against-all BLASTP results using MCL[57] with an inflation factor of 1.1 and a bit-score cutoff of 60.

2.1.14 Rarefaction analysis

Estimation of total gene richness was done using EstimateS on 100 randomly picked samples due to memory limitations. Because the CV value was >0.5, both chao2 (classic) and ICE richness estimators were calculated and the larger estimate of the two (ICE) was used. The estimate for this sample size was 3,621,646 genes (ICE) whereas S_{obs} (Mao Tau) was 3,090,575 genes, or 85.3%. The ICE estimator curve did not completely saturate, (data not shown) indicating that additional samples will need to be added to achieve a final, conclusive estimate.

2.1.15 Common bacterial core

To eliminate the influence of very similar strains and assess the presence of known microbial species among the individuals of the cohort, we used 650 sequenced bacterial and archaeal genomes as a reference set. The set was composed from 932 publicly available genomes, which were grouped by similarity, using a 90% identity cutoff and the similarity over at least 80% of the length. From each group only the largest genome was used. Illumina reads from 124 individuals were mapped to the set, for species profiling analysis and the genomes originating from the same species (by differing in size >20%) curated by manual inspection and by using the 16S-based clustering when the sequences were available.

2.1.16 Relative abundance of microbial genomes among individuals

We computed the genome coverage by uniquely mapping Illumina reads and normalized it to 1 Gb of sequence, to correct for different sequencing levels in different individuals. The coverage was summed over all species of the non-redundant bacterial genome set for each individual and the proportion of each species relative to the sum calculated.

2.1.17 Species co-existence network

For the 155 species that had genome coverage by the Illumina reads ≥1% in at least one individual we calculated the pair-wise inter-species Pearson correlations between sequencing depths (abundance) throughout the entire cohort of 124 individuals. From the resulting 11,175 inter-species correlations, correlations less than−0.4 or above 0.4 ($n = 342$) were visualized in a graph using Cytoscape[58] displaying the average genome coverage of each species as node size in the graph.

2.2 Comparative metagenomic analysis of 39 human gut microbiomes

A large majority of this comparative analysis was performed using SMASH. These steps, which are now integral parts of SMASH, are marked with an asterisk.

2.2.1 Sample collection and sequencing

Human fecal samples from European individuals were collected and frozen immediately, and DNA was purified as described previously[59]. Sanger sequencing was performed using standard protocols. Shotgun randomly shared DNA libraries were constructed using low copy plasmid (pCNS, 3 kb insert). Terminal clone end sequences were determined using BigDye terminator chemistry and capillary DNA sequencers (3730XL, Applied Biosystems) according to standard protocols established at Genoscope.

2.2.2 Sequence processing*

European samples: Cloning vector and sequencing primer were removed from raw reads after aligning reads to the vector/primer sequences using BLASTN. Reads were quality trimmed by removing bases in either end with phred quality under 15, which translates to 97% accuracy. Lastly, reads shorter than 300bp were removed since the average read length was approximately 650bp, and the objective was to generate high quality reference metagenomes. **American samples:** Sanger reads for two American adult human gut metagenomes[28] were downloaded from NCBI Trace Archive. The vector and sequence trimming coordinates from the trace information were used to remove the cloning vector and sequencing primer. 454-Titanium reads for two American female obese individuals[30] were downloaded from the NCBI Short Read Archive and used without any further processing. **Japanese samples:** By comparing the terminal sequences of all reads with each other using BLASTN, we identified the following unclipped vector/linker sequences in the Japanese samples:

1. 5'- GAGAGCTCCTGCAGGCTAGCTTGCGCAAGGATCCTAGGCCTGAAGCTTGTC - 3'
2. 5'- GCATGGTACCACGCGTACGTAAGCAAGATCTTCCCGGGTGAATTCGTC - 3'

These sequences from the pTS1 cloning vector (Ken Kurokawa and Tetsuya Hayashi, personal communication) were clipped from the 13 Japanese samples using the makeClip program from Forge assembler[60] with default parameters. **Further trimming:** All Sanger reads were finally trimmed for low quality regions in the ends using makeClip with default parameters.

2.2.3 Assembly and gene prediction*

Assembly and gene prediction were performed using the SMASH comparative metagenomics pipeline[61]. To obtain contigs and scaffolds from the reads, we employed SMASH's iterative assembly procedure using Arachne software[62-63]. This procedure iteratively assembles unassembled reads (singletons) from the previous iteration until no more assembly is possible. Protein coding genes were predicted using GeneMark[64-65] (v 2.6p) by the SMASH pipeline. GeneMark uses heuristic Markov models of coding and non-coding sequences to identify coding sequences, and a

* Methods in Section 2.2 that are marked with an asterisk are integral part of the SMASH pipeline.

ribosome binding site model to identify translation start sites. SMASH uses the GC-content based heuristic models (provided with GeneMark software) to predict genes on scaffolds shorter than 200kb as well as unassembled reads, and a self-trained hybrid model using both GC-content and sequence content on scaffolds longer than 200kb.

2.2.4 Phylogenetic annotation*

Phylogenetic annotation of each metagenome sample was performed using the SMASH pipeline as follows.

Reference genome mapping: Sequence reads were aligned to a set of reference microbial genomes[23] obtained from NCBI[66] and other human microbiome sequencing centers[67-70] using BLASTN (WU-BLAST 2.0, default parameters except E=1e-20 Z=4000000000 B=5). Each read was assigned the taxonomy of the highest scoring hit(s) above the similarity threshold for the taxonomic rank (>65% for phylum >85% for genus, >95% for species). Alignments were also required to span over 75bp covering >80% of the read length. Since paired-end reads are from two ends of a cloned DNA fragment, two reads from such a fragment represent only one physical DNA fragment. Hence taxonomy assignments of reads were transferred to the corresponding fragments. The numbers of fragments assigned to each reference genome were counted. (A fragment assigned to N different reference genomes contributes $1/N$ to each genome). These counts were normalized by the sizes of these genomes to obtain the quantitative relative abundance (relative number of individuals) of each genome in the sample. Number of unassigned fragments was normalized by the average genome size in the reference set (3.54Mb) to calculate the approximate abundance of unknown genomes. Phylogenetic abundances at various phylogenetic ranks (species, genus, phylum etc) were calculated by adding the abundances of genomes under that rank.

Assignment through 16S rRNA molecules: 16S rRNA sequences were identified from metagenomics reads using an HMM based algorithm[71]. Reads were phylogenetically classified using the RDP classifier[72], and the genus level assignment was recorded if the read was longer than 250bp and the confidence score was higher than 50%[73].

2.2.5 Functional annotation*

Functional annotation of each sample was performed using the SMASH pipeline. Abundance of each predicted gene from a sample was estimated analogous to the contig coverage in sequence assembly. If $R = \{r\}$ is the set of assembled reads overlapping the locus of predicted gene g in a contig, abundance of g was calculated as

$$abundance(g) = \sum_{r \in R} \frac{base\_overlap(g, r)}{base\_length(g)} \quad (1)$$

Genes on a singleton read thus have an abundance of 1. Predicted proteins were aligned to proteins from the eggNOG v2 database[74] using BLASTP (WU-BLAST 2.0, default parameters except E=1e-5 B=10000) and were assigned to an orthologous group as described elsewhere[26]. From these alignments between the set of predicted proteins $G = \{g\}$ from a sample and the set of eggNOG reference proteins $K = \{k\}$, the abundance of each reference protein k in the sample was calculated as

$$abundance(k) = \sum_{g \in G} \frac{aa\_overlap(k, g) * abundance(g)}{aa\_length(k)} \quad (2)$$

Functional abundances at the OG level were calculated by adding abundances of reference proteins under each OG.

Predicted proteins were also aligned to proteins from Kyoto Encyclopedia of Genes and Genomes (KEGG) database[75] as before. Each protein was assigned orthology to the highest scoring hit(s) with an annotated KEGG orthologous group (KO) and at least one HSP scoring over 60 bits. The abundance of each KEGG protein was calculated as in Equation (2). Functional abundances at KO, KEGG module and KEGG pathway levels were calculated by adding abundances of KEGG proteins under each KO, module and pathway, respectively.

2.2.6 Highly abundant functions from low-abundance microbes*

To identify functions that are predominantly from low-abundance microbes, we estimated the phylogenetic origin of each function, by combining phylogenetic assignment of reads to genera/phyla and functional annotation of genes to orthologous groups, and assigned orthologous groups to genera/phyla through the reads that constitute genes. We then looked for highly abundant functions (among the top 20% = above 80th percentile) that are primarily contributed by low-abundance genera

(<2.5%), and found 109 such orthologous groups in all samples. Since we only chose functions that received more than 50% contribution from such genera, our observations will still be valid even if the unmapped portions of the genes are mapped to their rightful genera.

2.2.7 Clustering*

Genus abundance profiles (phylogenetic) and OG abundance profiles (functional) were normalized to generate probability distributions (called abundance distributions hereafter). We used a probability distribution distance metric[76-77] related to Jensen-Shannon divergence to cluster the samples. The distance $D(a,b)$ between samples a and b is defined as

$$D(a,b) = \sqrt{JSD(p_a, p_b)} \tag{3}$$

where p_a and p_b are the abundance distributions of samples a and b and $JSD(x,y)$ is the Jensen-Shannon divergence between two probability distributions x and y defined as

$$JSD(x,y) = \frac{1}{2}KLD(x,m) + \frac{1}{2}KLD(y,m) \tag{4}$$

where $m = \frac{x+y}{2}$ and $KLD(x,y)$ is the Kullback-Leibler divergence between x and y defined as

$$KLD(x,y) = \sum_i x_i \log\frac{x_i}{y_i} \tag{5}$$

We added a pseudocount of 0.000001 to the abundance distributions and renormalized them to avoid zero in the numerator and/or denominator of equation (5).

To test the robustness of the clusters we obtain, we generated 100 bootstrap replicates of the raw counts (number of templates for phylogenetic comparison and KEGG protein abundance for functional comparison) by resampling (with replacement) the counts using the appropriate probabilities. We calculated abundance profiles from these replicates as mentioned in Sections 2.2.4 and 2.2.5 and generated 100 trees with

the neighbor-joining approach using the neighbor program from phylip package[78]. We then created a consensus tree using the consense program from phylip package.

2.2.8 Principal component analysis

Principal component analysis (PCA) was performed to support the clustering (Section 2.2.7) and to identify drivers for the enterotypes. The analysis was done using R. Prior to the analysis the data was sample size normalized and very low abundant genera / orthologous groups were removed to decrease noise if their average abundance across all samples was below 0.01%. The PCA based on genus abundances was done using standard parameters, while the PCA based on orthologous groups was additionally scaled (the standard deviation of each OG was scaled to 1).

2.2.9 Statistical treatment of over-/under-representation

Over- and underrepresented features (eggNOG and KEGG orthologous groups, KEGG modules, KEGG pathway maps, genera and phyla) were identified using Fisher's exact test on pooled counts depending on the sample groups compared. Correction for multiple testing was done based on the Benjamini-Hochberg False Discovery Rate (corrected p-value <0.05). To avoid artefacts, we only took those features into account that were specifically overrepresented in onlCase studies described in the main text were further manually scrutinized to avoid artefacts.

2.2.10 Correlations with host properties

Correlation analysis between host metadata (Supplementary Table 1) and feature (OG, module, pathway, genus, phylum) frequencies was done as described previously[27]. In short, Spearman pairwise correlations between continuous metadata variables (age, bmi) were calculated and p-values were corrected for multiple testing using Benjamini-Hochberg False Discovery Rate correction. Significant features were used as input for building linear models using stepwise regression (top-down and bottom-up feature selection) based on the Akaike Information Criterion. For categorical metadata, samples were pooled into bins (male/female, obese/lean, specific nationality/rest) and treated as in Section 2.2.9. For nationality analysis, also the general variability of features across nations was investigated. For each nationality, we calculated the standard deviation (SD) of investigated features (relative abundance of

OGs, genera) across samples, and compared this to the SD of the distribution of mean relative abundances of each nationality (to measure across-nationality variation). Examples with a across SD/within SD ratio >1 are discussed in the main text.

2.2.11 Estimating sequence similarity barriers across phylogenetic ranks

We estimated the phylogenetic composition of the samples using sequence similarity between metagenomics reads and a reference genome set consisting 1152 completely sequenced microbial genomes. Since there are no established sequence similarity barriers to differentiate genomes from different phylogenetic ranks, we estimated the sequence similarity cutoffs to safely assign a sequence to either a genus or a phylum. For this purpose, we retrieved 40 single copy marker genes[79] from a subset of 835 genomes (after removing some redundancy at species level) and generated 40 sets of pairwise alignments using BLASTN. These marker genes are highly representative of the reference genome set, and hence of at least the sequenced microbial species, since 801 of the 853 genomes (94.6%) contained at least 38 out of the 40 genes. Figure 2-1a shows the distribution of sequence similarity levels between genomes from the same phylum (green) and different phyla (red). Figure 2-1b shows the same distribution at genus level. We estimated the false positive rates at different similarity levels (Figure 2-2) and chose a sequence similarity cutoff of >65% to assign a read to the phylum of the best hit, and >85% to assign it to the genus to minimize false positive assignments. This is a rather conservative cutoff, since the marker genes are among the genes under the highest levels of selective constraint[79].

2.2.12 Non bacterial DNA content

For this part of the analysis only, we counted the number of reads per samples, since this is not used in a quantitative manner for comparative analysis. This is different from the quantitative abundance estimation by counting the mate-paired reads as a single DNA fragment, as in Sections 2.2.4 and 2.2.5. **Eukaryotic DNA:** Sequence reads were aligned against human genome assembly hg18 obtained from UCSC Genome Browser[80] using BLAT[54] (gfClient v 31, default parameters). Possible human DNA sequences were identified with a very low alignment threshold to maximize true

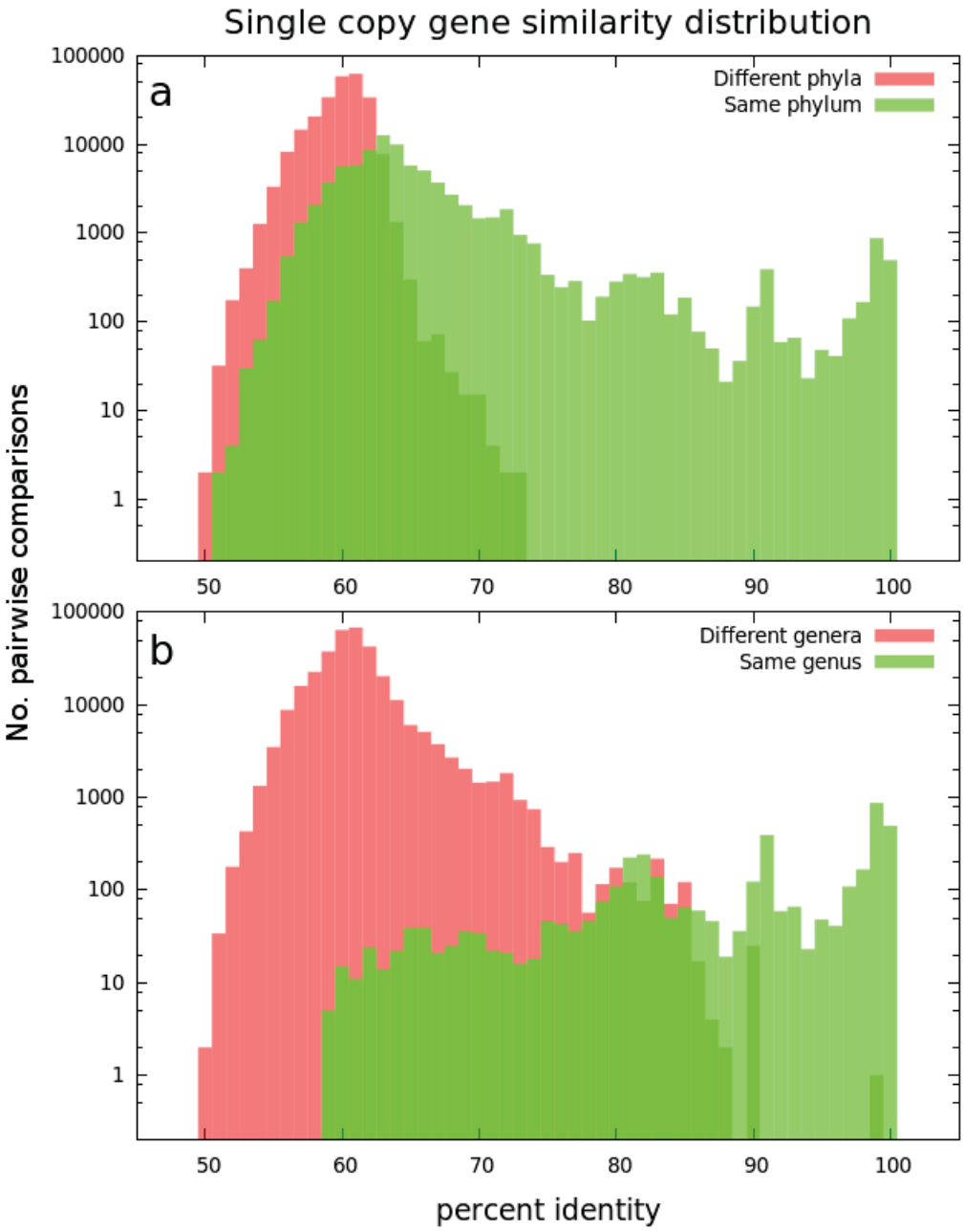


Figure 2-1. Establishing DNA sequence similarity thresholds for phylum and genus levels.

Sequence similarity distributions of pairwise alignments of 40 universal single copy genes from 835 microbial genomes reveal that (a) 65% DNA sequence similarity threshold accurately groups genomes within the same phylum (with 31.1% sensitivity and 0.77% false positive rate) and (b) 85% threshold accurately groups genomes within the same genera (with 63.23% sensitivity and 5.1% false positive rate). Pairwise comparisons of genomes within the same phylum (genus) are colored green and different phyla (genera) are colored red.

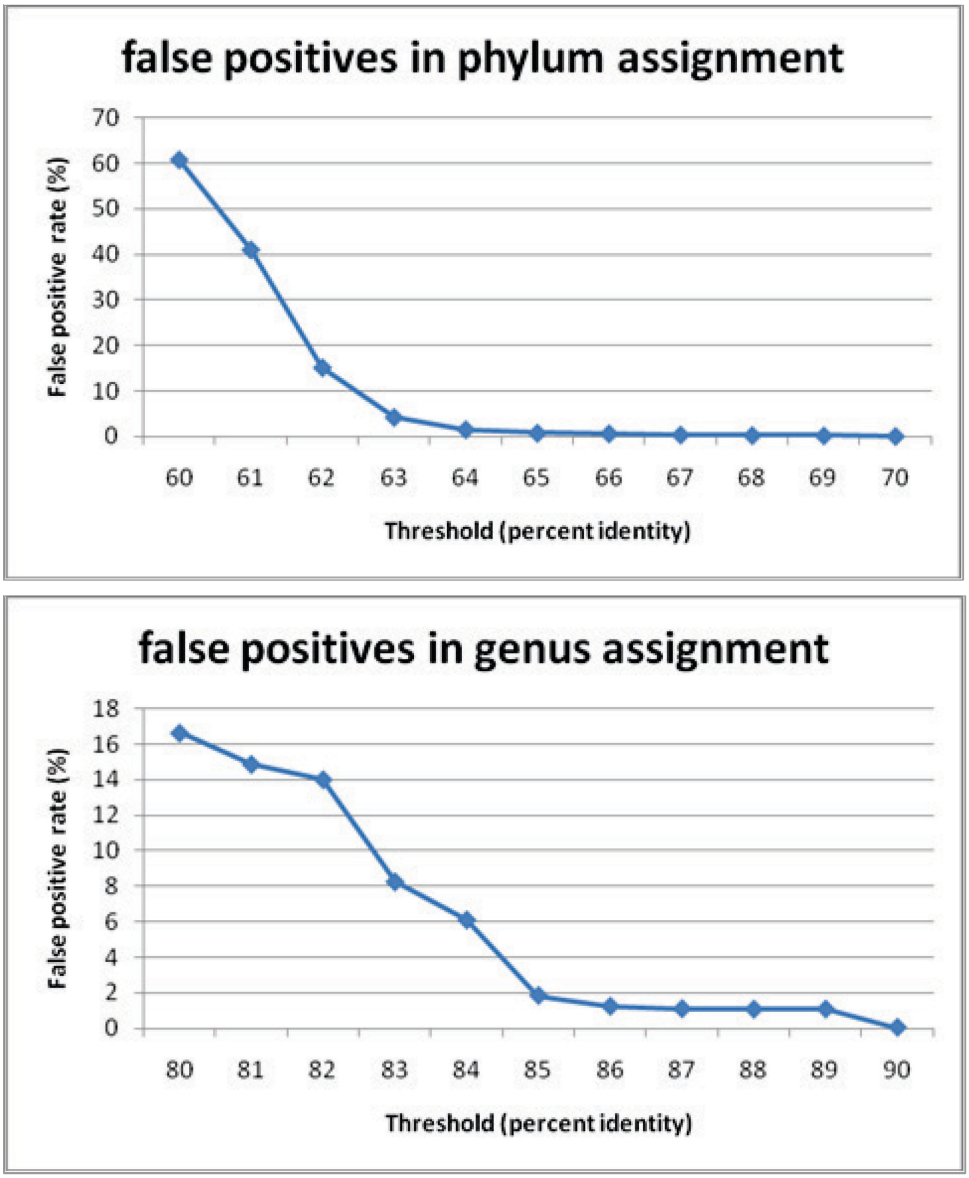


Figure 2-2. False positive rates at the phylum and genus levels estimated by pairwise comparisons of 40 marker genes for different sequence similarity thresholds.

positives and minimize false negatives ('pslFilter -minMatch=50' from the BLAT package), and were removed. Other eukaryotic DNA fraction was estimated by identifying metagenome proteins whose best hit in STRING v8 database comes from a eukaryote. **Prophage sequences:** To assess the fraction of prophage sequences, we performed a BLAST search of all reads of our samples against the ACLAME mobile genetic elements database[81] as well as 2969 viral and 579 phage genomes from NCBI[82-83]. We also estimated the lower bound of the prophage fraction using the

number of these reads that had a significantly better hit to a viral sequence than to bacterial genome.

2.2.13 Deriving enterotypes*

Samples were clustered using genus abundance profiles in Fig. 1b as explained in Section 2.2.7. Enterotypes were derived from the clustering tree structure (Figure 6-2a). We derived four enterotypes (A, B, C and D) by splitting the tree at the branching point with low bootstrap support (< 80). Bootstrap support values just above the leaves of the tree are ignored (e.g., JP-AD-6 and ES-AD-1 are in the same enterotype even though the bootstrap support for their parent node is below 80). For enterotypes A and B, cluster analysis (Figure 6-2) and PCA analysis (Figure 6-3) indicated a more coherent core of these clusters (Acore and Bcore hereafter) as well as some more peripheral samples that gave distinct clusters in the function-based analysis (Aper and Bper hereafter). We thus considered them independently.

2.2.14 Jackknife test for robustness of enterotypes*

To check the robustness of the enterotypes, we performed 50% jack-knife tests on the clusters in Figure 6-2a and Figure 6-2b. We separated the samples into two almost equal-sized sets, and performed the clustering procedure on each half. We then compared the grouping tendency of the reduced sets to the enterotypes in Figure 6-2a. We repeated this five times each for the genus profile and orthologous group profile based clusters.

2.2.15 Independent experimental verification of enterotypes

DNA microarray: A 2.1 million feature Roche NimbleGen microarray targeting a 700,000 subset of potential coding regions (CDS) from gut microbiota was designed based on public data including the most widespread CDS sequenced from the 124 stool samples from an earlier study[84]. The samples were prepared and hybridized according to standard NimbleGen protocols. Data was preprocessed using RMA[85] implementation under the 'oligo' package available in the R statistical programming environment. Species abundance levels were estimated as the difference between the mean of the probe-set signals targeting a given species and the mean of background probe-set signals. The distance matrix was calculated using Kullback–Leibler

divergence using the 'flexmix' R package[86] and the hierarchical cluster was generated and drawn with the 'hclust' R package[87]. **HITChip:** A phylogenetic analysis of the DNA extracts of the samples was performed with the Human Intestinal Tract Chip (HITChip)[88]. This phylogenetic microarray has over 4,800 oligonucleotide probes, which target the 16S rRNA genes of more than 1,100 intestinal bacterial phylotypes. Hybridization and analysis were performed as described before[88]. 10 ng from the fecal DNA extract was used to amplify the 16S rRNA genes with the *T7prom*-Bact-27-for and Uni-1492-rev primers. Subsequently, an *in vitro* transcription and labeling with Cy3 and Cy5 dyes, was performed. Fragmentation of Cy3/Cy5 labeled target mixes was followed by hybridization on the arrays at 60°C for 16h in a rotation oven (Agilent Technologies, Amstelveen, The Netherlands). The slides were washed and dried before scanning. Signal intensity data was obtained from the microarray images using the Agilent Feature Extraction software, version 9.1 (`http://www.agilent.com`). Microarray data normalization and further analysis was performed using a set of R-based scripts (`http://r-project.org`) in combination with a custom designed relational database[88] which operates under the MySQL database management system (`http://www.mysql.com`). Hierarchical clustering of probe profiles was carried out by converting Pearson's product-moment correlation coefficient into a distance (e.g., 1 - r) combined with Ward's minimum variance method.

Chapter 3

SMASH-Community: Simple Metagenomic Analysis Shell for Metagenomic Sequences

Manimozhiyan Arumugam, Eoghan Harrington, Jeroen Raes, Peer Bork

Manuscript in preparation.

SMASH is a stand-alone comparative metagenomic analysis pipeline suitable for data from Sanger and 454 Titanium sequencing technologies. It comes with built-in support for state-of-the-art software programs and can be easily extended to support additional software. SMASH also provides tools to produce easily comprehended visual representation of results. SMASH is available at `http://www.bork.embl.de/software/smash/`.

Introduction

Metagenomics allows the culture-free characterization of natural and host-associated microbial communities and enables understanding of their structure, dynamics and functionality as well as the investigation of the environmental factors that shape them[32,42-43]. Early metagenomic studies characterized the microbial communities of a specific environment, e.g. by reconstructing genomes of dominant species[24], and identifying 148 previously unidentified bacterial phylotypes as well as 1.2 million previously unknown genes[25]. Later studies compared different environments and identified habitat specific fingerprints of gene content[26] and metabolic footprints that co-varied with combination of environmental variables[27]. These later studies ushered the field of comparative metagenomics that helps us to understand how microbial communities respond to environmental pressure. However, different groups have used different sets of methods (experimental procedures as well as computational pipelines) to obtain their results, restricting direct comparison of results from multiple studies[89]. Most of these pipelines are not publicly available to either reproduce the results on published datasets, or analyze novel datasets. With exponentially increasing amount of data from metagenomic studies, and even more large scale studies planned and in progress[90], there is an imminent demand for a publicly available pipeline to analyze metagenomic datasets from different environments, obtained using different technologies. Here we present SMASH, a metagenomic analysis pipeline designed to leverage the power of comparative metagenomics (see Figure 3-1 for an overview). SMASH was developed as part of the MetaHIT study and has been used in analyzing 39 human gut metagenome samples from published[28-30] and unpublished studies (See Chapter 6).

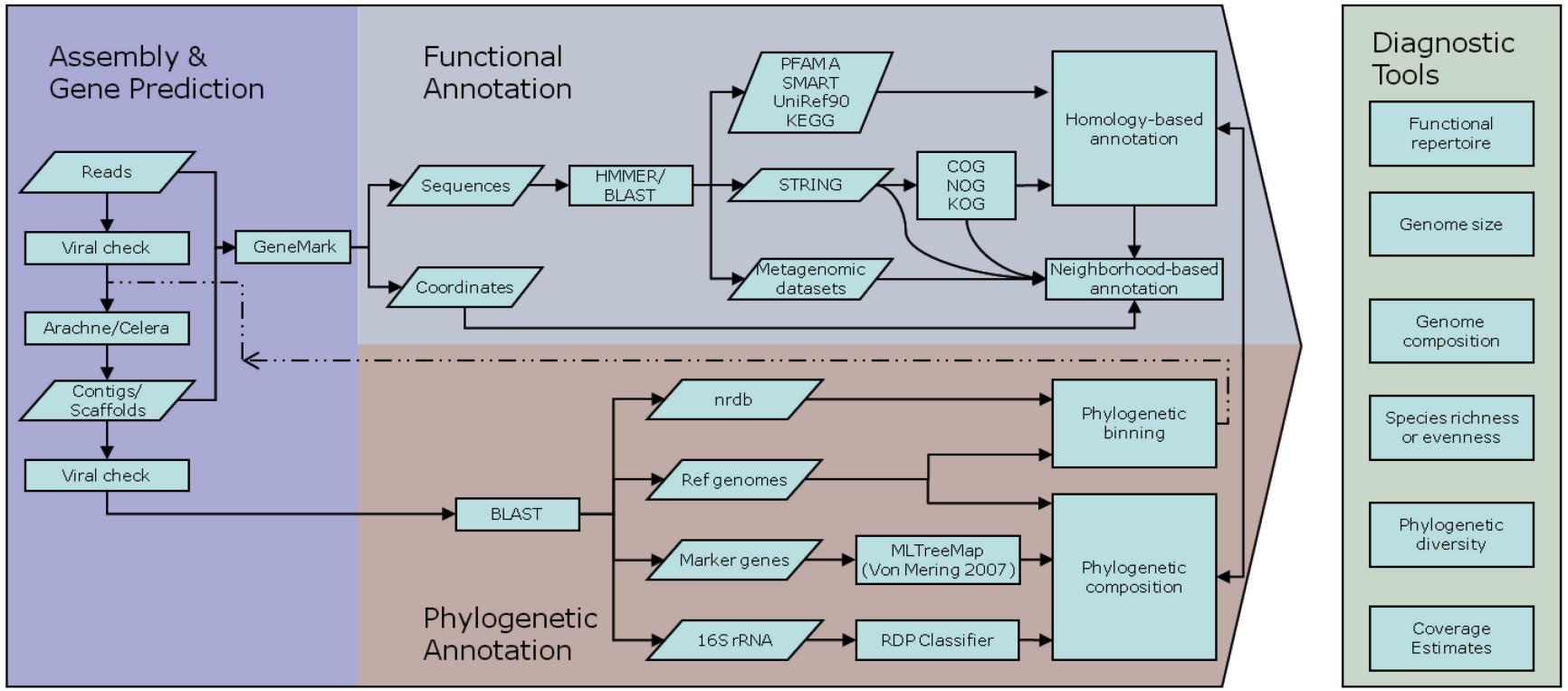


Figure 3-1. Schematic design of SMASH.

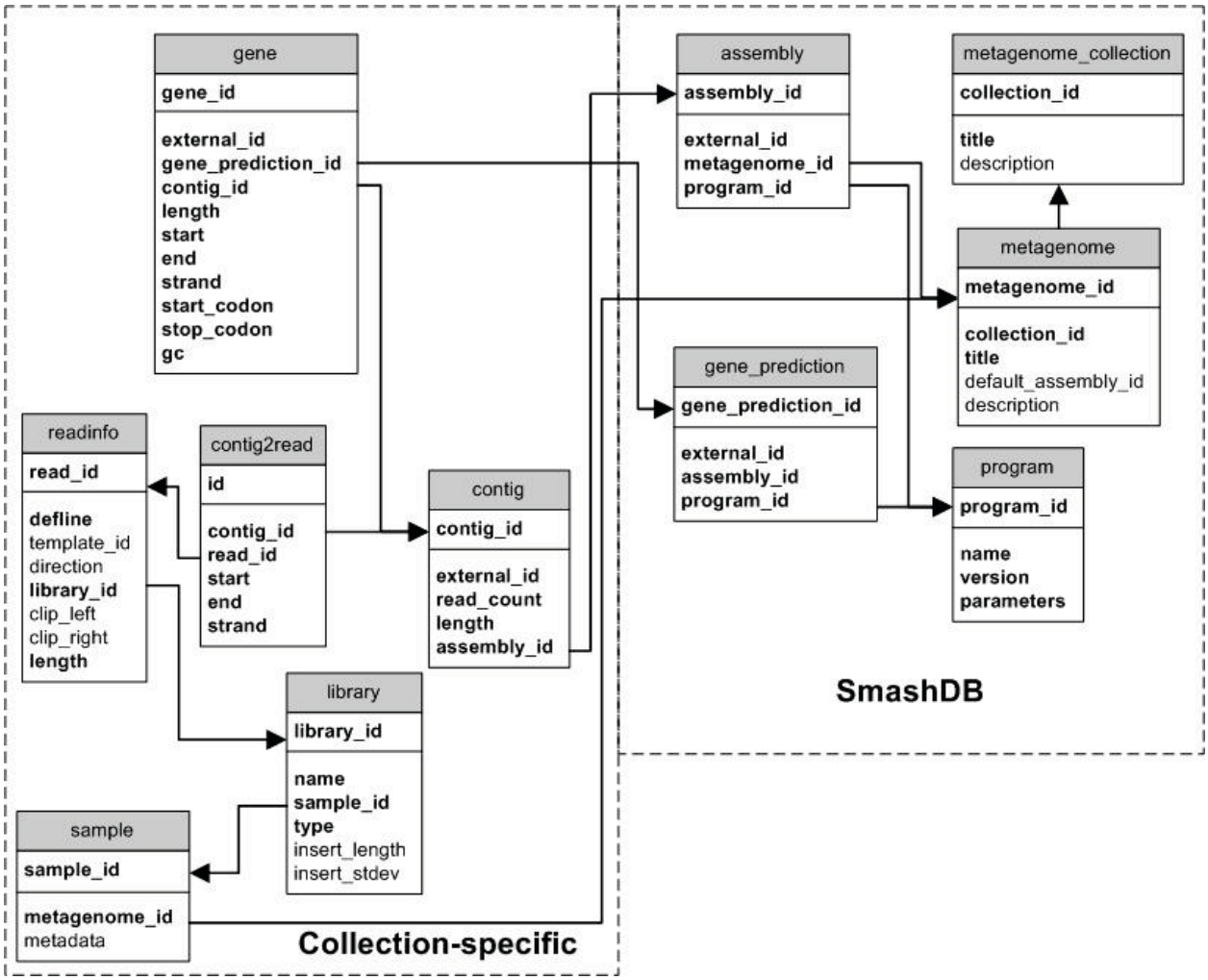


Figure 3-2. Database schema of SMASH.

A higher level database (SmashDB, right) contains sample metadata about each metagenome, and each collection of metagenomes forms its own database.

Pipeline features

The schematic design of SMASH is shown in Figure 3-1. SMASH is written in Perl with well-defined modular architecture and intermodular interfaces between the components shown in Figure 3-1. All the information except the sequences is stored in a MySQL database (Figure 3-2). Each task in metagenomics analysis, such as sequence assembly or gene prediction, is implemented in SMASH as a module that is a wrapper around a software program that performs such task. This design of independent modules with common interfaces enables replacement of software programs with better alternatives in the future. SMASH comes with built-in support for popular state-of-the-

art programs that are publicly available, and adding support for additional programs requires minimal effort. These programs must be obtained from the respective sources and installed as specified in the user manual[91]. The common workflow in SMASH is as follows (Please see the user manual[91] for details):

Data processing, assembly and gene prediction: Sequence data generated from a sample should first be stored in the SMASH repository. Sample and sequence metadata are loaded into the MySQL database while sequences themselves are stored in fasta format. SMASH can analyze sequence data generated by Sanger and 454 Titanium sequencing technologies. SMASH has built-in support for Arachne assembler[63] for Sanger sequences and Celera assembler[92] for Sanger and/or 454 Titanium sequences. After a successful assembly, assembly coordinates are loaded into the database and assembled contigs and scaffolds are stored in fasta format. SMASH predicts protein coding genes on these contigs using GeneMark[65] or MetaGene[53]. Data from external assembly or gene prediction (performed outside of SMASH) can also be loaded into the repository using GFF files[93] following SMASH-specific format.

Phylogenetic and functional annotation: Samples can be phylogenetically characterized (a) using best BLAST hits of sequence reads to microbial reference genomes, and (b) by identifying reads containing 16S rDNA sequences[71] and classifying them using RDP classifier[72] (see Section 2.2.4 for more details). Predicted genes are annotated through BLAST-based homology to orthologous groups from eggNOG[74] database and KEGG pathway database[75] (see Section 2.2.5 for more details).

Analysis tools and visualization: SMASH includes scripts for downstream analysis of datasets. They can generate easily comprehended visual tree-based representations of the results through the batch access API of the interactive Tree of Life web-tool (iTOL)[94]. For example, SMASH can generate the phylogenetic or functional profiles of one or more metagenomes. These quantitative profiles (relative abundances) are calculated more accurately at the read level and are corrected for sample size, copy number variation of the 16S rRNA gene and genome size variation. These profiles can be exported into tables that can be manipulated by the R programming environment, or could be uploaded and browsed on the iTOL website. SMASH also downloads these profiles as images and provides a useful visual representation by slightly modifying the original images from iTOL. It can also compare multiple metagenomes using these

profiles, cluster them based on a relative entropy-based distance measure suitable for comparing such quantitative profiles, perform bootstrap analysis of the clustering and generate visual representation of the clustering results.

SMASH was used in analyzing the human gut microbiomes of 39 individuals, as reported in Chapter 3. Details of the analysis tools available in SMASH and their application to these samples are explained in Section 2.2. Supplementary Figure C-3a shows a sample visual representation of the phylogenetic profiles of two human gut metagenomes generated using Sanger and 454-Titanium technologies (see Chapter 6 for more details). This provides a preliminary glimpse of the species richness and evenness of the microbial communities at the given sampling depth, while highlighting the differences in phylogenetic composition between the two samples. It also shows that the compositions of the metagenomes obtained from the same individual using two different technologies are similar. Supplementary Figure C-3b shows the relative abundance of the 50 most abundant eggNOG orthologous groups in the same samples. Figure 6-2 shows the clustering of 39 datasets based on 16S phylogenetic profiles, reference genome based phylogenetic profiles and functional profiles with bootstrap analysis. Such visual representations of comparative analysis of metagenomic data provide a better understanding of the data and easier interpretations of the results.

Limitations and future work

Several new research initiatives, especially the ones exploring the human microbiome, plan to use next generation high throughput sequencing technologies such as Roche 454 and Illumina Solexa technologies. For example, Chapter 5 discusses the human gut microbial gene catalogue generated from Illumina Solexa sequences from 124 individuals. With even more next generation sequencing technologies promising to be available in the near future, metagenomic sequencing will see a new era. One limitation of SMASH is that currently it can only analyze raw data produced using Sanger and 454-Titanium sequencing technologies. This is primarily due to the efforts it takes to evaluate multiple sequence assembly software and choosing one that best fits our needs. As far as we know, a computational metagenomic analysis pipeline that supports all currently available sequencing technologies does not exist. However, assembled sequences (contigs) from any technology can be analyzed by SMASH. While

we work on adding support for new technologies, this is a viable option for analyzing metagenomes generated using other next generation sequencing technologies.

Chapter 4

Transcriptome complexity in a genome-reduced bacterium

Marc Güell, Vera van Noort, Eva Yus, Wei-Hua Chen, Justine Leigh-Bell, Konstantinos Michalodimitrakis, Takuji Yamada, Manimozhiyan Arumugam, Tobias Doerks, Sebastian Kühner, Michaela Rode, Mikita Suyama, Sabine Schmidt, Anne-Claude Gavin, Peer Bork, Luis Serrano

Science **326**(5957):1268-1271. doi: 10.1126/science.1176951

To study basic principles of transcriptome organization in bacteria, we analyzed one of the smallest self-replicating organisms, *Mycoplasma pneumoniae*. We combined strand-specific tiling arrays, complemented by transcriptome sequencing, with more than 252 spotted arrays. We detected 117 previously undescribed, mostly noncoding transcripts, 89 of them in antisense configuration to known genes. We identified 341 operons, of which 139 are polycistronic; almost half of the latter show decaying expression in a staircase-like manner. Under various conditions, operons could be divided into 447 smaller transcriptional units, resulting in many alternative transcripts. Frequent antisense transcripts, alternative transcripts, and multiple regulators per gene imply a highly dynamic transcriptome, more similar to that of eukaryotes than previously thought.

Although large-scale gene expression studies have been reported for various bacteria[95-101], comprehensive strand-specific data sets are still missing, limiting our understanding of operon structure and regulation. Similarly, the number of classified non-coding RNAs in bacteria has recently been expanded[102], but a complete and unbiased repertoire is still not available. To obtain a blueprint of bacterial transcription, we combined the robustness and versatility of spotted arrays (62 independent conditions and 252 array experiments[103]), the superior resolution of strand specific tiling arrays (Figure 4-1A)(designed after genome re-sequencing, Supplementary Table A-1) and the mapping capacity of RNA deep sequencing (Direct Strand Specific Sequencing, DSSS)(Figure 4-1A and Supplementary Figure A-1) in one of the smallest bacteria that can live outside a host cell, *Mycoplasma pneumoniae*, with annotated 689 protein coding genes and 44 non coding RNAs (ncRNAs).

Considering DSSS under reference conditions[103] and 43 tiling arrays from four time series (growth-curve, heat shock, DNA damage and cell cycle arrest; Supplementary Table A-8), we observed the expression of all genes. Using a segmentation algorithm for the tiling arrays[103], we identified an additional 117 regions with no previous annotation (Supplementary Table A-2)[103]. These regions were further confirmed by DSSS (Figure 4-1B and Supplementary Figure A-1) and in four cases by quantitative polymerase chain reaction (Supplementary Table A-3). Sequence similarity with known proteins revealed the presence of two new protein-coding genes, a pseudogene, one

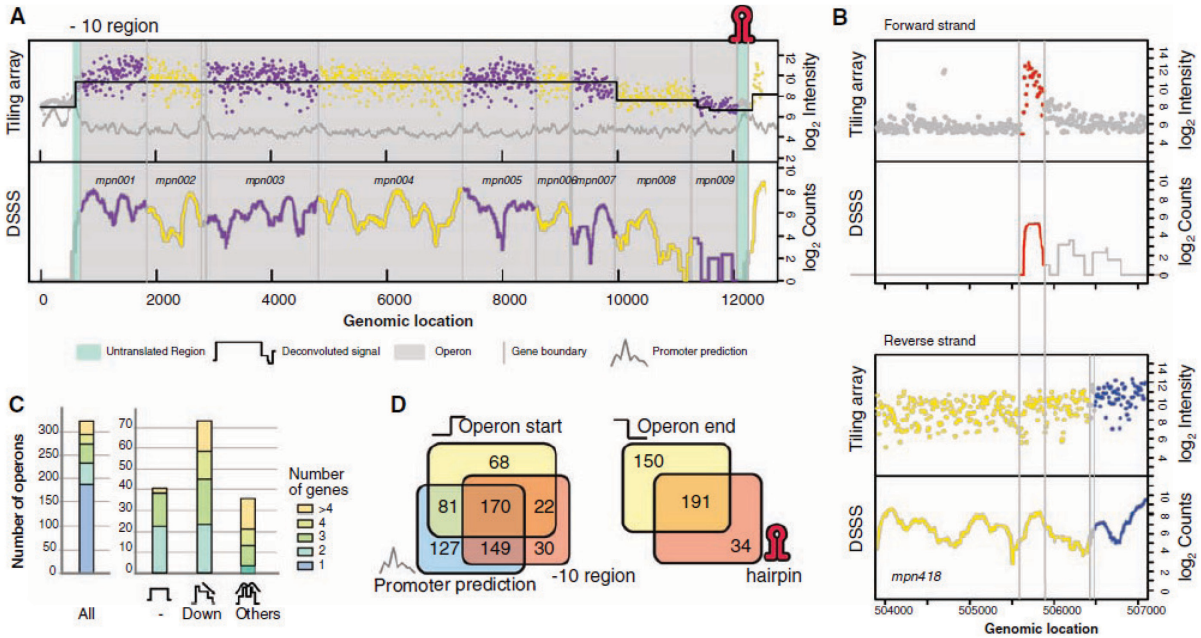


Figure 4-1. Transcriptome feature in the reference condition.

(**A**) The first operon in the genome on the forward strand has a staircase behavior, meaning that the consecutive genes have lower and steady expression levels. (**B**) Example of an anti-sense RNA transcript. (**C**) Analysis of staircase operons. Left: all reference operons subdivided by the number of protein coding genes they contain. Right: all reference operons subdivided by their staircase behavior (see bottom graphs). (**D**) Left, Overlap of operon starts and single gene starts with previously identified -10 promoter sequence motifs in *M. pneumoniae* [104] and predicted promoters based on hexamers. Right, Overlap of operon ends and single gene ends with predicted transcription termination hairpins.

N-terminal truncation, and five 5'-extensions of known genes (Supplementary Table A-2). The remaining 108 transcripts are probably regulatory rather than structural RNAs, because comparison of their predicted secondary structures with the ones of coding genes does not show any substantial difference[103]. Eighty-nine of them are antisense with respect to previously annotated genes. Out of the non-overlapping ones, two of them (NEW87 and NEW8) are conserved in *M. genitalium* and could be involved in DNA replication and repair, and in peptide transport, respectively[103] (Supplementary Figure A-3, Supplementary Figure A-4, Supplementary Figure A-5).

In total, 13% of the coding genes are covered by antisense; this is twice more than in yeast (7%)[105], and about half of what was reported for plants (22.2%)[106-107], or humans (22.6%)[108]. Antisense transcripts may affect expression of the overlapping functional sense transcripts through several mechanisms[109]: Double-stranded RNA-

dependent mechanisms require coexpression with their target[110], whereas transcriptional interference rather implies mutual exclusion of sense and antisense transcripts[111-112]. In *M. pneumoniae,* we observed a predominance of double-stranded RNA mechanism as in mammals[113] (47% positive correlation versus 2% negative correlation). In addition, we detected a reduced expression level of genes targeted by antisense transcripts, as reported in some prokaryotes[103,112] (Supplementary Figure A-6).

We identified operon boundaries through sharp transcription changes in the tiling reference condition by using local convolution methods[103,114](Figure 4-1A). More than 90% of the operons (139 polycistronic and 202 monocistronic operons; Supplementary Table A-4) were well supported by DSSS reads (DSSS alone was not sufficient to unambiguously characterize operons; Supplementary Figure A-2)[103]. Most polycistronic operons contain two or three genes (Figure 4-1C and Supplementary Figure A-7, see Supplementary Table A-4); the largest one is the ribosomal operon containing 20 genes. For the majority of operons, we observed a canonical or slightly altered version of a standard sigma 70 promoter region (Supplementary Figure A-8), with transcription starts located within 60 bp (Supplementary Figure A-9) upstream of the translation start[100]. In contrast to previous suggestions[115], we observe, as proposed by others[116], a preferential use of termination hairpins for tight regulation of gene expression (Figure 4-1, A and D and Supplementary Table A-5). Moreover, we found that almost half of the consecutive genes within polycistronic operons show a decay behavior Figure 4-1A and Supplementary Figure A-1), indicating that such 'staircase'-like expression is a widespread phenomenon in bacteria[103].

Analysis of the 43 tiling arrays and integration with 252 spotted arrays representing 173 independent conditions, some of them from time-series, revealed context-dependent modulation of operon structures involving repression or activation of operon internal genes, as well as of genes located at the beginning, or end (Figure 4-2 and Figure 4-3; see also Supplementary Figure A-10 and Supplementary Table A-5). In some cases this modulation can be assigned to specific environmental changes. Down regulation of the first four genes of the *ftsZ* operon involved in initiation of cell division corresponds to

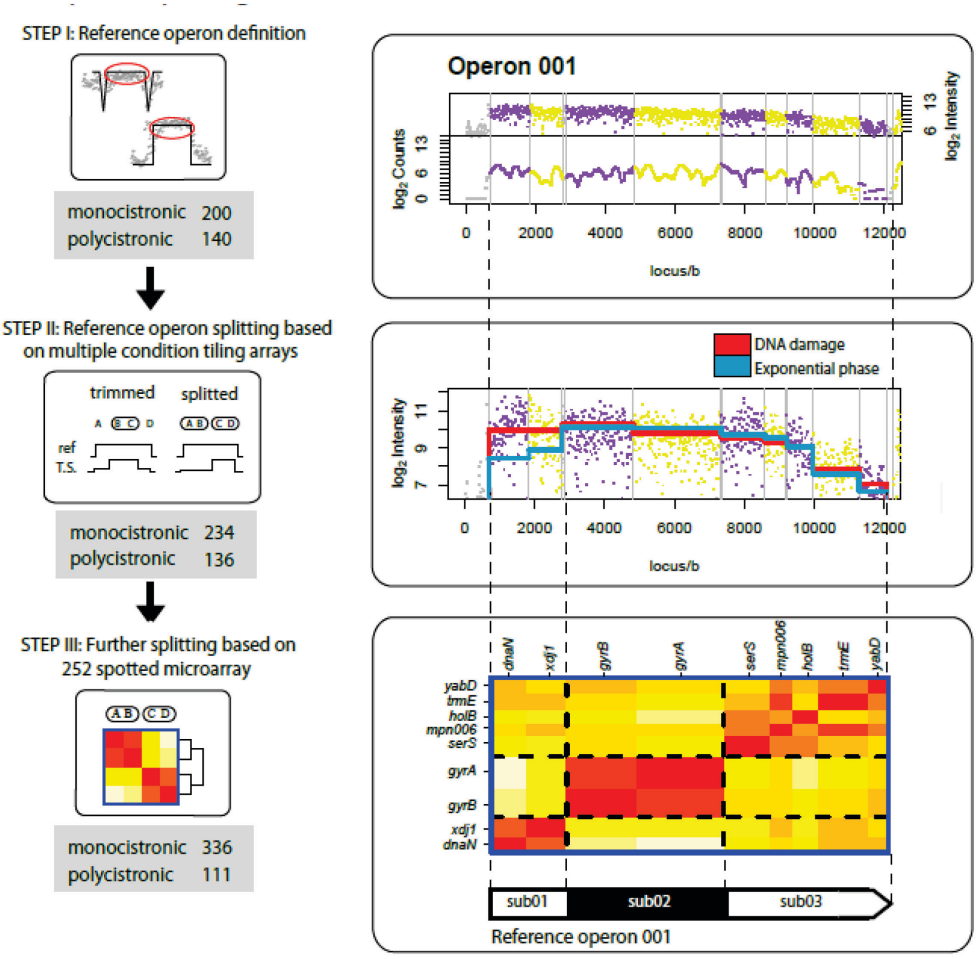


Figure 4-2. Operon splitting.

Left: Alternative transcripts discovery pipeline[103]. Right: Reference operon 001 is split into 3 suboperons. (Top) Tiling and DSSS under reference conditions. (Middle) Specific expression changes for genes dnaA and xdj1 involved in DNA repair and replication. (Bottom) The coexpression matrices correspond to the final conditional operon splitting by 252 arrays. Continuous lines indicate expression level measured with tiling arrays.

entry into stationary phase (Figure 4-3, lower panel). Increase in expression of arginine fermentation genes (*arcA*, a*rcI*, *arcC)* (Figure 4-3) in stationary phase could be a mechanism to cope with acidification[117]. We found formal evidence for a total of 47 transcriptional units (336 monocistronic and 111 polycistronic), implying a high rate of alternative transcripts (42%) in this bacterium in the conditions studied, similar to that in eukaryotes (40%, although still under debate)[118] and archaea (40% in *H.*

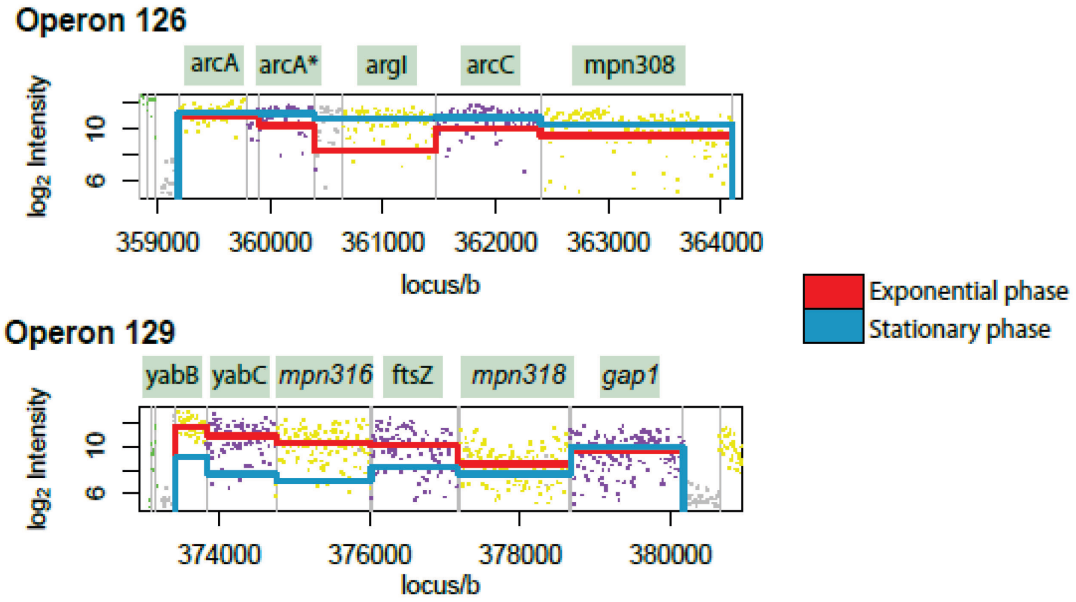


Figure 4-3. Examples of suboperon dynamics.

Two examples of conditional operons are presented. Continuous lines indicate expression level measured with tiling arrays. (Top) Specific induction of the middle genes in the operon 126 when the cells reach stationary phase. (Bottom) Repression of the first 4 genes of the operon 129 involved in cell division when the cells reach stationary phase.

salinarum)[119]. Interestingly we found that genes that are split into different suboperons tend to belong to different functional categories[103]. Thus, although genome reduction leads to longer operons accommodating genes with different functions[120], the latter can still retain internal transcription and termination sites under certain conditions.

The high frequency of alternative transcripts of *M. pneumoniae* genes hints at a situation similar to that in eukaryotes where many factors contribute to the regulation of gene expression. To further support this hypothesis, we used gene expression clustering under the 62 distinct conditions (Supplementary Table A-7) to identify groups of co-expressed genes and their possible common regulatory motifs. Using a correlation cutoff of 0.65 we identified 94 co-expression groups (Supplementary Table A-6 and Supplementary Figure A-11), encompassing 416 genes. Thirty of the clusters contained genes from more than two operons. Of these, 14 share a unique sequence

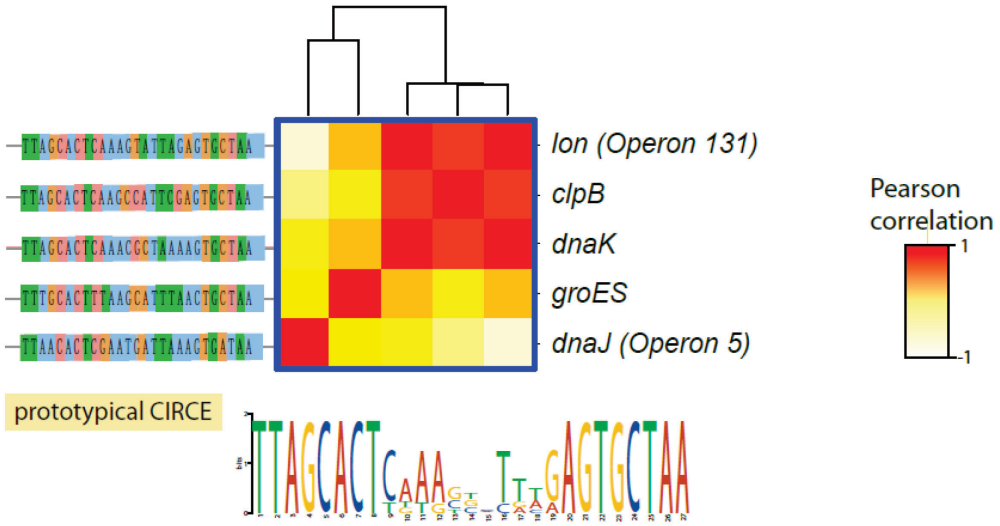


Figure 4-4. Differential expression of dnaJ and groES despite CIRCE element.
Example of heat shock induced genes sharing the known CIRCE element. The calculated sequence consensus is represented below.

motif in their upstream region and another 8 have a unique combination of motifs (Supplementary Figure A-12), which might drive the co-expression (For example, 4 of the 14 motifs are found at splitting sites inside operons). This is exemplified by the five heat shock induced genes containing a regulatory CIRCE (controlling inverted repeat of chaperone expression) element[121] (Figure 4-4). Not all of them clustered together indicating at least another regulatory element. Similarly, overexpression of a transcription factor (mpn329; Fur, ferric uptake regulator) reveals a common motif in all genes significantly changing expression, although they belong to different coexpression clusters (Supplementary Figure A-13 and Supplementary Table A-6).

Our work revealed an unanticipated complexity in the transcriptome of a genome-reduced bacterium. This complexity cannot be explained by the presence of eight predicted transcription factors[120]. Furthermore, the fact that the proteome organization is not explainable by the genome organization[122] indicates the existence of other regulatory processes. The surprisingly frequent expression heterogeneity within operons, the change of operon structures leading to alternative transcripts in response to environmental perturbations and the frequency of antisense RNA which might explain some of these expression changes suggest that transcriptional regulation in bacteria resemble that of eukaryotes more than previously thought.

Chapter 5

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, et al.

Nature **464**(7285):59-65. doi:10.1038/nature08821

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from fecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbors between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

Introduction

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome[123]. The majority of these microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life[6,124]. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity[15,18,30,125-126].

To understand and exploit the impact of the gut microbes on human health and well-being it is necessary to decipher the content, diversity and functioning of the microbial gut community. 16S ribosomal RNA gene (rRNA) sequence-based methods[127] revealed that two bacterial divisions, the Bacteroidetes and the Firmicutes, constitute over 90% of the known phylogenetic categories and dominate the distal gut microbiota[9]. Studies also showed substantial diversity of the gut microbiome between healthy individuals[9,15,30,47]. Although this difference is especially marked among infants[128], later in life the gut microbiome converges to more similar phyla.

Metagenomic sequencing represents a powerful alternative to rRNA sequencing for analyzing complex microbial communities[43,129-130]. Applied to the human gut, such

studies have already generated some 3 gigabases (Gb) of microbial sequence from fecal samples of 33 individuals from the United States or Japan[28-30]. To get a broader overview of the human gut microbial genes we used the Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from fecal samples of 124 European adults. We generated 576.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut and indicates that many bacterial species are shared by different individuals. Our results also show that short-read metagenomic sequencing can be used for global characterization of the genetic potential of ecologically complex environments.

Metagenomic sequencing of gut microbiomes

As part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we collected fecal specimens from 124 healthy, overweight and obese individual human adults, as well as inflammatory bowel disease patients, from Denmark and Spain (Supplementary Table B-1). Total DNA was extracted from the fecal specimens[131] and an average of 4.5 Gb (ranging between 2 and 7.3 Gb) of sequence was generated for each sample, allowing us to capture most of the novelty. In total, we obtained 576.7 Gb of sequence (Supplementary Table B-3).

Wanting to generate an extensive catalogue of microbial genes from the human gut, we first assembled the short Illumina reads into longer contigs, which could then be analyzed and annotated by standard methods. Using SOAPdenovo[52], a de Bruijn graph-based tool specially designed for assembling very short reads, we performed *de novo* assembly for all of the Illumina GA sequence data. Because a high diversity between individuals is expected[28-30], we first assembled each sample independently (Supplementary Figure B-1). As much as 42.7% of the Illumina GA reads was assembled into a total of 6.58 million contigs of a length >500 bp, giving a total contig length of 10.3 Gb, with an N50 length of 2.2 kb (Supplementary Figure B-4) and the range of 12.3 to 237.6 Mb (Supplementary Table B-4). Almost 35% of reads from any

one sample could be mapped to contigs from other samples, indicating the existence of a common sequence core.

To assess the quality of the Illumina GA-based assembly we mapped the contigs of samples MH0006 and MH0012 to the Sanger reads from the same samples (Supplementary Table B-2). A total of 98.7% of the contigs that map to at least one Sanger read were collinear over 99.6% of the mapped regions. This is comparable to the contigs that were generated by 454 sequencing for one of the two samples (MH0006) as a control, of which 97.9% were collinear over 99.5% of the mapped regions. We estimate assembly errors to be 14.2 and 20.7 per megabase (Mb) of Illumina- and 454-based contigs, respectively (see Section 2.1.6 and Supplementary Figure B-5), indicating that the short- and long-read-based assemblies have comparable accuracies.

To complete the contig set we pooled the unassembled reads from all 124 samples, and repeated the *de novo* assembly process. About 0.4 million additional contigs were thus generated, having a length of 370 Mb and an N50 length of 939 bp. The total length of our final contig set was thus 10.7 Gb. Some 80% of the 576.7 Gb of Illumina GA sequence could be aligned to the contigs at a threshold of 90% identity, allowing for accommodation of sequencing errors and strain variability in the gut (Figure 5-1). This is almost twice the 42.7% of sequence that was assembled into contigs by SOAPdenovo, because assembly uses more stringent criteria. This indicates that a vast majority of the Illumina sequence is represented by our contigs.

To compare the representation of the human gut microbiome in our contigs with that from previous work, we aligned them to the reads from the two largest published gut metagenome studies (1.83 Gb of Roche/454 sequencing reads from 18 US adults[30], and 0.79 Gb of Sanger reads from 13 Japanese adults and infants[29]), using the 90% identity threshold. A total of 70.1% and 85.9% of the reads from the Japanese and US samples, respectively, could be aligned to our contigs (Figure 5-1), showing that the contigs include a high fraction of sequences from previous studies. In contrast, 85.7% and 69.5% of our contigs were not covered by the reads from the Japanese and US samples, respectively, highlighting the novelty we captured.

Only 31.0–48.8% of the reads from the two previous studies and the present study

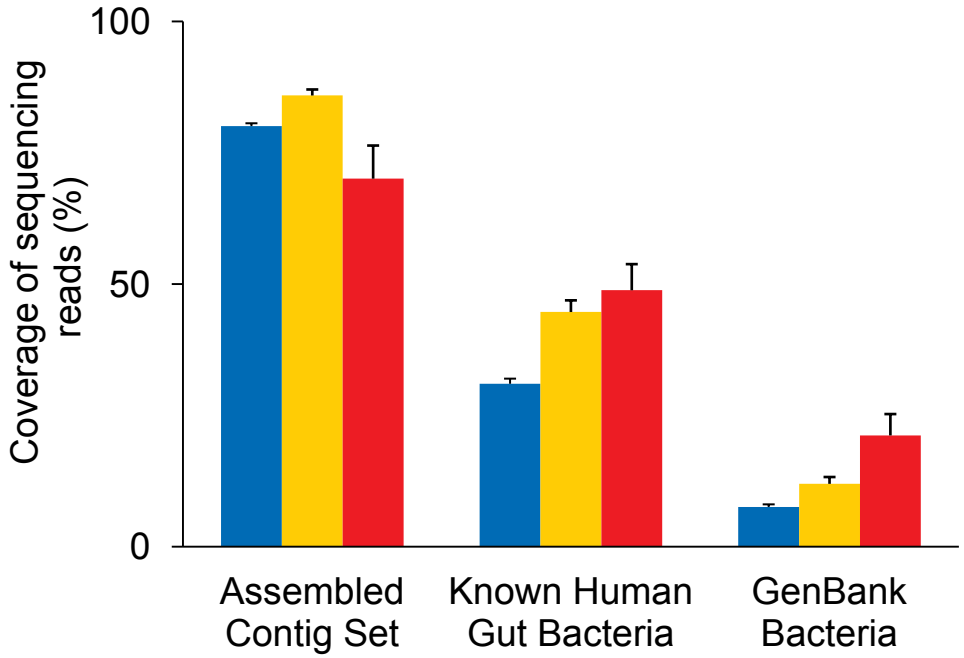


Figure 5-1. Coverage of human gut microbiome.

The three human microbial sequencing read sets, Illumina GA reads generated from 124 individuals in this study (blue; n=124), Roche/454 reads from 18 human twins and their mothers (yellow; n=18), and Sanger reads from 13 Japanese individuals (red; n=13) were aligned to each of the reference sequence sets. Mean values + s.e.m. are plotted.

could be aligned to 194 public human gut bacterial genomes (Supplementary Table B-5 available online[132]), and 7.6–21.2% to the bacterial genomes deposited in GenBank (Figure 5-1). This indicates that the reference gene set obtained by sequencing genomes of isolated bacterial strains is still limited.

A gene catalogue of the human gut microbiome

To establish a non-redundant human gut microbiome gene set we first used the MetaGene[53] program to predict ORFs in our contigs and found 14,048,045 ORFs longer than 100 bp (Supplementary Table B-6). They occupied 86.7% of the contigs, comparable to the value found for fully sequenced genomes (~86%). Two-thirds of the ORFs appeared incomplete, possibly due to the size of our contigs (N50 of 2.2 kb). We next removed the redundant ORFs, by pair-wise comparison, using a very stringent

Table 5-1. Non-redundant genes.

Genes were compared at 95 % identity cut-off. Those that were overlapped over 90% length were considered redundant and removed. Common and rare genes were present in >50% and < 20% of individuals, respectively.

	# of genes	Total length (bp)	Mean length (bp)
Non-redundant gene set	3,299,822	2,323,171,095	704.03
Common	294,110	292,960,308	996.09
Rare	2,375,655	1,510,527,924	635.84

criterion of 95% identity over 90% of the shorter ORF length, which can fuse orthologs but avoids inflation of the data set due to possible sequencing errors (see Section 2.1.8). Yet, the final non-redundant gene set contained as many as 3,299,822 ORFs with an average length of 704 bp (Table 5-1).

We term the genes of the non-redundant set 'prevalent genes', as they are encoded on contigs assembled from the most abundant reads (see Section 2.1.8). The minimal relative abundance of the prevalent genes was $\sim 6 \times 10^{-7}$, as estimated from the minimum sequence coverage of the unique genes (close to 3), and the total Illumina sequence length generated for each individual (on average, 4.5 Gb), assuming the average gene length of 0.85 kb (that is, $3 \times 0.85 \times 10^3 / 4.5 \times 10^9$).

We mapped the 3.3 million gut ORFs to the 319,812 genes (target genes) of the 89 frequent reference microbial genomes in the human gut. At a 90% identity threshold, 80% of the target genes had at least 80% of their length covered by a single gut ORF (Figure 5-2). This indicates that the gene set includes most of the known human gut bacterial genes.

We examined the number of prevalent genes identified across all individuals as a function of the extent of sequencing, demanding at least two supporting reads for a gene call (Figure 5-2). The incidence-based coverage richness estimator (ICE), determined at 100 individuals (the highest number the EstimateS[133] program could accommodate), indicates that our catalogue captures 85.3% of the prevalent genes. Although this is probably an underestimate, it nevertheless indicates that the catalogue contains an overwhelming majority of the prevalent genes of the cohort.

Each individual carried $536{,}112 \pm 12{,}167$ (mean $\pm$ s.e.m.) prevalent genes (see

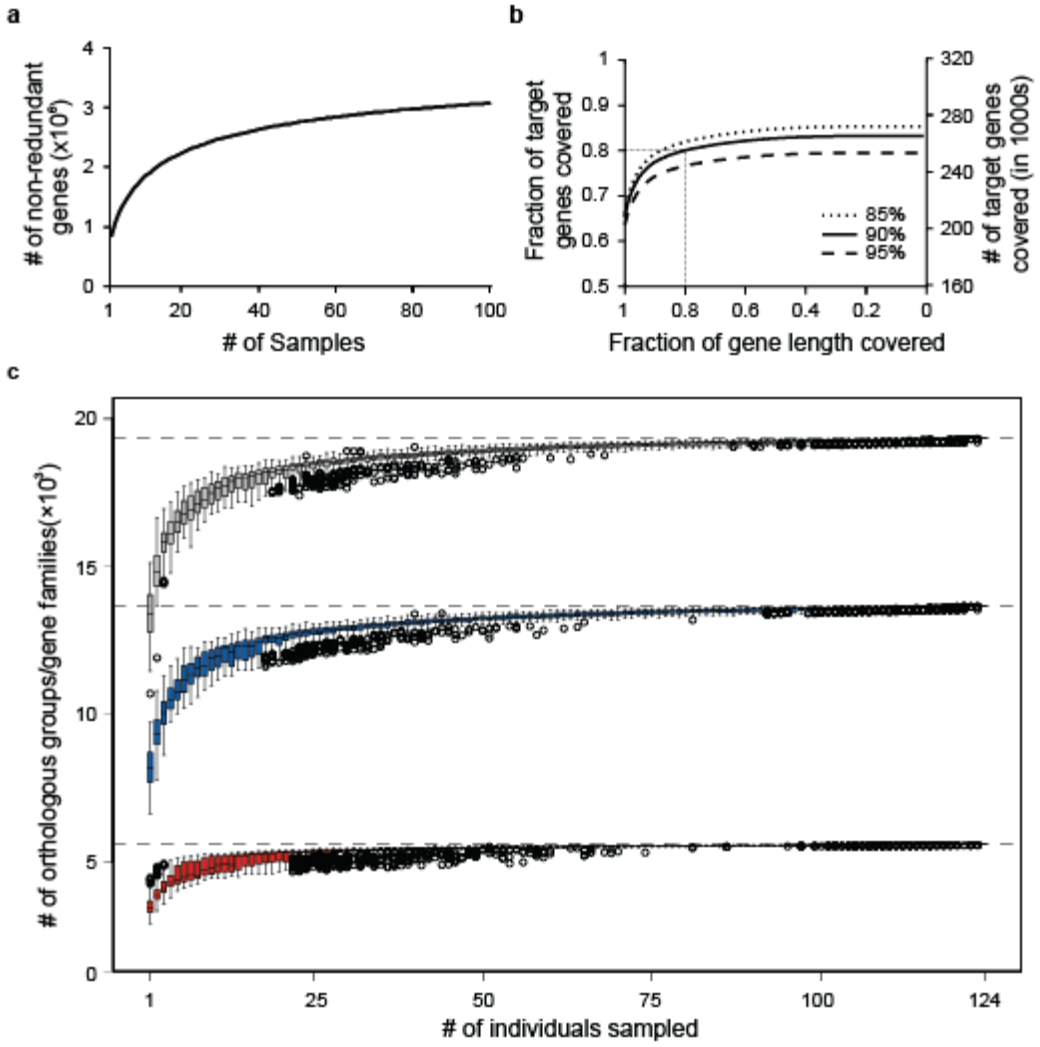


Figure 5-2. Predicted ORFs in the human gut microbiomes.

a) Number of unique genes as a function of the extent of sequencing. The gene accumulation curve corresponds to the S_{obs} (Mao Tau) values, calculated using EstimateS[133] (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). **b)** Coverage of genes from 89 frequent gut microbial species (Supplementary Table B-10). At 90% similarity threshold, 65% of the target genes are fully covered and 80% of the target genes have at least 80% of their length covered. **c)** Number of functions captured by number of samples investigated, based upon known (well characterized) orthologous groups (red), known and unknown orthologous groups (including e.g. putative, predicted, conserved hypothetical functions; blue) and orthologous groups plus novel gene families (>20 proteins; grey) recovered from the metagenome. Boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers.

Supplementary Figure B-6b), indicating that most of the 3.3 million gene pool must be shared. However, most of the prevalent genes were found in only a few individuals: 2,375,655 were present in less than 20%, whereas 294,110 were found in at least 50% of individuals (we term these 'common' genes). These values depend on the sampling depth; sequencing of MH0006 and MH0012 revealed more of the catalogue genes, present at a low abundance (Supplementary Figure B-7). Nevertheless, even at our routine sampling depth, each individual harbored 204,056 ± 3,603 (mean ± s.e.m.) common genes, indicating that about 38% of an individual's total gene pool is shared. Interestingly, the IBD (Inflammatory Bowel Disease) patients harbored, on average, 25% fewer genes than the individuals not suffering from IBD (Supplementary Figure B-8), consistent with the observation that the former have lower bacterial diversity than the latter[19].

Common bacterial core

Deep metagenomic sequencing provides the opportunity to explore the existence of a common set of microbial species (common core) in the cohort. For this purpose, we used a non-redundant set of 650 sequenced bacterial and archaeal genomes (see Section 2.1.15). We aligned the Illumina GA reads of each human gut microbial sample onto the genome set, using a 90% identity threshold, and determined the proportion of the genomes covered by the reads that aligned onto only a single position in the set. At a 1% coverage, which for a typical gut bacterial genome corresponds to an average length of about 40 kb, some 25-fold more than that of the 16S gene generally used for species identification, we detected 18 species in all individuals, 57 in ≥90% and 75 in ≥50% of individuals (Supplementary Table B-7). At 10% coverage, requiring ~10-fold higher abundance in a sample, we still found 13 of the above species in ≥90% of individuals and 35 in ≥50%.

When the cumulated sequence length increased from 3.96 Gb to 8.74 Gb and from 4.41 Gb to 11.6 Gb, for samples MH0006 and MH0012, respectively, the number of strains common to the two at the 1% coverage threshold increased by 25%, from 135 to 169. This indicates the existence of a significantly larger common core than the one we could observe at the sequence depth routinely used for each individual.

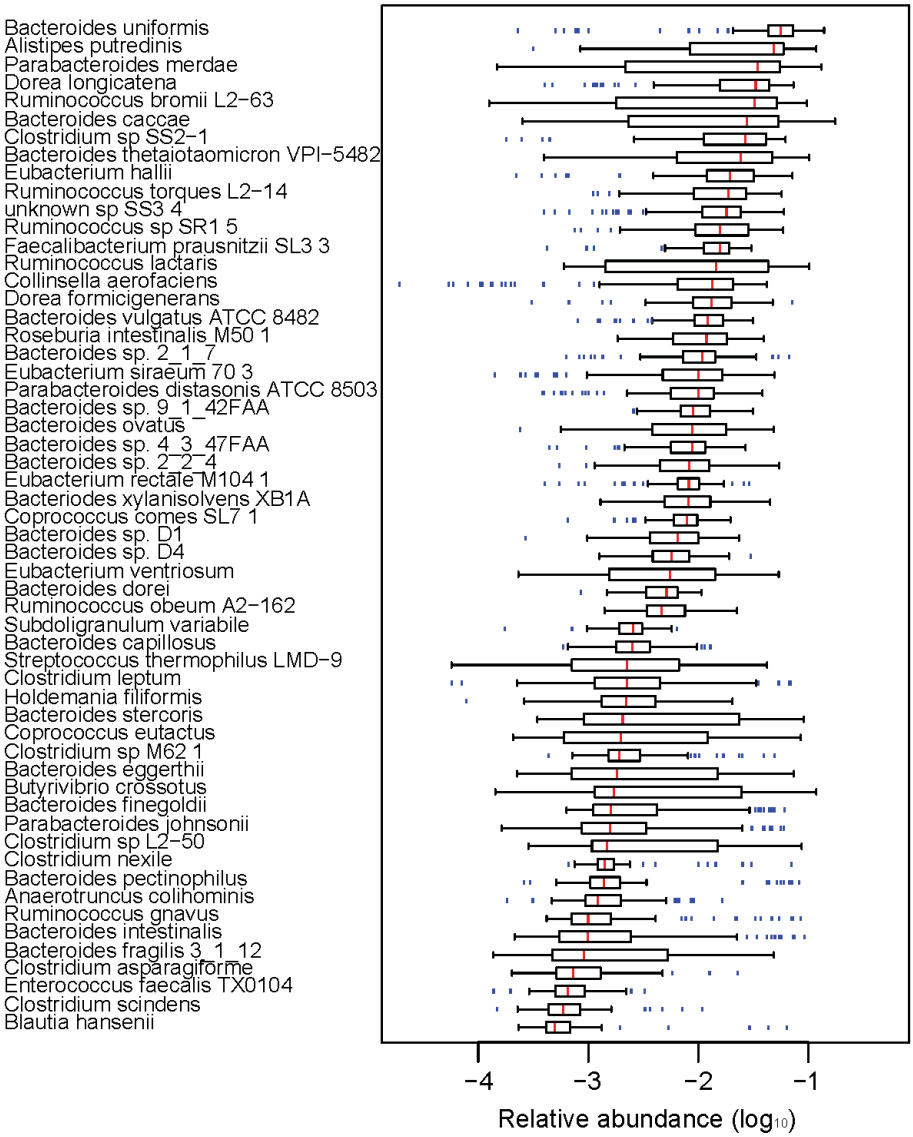


Figure 5-3. Relative abundance of 57 frequent microbial genomes among individuals of the cohort.

See Figure 5-2c for definition of box and whisker plot. See Section 2.1.7 for computation.

The variability of abundance of microbial species in individuals can greatly affect identification of the common core. To visualize this variability, we compared the number of sequencing reads aligned to different genomes across the individuals of our cohort. Even for the most common 57 species present in ≥90% of individuals with genome coverage >1% (Supplementary Table B-7), the inter-individual variability was between 12- and 2,187-fold (Figure 5-3). As expected[9,134], Bacteroidetes and Firmicutes had the highest abundance.

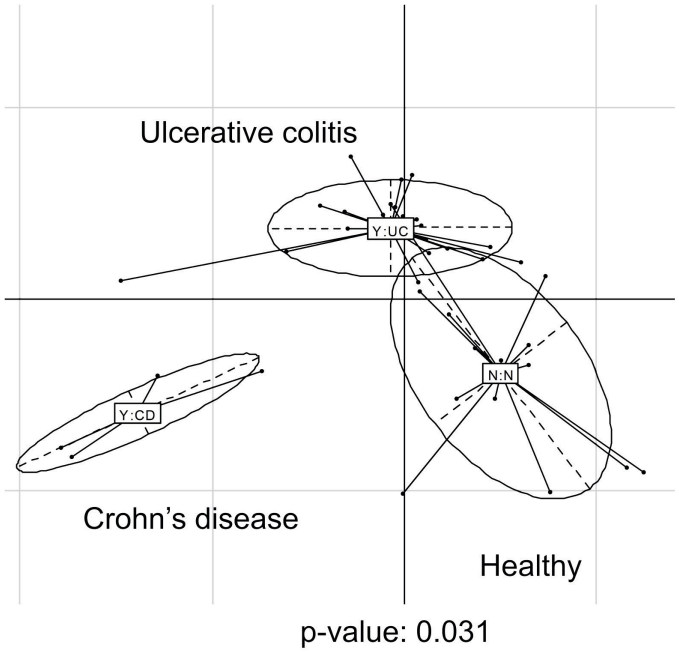


Figure 5-4. Species abundance differentiates IBD patients and healthy individuals.

Principal component analysis with health status as instrumental variables, based on the abundance of 155 species with ≥1% genome coverage by the Illumina reads in ≥1 individual of the cohort, was carried out with 14 healthy individuals and 25 IBD patients (21 UC and 4 CD) from Spain (Supplementary Table B-1). Two first components (PC1 and PC2) were plotted and represented 7.3% of whole inertia. Individuals (represented by points) were clustered and centre of gravity computed for each class; P-value of the link between health status and species abundance was assessed using a Monte-Carlo test (999 replicates).

A complex pattern of species relatedness, characterized by clusters at the genus and family levels, emerges from the analysis of the network based on pair-wise Pearson correlation coefficients of 155 species present in at least one individual at ≥1% coverage (Supplementary Figure B-9). Prominent clusters include some of the most abundant gut species, such as Bacteroidetes and *Dorea/Eubacterium/Ruminococcus* groups and also bifidobacteria, proteobacteria and streptococci/lactobacilli groups, indicating that similar constellations of bacteria may be present in different individuals of our cohort, for reasons that remain to be established.

The above result indicates that the Illumina-based bacterial profiling should reveal differences between the healthy individuals and patients. To test this hypothesis we compared the IBD patients and healthy controls (Supplementary Table B-1), as it was previously reported that the two have different microbiota[19]. (Figure 5-4) shows that

Table 5-2. Number of genes classified.

The predicted genes were aligned to the known microbial genes, the genes in KEGG orthology database and in COG database. Blastp software was used to align genes with E-value <1E-5, and the best hit was selected. LCA-based algorithm was used to assign gene sequences to taxa. When a gene was conserved in many species, it was assigned to the lowest common ancestor (LCA). However, if the LCA is at phylum-level or below, it was considered to be "Classified" all the same. If not, it was treated as "Unclassified".

		Common genes	Rare genes	All genes
Total		**294,110**	**2,375,655**	**3,299,822**
Phylotype	Unknown	5.99%	27.64%	22.93%
	Unclassified	4.31%	3.82%	3.88%
	Classified	89.70%	68.54%	73.19%
eggNOG	Unannotated	23.88%	47.36%	42.46%
	Annotated	76.12%	52.64%	57.54%
COG	Unannotated	31.63%	54.89%	49.99%
	Annotated	68.37%	45.11%	50.01%
KEGG orthology	Unannotated	34.73%	57.84%	52.97%
	Annotated	65.27%	42.16%	47.03%
KEGG pathway	Unannotated	73.81%	83.19%	81.26%
	Annotated	26.19%	16.81%	18.74%

the principal component analysis based on the same 155 species clearly separates patients from healthy individuals and the UC (Ulcerative Colitis) from the CD (Crohn's Disease) patients, confirming our hypothesis.

Functions encoded by the prevalent gene set

We classified the predicted genes by aligning them to the integrated NCBI-NR database of non-redundant protein sequences, the genes in the KEGG (Kyoto Encyclopedia of Genes and Genomes)[56] pathways, and COG (Clusters of Orthologous Groups)[35] and eggNOG[55] databases. There were 77.1% genes classified into phylotypes, 57.5% to eggNOG clusters, 47.0% to KEGG orthology and 18.7% genes assigned to KEGG pathways, respectively (Table 5-2). Almost all (99.96%) of the phylogenetically assigned genes belonged to the Bacteria and Archaea, reflecting their predominance in the gut. Genes that were not mapped to orthologous

groups were clustered into gene families (see Section 2.1.13). To investigate the functional content of the prevalent gene set we computed the total number of orthologous groups and/or gene families present in any combination of n individuals (with $n = 2$–124; see Figure 5-2). This rarefaction analysis shows that the 'known' functions (annotated in eggNOG or KEGG) quickly saturate (a value of 5,569 groups was observed): when sampling any subset of 50 individuals, most have been detected. However, three-quarters of the prevalent gut functionalities consists of uncharacterized orthologous groups and/or completely novel gene families (Figure 5-2). When including these groups, the rarefaction curve only starts to plateau at the very end, at a much higher level (19,338 groups were detected), confirming that the extensive sampling of a large number of individuals was necessary to capture this considerable amount of novel/unknown functionality.

Bacterial functions important for life in the gut

The extensive non-redundant catalogue of the bacterial genes from the human intestinal tract provides an opportunity to identify bacterial functions important for life in this environment. There are functions necessary for a bacterium to thrive in a gut context (that is, the 'minimal gut genome') and those involved in the homeostasis of the whole ecosystem, encoded across many species (the 'minimal gut metagenome'). The first set of functions is expected to be present in most or all gut bacterial species; the second set in most or all individuals' gut samples.

To identify the functions encoded by the minimal gut genome we use the fact that they should be present in most or all gut bacterial species and therefore appear in the gene catalogue at a frequency above that of the functions present in only some of the gut bacterial species. The relative frequency of different functions can be deduced from the number of genes recruited to different eggNOG clusters, after normalization for gene length and copy number (Supplementary Figure B-10). We ranked all the clusters by gene frequencies and determined the range that included the clusters specifying well-known essential bacterial functions, such as those determined experimentally for a well-studied firmicute, *Bacillus subtilis*[135], hypothesizing that additional clusters in this range are equally important. As expected, the range that included most of

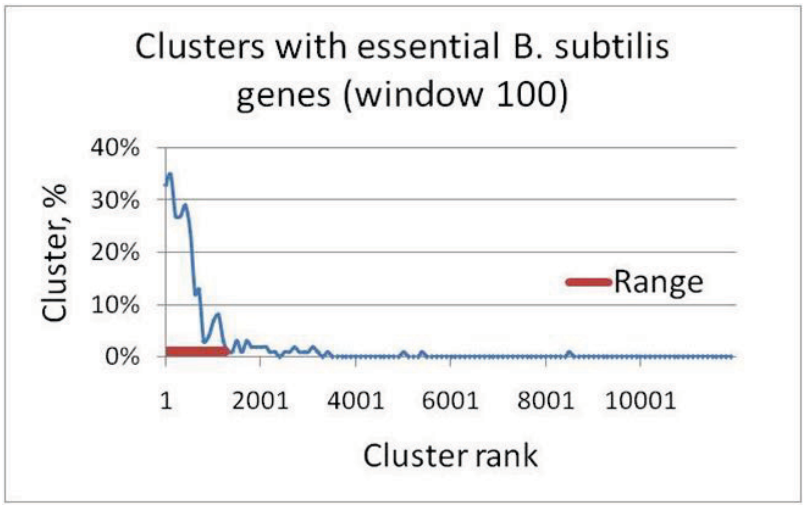


Figure 5-5. Clusters that contain the *B. subtilis* essential genes.

The clusters were ranked by the number of genes they contain, normalized by average length and copy number (see Supplementary Figure B-10) and the proportion of clusters with the essential *B. subtilis* genes was determined for successive groups of 100 clusters (window 100). Range indicates the part of the cluster distribution that contains 86 % of the *B. subtilis* essential genes.

B. subtilis essential clusters (86%) was at the very top of the ranking order (Figure 5-5). Some 76% of the clusters with essential genes of *Escherichia coli*[136] were within this range, confirming the validity of our approach. This suggests that 1,244 metagenomic clusters found within the range (Supplementary Table B-8; termed 'range clusters' hereafter) specify functions important for life in the gut.

We found two types of functions among the range clusters: those required in all bacteria (housekeeping) and those potentially specific for the gut. Among many examples of the first category are the functions that are part of main metabolic pathways (for example, central carbon metabolism, amino acid synthesis), and important protein complexes (RNA and DNA polymerase, ATP synthase, general secretory apparatus). Not surprisingly, projection of the range clusters on the KEGG metabolic pathways gives a highly integrated picture of the global gut cell metabolism (Figure 5-6a).

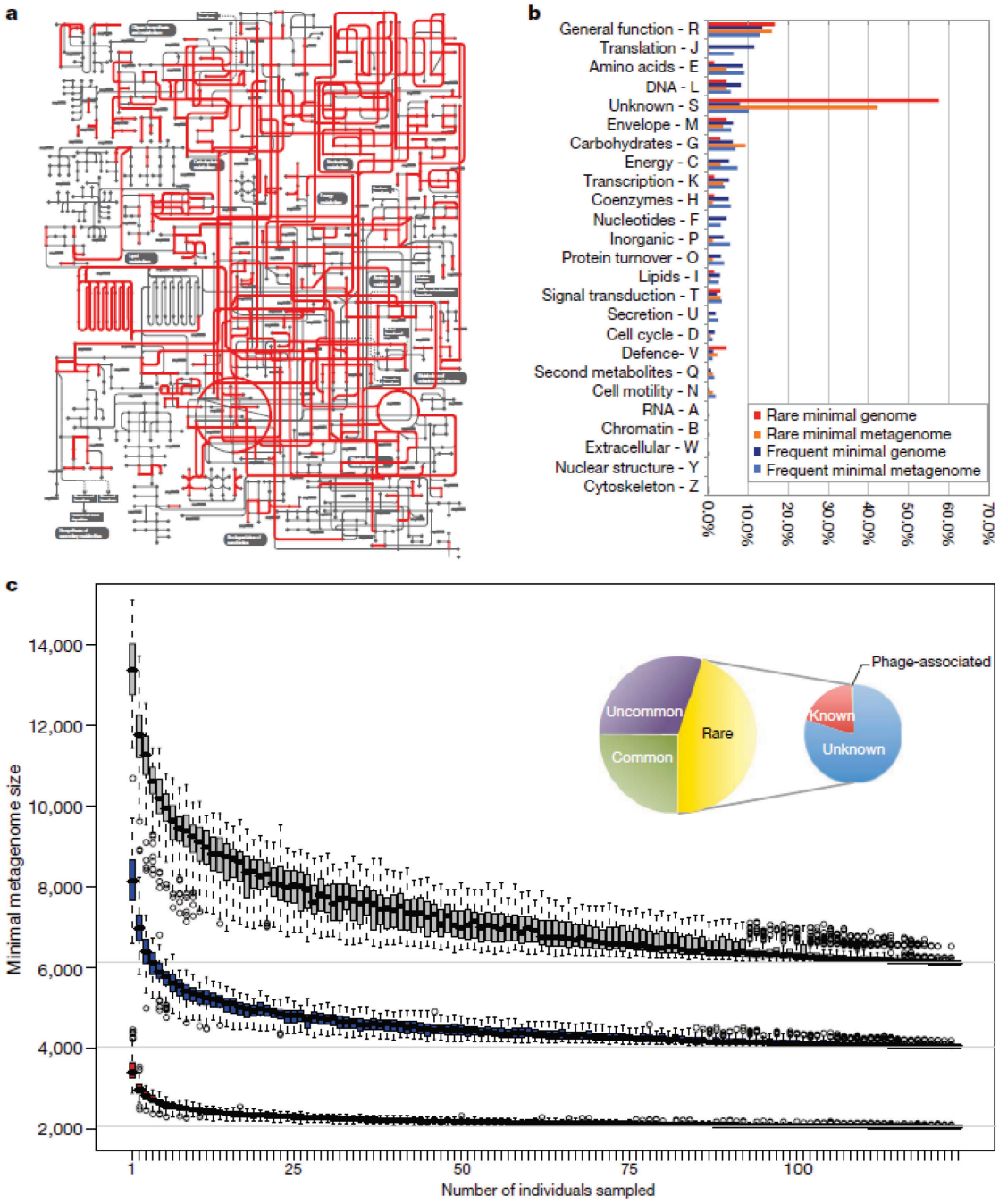


Figure 5-6. Characterization of the minimal gut genome and metagenome.

a) Functional composition of the minimal gut genome and metagenome, separated by their frequency in sequenced genomes. **b)** projection of the minimal gut genome on the KEGG pathways using the iPath tool[137]. **c)** Estimation of the minimal gut metagenome size based upon known (well characterized) orthologous groups (red), known and unknown orthologous groups (including e.g. putative, predicted, conserved hypothetical functions; blue) and orthologous groups with novel gene families (>20 proteins) recovered from the metagenome. Inset: General composition of the gut

minimal microbiome. Large circle: classification in the minimal metagenome according to orthologous group occurrence in STRING7[138] bacterial genomes. Common (green, 25%), uncommon (purple, 35%) and rare (yellow, 45%) refer to functions that are present in >50%, <50% but >10%, and <10% of STRING bacterial genomes, respectively. Small circle: composition of the rare orthologous groups. Unknown (blue, 80%) have no annotation or are poorly characterized, whereas known bacterial (red, 19%) and phage-related (light green, 1%) orthologous groups have functional description.

The putative gut-specific functions include those involved in adhesion to the host proteins (collagen, fibrinogen, fibronectin) or in harvesting sugars of the globoseries glycolipids, which are carried on blood and epithelial cells. Furthermore, 15% of range clusters encode functions that are present in <10% of the eggNOG genomes (Supplementary Figure B-11) and are largely (74.3%) not defined (Figure 5-6b). Detailed studies of these should lead to a deeper comprehension of bacterial life in the gut.

To identify the functions encoded by the minimal gut metagenome, we computed the orthologous groups that are shared by individuals of our cohort. This minimal set, of 6,313 functions, is much larger than the one estimated in a previous study[30]. There are only 2,069 functionally annotated orthologous groups, showing that they gravely underestimate the true size of the common functional complement among individuals (Figure 5-6c). The minimal gut metagenome includes a considerable fraction of functions (~45%) that are present in <10% of the sequenced bacterial genomes (Figure 5-6c, inset). These otherwise rare functionalities that are found in each of the 124 individuals may be necessary for the gut ecosystem. Eighty per cent of these orthologous groups contain genes with at best poorly characterized function, underscoring our limited knowledge of gut functioning.

Of the known fraction, about 5% codes for (pro)phage-related proteins, implying a universal presence and possible important ecological role of bacteriophages in gut homeostasis. The most striking secondary metabolism that seems crucial for the minimal metagenome relates, not unexpectedly, to biodegradation of complex sugars and glycans harvested from the host diet and/or intestinal lining. Examples include degradation and uptake pathways for pectin (and its monomer, rhamnose) and sorbitol, sugars which are omnipresent in fruits and vegetables, but which are not or poorly absorbed by humans. As some gut microorganisms were found to degrade both of them[139-140], this capacity seems to be selected for by the gut ecosystem as a non-

competitive source of energy. Besides these, capacity to ferment, for example, mannose, fructose, cellulose and sucrose is also part of the minimal metagenome. Together, these emphasize the strong dependence of the gut ecosystem on complex sugar degradation for its functioning.

Functional complementarities of the genome and the metagenome

Detailed analysis of the complementarities between the gut metagenome and the human genome is beyond the scope of the present work. To provide an overview, we considered two factors: conservation of the functions in the minimal metagenome and presence/absence of functions in one or the other (Supplementary Table B-9). Gut bacteria use mostly fermentation to generate energy, converting sugars, in part, to short-chain fatty acids, that are used by the host as energy source. Acetate is important for muscle, heart and brain cells[141], propionate is used in host hepatic neoglucogenic processes, whereas, in addition, butyrate is important for enterocytes[142]. Beyond short-chain fatty acid, a number of amino acids are indispensable to humans[143] and can be provided by bacteria[144]. Similarly, bacteria can contribute certain vitamins[6] (for example, biotin, phylloquinone) to the host. All of the steps of biosynthesis of these molecules are encoded by the minimal metagenome.

Gut bacteria seem to be able to degrade numerous xenobiotics, including non-modified and halogenated aromatic compounds (Supplementary Table B-9), even if the steps of most pathways are not part of the minimal metagenome and are found in a fraction of individuals only. A particularly interesting example is that of benzoate, which is a common food supplement, known as E211. Its degradation by the coenzyme-A ligation pathway, encoded in the minimal metagenome, leads to pimeloyl-coenzyme-A, which is a precursor of biotin, indicating that this food supplement can have a potentially beneficial role for human health.

Discussion

We have used extensive Illumina GA short-read-based sequencing of total fecal DNA from a cohort of 124 individuals of European (Nordic and Mediterranean) origin to

establish a catalogue of non-redundant human intestinal microbial genes. The catalogue contains 3.3 million microbial genes, 150-fold more than the human gene complement, and includes an overwhelming majority (>86%) of prevalent genes harboured by our cohort. The catalogue probably contains a large majority of prevalent intestinal microbial genes in the human population, for the following reasons: (1) over 70% of the metagenomic reads from three previous studies, including American and Japanese individuals[28-30], can be mapped on our contigs; (2) about 80% of the microbial genes from 89 frequent gut reference genomes are present in our set. This result represents a proof of principle that short-read sequencing can be used to characterize complex microbiomes.

The full bacterial gene complement of each individual was not sampled in our work. Nevertheless, we have detected some 536,000 prevalent unique genes in each, out of the total of 3.3 million carried by our cohort. Inevitably, the individuals largely share the genes of the common pool. At the present depth of sequencing, we found that almost 40% of the genes from each individual are shared with at least half of the individuals of the cohort. Future worldwide studies, envisaged within the International Human Microbiome Consortium, will complete, as necessary, our gene catalogue and establish boundaries to the proportion of shared genes.

Essentially all (99.1%) of the genes of our catalogue are of bacterial origin, the remainder being mostly archaeal, with only 0.1% of eukaryotic and viral origins. The gene catalogue is therefore equivalent to that of some 1,000 bacterial species with an average-sized genome, encoding about 3,364 non-redundant genes. We estimate that no more than 15% of prevalent genes of our cohort may be missing from the catalogue, and suggest that the cohort harbors no more than ~1,150 bacterial species abundant enough to be detected by our sampling. Given the large overlap between microbial sequences in this and previous studies we suggest that the number of abundant intestinal bacterial species may be not much higher than that observed in our cohort. Each individual of our cohort harbours at least 160 such bacterial species, as estimated by the average prevalent gene number, and many must thus be shared.

We assigned about 12% of the reference set genes (404,000) to the 194 sequenced intestinal bacterial genomes, and can thus associate them with bacterial species. Sequencing of at least 1,000 human-associated bacterial genomes is foreseen within the

International Human Microbiome Consortium, via the Human Microbiome Project and MetaHIT. This is commensurate with the number of dominant species in our cohort and expected more broadly in human gut, and should enable a much more extensive gene to species assignment. Nevertheless, we used the presently available sequenced genomes to explore further the concept of largely shared species among our cohort and identified 75 species common to >50% of individuals and 57 species common to >90%. These numbers are likely to increase with the number of sequenced reference strains and a deeper sampling. Indeed, a 2–3-fold increase in sequencing depth raised the number of species that we could detect as shared between two individuals by 25%. A large number of shared species supports the view that the prevalent human microbiome is of a finite and not overly large size.

How can this view be reconciled with that of a considerable inter-personal diversity of innumerable bacterial species in the gut, arising from most previous studies using the 16S RNA marker gene[9,15,30,47]? Possibly the depth of sampling of these studies was insufficient to reveal common species when present at low abundance, and emphasized the difference in the composition of a relatively few dominant species. We found a very high variability of abundance (12- to 2,200-fold) for the 57 most common species across the individuals of our cohort. Nevertheless, a recent 16S rRNA-based study concluded that a common bacterial species 'core', shared among at least 50% of individuals under study, exists[145].

Detailed comparisons of bacterial genes across the individuals of our cohort will be carried out in the future, within the context of the ongoing MetaHIT clinical studies of which they are part. Nevertheless, clustering of the genes in families allowed us to capture a virtually full functional potential of the prevalent gene set and revealed a considerable novelty, extending the functional categories by some 30% in regard to previous work[30]. Similarly, this analysis has revealed a functional core, conserved in each individual of the cohort, which reflects the full minimal human gut metagenome, encoded across many species and probably required for the proper functioning of the gut ecosystem. The size of this minimal metagenome exceeds several-fold that of the core metagenome reported previously[30]. It includes functions known to be important to the host–bacterial interaction, such as degradation of complex polysaccharides, synthesis of short-chain fatty acids, indispensable amino acids and vitamins. Finally,

we also identified functions that we attribute to a minimal gut bacterial genome, likely to be required by any bacterium to thrive in this ecosystem. Besides general housekeeping functions, the minimal genome encompasses many genes of unknown function, rare in sequenced genomes and possibly specifically required in the gut.

Beyond providing the global view of the human gut microbiome, the extensive gene catalogue we have established enables future studies of association of the microbial genes with human phenotypes and, even more broadly, human living habits, taking into account the environment, including diet, from birth to old age. We anticipate that these studies will lead to a much more complete understanding of human biology than the one we presently have.

Methods summary

Human fecal samples were collected, frozen immediately and DNA was purified by standard methods[19]. For all 124 individuals, paired-end libraries were constructed with different clone insert sizes and subjected to Illumina GA sequencing. All reads were assembled using SOAPdenovo[52], with specific parameter '-M 3' for metagenomics data. MetaGene was used for gene prediction. A non-redundant gene set was constructed by pair-wise comparison of all genes, using BLAT[54] under the criteria of identity >95% and overlap >90%. Gene taxonomic assignments were made on the basis of BLASTP[146] search (e-value $<1 \times 10^{-5}$) of the NCBI-NR database and 126 known gut bacteria genomes. Gene functional annotations were made by BLASTP search (e-value $<1 \times 10^{-5}$) with eggNOG and KEGG (v48.2) databases. The total and shared number of orthologous groups and/or gene families were computed using a random combination of n individuals (with $n = 2$ to 124, 100 replicates per bin). The raw Illumina reads of all 124 samples were submitted to EBI, with the SUBMISSION_ACCOUNT_ID: ERA000116. The contigs and gene set are available to download from BGI-Shenzhen (`http://gutmeta.genomics.org.cn`) and EMBL (`http://www.bork.embl.de/~arumugam/Qin_et_al_2009/`) websites.

Chapter 6

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, et al.

Submitted to *Nature*.

Our knowledge on species and function composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about their variation across the world. In the fecal metagenomes of 39 individuals from 6 countries we identified several robust clusters (enterotypes hereafter) that are not nation or continent-specific and suggest that the intestinal microbiota variation is stratified, not continuous. This further indicates the existence of a limited number of well-balanced host-microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but the abundance of molecular functions detected therein does not necessarily correlate with the known abundant species, highlighting the importance of a functional analysis for a community understanding. While individual host properties such as disease state, age, or gender cannot explain the observed enterotypes, data-driven marker species, genes or pathways can be identified for each of these host properties. For example, *Eubacterium* abundance is linked to nationality, three functional modules significantly correlate with the body mass index and 11 genes change in abundance with age, hinting at a diagnostic potential of microbial markers.

Introduction

Various studies of the human intestinal tract microbiome, based on the 16S ribosomal RNA-encoding gene, reported species diversity within and between individuals[9-12,30] and first metagenomics studies characterized the functional repertoire of the microbiome of several American[28,30] and Japanese[29] individuals. Although a general consensus about the phylum level composition in the human gut is emerging[9,11,13-14], the amount of similarity of shared species composition[9-10] and the similarity of the gene pools[30,84] is less clear. Furthermore, nothing is known about the nature of the expected variation of the gut microbiota in the human population, whether it is a continuum of different community compositions or whether there is a dominance of discrete, balanced and stable microbiomes that can be classified. Studying such questions is complicated by the complexity of sampling, DNA preparation, processing, sequencing and analysis protocols[32] as well as to varying physiological, nutritional and environmental conditions. To analyze the feasibility of comparative metagenomics of the human gut and to obtain first insights in

commonalities and differences between gut microbiomes across different populations, we Sanger-sequenced 22 fecal samples from European individuals at an average depth of 105 Mb each (Table 6-1 and Table 6-2) and compared them among each other and with published data from 17 individuals from two other continents (13 from Japan[29] and 4 from America[28,30], sequenced at an average depth of 61 Mb and 92 Mb, respectively). To capture diverse gut microbial communities, the European individuals were drawn from four different nations (Denmark, France, Italy and Spain), were enriched in expected disease-associated microbiota (6 obese individuals and 2 inflammatory bowel disease patients; see Table 6-1) and were selected for a broad range of microbiota (8 of a larger group of 40 were found by a HITchip[88] to be particularly divergent; 6 were over 70 years old, as it was reported that the diversity of the microbiota increases with age[147-148]).

Feasibility of comparative gut metagenomics

In order to maximize comparability of the diverse datasets, which were derived from varying sample preparation protocols, generated by distinct sequencing technologies (Sanger and pyrosequencing) and filtered using diverse pipelines in different sequencing centers (Table 6-1), we newly developed unified species and function annotation protocols (see Section 2.2) that we analyzed for possible biases (for example, with regard to functional composition when compared to STRING[22] genomes; for more details see Supplementary Figure C-1, Supplementary Figure C-2). To investigate the phylogenetic composition of samples, we mapped metagenomic reads, using DNA sequence similarity, to 1152 reference genomes[23] including almost 300 publicly available human microbiome genomes generated through the NIH Human Microbiome Project[90] and the European MetaHIT consortium[69]. As this procedure is novel, we ensured via a parameter exploration (Figure 2-1) that sequences that match a reference genome with >65% sequence identity can be safely assigned to the respective phylum and above 85% to the correct genus (see Section 2.2.11).

To prove that the approach can be generalized beyond the Sanger sequencing technology, we also generated 490 Mb of sequence using pyrosequencing[149] for two Danish samples and found them to agree very well with the Sanger data with regard to

Table 6-1. Details of the human subjects.

Details of the human subjects from three cohorts investigated in this study and three published studies: kurokawa07[7], gill06[6] and turnbaugh09[4]. *,†,‡,& - subjects with the same sign are familially related to each other.

Internal ID	Sample ID	Project	Sample Name	Nationality	Gender	Age	Clinical Status	BMI
MC20.MG29	DA-AD-1	MetaHIT	MH6	danish	F	59	healthy	22.38
MC20.MG30	DA-AD-2	MetaHIT	MH13	danish	M	54	healthy	20.46
MC20.MG31	DA-AD-3	MetaHIT	MH12	danish	F	49	obese	32.1
MC20.MG32	DA-AD-4	MetaHIT	MH30	danish	M	59	obese	35.21
MC20.MG33	ES-AD-1	MetaHIT	CD1	spanish	F	25	CD*	17.9
MC20.MG34	ES-AD-2	MetaHIT	CD2	spanish	M	49	healthy*	27.8
MC20.MG35	ES-AD-3	MetaHIT	UC4	spanish	F	47	UC	26.37
MC20.MG36	ES-AD-4	MetaHIT	UC6	spanish	F	38	healthy	23.18
MC20.MG22	FR-AD-1	MicroObes	NO1	french	M	63	healthy	23.1
MC20.MG23	FR-AD-2	MicroObes	NO3	french	M	61	healthy	22.0
MC20.MG24	FR-AD-3	MicroObes	NO4	french	M	60	healthy	23.8
MC20.MG25	FR-AD-4	MicroObes	NO8	french	M	60	healthy	21.9
MC20.MG27	FR-AD-5	MicroObes	OB2	french	M	64	obese	30.8
MC20.MG26	FR-AD-6	MicroObes	OB1	french	M	63	obese	33.7
MC20.MG28	FR-AD-7	MicroObes	OB6	french	M	62	obese	28.9
MC20.MG37	FR-AD-8	MicroObes	OB8	french	M	60	obese	32.0
MC20.MG16	IT-AD-1	MicroAge	A	italian	F	84	elderly	
MC20.MG17	IT-AD-2	MicroAge	B	italian	M	87	elderly	
MC20.MG18	IT-AD-3	MicroAge	C	italian	F	77	elderly	
MC20.MG19	IT-AD-4	MicroAge	D	italian	M	80	elderly	
MC20.MG20	IT-AD-5	MicroAge	E	italian	M	70	elderly	
MC20.MG21	IT-AD-6	MicroAge	G	italian	F	72	elderly	
MC20.MG3	JP-AD-1	kurokawa07	F1-S	japanese	M	30	healthy†	
MC20.MG4	JP-AD-2	kurokawa07	F1-T	japanese	F	28	healthy†	
MC20.MG5	JP-AD-3	kurokawa07	F2-V	japanese	M	37	healthy‡	
MC20.MG6	JP-AD-4	kurokawa07	F2-W	japanese	F	36	healthy‡	
MC20.MG7	JP-AD-5	kurokawa07	F2-X	japanese	M	3	healthy‡	
MC20.MG8	JP-AD-6	kurokawa07	F2-Y	japanese	F	1.5	healthy‡	
MC20.MG9	JP-AD-7	kurokawa07	In-A	japanese	M	45	healthy	
MC20.MG10	JP-AD-8	kurokawa07	In-D	japanese	M	35	healthy	
MC20.MG11	JP-AD-9	kurokawa07	In-R	japanese	F	24	healthy	
MC20.MG12	JP-IN-1	kurokawa07	F1-U	japanese	F	0.58	healthy†	
MC20.MG13	JP-IN-2	kurokawa07	In-B	japanese	M	0.5	healthy	
MC20.MG14	JP-IN-3	kurokawa07	In-E	japanese	M	0.25	healthy	
MC20.MG15	JP-IN-4	kurokawa07	In-M	japanese	F	0.33	healthy	
MC20.MG1	AM-AD-1	gill06	Subject7	american	F	28	healthy	
MC20.MG2	AM-AD-2	gill06	Subject8	american	M	37	healthy	
MC16.MG13	AM-F10-T1	turnbaugh09	F10T1Ob1	american	F		obese&	
MC16.MG14	AM-F10-T2	turnbaugh09	F10T2Ob1	american	F		obese&	

Table 6-2. Summary statistics of 39 samples.

Summary statistics of the metagenome sequences used: raw sequence details, assembled contigs and predicted protein coding genes. NR sequences: non-redundant sequences after merging contigs and unassembled reads.

Sample ID	Sample size (Mb)	Reads	Singleton reads	Contigs	Contig length (Mb)	NR sequences	NR sequence length (Mb)	Genes	Coding length (Mb), fraction
DA-AD-1	156.96	237710	85700 (36.1%)	19816	31.36	105516	86.75	152959	76.29 (87.94%)
DA-AD-2	146.77	224711	80256 (35.7%)	18910	32.89	99166	83.35	147519	73.46 (88.13%)
DA-AD-3	154.69	231024	88736 (38.4%)	21465	36.11	110201	93.51	162534	84.34 (90.20%)
DA-AD-4	150.17	227411	91405 (40.2%)	22135	37.52	113540	96.15	167530	84.80 (88.19%)
ES-AD-1	144.87	223746	50190 (22.4%)	14898	32.23	65088	63.58	102806	56.00 (88.08%)
ES-AD-2	151.91	230738	69752 (30.2%)	15257	26.1	85009	70.6	122628	61.74 (87.45%)
ES-AD-3	147.49	236855	78396 (33.1%)	20260	32.84	98656	79.89	140465	70.26 (87.95%)
ES-AD-4	144.35	229783	90695 (39.5%)	24863	38.63	115558	94.54	166469	83.80 (88.63%)
FR-AD-1	85.90	125260	66486 (53.1%)	15390	22.43	81876	67.37	118183	59.53 (88.37%)
FR-AD-2	73.83	113507	61151 (53.9%)	12439	18.1	73590	57.13	103732	50.48 (88.37%)
FR-AD-3	75.06	115862	55637 (48.0%)	14694	21.46	70331	57.19	100309	50.49 (88.29%)
FR-AD-4	79.60	120268	72738 (60.5%)	14808	19.41	87546	67.19	122497	59.14 (88.03%)
FR-AD-5	85.84	129745	70637 (54.4%)	13294	20.11	83931	66.17	119784	58.63 (88.60%)
FR-AD-6	75.84	118423	64043 (54.1%)	14112	19.33	78155	59.65	109207	52.21 (87.53%)
FR-AD-7	76.44	118172	56166 (47.5%)	14994	21.57	71160	57.08	101769	50.44 (88.37%)
FR-AD-8	71.98	112592	64959 (57.7%)	12266	16.38	77225	57.17	106497	50.42 (88.19%)
IT-AD-1	76.65	116244	43644 (37.5%)	13489	21.22	57133	49.2	84781	43.91 (89.25%)
IT-AD-2	79.21	115636	47103 (40.7%)	12461	21.52	59564	53.02	90859	47.54 (89.66%)
IT-AD-3	78.98	116746	57795 (49.5%)	16029	22.92	73824	61.49	107924	54.55 (88.71%)
IT-AD-4	80.28	116891	31691 (27.1%)	6606	15.12	38297	36.28	58967	31.47 (86.74%)
IT-AD-5	80.80	118227	62846 (53.2%)	14236	20.94	77082	63.14	111891	56.13 (88.90%)
IT-AD-6	80.39	116085	61669 (53.1%)	13766	20.16	75435	62.14	108567	55.04 (88.58%)
JP-AD-1	59.27	78123	16561 (21.2%)	14535	24.1	31096	35.58	54856	30.43 (85.54%)
JP-AD-2	59.94	80477	22788 (28.3%)	14961	22.99	37749	39.02	63230	33.95 (87.00%)
JP-AD-3	60.80	79846	20442 (25.6%)	17351	27.1	37793	41.31	64201	35.79 (86.63%)
JP-AD-4	60.50	78670	17634 (22.4%)	13537	23.51	31171	36.16	55693	31.17 (86.20%)
JP-AD-5	61.24	79773	19383 (24.3%)	12302	20.54	31685	34.47	54699	29.86 (86.64%)
JP-AD-6	61.43	79357	21669 (27.3%)	15134	26.06	36803	41.72	63735	36.10 (86.54%)
JP-AD-7	53.29	75532	15765 (20.9%)	5327	14.35	21092	24.49	37212	21.09 (86.11%)
JP-AD-8	60.41	80627	28252 (35.0%)	10390	19.86	38642	39.89	64333	34.23 (85.82%)
JP-AD-9	61.02	81346	17969 (22.1%)	16420	25.39	34389	38.23	59820	32.52 (85.08%)
JP-IN-1	59.63	80796	11452 (14.2%)	6136	15.35	17588	22.82	33993	19.10 (83.71%)
JP-IN-2	66.43	79972	5120 (6.4%)	1671	6.75	6791	10.69	14334	9.14 (85.51%)
JP-IN-3	62.21	79787	10324 (12.9%)	5647	12.22	15971	19.47	29305	16.57 (85.09%)
JP-IN-4	62.86	87324	11137 (12.8%)	6665	18.19	17802	23.88	34732	20.62 (86.34%)
AM-AD-1	58.66	65042	34718 (53.4%)	7113	15.16	41831	46.14	72772	40.38 (87.52%)
AM-AD-2	68.39	74452	27947 (37.5%)	9501	20.65	37448	46	69574	40.50 (88.03%)
AM-F10-T1	87.04	248939	117041 (47.0%)	33379	14.52	150420	52.68	152956	39.09 (74.21%)
AM-F10-T2	153.13	435911	132093 (30.3%)	46287	28.94	178380	70.44	188665	53.24 (75.58%)

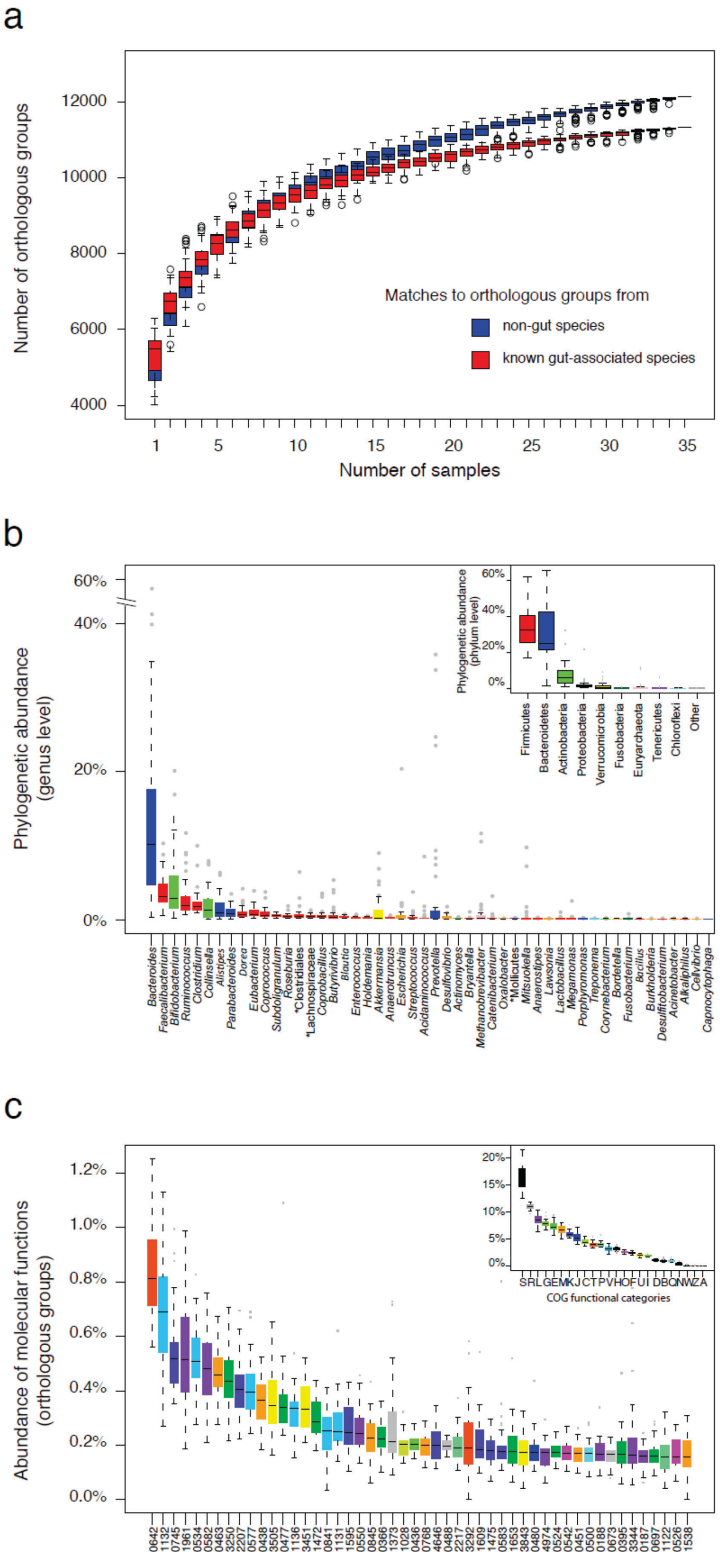


Figure 6-1. Functional and phylogenetic profiles of human gut microbiome.

a) Simulation of the detection of distinct orthologous groups when increasing the number of individuals (samples). Complete genomes were classified by habitat-information and the orthologous

groups divided into those that occur in known gut-species (red) and those that have not yet associated to gut (blue). The former are close to saturation when sampling 35 individuals (infants were excluded) whereas functions from non-gut (probably rare and transient) species are not. **b)** Genus abundance variation plot for the 50 most abundant genera and 5 phyla as determined by read abundance. Box-plots of the abundance of each genus in the samples, deduced from read abundance. Genera are colored by their respective phylum (see inset for color key). Inset: box-plot of the abundance of each phylum in the samples. Genus and phylum level abundances were measured using 85% and 65% sequence similarity cutoffs (Figure 2-1). Unclassified genera under a higher rank are marked by asterisks. **c)** Orthologous group (OG) abundance variation as a box plot for the 50 most abundant OGs and 24 functional categories as determined by assignment to eggNOG[74]. OGs are colored by their respective functional category (see inset for color key). Inset: box-plot of the abundance of each functional category in the samples.

both phylogenetic and functional composition (Pearson correlation coefficient r >0.988 and r >0.91, respectively; Supplementary Figure C-3, Supplementary Table C-1). These results imply that future samples from different sequencing technologies can be integrated and compared, provided that the sequencing coverage is sufficient to discriminate between meaningful and random variation.

As the sequence amounts per sample in our dataset (between 53 and 295 Mb) are somewhat arbitrary, they might bias the outcome of a comparison. We therefore simulated the total number of orthologous groups (OGs) that could be functionally assigned in relation to the number of sequenced samples (Figure 6-1a). As many genes might be 'bystanders' i.e. genes from transient, perhaps food-associated microbiota that just passage through the gut, we assigned habitat information to 1106 out of the 1152 reference genomes and distinguished between eggNOG[74] orthologous groups from gut and non-gut species. As expected, OGs found in known gut species seem to be close to saturation while functions from 'non-gut' species still accumulate with each sample at our given coverage of 53 to 295 Mb per individual (Figure 6-1a). Thus, although the coverage at hand will miss rare gut species and genes from these, the coverage seems sufficient to cover major trends caused by resident gut species and to robustly identify species and functionalities that are common and different between samples.

Global phylogenetic and functional variation of intestinal metagenomes

Only 0.14% of the reads could be classified as potential human contamination in the newly sequenced European samples (Supplementary Table C-2). This is very low considering that we used very lenient criteria to capture as many human sequences as possible (see Section 2.2.12). All other reads with best hits to eukaryotes together contribute 0.5%, with metazoan and fungal organisms accounting for more than half of these (Supplementary Table C-2 and Supplementary Table C-3). As on average only 0.5% of the fragments were of archaeal origin, the vast majority of the data is contributed by bacteria.

To characterize the phylogenetic and functional spectrum in all samples we measured the phylogenetic variation at the genus and phylum levels and the functional variation at the gene and functional class levels. As infants are known to have very heterogeneous and distinctive microbiota[29,128], we considered the four Japanese samples from infants as outliers and did not include them in the analysis. Using calibrated similarity cutoffs (Figure 2-1), on average, 48.1% of the fragments in each sample can be robustly assigned to a genus in our reference genome set (ranging from 19.1% to 88.8%), and 75.6% can be assigned to a phylum (ranging from 45.8% to 94.3%) implying that the trends observed (Figure 6-1b) represent the majority of the metagenome.

The phylum mapping agrees with earlier observations[14] that Firmicutes and Bacteroidetes constitute the vast majority of the dominant human gut microbiota (Figure 6-1b, inset) – these phyla make up 33.7% and 29.3% of the metagenome sequences, respectively. *Bacteroides* was the most abundant genus (Figure 6-1b), accounting on average for 14.2% of the sequences; it also had the highest variation among samples (ranging from 0.1% to 55%), agreeing with previous observations[29,145]. *Prevotella* showed the next highest variation (0.01% to 41.1%). *Ruminococcus*, *Clostridium*, *Faecalibacterium*, *Bacteroides*, *Parabacteroides*, *Alistipes*, *Bifidobacterium* and *Collinsella* were among the five most abundant genera from their respective phyla in at least 33 out of 35 samples (excluding infants) and five of these are among the 10 most abundant genera in a subset of 8 European samples analyzed

Table 6-3. Consistently top genera.

Genera that are almost always among the five most abundant from their respective phyla. Top 5 occurrences are counted in 35 samples. * - These genera are also among the 10 most abundant genera overall, in HITChip analysis of 8 MetaHIT samples (See Section 2.2.15).

Genus	Phylum	Top 5 occurrences
*Ruminococcus**	Firmicutes	34
Clostridium	Firmicutes	34
*Faecalibacterium**	Firmicutes	33
*Bacteroides**	Bacteroidetes	35
*Parabacteroides**	Bacteroidetes	33
*Alistipes**	Bacteroidetes	33
Bifidobacterium	Actinobacteria	35
Collinsella	Actinobacteria	34

by HITChip (Table 6-3). This implies that species from these genera are predisposed and/or selected to be among the abundant species in the gut environment regardless of geographic location.

Our protocol led to a high functional assignment rate: 64.4% of all predicted genes in the 33 Sanger-sequenced samples analyzed (41% of all predicted genes in two samples obtained by pyrosequencing; Supplementary Table C-4) can be assigned to OGs, although some of these have no or only loose functional descriptions ('unknown' and 'general functions' account for 16.2% and 11%, respectively) comparable to other metagenomic samples from diverse habitats[31]. However, functionally uncharacterized genes usually form small OGs (Supplementary Figure C-1) while large OGs with many genes are usually well-characterized[31] and their variation can be interpreted. The most frequent OG is formed by histidine kinases (COG0642) contributing on average 0.8% of all assigned genes in each sample and implying intensive signaling in this community, for example, triggered by environmental (nutritional or stress-related) compounds or in the context of specific quorum sensing communication[150]. This is in accordance with the observed expansion of signaling genes in *Bacteroides thetaiotaomicron* in mouse gut[151], but it clearly extends also to other genera such as *Prevotella* (Section C.3.1). The most variable OG is the ATPase component (COG1132) of ABC-type transporters (ranging from 0.2% to 1.1%, Figure 6-1c) which, as the most conserved domain in this type of transport systems, indicates different

levels of exchange within the microbial communities of the different individuals[152] (this OG is also variable in the genomes annotated in STRING; Section C.3.1).

Highly abundant functions from low-abundance microbes

Microbes in the human gut undergo selective pressure from the host as well as from microbial competitors. This typically leads to a homeostasis of the ecosystem in which some species occur in high and many in low abundance[153] (the "long-tail" effect, as seen in Figure 6-1b), with some low-abundance species, like methanogens[154], performing specialized functions beneficial to the host. The presence of abundant functions shared by several low-abundance species could shed light on their survival strategies in the human gut. To identify such functions in the samples, we used genus mappings of the constitutive reads of each gene to estimate the contributions of different genera to the functions retrieved (Section 2.2.6). In general, the most abundant molecular functions come from the most dominant species. However, we identified some abundant orthologous groups that are contributed primarily by low abundant genera (see Supplementary Table C-5). For example, the growth inhibitor of the mazE/mazF toxin-antitoxin (TA) system (COG2337) and a DNA-damage-inducible antitoxin of the yafQ/dinJ TA system (COG3077) are among the top 20% abundant functions in some European samples (Supplementary Table C-6). TA systems help bacteria in nutritional stress response by tightly controlling reproduction and thus the population size[155]. This could be a self regulatory mechanism carried out by these species in the human gut to maintain a sustainable population size under continuously changing resource levels. A few low-abundance (<2.5%) genera from Firmicutes contribute more than 50% of these two OGs in these samples, implying that hitherto unidentified, potential high-abundant species will not be able to change this observation. Low-abundance genera belonging to Enterobacteriales, including *Escherichia*, contribute even over 90% of two other abundant proteins associated with bacterial pilus assembly, FimA (COG3539) and PapC (COG3188), found in one individual (IT-AD-5). Pili are hair-like structures on the surface of microbes that enable them to colonize the epithelium of specific host organs; they help microbes to stay longer in the human intestinal tract by binding to the human mucus or mannose sugars present on intestinal surface structures[156-157]. They are also key components in the transfer of plasmids between bacteria through conjugation, which often

exchanges beneficial functions such as antibiotic resistance[158]. Pili can thus provide multiple benefits to these low-abundance microbes in their efforts to survive and persist in the human gut. Such examples imply that abundant species or genera can not reveal the entire functional complexity of the gut microbiota. More reference genomes will facilitate better gene assignments from samples and thus the detection of more low abundance species. However, there is not much room for as yet undetected, abundant genera. Even with our limited genus assignment rate of 48.1% of all reads, we estimate that we miss another 28% of the already classified genera due to our strict assignment criteria (Figure 2-1), i.e. only 23.9% of all reads are likely to belong to hitherto unknown genera.

Robust clustering of samples across nations: Identification of enterotypes

To compare the individual samples phylogenetically and functionally, we used several independent metrics for both clustering and principal component analysis (PCA). First, to get an overview of the species variation we used phylogenetic profile similarities obtained from the mapping of reads to the 1152 reference genomes (Figure 6-2a, see Section 2.2.4). Independently, we also extracted 16S genes in the metagenomes to cluster the samples (Figure 6-2b). Both conceptually different methods reveal very similar groupings of the samples (Figure 6-2) that extend across nations and even continents, despite the obvious differences in environment, food habits and genetic background. The consistent results also show the robustness of these clusters and establish the comparability of data from different sequencing centers as well as the sufficiency of coverage for classifying samples based on these profiles, when treated in a computationally consistent manner[49] even at the limited sequencing depth of 105 Mb per sample. Second, we clustered the samples using a purely functional metric: the abundance of the assigned orthologous groups (Figure 6-2c). Remarkably, this clustering also showed a similar grouping of the samples with only minor differences indicating that function and species composition roughly coincide. Sanger- and pyrosequencing-based sequences from the same samples cluster

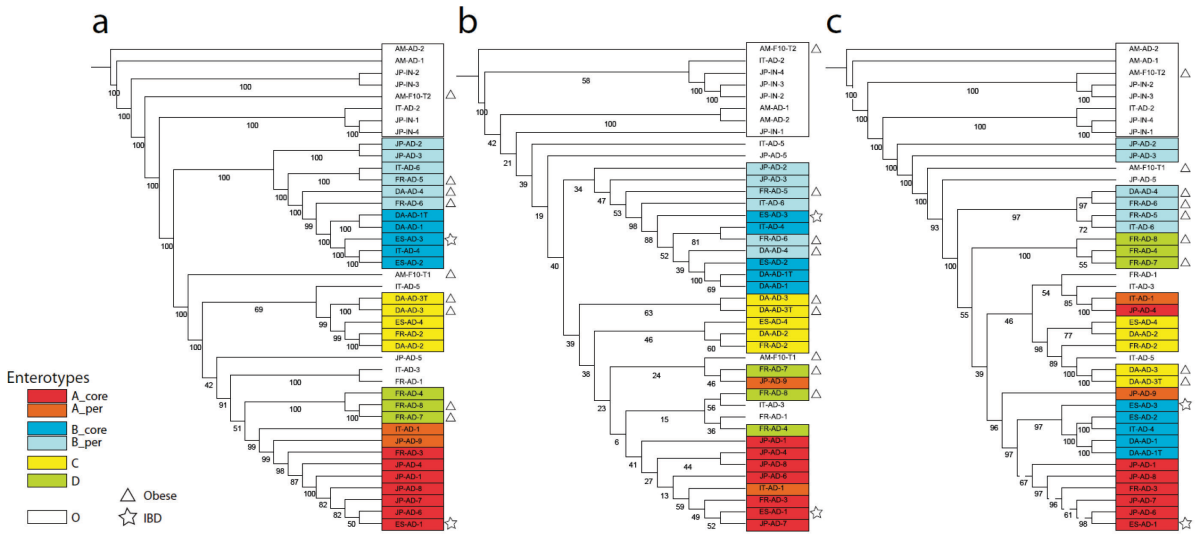


Figure 6-2. Clustering of Enterotypes

Consistent inter-ethnic/inter-sequencing center subclusters from different methods. 41 samples were clustered using different features of the sample metagenomes: (a) genus abundance estimated by mapping the metagenome reads to 1152 reference genome sequences using an 85% similarity threshold, (b) genus abundance estimated using 16S rRNA reads identified from the metagenome reads and (c) eggNOG orthologous group abundance (see Sections 2.2.4 and 2.2.5). These resulted in consistent subclusters that we call enterotypes. Identified enterotypes are colored (see figure for color key) and outliers are marked by an empty box around them. Obese and IBD individuals are marked by triangles and stars, respectively. Clusters also show bootstrap values at each node.

together in all cases, reinforcing the feasibility of comparisons across sequencing platforms. In 10 independent 50% jack-knife tests, the clustering of individuals remains largely the same implying that the grouping is robust towards the addition of more samples (Supplementary Figure C-5 and Section 2.2.14). A few samples clustered with the infants and were therefore also considered as outliers and not analyzed further (e.g., IT-AD-2 is dominated by *Klebsiella* and the two US samples AM-AD-1 and AM-AD-2 have an extremely low fraction of Bacteroidetes, Supplementary Figure C-6). Within the remaining 31 samples we defined four robust groups of at least three samples (Figure 6-2). We interpret this clustering as being the result of well-balanced, defined microbial community compositions, suggesting that there exists a limited number of types rather than a 'continuum' of different gut microbiota constellations

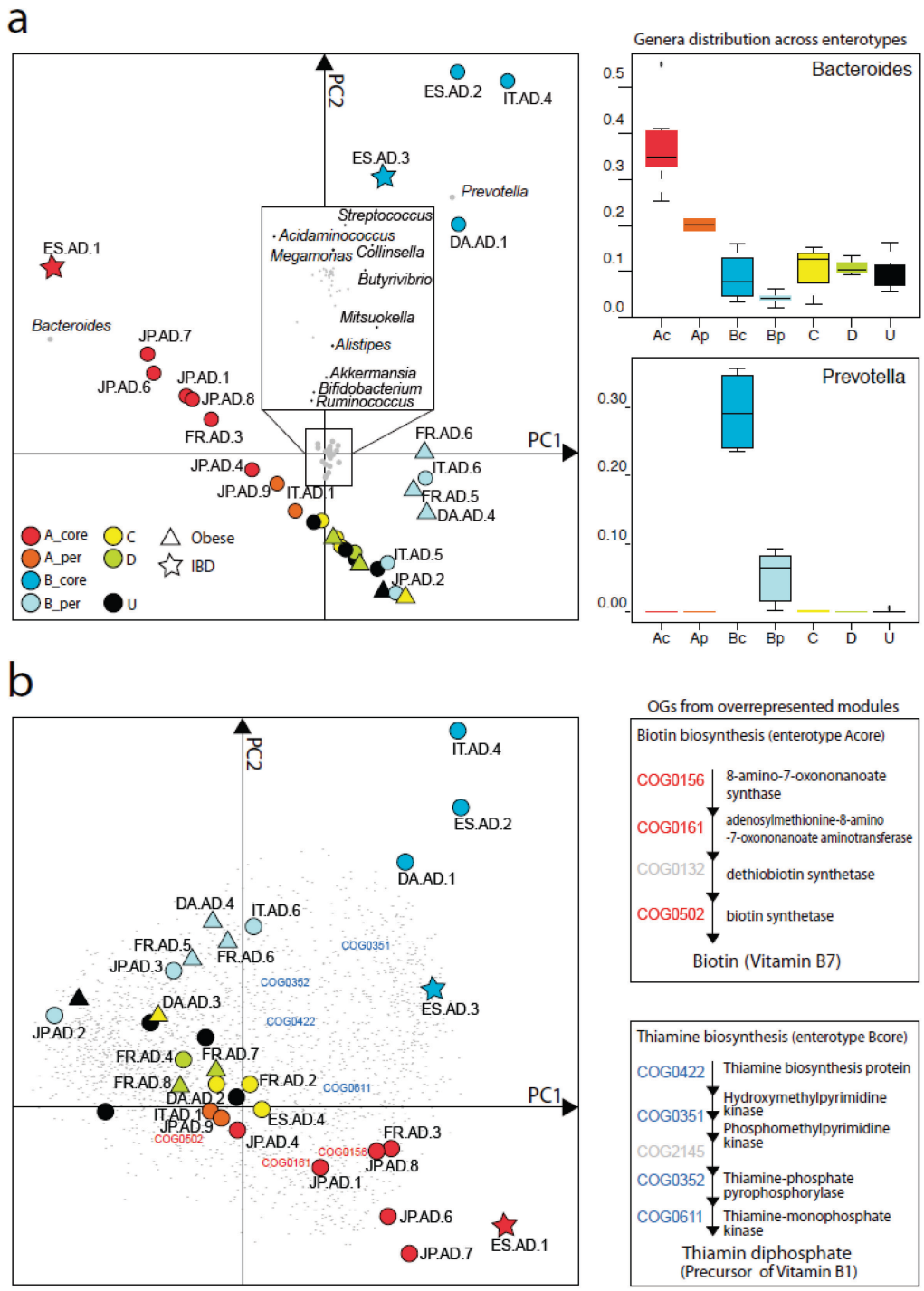


Figure 6-3. Principal component analysis of genus and orthologous group (OG) profiles.
PCA of genus abundances and OG abundances show that the first two components clearly separate the enterotypes. Healthy, obese and IBD individuals are represented by filled circles, triangles and stars, respectively, and colored by their enterotype affiliation. **(a)** Separation of enterotypes seen by the first two components (accounting for 60.5% and 26.4% of the variance) of the genus abundance

PCA. *Bacteroides* and *Prevotella* variations are clearly the major drivers for enterotypes A and B. Inset: combination of several other genera contributes to the separation of the other enterotypes. Top right: *Bacteroides* is overrepresented in Acore and Aper. Bottom right: *Prevotella* is overrepresented in Bcore and Bper. **(b)** Separation of enterotypes seen by the first two components (accounting for 20.8% and 9.4% of the variance) of the OG abundance PCA. Combination of several functions contributes to the separation of Acore and Bcore. Top right: Steps involved in biotin biosynthesis. Three out of the four enzymes (belonging to orthologous groups COG0156, COG0161 and COG0502) are overrepresented in Acore. Bottom right: Steps involved in thiamin biosynthesis. Four out of the five enzymes (belonging to orthologous groups COG0422, COG0351, COG0352 and COG0611) are overrepresented in Bcore.

across individuals. Such 'enterotypes', as we define them here, are in line with previous reports that gut microbiota is rather stable in individuals and can even be restored after perturbation [12,159-161]. As our current data do not reveal which environmental or even genetic factors are causing the clustering, and as fecal samples might not be representative of the entire intestine, we anticipate that the enterotypes introduced here will be refined with deeper and broader analysis of individuals' microbiomes. Our operational definition of the enterotypes is based on the most robust reference-genome clustering (see Section 2.2.13); two enterotypes, A and B, were further subdivided into core and periphery (Acore, Bcore, Aper, Bper hereafter) to accommodate the differences in the functional clustering (Figure 6-2).

As this is a far reaching concept we further validated the enterotypes by two independent experimental approaches: we analyzed 8 of the European samples (4 Danish and 4 Spanish ones) using customized ultra-high density DNA microarrays and HITChip[88] (see Section 2.2.15), and both approaches robustly agreed with the enterotypes defined above (Supplementary Figure C-7).

Phylogenetic and functional variation between enterotypes

To determine the phylogenetic and functional basis of the enterotypes, we investigated, using PCA, which species and functions are commonly over- or underrepresented in each enterotype and what their differences are (Figure 6-3; also see Sections 2.2.8 and 2.2.9). The largest enterotype (Acore) comprising 7 samples is enriched in the genera Bacteroides and Acidaminococcus ($p<0.01$; Figure 6-3a) and is functionally characterized by a specific overrepresentation of genes involved in vitamin biosynthesis (biotin, riboflavin, pantothenate, pyridoxal phosphate and cobalamin) and degradation

of (complex) carbohydrates and proteins (galactosidases, hexosaminidase and peptidases). The latter suggests that it constitutes a cluster in which the early stage of the bacterial food chain is most prominent[46] (see Supplementary Table C-7 and Supplementary Table C-8 for a full list of functions enriched in each enterotype). Enterotype Bcore (6 samples) is highly enriched in the genera *Prevotella* and *Megamonas* ($p<0.01$; Figure 6-3a and Supplementary Figure C-8), and functional analysis shows that vitamin biosynthesis pathways different from those in Acore, thiamine and folate, are enriched, as are several amino acid biosynthesis / degradation pathways (e.g., tryptophan, proline and leucine). Enterotypes C and D contain only 4 and 3 samples respectively. C is enriched in *Akkermansia* and *Alistipes* ($p<0.01$; Supplementary Figure C-9), as well as in replication and translation functions. D is enriched in *Ruminococcus* species ($p<0.01$; Supplementary Figure C-10), which are known to degrade mucins[162]. It is also enriched in membrane transporters, mostly of sugars, suggesting the efficient binding of mucin and its subsequent hydrolysis as well as uptake of the resulting simple sugars by these genera.

We also found high-abundant functions in enterotypes that were encoded in low-abundance genera. For example, carbon-nitrogen hydrolases (COG3049) involved in conjugated bile acid biosynthesis or degradation of penicillin are only encoded by known low abundance genera. These functions are only abundant in Bper but not in Bcore thus contributing to their separation in the functional clustering (Figure 6-2).

Phylogenetic and functional biomarkers for host properties

The general functional composition of the enterotypes does not correlate with any of several measured host properties, namely nationality, gender, age or body mass index (BMI). However, some strong correlations to these properties occur for particular phylogenetic groups, genes or functional modules (See Section 2.2.10).

With regard to nationality, the abundance of the genus *Eubacterium* (Firmicutes) varies more between countries than within (Figure 6-4a) as do those of *Dorea* and *Coprobacillus* (both Firmicutes), albeit to a lesser extent. As the abundance of members of *Eubacterium* is known to be influenced with a change in diet[163-164], the differences in food habits by the different ethnic groups in our study likely explain this

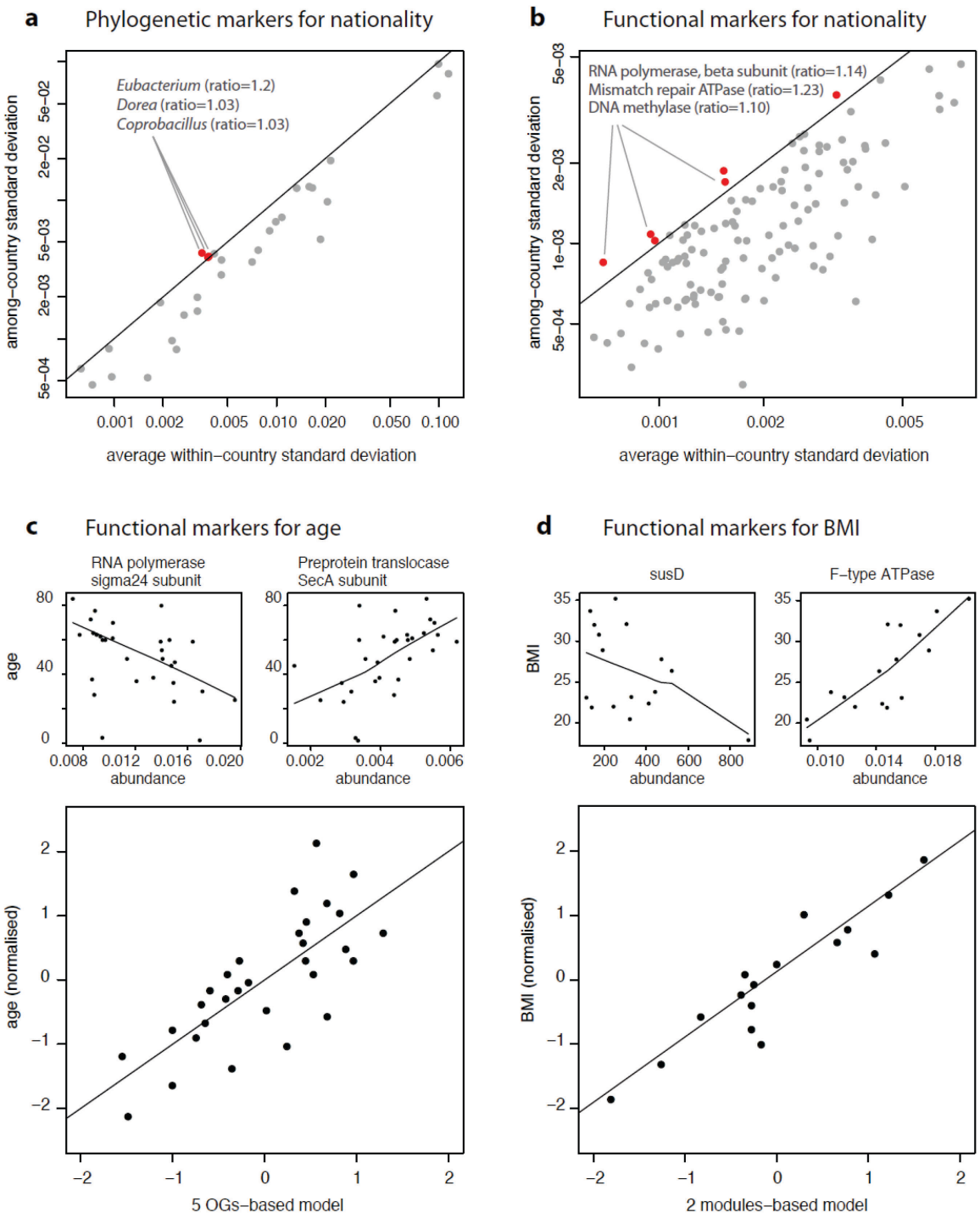


Figure 6-4. Correlations with host properties.

(a) Variation of genus abundance with nationality. The plot compares the among-nationality standard deviation (SD) with the within-nationality SD. Points above the diagonal (red, discussed in text) represent genera whose abundance varies more among than within nationalities. (b) Variation of orthologous groups (OGs) with nationality. See Table 6-4 for a full list. (c) Selected orthologous

groups whose abundance correlates with host age. Top: pairwise correlations (left, RNA polymerase facultative sigma24 subunit (COG1595), p=0.04, rho=-0.56; right, preprotein translocase secA subunit (COG0653), p=0.01, rho=0.65). Bottom: multiple significantly correlating OGs (COG0085, COG0086, COG0205, COG0739 and COG1595; see Supplementary Table C-11) combined into a linear model (see Section 2.2.10 and[27] for details; p= 6.27e-04, adjusted R2=0.48). (d) Selected genes modules whose abundance correlate with host body mass index (BMI). Top: pairwise correlations (left: SusD, a family of proteins that bind glycan molecules before they are transported into the cell, which correlates weakly, p=0.27, rho=-0.29; right: F-type ATPase (M00286), p=0.04, rho=0.78). Bottom: 2 modules, ATPase and ectoine biosynthesis (M00051), combined into a linear model (p= 6.786e-06, adjusted R2=0.82).

observation and also suggest that other Firmicutes might be affected. The phylogenetic differences were paralleled by functional ones: the abundance of several OGs (mismatch repair ATPases, DNA methylases and DNA polymerases; Table 6-4) varies more between than within nationalities (Figure 6-4b) and several functionalities are specifically overrepresented in different ethnic groups (e.g., a polar amino acid transporter module in Japanese individuals; see Supplementary Table C-9 and Supplementary Table C-10), again potentially linked to nutrition (e.g., the strong presence of glutamate in Japanese diet[165]). Generally speaking, however, the phylogenetic and functional compositions of the metagenomes of individuals from different nations were similar at the given sequencing depth. For example, the core metagenome (the set of functions present in all individuals) has a similar size in each nation, suggesting that the core functioning of the human intestine is similar in different ethnicities (Supplementary Figure C-11). This also confirms the similarities of gut microbiomes from a large cohort of Danish and Spanish individuals[84].

For the other host properties tested, no significant phylogenetic markers could be found; our data did also not show any correlation between BMI and the Firmicutes/Bacteroidetes ratio and we thus cannot contribute to the ongoing debate on the relation between this ratio and obesity[15-17]. However, several functional markers for host properties were identified after correcting for multiple testing to avoid artefacts (Section 2.2.10). For example, five functional modules and three orthologous groups (OGs) significantly correlate with gender ($p<0.05$) with its only two distinct states (e.g., enriched aspartate biosynthesis modules in males; see Supplementary Table C-10). 11 OGs significantly correlate with age (Table 6-5), some of which

Table 6-4. Functions varying more between than within countries.

OG	Description
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit
COG0249	Mismatch repair ATPase (MutS family)
COG0587	DNA polymerase III, alpha subunit
COG2207	AraC-type DNA-binding domain-containing proteins
COG3291	FOG: PKD repeat
COG4646	DNA methylase

Table 6-5. Orthologous groups significantly correlating with age.

Correlating orthologous groups with the R^2 and P-values for each correlation. Functions mentioned in the main text are emphasized in bold text.

OG	Description	R^2	P-value
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit	0.609151	0.014845
COG0086	DNA-directed RNA polymerase, beta subunit/160 kD subunit	0.510297	0.045398
COG0187	Type IIA topoisomerase	0.517645	0.047501
COG0205	6-phosphofructokinase	-0.5526	0.032389
COG0493	NADPH-dependent glutamate synthase beta chain and related	0.527886	0.042783
COG0653	Preprotein translocase subunit SecA	0.645664	0.01469
COG0739	Membrane proteins related to metalloendopeptidases	-0.54904	0.030202
COG1595	DNA-directed RNA polymerase specialized sigma subunit, sigma-24 homolog	-0.55972	0.040933
COG2207	AraC-type DNA-binding domain-containing proteins	-0.5595	0.032924
COG3291	FOG: PKD repeat	0.513414	0.046772
COG4646	DNA methylase	0.614272	0.019217

Table 6-6. Functional modules significantly correlating with the body mass index of individuals.

Module	Description	R^2	P-value
M00051	Ectoine biosynthesis	-0.8	0.050253
M00286	F-type ATPase (Bacteria)	0.779412	0.036732
M00293	ATP synthase	0.779412	0.036732

increase in abundance with age (e.g., the secA preprotein translocase; see Figure 6-4c), others decrease. An example of the latter is an OG coding for the facultative sigma-24 subunit of RNA polymerase, which drives expression under various stress responses and is linked to intestinal survival[166]. One explanation for its decrease with age could be the reduced need for stress response in the gut due to the age-associated decline in host immune response[167] (immunosenescence). Our analyses also identified three marker modules that correlate strongly with the hosts' BMI (Table 6-6), two of which are ATPases, confirming the link found between the gut microbiota's capacity for energy harvest and host's obesity[18], although causality cannot be determined at this point. Interestingly, functional markers found by a data-driven approach gave much stronger correlations than genes that would be suspected to show an effect such as SusC/SusD, the starch binding components of the starch utilization system[168-169] (Figure 6-4d).

With the current sequencing depth in our data set, additional phenotypic classification attempts such as those based on hydrogenotrophic microorganisms (methanogens, reductive acetogens or sulphate reducers) could not be verified, as the respective marker genes (e.g., coenzyme-M reductase mcrA[170], formyltetrahydrofolate synthetase, or dissimilatory sulphite reductase dsrA/dsrB) from these less abundant microbes could barely be identified. For example, mcrA was only found in 3 out of the 22 European samples, although 30-50% of the western population are estimated to have methanogenic bacteria in their faeces[154,171]. We see this rather as a strength of our unbiased approach as the hydrogenotrophy-based classification scheme is arbitrary and the enterotypes seem to be functionally mostly driven by complex community properties which could be caused by hitherto unexplored physiological conditions such as transit time of luminal context.

Taken together, we have demonstrated the existence of enterotypes across several nations and continents that vary in species and functional composition. Although more and deeper sequencing will certainly lead to a more fine-grained classification of the enterotypes, their existence implies a limited number of balanced and reasonably stable symbiotic host-microbe interaction states in the human population, which is unprecedented and unexpected. Presumably, enterotypes are not limited to humans but also occur in animals. Their future investigations might well reveal novel facets of

the human and animal symbiotic biology and lead to the discovery of the microbial properties correlated with the health status of individuals. We anticipate that they might allow classification of human groups that respond differently to diet or drug intake. The enterotypes appear complex, are probably not driven by nutritional habits and cannot simply be explained by host properties such as age or BMI, although there are functional markers such as genes or pathways that correlate remarkably well with individual features. We have shown that unbiased, data-driven approaches can outperform the usage of knowledge-based molecules such as the starch-binder SusD (Figure 6-4d), and can lead to the discovery of novel functional markers in the human fecal microbiota that correlate well with heterogeneous host properties. This augurs the possibility to use such markers as diagnostic and perhaps even prognostic tools for numerous human disorders, for instance obesity-linked co-morbidities such as metabolic syndrome, diabetes and cardio-vascular pathologies.

Methods summary

Sample collection: Human fecal samples from European individuals were collected and frozen immediately, and DNA was purified as described previously[59]. **Sequencing:** Sanger sequencing was performed using standard protocols. Shotgun randomly shared DNA libraries were constructed using low copy plasmid (pCNS, 3 kb insert). Terminal clone end sequences were determined using BigDye terminator chemistry and capillary DNA sequencers (3730XL, Applied Biosystems) according to standard protocols established at Genoscope. **Sequence processing:** Cloning vector and sequencing primers were removed from raw Sanger reads using BLASTN identification. Reads were then quality trimmed by removing low quality bases in either end. Possible human DNA sequences were then removed by aligning reads against human genome. Reads were then processed by the SMASH comparative metagenomics pipeline[61] to assemble them and to predict genes on assembled contigs and singletons. **Phylogenetic annotation:** Phylogenetic annotation of samples was performed (1) using BLASTN of reads against a database of 1152 reference genomes[23] and (2) identifying 16S rRNA gene containing reads and classifying them using RDP classifier[72]. Species abundance was estimated after normalizing for genome size for the former, and for 16S gene copy number for the latter. Two different genus abundance profiles of samples were created

from phylogenetic annotation. **Functional annotation:** Genes were functionally annotated using BLASTP against KEGG (v50) and eggNOG (v2) databases. Protein abundances were estimated after normalizing for protein length. Functional abundance profiles of samples were created at the KEGG and eggNOG orthologous group levels as well as functional modules and pathways. **Clustering:** Samples were clustered using Jensen-Shannon distance, a relative entropy based distance measure, and neighbor-joining method. 100 bootstrap replicates were generated from each of the three profiles. Enterotypes were defined by identifying bottom-up clusters with bootstrap support over 80% in the reference genome mapping tree (Fig. 2a). Only some samples of two clusters differ in the functional mapping and were partially separated. **Statistics and correlations with host properties:** Correlations between metadata and feature abundances were done as described previously[27], based on multiple-testing corrected pairwise Spearman correlation analysis and stepwise regression for multi-feature model building. For categorical metadata and enterotype comparisons, samples were pooled into bins (male/female, obese/lean, enterotype/rest, specific nationality/rest etc.) and significant features were identified using Fisher's exact test with multiple testing correction of p-values. We only took those features into account that were specifically overrepresented in only one of the groupings tested (e.g., only in one enterotype) and highlighted case studies were manually scrutinized to avoid artefacts.

Chapter 7

Discussion

"Ah! My child, how I would like to have a new life in front of me. With what pleasure I would take up again my crystallographic studies. I should never have abandoned my crystals."

Louis Pasteur to his grandson, ca. 1895 [172]

This cumulative thesis focuses on two strong publications that appeared in Nature and Science, and a third paper (my first author paper) that has been submitted to Nature. Analysis of the human gut microbiome, the microbial community that resides in our gut, is a formidable task since the gut microbiome is estimated to harbor 500 to 1000 different species[3] and 100 times as many genes as in the human genome[123]. Such analysis would not have been possible just a few years ago, and is now within our reach with the advent of high-throughput sequencing technologies and the next generation sequencing technologies. The bioinformatic analysis of human gut metagenomes is no less challenging with hundreds of genomes and millions of genes mixed together in each metagenomic sample. The basic analysis and functional annotation alone of a Sanger-sequencing based metagenome containing 100Mb takes close to 300 CPU hours, and the available metagenomic analysis tools have limitations.

To enable the detailed analysis of the human gut microbiome, I have developed SMASH, a new computational pipeline for comparative metagenomic analysis. It is the first stand-alone metagenomic analysis tool that provides efficient metagenome assembly thereby improving the functional annotation of metagenomes. It is also the first tool to estimate quantitative phylogenetic composition of metagenomes after correcting for genome size or copy number of the 16S rRNA gene. SMASH also provides a novel quantitative functional characterization of metagenomes at the orthologous group and functional module levels. It can use these phylogenetic and functional profiles to cluster multiple samples with bootstrap analysis. Although developed for metagenomic analysis, SMASH has also been successfully applied in analyzing the bacterium *Mycoplasma pneumoniae* and the thermophilic fungus *Chaetomium thermophilum*.

In the context of MetaHIT, I used SMASH to validate the assembly of metagenomes derived from the total fecal DNA from a cohort of 124 individuals of European (Nordic and Mediterranean) origin using Illumina Solexa technology. I estimated the assembly

error rate and showed it to be within acceptable limits and comparable to that of metagenomes obtained using 454 Titanium technology. These metagenomes were used to derive a gene catalogue containing 3.3 million microbial genes, 150-fold more than the human gene complement, which includes an overwhelming majority (>86%) of prevalent genes harbored by our cohort. The full bacterial gene complement of each individual was not sampled in our work. Nevertheless, we have detected some 536,000 prevalent unique genes in each, out of the total of 3.3 million carried by our cohort. Inevitably, the individuals largely share the genes of the common pool. At the present depth of sequencing, we found that almost 40% of the genes from each individual are shared with at least half of the individuals of the cohort. The gene catalogue is equivalent to that of some 1,000 bacterial species with an average-sized genome, encoding about 3,364 non-redundant genes. We estimate that no more than 15% of prevalent genes of our cohort may be missing from the catalogue, and suggest that the cohort harbors no more than ~1,150 bacterial species abundant enough to be detected by our sampling. A large number of shared species supports the view that the prevalent human microbiome is of a finite and not overly large size. Detailed comparisons of bacterial genes across the individuals of our cohort will be carried out in the future, within the context of the ongoing MetaHIT clinical studies of which they are part. Nevertheless, clustering of the genes in families allowed us to capture a virtually full functional potential of the prevalent gene set and revealed a considerable novelty, extending the functional categories by some 30% in regard to previous work[30]. Similarly, this analysis has revealed a functional core, conserved in each individual of the cohort, which reflects the full minimal human gut metagenome, encoded across many species and probably required for the proper functioning of the gut ecosystem. The size of this minimal metagenome exceeds several-fold that of the core metagenome reported previously[30]. By analyzing the largest metagenomic dataset, we provide a proof of principle that short-read sequencing can be used to characterize complex microbiomes.

I analyzed the Sanger-sequencing and pyrosequencing based human gut metagenomes of 39 individuals from 6 countries using SMASH, and showed that comparative metagenomics of samples of diverse origins – different nationalities, sequencing centers and sequencing technologies – is feasible when they are treated in a computationally

consistent manner. I estimated the phylogenetic composition of the samples at various taxonomic ranks after establishing a rank-specific sequence similarity threshold for accurate phylogenetic assignment of reads. I also estimated the functional repertoire of each microbiome by a novel quantitative functional assignment method using homology to known orthologous groups. I identified that some highly abundant functions are primarily contributed by lower abundance species, reinforcing the need for molecular functional characterization of the gut microbiome for a community understanding. Using these 39 samples, we have demonstrated the existence of enterotypes across several nations and continents that vary in species and functional composition. Although more and deeper sequencing will certainly lead to a more fine-grained classification of the enterotypes, their existence implies a limited number of balanced and reasonably stable symbiotic host-microbe interaction states in the human population, which is unprecedented and unexpected. Presumably, enterotypes are not limited to humans but also occur in animals. Their future investigations might well reveal novel facets of the human and animal symbiotic biology and lead to the discovery of the microbial properties correlated with the health status of individuals. We anticipate that they might allow classification of human groups that respond differently to diet or drug intake. The enterotypes appear complex, are probably not driven by nutritional habits and cannot simply be explained by host properties such as age or body mass index, although there are functional markers such as genes or pathways that correlate remarkably well with individual features. We have shown that unbiased, data-driven approaches can outperform the usage of knowledge-based molecules, and can lead to the discovery of novel functional markers in the human fecal microbiota that correlate well with heterogeneous host properties. This augurs the possibility to use such markers as diagnostic and perhaps even prognostic tools for numerous human disorders, for instance obesity-linked co-morbidities such as metabolic syndrome, diabetes and cardio-vascular pathologies.

Understanding the human-microbial crosstalk in the human gut environment as well as other body sites will lead to improved diagnosis of and specialized treatments to microbe-associated human disorders. The International Human Microbiome Consortium, officially launched in Heidelberg in October 2008, aims to “work under a common set of principles and policies to study and understand the role of the human

microbiome in the maintenance of health and causation of disease and to use that knowledge to improve the ability to prevent and treat disease"[173]. Such studies with common, streamlined and consistent working principles will provide comprehensive resources for researchers to understand the human-microbial interactions and enable the elucidation of their relationship with human health and well-being.

Appendix A

Supplementary information for Chapter 4

A.1 Supplementary Figures

The following figures are available as supporting material on Science online[103].

Supplementary Figure A-1. High resolution transcriptome mapping data.
Figure S1 online.

Supplementary Figure A-2. Tiling array and DSSS data comparison.
Figure S2 online.

Supplementary Figure A-3. Expression of some small RNA were sometimes independent from that of neighboring genes.
Figure S3 online.

Supplementary Figure A-4. Conservation of a new ncRNA, NEW8, that is specifically regulated under different conditions.
Figure S4 online.

Supplementary Figure A-5. Conservation and dynamics of NEW87.
Figure S5 online.

Supplementary Figure A-6. Gene expression distribution.
Figure S6 online.

Supplementary Figure A-7. Distribution of polycistronic operon sizes.
Figure S7 online.

Supplementary Figure A-8. Promoter architecture.
Figure S8 online.

Supplementary Figure A-9. Distribution of the 5'UTR sizes in *M. pneumoniae*.
Figure S9 online.

Supplementary Figure A-10. Examples of operons having alternative transcripts.
Figure S10 online.

Supplementary Figure A-11. Graph representation of the co-expression matrix obtained from the analysis of all arrays described in this work.
Figure S11 online.

Supplementary Figure A-12. Coexpressed genes.
Figure S12 online.

Supplementary Figure A-13. Overexpression of the transcription factor *fur* (ferric uptake regulator) shows evidence of complex regulation.
Figure S13 online.

Supplementary Figure A-14. Noise assessment of genes not clustering together.
Figure S14 online.

A.2 Supplementary Tables

The following tables are available as supporting material on Science online[103].

Supplementary Table A-1. Changes in the *M. pneumoniae* M129 genome after 33 passages revealed by re-sequencing.
Table S1 online.

Supplementary Table A-2. New genes detected.
Table S2 online.

Supplementary Table A-3. qPCR results. Ctl1 and Ctl2 are regions without any expression detected by tiling arrays.
Table S3 online.

Supplementary Table A-4. Reference operons.
Table S4 online.

Supplementary Table A-5. Operon and alternative transcript details.
Table S5 online.

Supplementary Table A-6. Gene expression clustering.
Table S6 online.

Supplementary Table A-7. Detail of array experiments carried out.
Table S7 online.

Supplementary Table A-8. Detail of tiling array experiments carried out.
Table S8 online.

Appendix B

Supplementary information for Chapter 5

B.1 Supplementary Figures

The following figures are available as supplementary information on Nature online[132]

Supplementary Figure B-1. Coverage of Sanger sequencing reads by Illumina GA reads.
Supplementary Figure 1 online.

Supplementary Figure B-2. Distribution of Illumina GA sequencing read coverage of each Sanger read.
Supplementary Figure 2 online.

Supplementary Figure B-3. Flowchart of human gut microbiome data analysis process.
Supplementary Figure 3 online.

Supplementary Figure B-4. Length distribution of assembled contigs.
Supplementary Figure 4 online.

Supplementary Figure B-5. Validating Illumina contigs using Sanger reads.
Supplementary Figure 5 online.

Supplementary Figure B-6. Unique Genes.
Supplementary Figure 6 online.

Supplementary Figure B-7. Number of unique genes identified with increase of sequencing depth in sample MH0006 and MH0012.
Supplementary Figure 7 online.

Supplementary Figure B-8. Distribution of non-redundant bacterial genes in IBD patients and healthy controls.
Supplementary Figure 8 online.

Supplementary Figure B-9. Relations between the most abundant bacterial species.

Supplementary Figure 9 online.

Supplementary Figure B-10. a, The number of genes assigned to different clusters is correlated with the protein length. b, The effect of copy number (CN) normalization to a single copy is illustrated for RNA polymerase.

Supplementary Figure 10 online.

Supplementary Figure B-11. Distribution of the range clusters across the eggNOG genomes.

Supplementary Figure 11 online.

B.2 Supplementary Tables

Supplementary Table B-1. DNA sample information.

All Danish individuals in the present subsample were originally recruited from a larger population-based sample of middle-aged people living in the northern part of Copenhagen region and sampled from the centralized personal number register. At the original recruitment the individuals included in the present study had normal fasting plasma glucose and normal 2 hour plasma glucose following an oral glucose tolerance test. At the time of fecal sampling all were examined in the fasting state and had non-diabetic fasting plasma glucose levels below 7,0 mmol/l. All of the IBD patients were in clinical remission at the time of fecal sampling. N refers to no IBD, CD & UC to Crohn's disease and ulcerative colitis, respectively.

Sample Name	Country	Gender	Age	BMI	IBD
MH0001	Denmark	female	49	25.55	N
MH0002	Denmark	female	59	27.28	N
MH0003	Denmark	male	69	33.19	N
MH0004	Denmark	male	59	31.18	N
MH0005	Denmark	male	64	21.68	N
MH0006	Denmark	female	59	22.38	N
MH0007	Denmark	male	69	33.60	N
MH0008	Denmark	male	59	24.35	N
MH0009	Denmark	male	64	29.04	N
MH0010	Denmark	male	64	33.27	N
MH0011	Denmark	female	0	22.31	N
MH0012	Denmark	female	42	32.10	N
MH0013	Denmark	male	54	20.46	N

MH0014	Denmark	female	54	38.49	N
MH0015	Denmark	male	59	25.47	N
MH0016	Denmark	female	49	30.50	N
MH0017	Denmark	male	64	21.81	N
MH0018	Denmark	male	49	31.37	N
MH0019	Denmark	female	44	20.01	N
MH0020	Denmark	female	63	33.23	N
MH0021	Denmark	female	49	25.42	N
MH0022	Denmark	male	64	24.42	N
MH0023	Denmark	male	69	31.74	N
MH0024	Denmark	female	59	22.72	N
MH0025	Denmark	female	49	34.20	N
MH0026	Denmark	female	49	37.32	N
MH0027	Denmark	female	59	23.07	N
MH0028	Denmark	female	44	22.70	N
MH0030	Denmark	male	59	35.21	N
MH0031	Denmark	male	69	22.34	N
MH0032	Denmark	male	69	35.28	N
MH0033	Denmark	female	59	31.95	N
MH0034	Denmark	male	54	39.97	N
MH0035	Denmark	male	49	22.66	N
MH0036	Denmark	male	64	30.74	N
MH0037	Denmark	male	44	24.02	N
MH0038	Denmark	female	54	21.97	N
MH0039	Denmark	male	58	23.07	N
MH0040	Denmark	female	67	20.87	N
MH0041	Denmark	male	59	23.17	N
MH0042	Denmark	male	49	24.46	N
MH0043	Denmark	male	69	23.72	N
MH0044	Denmark	male	64	24.48	N
MH0045	Denmark	male	59	25.11	N
MH0046	Denmark	male	54	23.74	N
MH0047	Denmark	female	69	30.40	N
MH0048	Denmark	female	54	19.40	N
MH0049	Denmark	female	44	35.52	N
MH0050	Denmark	male	49	25.08	N
MH0051	Denmark	female	69	23.15	N
MH0052	Denmark	female	49	33.18	N
MH0053	Denmark	female	49	32.70	N
MH0054	Denmark	male	49	20.31	N
MH0055	Denmark	male	59	30.29	N
MH0056	Denmark	male	54	25.35	N
MH0057	Denmark	female	54	32.98	N

MH0058	Denmark	female	54	22.04	N
MH0059	Denmark	male	59	33.27	N
MH0060	Denmark	male	54	23.52	N
MH0061	Denmark	female	69	30.12	N
MH0062	Denmark	female	49	37.54	N
MH0063	Denmark	male	59	30.23	N
MH0064	Denmark	female	54	23.18	N
MH0065	Denmark	male	59	28.23	N
MH0066	Denmark	female	44	20.79	N
MH0067	Denmark	male	54	21.07	N
MH0068	Denmark	female	54	28.97	N
MH0069	Denmark	female	59	36.71	N
MH0070	Denmark	male	49	22.69	N
MH0071	Denmark	female	44	25.37	N
MH0072	Denmark	female	64	40.21	N
MH0073	Denmark	male	54	32.49	N
MH0074	Denmark	female	49	20.46	N
MH0075	Denmark	male	64	30.55	N
MH0076	Denmark	female	69	34.78	N
MH0077	Denmark	female	49	24.92	N
MH0078	Denmark	female	49	36.90	N
MH0079	Denmark	female	64	19.97	N
MH0080	Denmark	female	59	18.59	N
MH0081	Denmark	female	49	37.95	N
MH0082	Denmark	female	59	22.56	N
MH0083	Denmark	female	54	30.59	N
MH0084	Denmark	male	64	31.67	N
MH0085	Denmark	female	59	36.46	N
MH0086	Denmark	female	59	21.59	N
O2.UC-1	Spain	male	37	31.02	Y
O2.UC-11	Spain	female	34	18.68	Y
O2.UC-12	Spain	male	43	21.60	Y
O2.UC-13	Spain	female	68	23.38	Y
O2.UC-14	Spain	male	31	32.65	Y
O2.UC-16	Spain	male	47	26.42	Y
O2.UC-17	Spain	male	56	21.87	Y
O2.UC-18	Spain	male	48	25.72	Y
O2.UC-19	Spain	male	42	24.15	Y
O2.UC-20	Spain	female	51	24.03	Y
O2.UC-21	Spain	female	49	30.46	Y
O2.UC-22	Spain	male	44	25.39	Y
O2.UC-23	Spain	female	44	28.16	Y
O2.UC-24	Spain	female	55	28.76	Y

O2.UC-4	Spain	female	57	28.53	Y
V1.CD-1	Spain	female	25	17.93	Y
V1.CD-11	Spain	female	62	35.46	N
V1.CD-12	Spain	female	41	20.20	Y
V1.CD-13	Spain	male	68	25.69	N
V1.CD-14	Spain	female	41	23.12	N
V1.CD-15	Spain	female	34	19.00	Y
V1.CD-2	Spain	male	49	27.76	N
V1.CD-3	Spain	female	18	21.51	N
V1.CD-4	Spain	female	46	29.69	N
V1.CD-6	Spain	female	36	18.52	Y
V1.CD-8	Spain	male	51	29.38	N
V1.CD-9	Spain	female	48	27.55	N
V1.UC-10	Spain	male	45	27.31	Y
V1.UC-13	Spain	female	51	28.51	Y
V1.UC-14	Spain	female	53	20.25	Y
V1.UC-15	Spain	female	25	22.77	Y
V1.UC-17	Spain	female	41	24.46	Y
V1.UC-18	Spain	female	63	28.67	N
V1.UC-19	Spain	female	37	21.19	N
V1.UC-21	Spain	male	62	25.21	Y
V1.UC-6	Spain	female	38	23.18	N
V1.UC-7	Spain	female	19	23.05	N
V1.UC-8	Spain	male	22	25.40	N
V1.UC-9	Spain	male	32	30.37	N

The following figures are available as supplementary information on Nature online[132]

Supplementary Table B-2. Summary of Sanger reads.

Supplementary Table 2 online.

Supplementary Table B-3. Summary of Illumina GA reads.

Supplementary Table 3 online.

Supplementary Table B-4. Summary of de novo assembly results.

Supplementary Table 4 online.

Supplementary Table B-5. List of 194 public human gut bacterial genomes.
Supplementary Table 5 online.

Supplementary Table B-6. ORF prediction in each sample.
Supplementary Table 6 online.

Supplementary Table B-7. Common species in human gut.
Supplementary Table 8 online.

Supplementary Table B-8. Range clusters.
Supplementary Table 10 online.

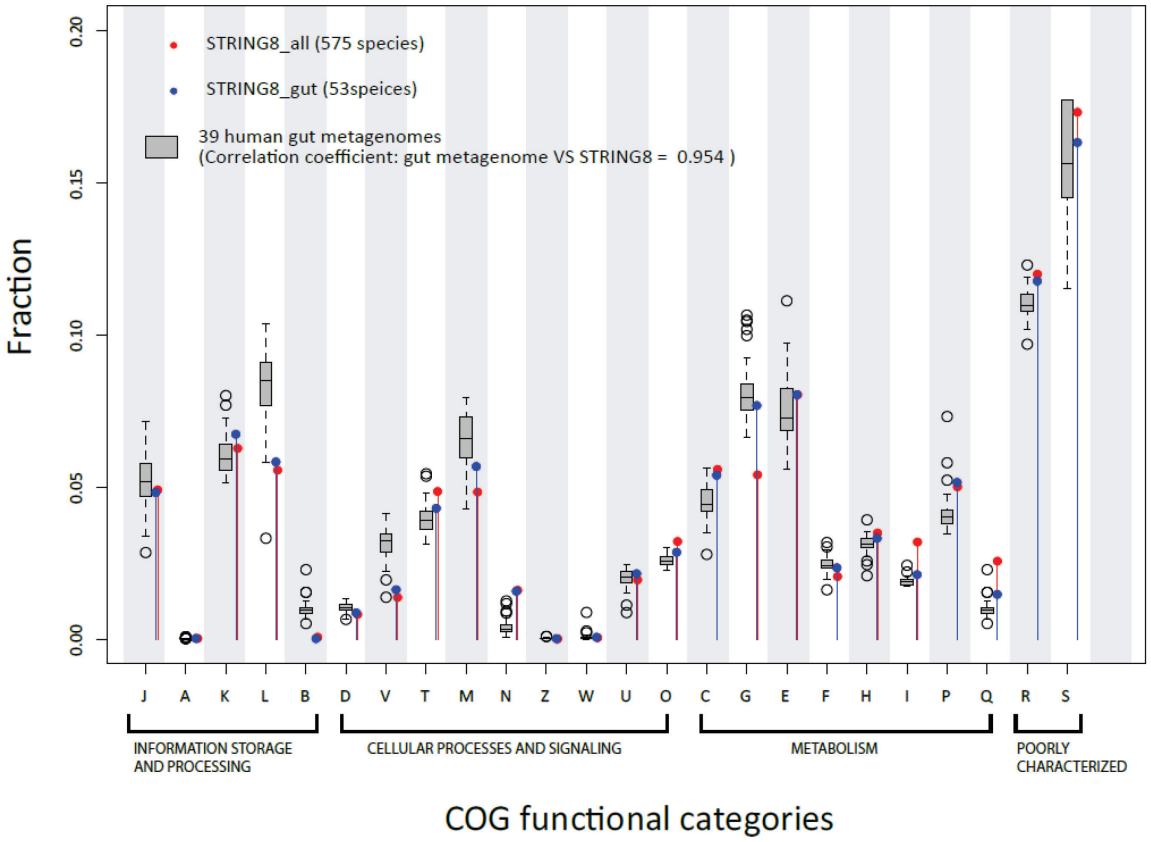
Supplementary Table B-9. Functions present in the human metagenome and genome.
Supplementary Table 11 online.

Supplementary Table B-10. 89 frequent microbial species/strains in human gut.
Supplementary Table 12 online.

Appendix C

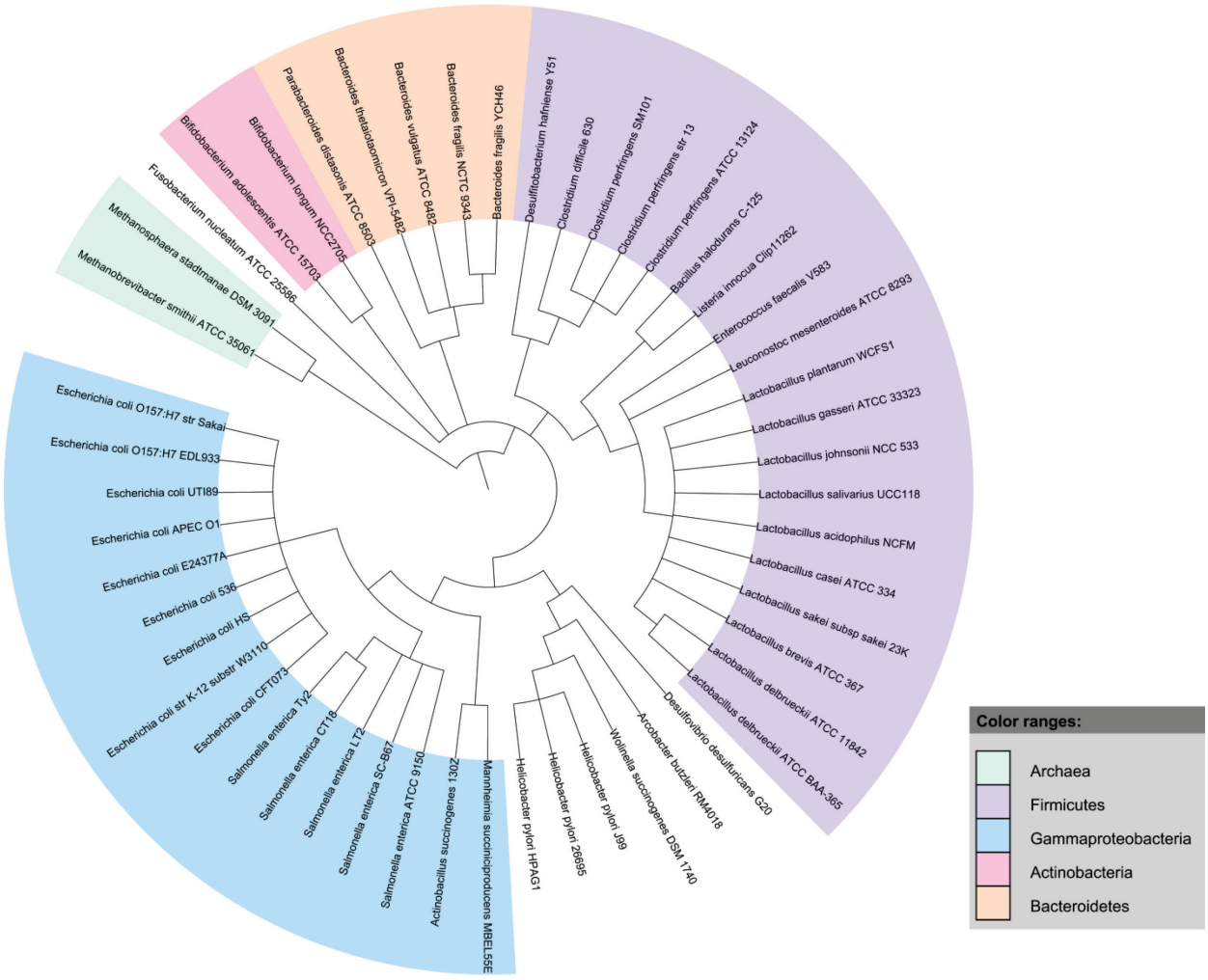
Supplementary information for Chapter 6

C.1 Supplementary Figures

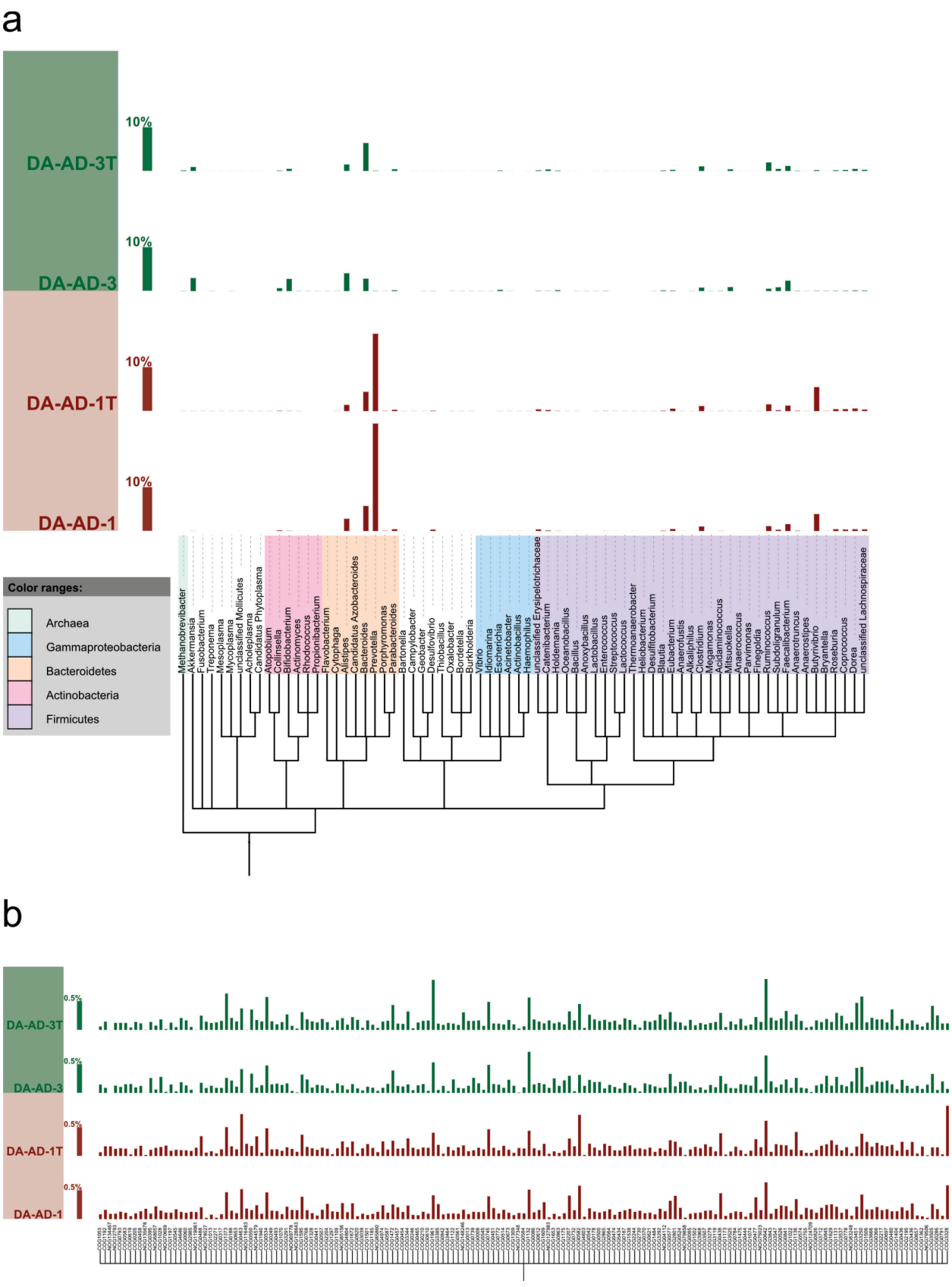


Supplementary Figure C-1. Abundance distribution of COG functional categories.

Abundance distribution of COG functional categories in 575 bacterial species in STRING v8.0 (red), 53 gut-associated species stored in the STRING v8.0 (blue) and 39 metagenomic samples (gray boxes). Distributions of metagenomics samples are similar to gut-associated bacteria. Functional category L (replication, recombination and repair), V (defense mechanisms), M (cell wall/membrane/envelope biogenesis) and G (carbohydrate transport and metabolism) are enriched in the gut metagenomes. In particular, enrichment of L and V is not supported by gut specific bacteria in STRING (blue). This implies metagenomic data includes several bacteria which are not in STRING, and have higher ratio of these functional categories.

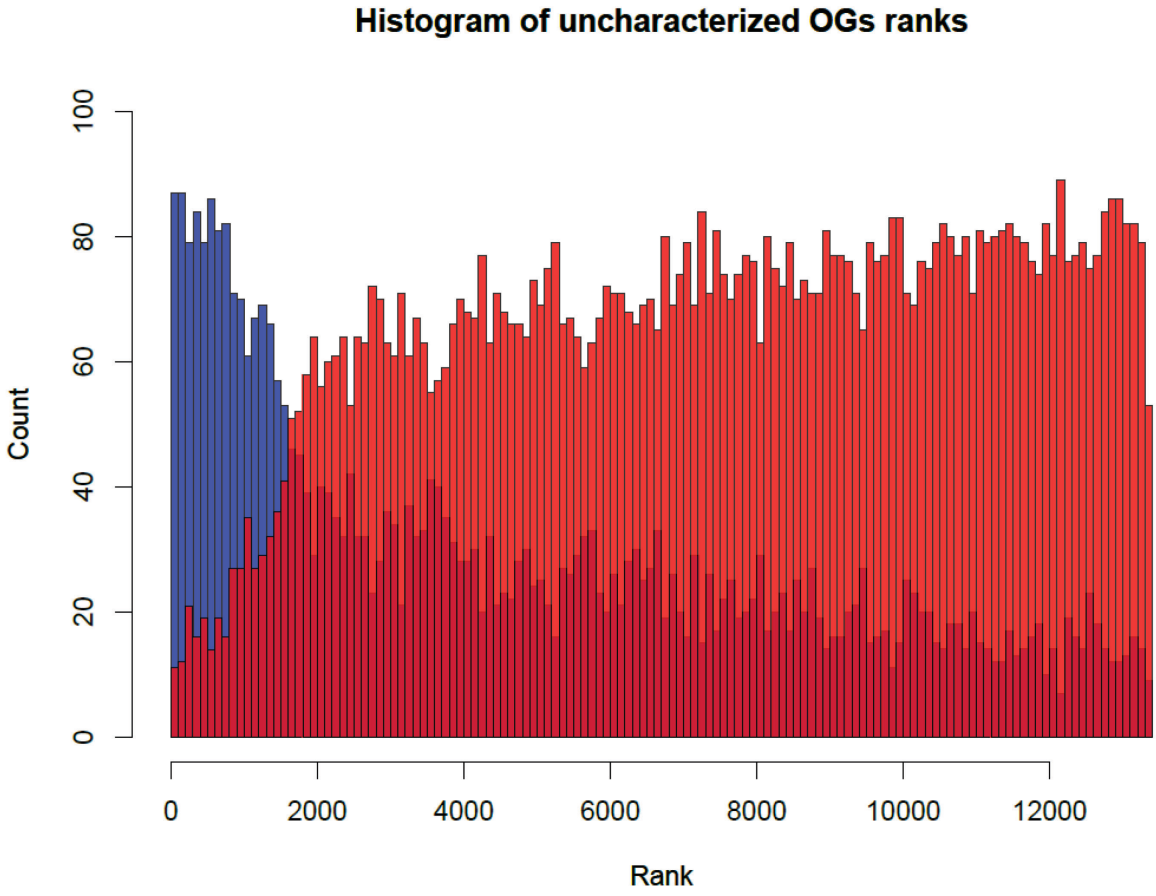


Supplementary Figure C-2. Phylogenetic tree of the 53 gut-specific genomes out of the 575 prokaryotic genomes in STRING.


a
DA-AD-3T
DA-AD-3
DA-AD-1T
DA-AD-1
10%
Color ranges:
Archaea
Gammaproteobacteria
Bacteroidetes
Actinobacteria
Firmicutes
Methanobrevibacter
Akkermansia
Fusobacterium
Treponema
Mesoplasma
Mycoplasma
unclassified Mollicutes
Acholeplasma
Candidatus Phytoplasma
Adlercreutzia
Collinsella
Bifidobacterium
Actinomyces
Rhodococcus
Propionibacterium
Flavobacterium
Cytophaga
Alistipes
Candidatus Azobacteroides
Bacteroides
Prevotella
Porphyromonas
Parabacteroides
Bartonella
Campylobacter
Geobacter
Desulfovibrio
Thiobacillus
Oxalobacter
Bordetella
Burkholderia
Vibrio
Idiomarina
Escherichia
Acinetobacter
Actinobacillus
Haemophilus
unclassified Erysipelotrichaceae
Catenibacterium
Holdemania
Dorea
Bacillus
Anoxybacillus
Lactobacillus
Enterococcus
Streptococcus
Lactococcus
Thermoanaerobacter
Heliobacterium
Desulfitobacterium
Eubacterium
Anaerofustis
Alkaliphilus
Clostridium
Megamonas
Acidaminococcus
Mitsuokella
Anaerococcus
Parvimonas
Finegoldia
Ruminococcus
Subdoligranulum
Faecalibacterium
Anaerotruncus
Anaerostipes
Butyrivibrio
Bryantella
Roseburia
Coprococcus
Dorea
unclassified Lachnospiraceae
b
DA-AD-3T
DA-AD-3
DA-AD-1T
DA-AD-1
0.5%

Supplementary Figure C-3. Genus and eggNOG orthologous group (OG) abundance distributions of Sanger and 454 Titanium based sequences from the same samples are similar.

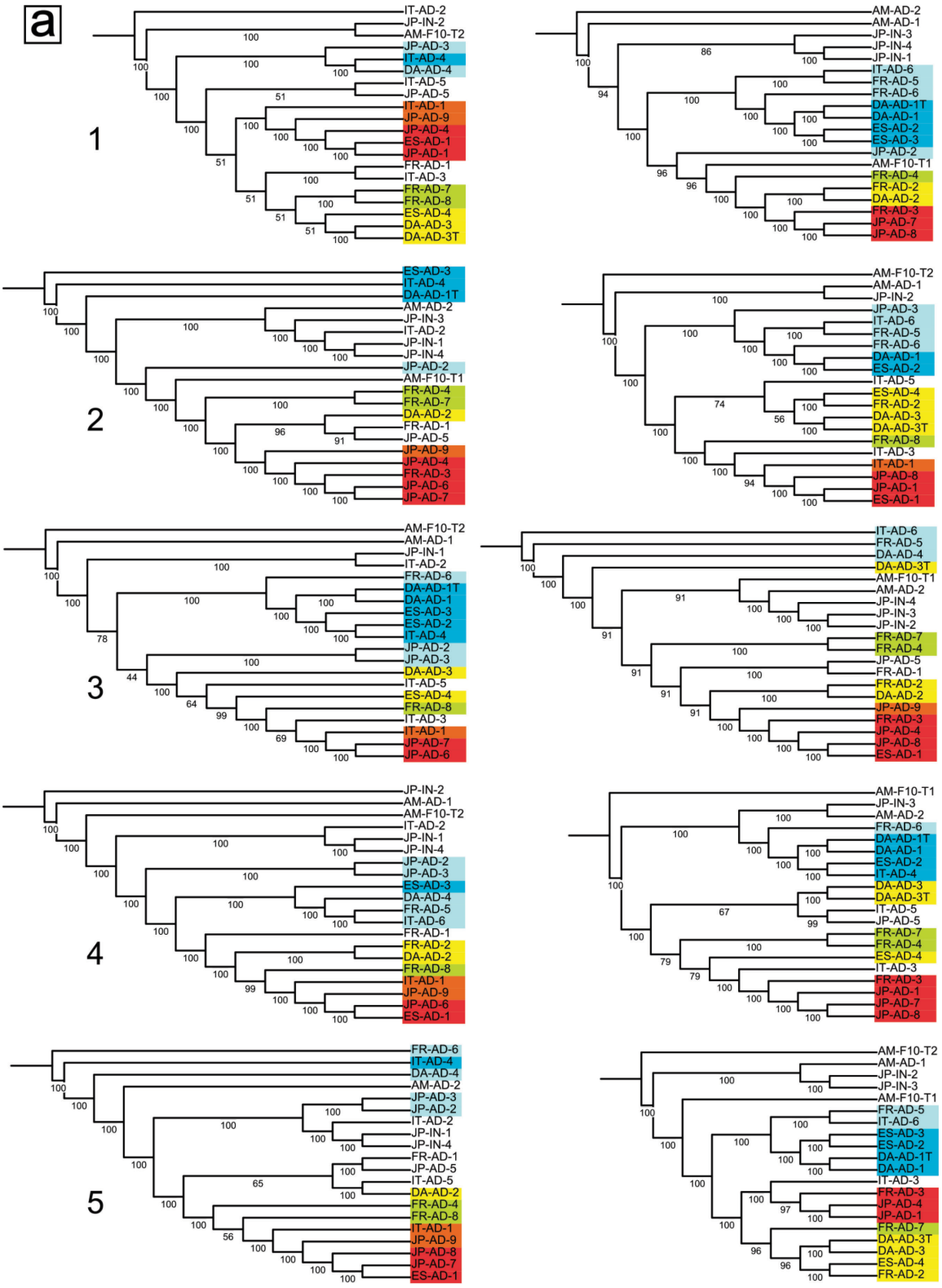
DA-AD-1 and DA-AD-1T are Sanger and 454 Titanium based sequences from MH6. DA-AD-3 and DA-AD-3T are from MH12. a) Genus abundance. Pearson correlation coefficients between DA-AD-1/DA-AD-1T is 0.9883, between DA-AD-3/DA-AD-3T is 0.9974. (Background for DA-AD-1/DA-AD-3 is 0.9075; DA-AD-1T/DA-AD-3T is 0.9594). b) Abundance of 50 most abundant orthologous groups. Pearson correlation coefficients between DA-AD-1/DA-AD-1T is 0.9482, between DA-AD-3/DA-AD-3T is 0.9153. (Background for DA-AD-1/DA-AD-3 is 0.8408; DA-AD-1T/DA-AD-3T is 0.8436).

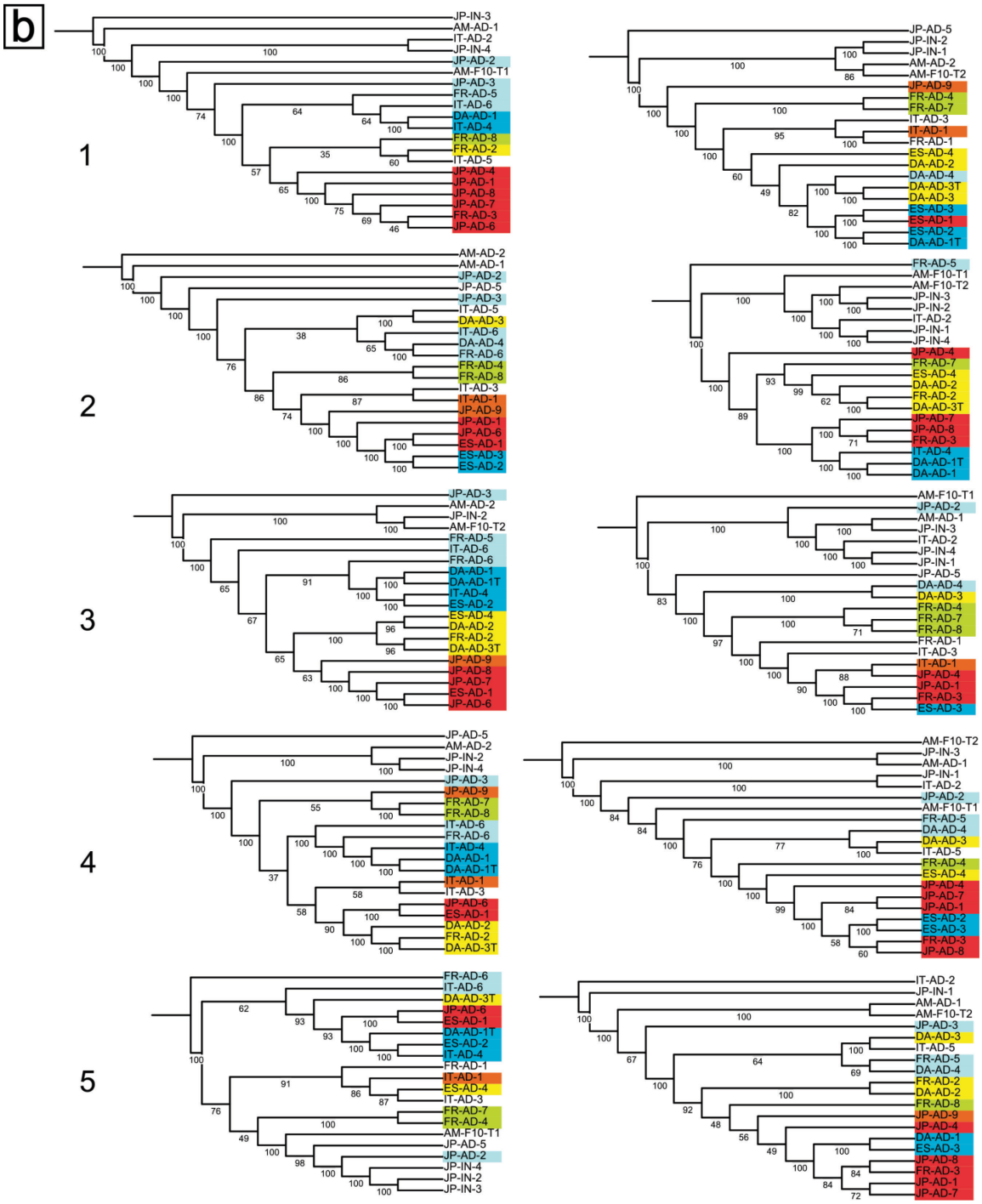


Supplementary Figure C-4. Lower ranked functions are enriched in uncharacterized orthologous groups (OGs).

A histogram of the number of functionally characterized (blue) and uncharacterized (red) OGs ranked by their average abundance in 35 metagenomes shows that uncharacterized proteins usually form small OGs (hence are predominantly ranked lower in abundance). In contrast, functionally

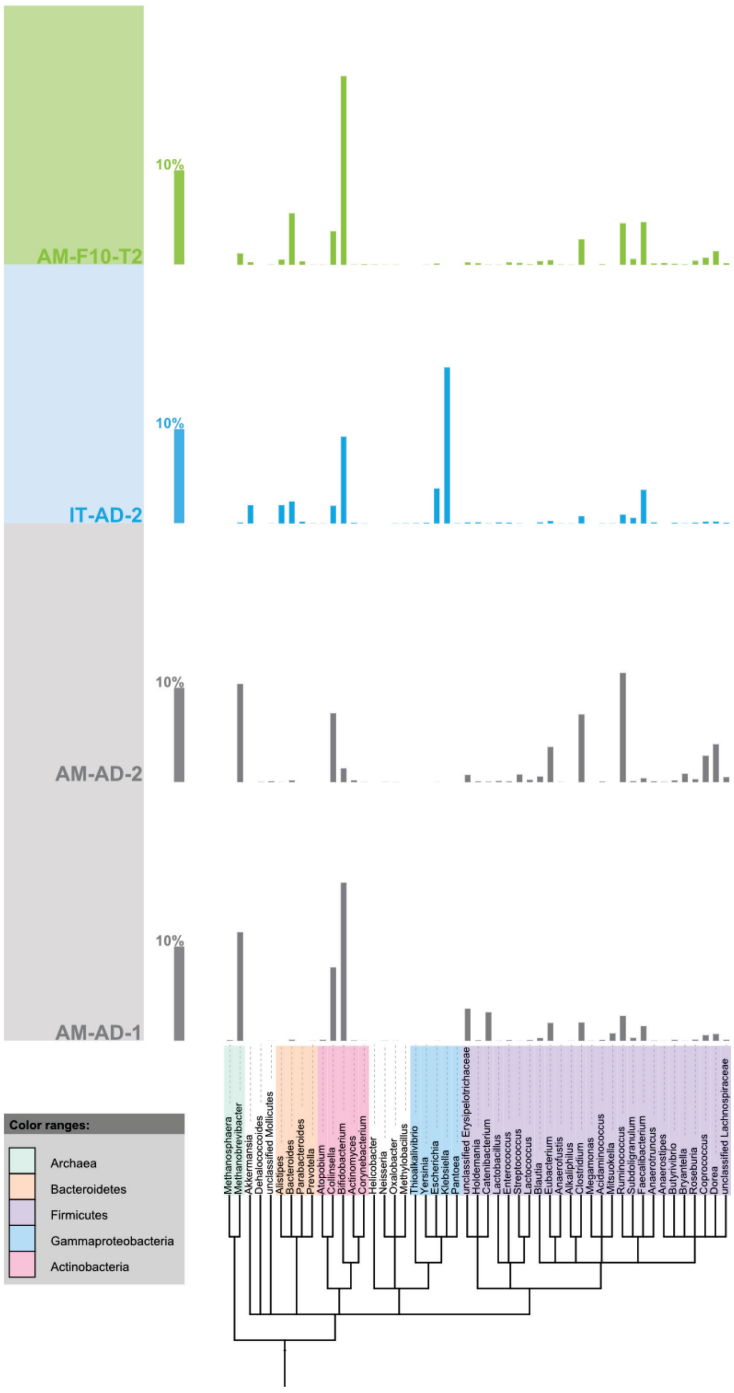
characterized OGs are large with many genes and are usually ranked higher in abundance. This agrees with the findings of [31].





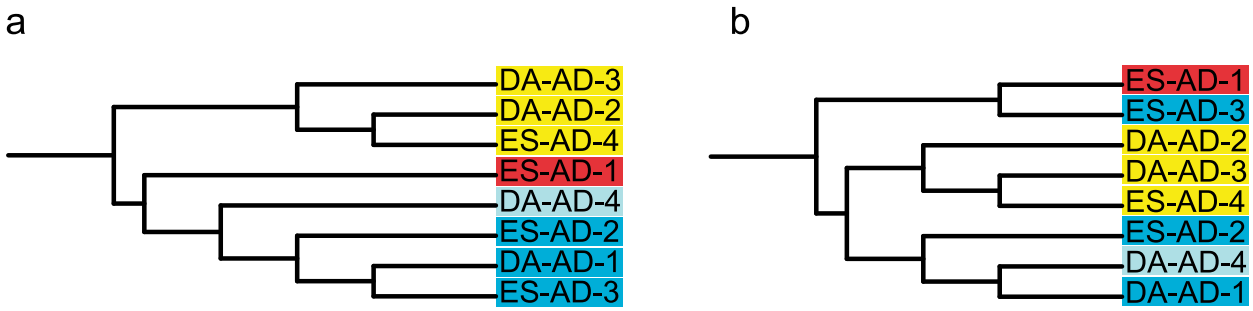
Supplementary Figure C-5. 50% jack-knife tests to test the robustness of enterotypes.

Samples were split into two halves and each half was bootstrapped and clustered separately. This was repeated five times. a) jack-knife test on reference genome mapping based clusters, b) jack-knife test on functional mapping based clusters.



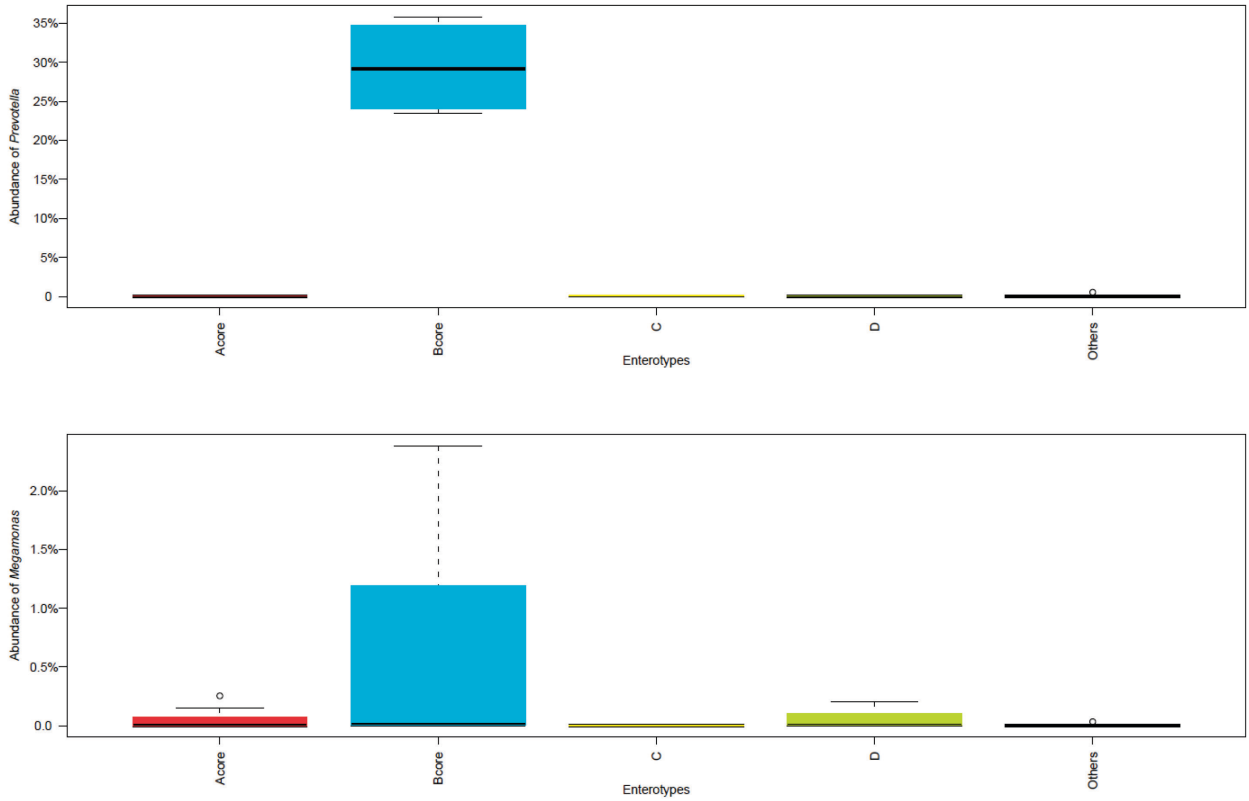
Supplementary Figure C-6. Genus abundance profiles of the outlier samples.

IT-AD-2 has an unexpectedly high abundance of *Klebsiella*; AM-AD-1 and AM-AD-2 have over 10% of *Methanobrevibacter*, an archaeal genus, and low abundance of *Bacteroides*. IT-AD-2, AM-AD-1, AM-AD-2 and AM-F10-T2 cluster with the four Japanese infant samples, which are known outliers, and hence are considered outliers themselves.

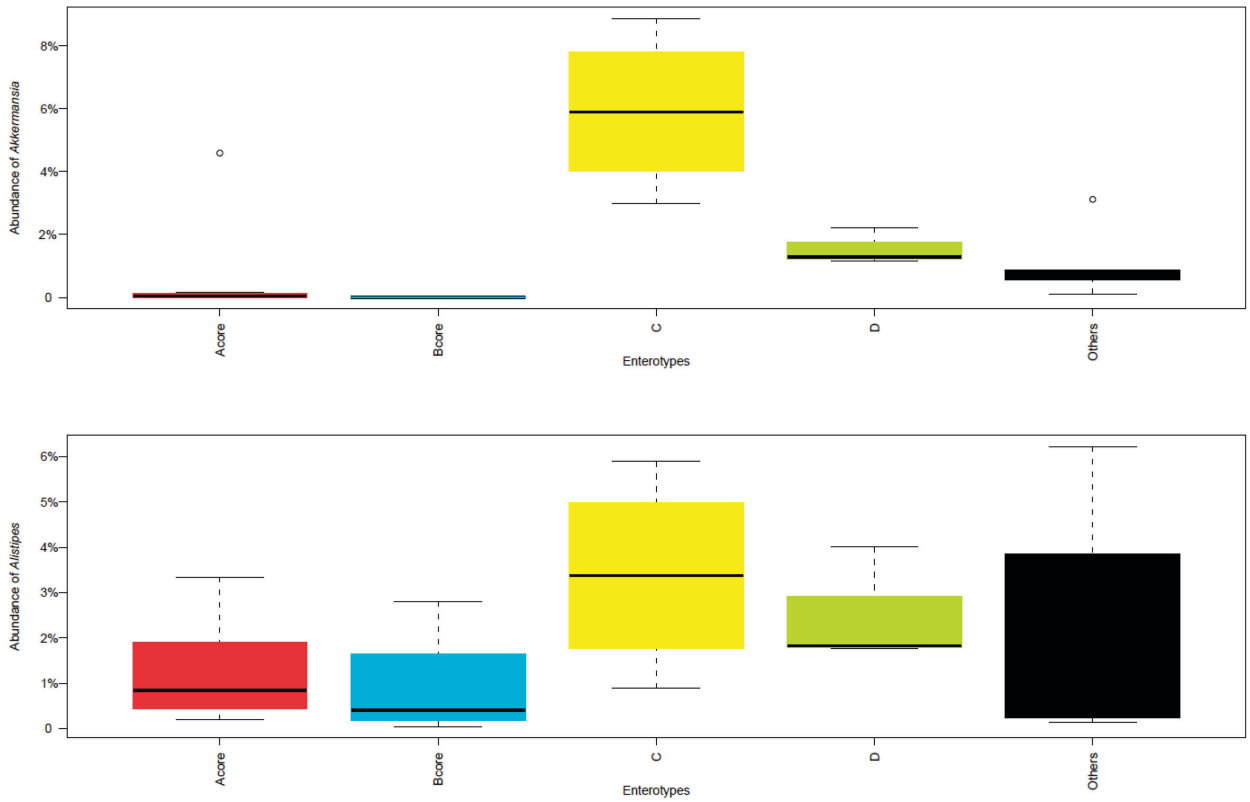


Supplementary Figure C-7. Clustering of 8 MetaHIT samples by other methods.

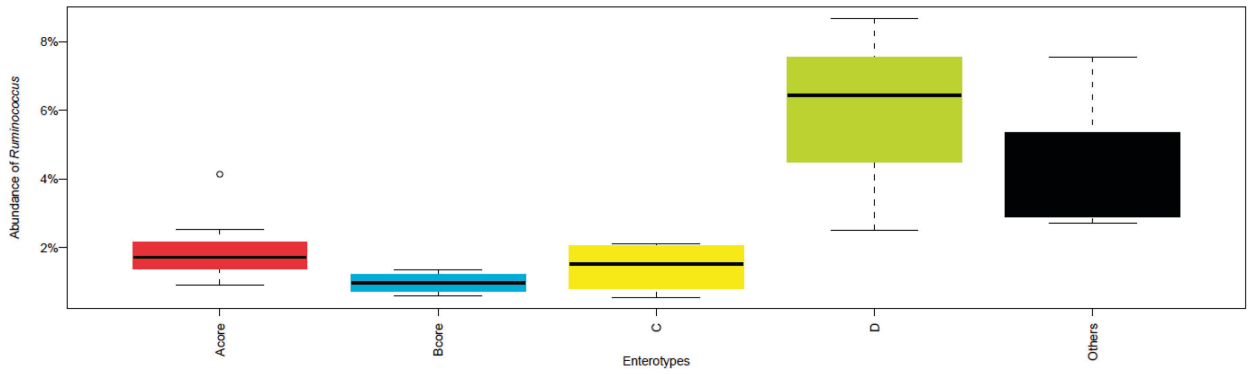
Clustering of Danish and Spanish samples by comparing phylogenetic abundance profiles at the genus level from (a) spotted array and (b) HIT-chip agree with the enterotypes in Figure 6-2.



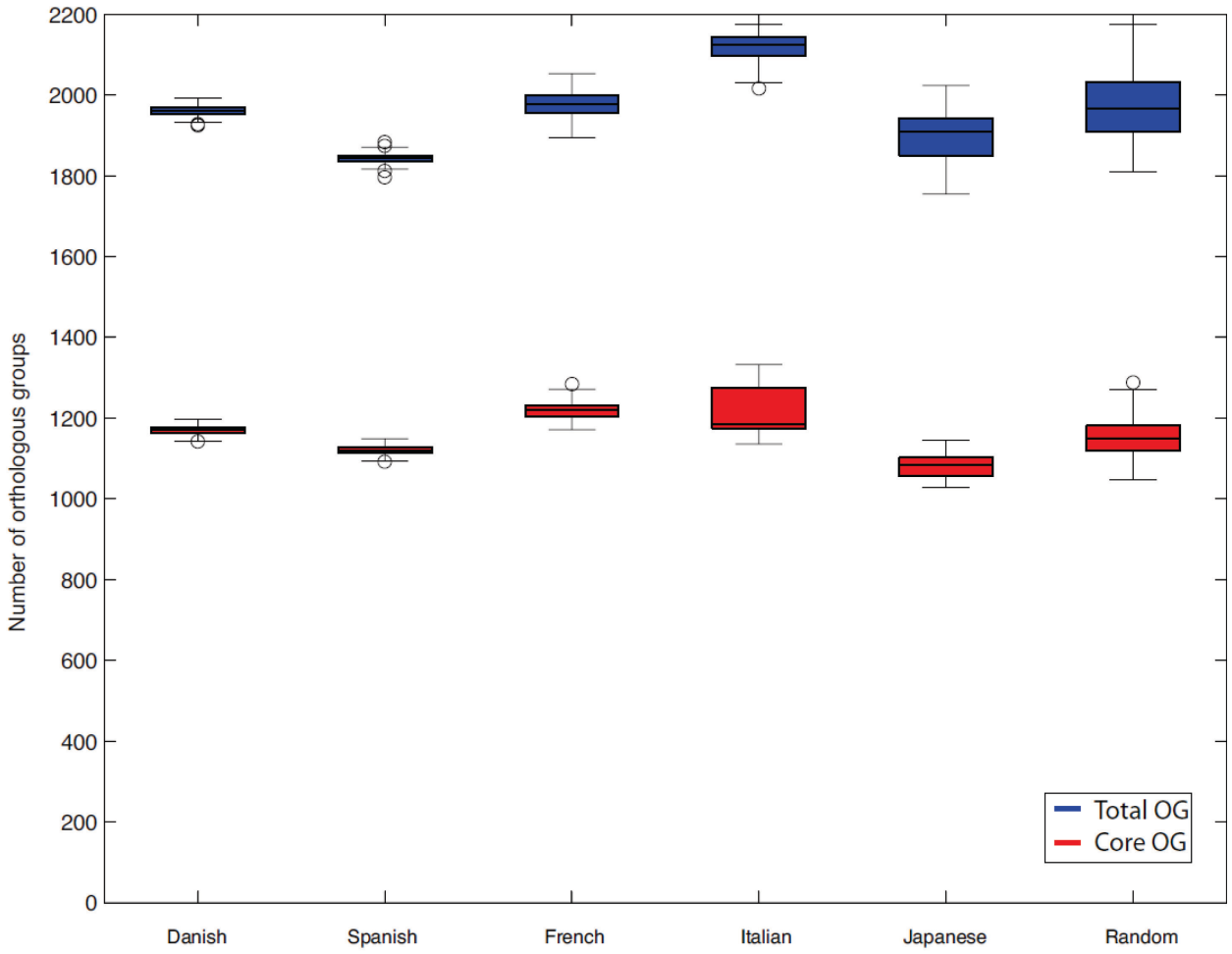
Supplementary Figure C-8. Enrichment of *Prevotella* and *Megamonas* in enterotype Bcore.



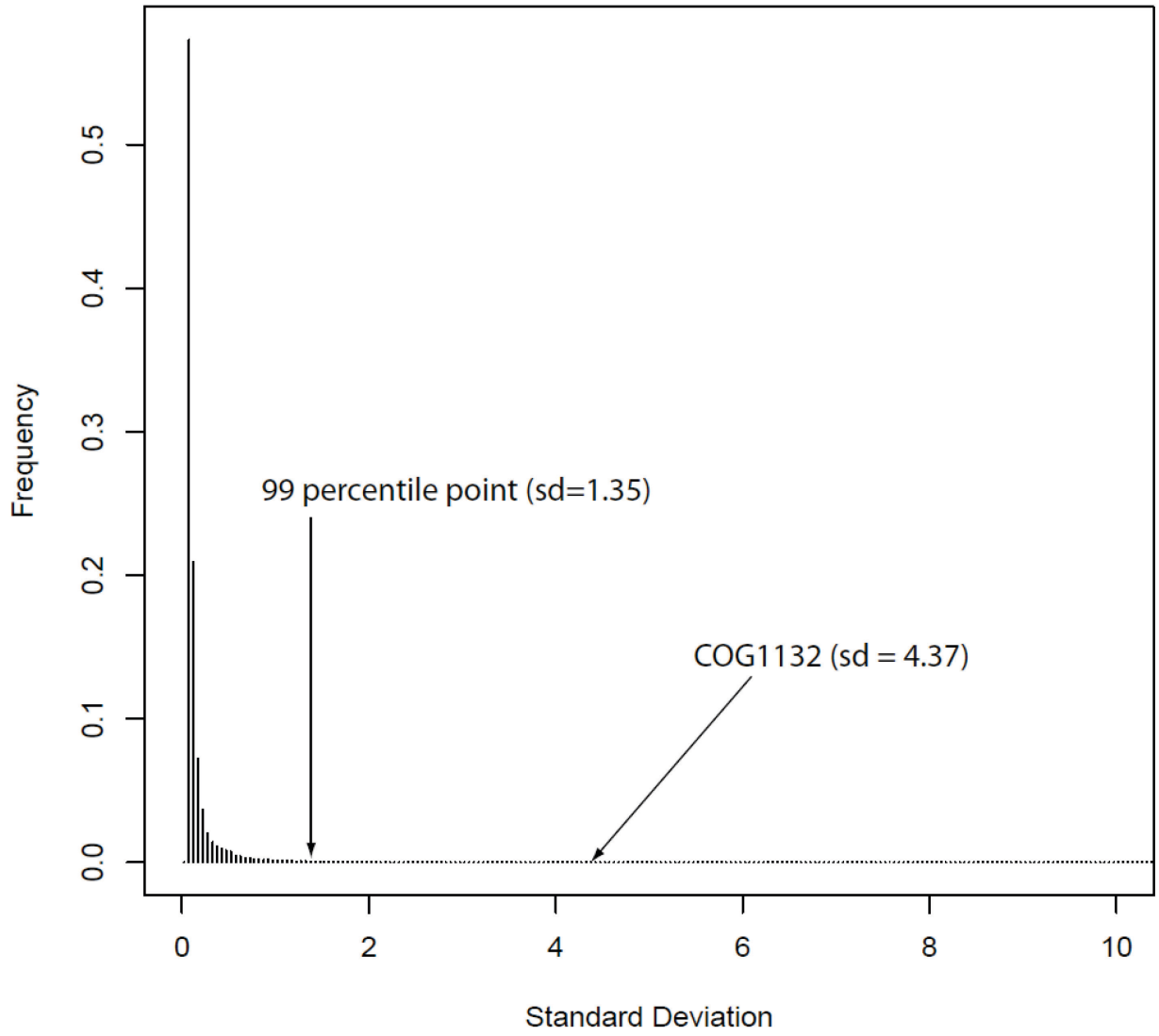
Supplementary Figure C-9. Enrichment of *Akkermansia* and *Alistipes* in enterotype C.



Supplementary Figure C-10. Enrichment of *Ruminococcus* in enterotype D.



Supplementary Figure C-11. Core metagenome sizes in metagenomes from different nations are similar.



Supplementary Figure C-12. Distribution of standard deviation of the number of genes in orthologous groups in the STRING database.

COG1132 is at 99.8 percentile, meaning only 0.2% of the orthologous groups in STRING have a higher variation.

C.2 Supplementary Tables

Supplementary Table C-1. Comparing Sanger and pyrosequencing technologies.

Amount of sequence generated and the number of genera (at 85% sequence identity) and eggNOG orthologous groups (OGs) retrieved from two Danish samples sequenced using Sanger and 454 Titanium technologies.

sample	DA-AD-1		DA-AD-3	
	Sanger	Titanium	Sanger	Titanium
amount of sequence	157Mb	295Mb	155Mb	195Mb
genera	56	121	58	117
eggNOG OGs	5863	5675	6015	5935

Supplementary Table C-2. Eukaryotic and viral sequences.

Potential human DNA fragments and fragments with best hits to eukaryotes. Values represent percentage of the metagenome sequence fragments from each sample. For the three published studies, these numbers do not include human/eukaryotic sequences removed by the authors of the respective publications before making the data publicly available. For the prophage fraction, we present an upper bound by counting all fragments with a BLASTN hit (>60 bits) to a viral and phage genome database, and a lower bound by considering only the fragments whose viral hit is significantly better than their best hit to a microbial sequence in the reference genome set of 1152 microbial genomes.

	Sample ID	% human	% other eukaryotes	% prophage	
				lowerbound	upperbound
European samples from this study	DA-AD-1	0.113	0.51	1.41	4.64
	DA-AD-2	1.354	1.22	6.87	14.66
	DA-AD-3	0.003	0.58	0.68	4.36
	DA-AD-4	0.401	1.01	3.7	8.81
	ES-AD-1	0.037	0.29	0.51	3.21
	ES-AD-2	0.166	0.39	0.85	4.15
	ES-AD-3	0.024	0.43	0.56	3.78
	ES-AD-4	0.831	0.5	0.8	4.26
	FR-AD-1	0.002	0.46	0.92	5.23
	FR-AD-2	0.068	0.99	2.83	6.74
	FR-AD-3	0.005	0.39	1.12	4.43
	FR-AD-4	0.024	0.59	1.71	5.96
	FR-AD-5	0.015	0.56	3.11	8.71
	FR-AD-6	0.008	0.43	1.33	5.05
	FR-AD-7	0.007	0.51	1.52	5.51
	FR-AD-8	0.011	0.49	1.39	5.2
	IT-AD-1	0.007	0.32	0.61	5.05
	IT-AD-2	0	0.35	0.68	6.4
	IT-AD-3	0.015	0.42	1.52	5.98
	IT-AD-4	0	0.28	0.71	4.22
	IT-AD-5	0.002	0.64	1.22	5.8

	IT-AD-6	0.011	0.41	0.83	5.12
	Average	**0.141**	**0.535**	**1.56**	**5.79**
Published samples	AM-AD-1	0.026	0.51	3.21	21.73
	AM-AD-2	0.013	0.8	2.35	16.43
	AM-F10-T1	0.004	0.76	0.25	1.36
	AM-F10-T2	0.001	0.28	0.09	1.28
	JP-AD-1	0.004	0.36	0.84	5.49
	JP-AD-2	0.014	0.54	1.06	6.9
	JP-AD-3	0.001	0.3	0.55	7.1
	JP-AD-4	0.005	0.44	0.41	4.93
	JP-AD-5	0.028	0.83	3.43	12.64
	JP-AD-6	0.006	0.38	0.61	4.78
	JP-AD-7	0.068	0.34	0.97	4.44
	JP-AD-8	0.024	0.35	1.19	5.77
	JP-AD-9	0.045	0.39	1.47	7.1
	JP-IN-1	0.005	0.13	1.41	11.27
	JP-IN-2	0	0.41	0.58	9.42
	JP-IN-3	0	0.43	1.09	11.02
	JP-IN-4	0.096	0.49	0.89	9.22
	Average	**0.02**	**0.455**	**1.2**	**8.3**
Overall average		**0.093**	**0.5**	**1.42**	**6.88**

Supplementary Table C-3. STRING eukaryotic kingdoms in gut metagenomes.

Average fraction of DNA fragments from the metagenome samples with best hits to each eukaryotic kingdom represented in the STRING database.

Kingdom	% fragments
Metazoa	0.1982
Fungi	0.1091
Amoebozoa	0.0693
Alveolata	0.0635
Viridiplantae	0.0339
Euglenozoa	0.0245
Fornicata	0.0012

Supplementary Table C-4. Functional assignment rate.

Orthologous group assignment rates, measured by the number of genes in each sample that can be assigned to an orthologous group in eggNOG database.

Subject ID	Genes	OG mapped genes (%)
DA-AD-1	152959	92517 (60.48%)
DA-AD-2	147519	89997 (61.01%)
DA-AD-3	162534	101144 (62.23%)

DA-AD-4	167530	96916 (57.85%)
ES-AD-1	102806	69312 (67.42%)
ES-AD-2	122628	75200 (61.32%)
ES-AD-3	140465	91523 (65.16%)
ES-AD-4	166469	104175 (62.58%)
FR-AD-1	118183	75343 (63.75%)
FR-AD-2	103732	62860 (60.6%)
FR-AD-3	100309	65051 (64.85%)
FR-AD-4	122497	75029 (61.25%)
FR-AD-5	119784	72765 (60.75%)
FR-AD-6	109207	66658 (61.04%)
FR-AD-7	101769	63132 (62.03%)
FR-AD-8	106497	66006 (61.98%)
IT-AD-1	84781	57173 (67.44%)
IT-AD-2	90859	59994 (66.03%)
IT-AD-3	107924	69317 (64.23%)
IT-AD-4	58967	38077 (64.57%)
IT-AD-5	111891	70115 (62.66%)
IT-AD-6	108567	68786 (63.36%)
JP-AD-1	54856	35601 (64.9%)
JP-AD-2	63230	39654 (62.71%)
JP-AD-3	64201	42916 (66.85%)
JP-AD-4	55693	36941 (66.33%)
JP-AD-5	54699	34814 (63.65%)
JP-AD-6	63735	43128 (67.67%)
JP-AD-7	37212	24829 (66.72%)
JP-AD-8	64333	40709 (63.28%)
JP-AD-9	59820	38200 (63.86%)
JP-IN-1	33993	27306 (80.33%)
JP-IN-2	14334	10400 (72.55%)
JP-IN-3	29305	19181 (65.45%)
JP-IN-4	34732	23957 (68.98%)
AM-AD-1	72772	45958 (63.15%)
AM-AD-2	69574	44750 (64.32%)
AM-F10-T1	152956	61800 (40.4%)
AM-F10-T2	188665	78441 (41.58%)

Supplementary Table C-5. Highly abundant functions from low abundance genera.

List of 109 functions which are abundant (among the top 20%, which is equivalent to above 80th percentile) and are primarily from low-abundance genera (<2.5% relative abundance). Samples where this was observed are also listed.

Orthologous group	Samples	Count
NOG139537	FR-AD-4, IT-AD-3, IT-AD-2, FR-AD-3, ES-AD-3, AM-AD-1, JP-AD-3, DA-AD-4, FR-AD-1, IT-AD-1, FR-AD-5, DA-AD-3, ES-AD-4, IT-AD-6, JP-AD-2, JP-AD-4, DA-AD-2, IT-AD-5, FR-AD-6, JP-AD-1	20
NOG79858	IT-AD-6, FR-AD-4, ES-AD-4, IT-AD-3, JP-AD-2, JP-AD-4, JP-AD-7, ES-AD-3, AM-AD-1, DA-AD-4, FR-AD-1, IT-AD-5, FR-AD-6, JP-AD-9, JP-AD-1, FR-AD-2, FR-AD-8	17
NOG137454	FR-AD-4, IT-AD-3, IT-AD-2, FR-AD-1, FR-AD-5, DA-AD-3, ES-AD-4, IT-AD-6, JP-AD-4, DA-AD-2, IT-AD-5, FR-AD-6, FR-AD-7, JP-AD-9, ES-AD-2, JP-AD-1	16
NOG79506	IT-AD-6, ES-AD-4, IT-AD-3, FR-AD-3, JP-AD-2, JP-AD-4, ES-AD-3, DA-AD-2, DA-AD-4, FR-AD-1, FR-AD-6, IT-AD-1, JP-AD-9, DA-AD-3, FR-AD-8	15
NOG131524	FR-AD-4, ES-AD-4, FR-AD-3, JP-AD-2, DA-AD-1, JP-AD-4, ES-AD-3, DA-AD-2, AM-AD-1, JP-AD-8, DA-AD-4, FR-AD-6, DA-AD-3, FR-AD-8	14
NOG68428	IT-AD-6, ES-AD-4, JP-AD-2, DA-AD-2, JP-AD-8, DA-AD-4, FR-AD-1, JP-AD-6, FR-AD-6, IT-AD-1, DA-AD-3, FR-AD-2, FR-AD-8	13
NOG81629	ES-AD-4, IT-AD-2, JP-AD-2, DA-AD-1, JP-AD-4, JP-AD-7, AM-AD-1, DA-AD-4, FR-AD-1, IT-AD-5, FR-AD-6, DA-AD-3, FR-AD-8	13
NOG83248	IT-AD-6, FR-AD-4, JP-AD-2, DA-AD-1, AM-AD-1, JP-AD-8, DA-AD-4, JP-AD-6, FR-AD-6, JP-AD-9, DA-AD-3, FR-AD-8, IT-AD-4	13
NOG116632	ES-AD-4, FR-AD-3, JP-AD-2, DA-AD-2, AM-AD-1, JP-AD-3, DA-AD-4, FR-AD-1, FR-AD-6, DA-AD-3, JP-AD-1, FR-AD-8	12
NOG86034	IT-AD-6, ES-AD-4, JP-AD-2, JP-AD-4, JP-AD-7, JP-AD-3, DA-AD-4, FR-AD-6, JP-AD-9, DA-AD-3, FR-AD-8, JP-AD-1	12
COG3598	IT-AD-6, ES-AD-4, FR-AD-3, JP-AD-2, DA-AD-4, IT-AD-5, FR-AD-6, FR-AD-5, DA-AD-3, FR-AD-8	10
COG3843	DA-AD-1, JP-AD-4, AM-AD-1, DA-AD-4, FR-AD-6, JP-AD-6, DA-AD-3, FR-AD-8	8
COG5545	ES-AD-4, JP-AD-5, IT-AD-2, JP-AD-2, JP-AD-4, AM-AD-1, DA-AD-4, FR-AD-6	8
NOG127983	FR-AD-6, FR-AD-3, JP-AD-2, ES-AD-3, DA-AD-3, DA-AD-2, DA-AD-4, FR-AD-8	8
NOG27013	ES-AD-4, JP-AD-2, JP-AD-4, JP-AD-8, DA-AD-4, FR-AD-6, IT-AD-1, FR-AD-2	8
COG4725	IT-AD-6, IT-AD-3, FR-AD-6, JP-AD-9, DA-AD-2, FR-AD-8	6
NOG07949	IT-AD-6, IT-AD-2, FR-AD-6, FR-AD-3, IT-AD-1, JP-AD-9	6

NOG120133	ES-AD-4, IT-AD-3, FR-AD-6, FR-AD-5, DA-AD-2, DA-AD-4	6
COG1289	IT-AD-5, FR-AD-6, JP-AD-2, JP-AD-8, DA-AD-4	5
COG3077	ES-AD-4, IT-AD-3, IT-AD-1, DA-AD-3, FR-AD-1	5
NOG44176	FR-AD-4, FR-AD-6, JP-AD-9, FR-AD-8, FR-AD-2	5
COG4422	JP-AD-2, JP-AD-4, FR-AD-5, DA-AD-4	4
NOG12663	FR-AD-4, JP-AD-2, IT-AD-1, FR-AD-7	4
NOG69420	IT-AD-1, DA-AD-3, AM-AD-2, FR-AD-2	4
COG1533	JP-AD-5, FR-AD-6, FR-AD-8	3
COG2088	JP-AD-2, JP-AD-7, DA-AD-3	3
COG2337	FR-AD-6, FR-AD-8	2
COG3041	IT-AD-3, IT-AD-1	2
COG3328	ES-AD-2, DA-AD-4	2
COG3340	JP-AD-5, JP-AD-9	2
COG3539	IT-AD-5, JP-AD-2	2
NOG120367	ES-AD-4, FR-AD-8	2
NOG139663	FR-AD-4, FR-AD-7	2
NOG14713	IT-AD-1, JP-AD-7	2
NOG16015	ES-AD-4, JP-AD-2	2
NOG20054	ES-AD-4, FR-AD-7	2
NOG25595	FR-AD-6, JP-AD-7	2
NOG40986	FR-AD-4, FR-AD-7	2
NOG45681	IT-AD-2, DA-AD-4	2
NOG69323	FR-AD-6, JP-AD-7	2
COG0137	JP-AD-5	1
COG0143	ES-AD-3	1
COG0372	JP-AD-7	1
COG0666	JP-AD-8	1
COG0675	FR-AD-5	1
COG0790	JP-AD-7	1
COG0791	FR-AD-6	1
COG0827	JP-AD-2	1
COG0863	JP-AD-3	1
COG1005	IT-AD-2	1
COG1009	IT-AD-5	1
COG1250	IT-AD-5	1
COG1277	FR-AD-6	1
COG1321	DA-AD-4	1
COG1528	IT-AD-5	1
COG1672	IT-AD-2	1
COG1846	FR-AD-8	1

COG2015	JP-AD-5	1
COG2252	IT-AD-5	1
COG2357	JP-AD-1	1
COG2452	FR-AD-6	1
COG2710	FR-AD-8	1
COG2723	JP-AD-7	1
COG2856	IT-AD-1	1
COG2932	FR-AD-6	1
COG2946	FR-AD-6	1
COG3022	JP-AD-9	1
COG3049	DA-AD-4	1
COG3188	IT-AD-5	1
COG3315	FR-AD-6	1
COG3378	DA-AD-4	1
COG3464	JP-AD-3	1
COG3711	FR-AD-8	1
COG3910	JP-AD-7	1
COG4200	FR-AD-6	1
NOG06430	JP-AD-9	1
NOG10311	IT-AD-5	1
NOG112926	DA-AD-4	1
NOG114060	JP-AD-9	1
NOG116483	DA-AD-4	1
NOG116612	JP-AD-7	1
NOG119061	DA-AD-4	1
NOG122382	FR-AD-8	1
NOG124981	FR-AD-4	1
NOG125084	FR-AD-6	1
NOG126306	JP-AD-5	1
NOG128643	DA-AD-4	1
NOG13016	JP-AD-8	1
NOG131675	DA-AD-4	1
NOG134467	DA-AD-4	1
NOG134563	JP-AD-7	1
NOG18439	DA-AD-4	1
NOG25785	DA-AD-4	1
NOG39150	DA-AD-4	1
NOG42453	IT-AD-2	1
NOG43858	DA-AD-4	1
NOG43914	IT-AD-2	1

NOG44566	JP-AD-7	1
NOG44869	DA-AD-4	1
NOG45139	DA-AD-4	1
NOG46999	JP-AD-2	1
NOG47313	DA-AD-4	1
NOG68338	IT-AD-1	1
NOG70669	DA-AD-4	1
NOG79696	JP-AD-5	1
NOG80481	DA-AD-4	1
NOG81060	DA-AD-4	1
NOG82576	DA-AD-4	1
NOG84056	IT-AD-1	1

Supplementary Table C-6. Abundant functions.

Abundant functions (among the top 20%, which is equivalent to above 80th percentile) from low-abundance genera (<2.5% relative abundance). Ranks of these functions are listed as percentiles. The abundance of each genus in the corresponding sample and its contribution to the listed function in that sample are also listed.

Function	**Sample**	**Rank of function in sample (percentile)**	**Genus / rank**	**abundance of genus in sample**	**contribution of genus to function**
COG2337	FR-AD-6	87.81	*Dorea*	1.74%	19.39%
			Clostridiales	0.41%	8.17%
			Butyrivibrio	1.52%	7.01%
			Eubacterium	1.37%	6.93%
			Roseburia	0.68%	6.77%
			Ruminococcus	1.34%	5.60%
			Faecalibacterium	2.29%	3.88%
	FR-AD-8	87.08	*Dorea*	2.18%	14.98%
			Clostridium	2.39%	12.64%
			Faecalibacterium	2.14%	9.74%
			Erysipelotrichaceae	0.52%	9.48%
			Enterobacteriales	0.16%	3.16%
COG3077	DA-AD-3	81.27	*Clostridium*	0.79%	21.64%
			Blautia	0.14%	20.25%
			Ruminococcus	0.55%	10.73%

			Faecalibacterium	2.38%	8.85%
			Collinsella	0.66%	2.95%
	ES-AD-4	80.22	*Blautia*	0.19%	40.31%
			Clostridium	2.14%	20.76%
			Ruminococcus	2.12%	8.76%
			Butyrivibrio	0.18%	4.11%
			Coprococcus	0.34%	4.11%
	FR-AD-1	83.56	*Clostridium*	1.83%	19.43%
			Coprococcus	1.22%	19.23%
			Blautia	0.13%	8.93%
			Catenibacterium	0.00%	4.47%
	IT-AD-1	89.35	*Clostridium*	1.92%	25.41%
			Blautia	0.15%	21.24%
			Enterobacteriales	1.67%	4.83%
	IT-AD-3	82.37	*Coprococcus*	1.08%	20.02%
			Catenibacterium	0.02%	14.57%
			Blautia	0.21%	10.96%
			Clostridium	1.54%	10.07%
COG3539	IT-AD-5	83.43	Enterobacteriales	2.10%	96.25%
COG3188	IT-AD-5	82.99	Enterobacteriales	2.10%	92.07%

Supplementary Table C-7. KEGG orthologous groups overrepresented in enterotypes.
KEGG orthologous groups (KO) overrepresented in enterotypes, and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text. eggNOG orthologous groups (COG) corresponding to these KO's are also listed when applicable.

Entero-type	**KO**	**Description**	**P-value**	**COG**
Acore	K12373	beta-hexosaminidase [EC:3.2.1.52]	5.91e-24	COG3525
	K03296	hydrophobic/amphiphilic exporter-1 (mainly G-bacteria), HAE1 family	8.64e-19	COG0841
	K03585	membrane fusion protein	5.42e-14	COG0845
	K00358		1.76e-13	COG2249
	K10819	histidine kinase	6.89e-10	-
	K03530	DNA-binding protein HU-beta	1.47e-09	COG0776
	K00850	6-phosphofructokinase [EC:2.7.1.11]	9.12e-08	COG0205 COG1105
	K03442	small conductance mechanosensitive ion channel, MscS family	1.26e-07	COG0668
	K00676	ribosomal-protein-alanine N-acetyltransferase [EC:2.3.1.128]	3.78e-07	-

	K00798	cob(I)alamin adenosyltransferase [EC:2.5.1.17]	1.68e-05	COG2109
	K00648	3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180]	3.04e-05	COG0332
	K00680		4.38e-05	-
	K03327	multidrug resistance protein, MATE family	5.83e-05	COG0534
	K12308	beta-galactosidase [EC:3.2.1.23]	0.0006	COG1874
	K00046	gluconate 5-dehydrogenase [EC:1.1.1.69]	0.0008	COG1028
	K03455	monovalent cation:H+ antiporter-2, CPA2 family	0.0012	COG0475 COG1226
	K00786		0.0013	-
	K00262	glutamate dehydrogenase (NADP+) [EC:1.4.1.4]	0.0018	COG0334
	K00638	chloramphenicol O-acetyltransferase [EC:2.3.1.28]	0.0031	COG0110
	K00026	malate dehydrogenase [EC:1.1.1.37]	0.0036	COG0039
	K03307	solute:Na+ symporter, SSS family	0.0038	COG0591
	K03294	basic amino acid/polyamine antiporter, APA family	0.0053	COG0531
	K00077	2-dehydropantoate 2-reductase [EC:1.1.1.169]	0.006	COG1893
	K11752	diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)uracil reductase [EC:3.5.4.26 1.1.1.193]	0.0076	COG0117 COG1985
	K00950	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase [EC:2.7.6.3]	0.0236	COG0801
	K00817	histidinol-phosphate aminotransferase [EC:2.6.1.9]	0.025	COG0079
	K03305	proton-dependent oligopeptide transporter, POT family	0.0253	COG3104
	K04041	fructose-1,6-bisphosphatase III [EC:3.1.3.11]	0.0297	COG3855
	K03100	signal peptidase I [EC:3.4.21.89]	0.03	COG0681
	K00566	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase [EC:2.1.1.61]	0.0303	COG0482
	K00845	glucokinase [EC:2.7.1.2]	0.031	COG0837
	K00767	nicotinate-nucleotide pyrophosphorylase (carboxylating) [EC:2.4.2.19]	0.0313	COG0157
	K11688	C4-dicarboxylate-binding protein DctP	0.0326	COG1638
	K00928	aspartate kinase [EC:2.7.2.4]	0.035	COG0527
	K00533	ferredoxin hydrogenase large subunit [EC:1.12.7.2]	0.0352	-
Bcore	K03543	multidrug resistance protein A	1.72e-20	COG1566
	K00971	mannose-1-phosphate guanylyltransferase [EC:2.7.7.22]	2.18e-18	COG0662 COG0836
	K03760	putative membrane protein	1.03e-17	COG2194
	K00973	glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24]	8.43e-14	COG1209
	K00640	serine O-acetyltransferase [EC:2.3.1.30]	4.92e-10	COG1045

	K03315	Na+:H+ antiporter, NhaC family	2.27e-09	COG1757
	K03453	bile acid:Na+ symporter, BASS family	3.15e-07	COG0385
	K00874	2-dehydro-3-deoxygluconokinase [EC:2.7.1.45]	4.22e-07	COG0524
	K03624	transcription elongation factor GreA	4.96e-07	COG0782
	K00847	fructokinase [EC:2.7.1.4]	1.18e-06	COG0524
	K03086	RNA polymerase primary sigma factor	1.67e-06	COG0568
	K00788	thiamine-phosphate pyrophosphorylase [EC:2.5.1.3]	1.9e-06	COG0352
	K03320	ammonium transporter, Amt family	3.6e-06	COG0004
	K03587	cell division protein FtsI (penicillin binding protein 3) [EC:2.4.1.129]	6.89e-06	COG0768
	K03630	DNA repair protein RadC	1.37e-05	COG2003
	K00784	ribonuclease Z [EC:3.1.26.11]	2.12e-05	COG1234
	K11068	hemolysin III	3.11e-05	COG1272
	K00941	phosphomethylpyrimidine kinase [EC:2.7.4.7]	6.1e-05	COG0351
	K03816	xanthine phosphoribosyltransferase [EC:2.4.2.22]	6.24e-05	COG0503
	K00877	hydroxymethylpyrimidine kinase [EC:2.7.1.49]	6.29e-05	-
	K00903	protein-tyrosine kinase [EC:2.7.10.-]	8.56e-05	-
	K03593	ATP-binding protein involved in chromosome partitioning	0.0001	COG0489
	K00278	L-aspartate oxidase [EC:1.4.3.16]	0.0001	COG0029
	K00924		0.0002	COG0515 COG1493
	K00842	aminotransferase [EC:2.6.1.-]	0.0002	COG1168
	K03298	drug/metabolite transporter, DME family	0.0006	COG0697
	K03321	sulfate permease, SulP family	0.0008	COG0659
	K00266	glutamate synthase (NADPH/NADH) small chain [EC:1.4.1.13 1.4.1.14]	0.0011	COG0493
	K00811	aspartate aminotransferase [EC:2.6.1.1]	0.0013	-
	K03546	exonuclease SbcC	0.0018	COG0419
	K00661	maltose O-acetyltransferase [EC:2.3.1.79]	0.0018	COG0110
	K03458	nucleobase:cation symporter-2, NCS2 family	0.0028	COG2233
	K03517	quinolinate synthase	0.0038	COG0379
	K03699	putative hemolysin	0.0039	COG1253
	K00611	ornithine carbamoyltransferase [EC:2.1.3.3]	0.0049	COG0078
	K03564	peroxiredoxin Q/BCP [EC:1.11.1.15]	0.0051	COG1225
	K00031	isocitrate dehydrogenase [EC:1.1.1.42]	0.0071	COG0538
	K00764	amidophosphoribosyltransferase [EC:2.4.2.14]	0.0087	COG0034
	K03427	type I restriction enzyme M protein [EC:2.1.1.72]	0.0116	COG0286
	K00970	poly(A) polymerase [EC:2.7.7.19]	0.0144	COG0617
	K03602	exodeoxyribonuclease VII small subunit [EC:3.1.11.6]	0.0156	COG1722
	K00865	glycerate kinase [EC:2.7.1.31]	0.0178	COG1929

	K03856	3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]	0.021	COG2876
	K03310	alanine or glycine:cation symporter, AGCS family	0.0211	COG1115
	K03308	neurotransmitter:Na+ symporter, NSS family	0.0233	COG0733
	K00818	acetylornithine aminotransferase [EC:2.6.1.11]	0.0239	COG4992
	K03531	cell division protein FtsZ	0.0242	COG0206
	K03092	RNA polymerase sigma-54 factor	0.0273	COG1508
	K00773	queuine tRNA-ribosyltransferase [EC:2.4.2.29]	0.0295	COG0343
	K03565	regulatory protein	0.0325	COG2137
	K03711	Fur family transcriptional regulator, ferric uptake regulator	0.0386	COG0735
	K00705	4-alpha-glucanotransferase [EC:2.4.1.25]	0.0394	COG1640
	K03551	holliday junction DNA helicase RuvB	0.0405	COG2255
	K00290	saccharopine dehydrogenase (NAD+, L-lysine forming) [EC:1.5.1.7]	0.0418	-
	K03561	biopolymer transport protein ExbB	0.0463	COG0811
	K12257	SecD/SecF fusion protein	0.0469	COG0342 COG0341
	K00287	dihydrofolate reductase [EC:1.5.1.3]	0.0469	COG0262
	K00703	starch synthase [EC:2.4.1.21]	0.0497	COG0297
C	K00571	site-specific DNA-methyltransferase (adenine-specific) [EC:2.1.1.72]	1.32e-06	-
	K03091	RNA polymerase sporulation-specific sigma factor	1.58e-06	-
	K00857	thymidine kinase [EC:2.7.1.21]	1.93e-06	COG1435
	K00336	NADH dehydrogenase I subunit G [EC:1.6.5.3]	7.29e-06	COG1034
	K00604	methionyl-tRNA formyltransferase [EC:2.1.2.9]	1.63e-05	COG0223
	K03111	single-strand DNA-binding protein	5.64e-05	COG0629
	K00852	ribokinase [EC:2.7.1.15]	5.92e-05	COG0524
	K03588	cell division protein FtsW	9.27e-05	COG0772
	K00939	adenylate kinase [EC:2.7.4.3]	0.0025	COG0563
	K03572	DNA mismatch repair protein MutL	0.0025	COG0323
	K00783	hypothetical protein	0.0025	COG1576
	K03110	signal recognition particle receptor	0.0025	COG0552
	K03925	MraZ protein	0.0035	COG2001
	K03709	DtxR family transcriptional regulator, Mn-dependent transcriptional regulator	0.0037	COG1321
	K03621	fatty acid/phospholipid synthesis protein	0.0042	COG0416
	K00942	guanylate kinase [EC:2.7.4.8]	0.0059	COG0194
	K00790	UDP-N-acetylglucosamine 1-carboxyvinyltransferase [EC:2.5.1.7]	0.006	COG0766
	K00943	dTMP kinase [EC:2.7.4.9]	0.0079	COG0125
	K03702	excinuclease ABC subunit B	0.0102	COG0556

	K03500	ribosomal RNA small subunit methyltransferase B [EC:2.1.1.-]	0.011	COG0144
	K03529	chromosome segregation protein	0.0111	COG1196
	K03617	electron transport complex protein RnfA	0.0112	COG2209
	K03581	exodeoxyribonuclease V alpha subunit [EC:3.1.11.5]	0.0113	COG0507
	K00248	butyryl-CoA dehydrogenase [EC:1.3.99.2]	0.0114	COG1960
	K03737	putative pyruvate-flavodoxin oxidoreductase [EC:1.2.7.-]	0.0143	COG0674 COG1013
	K03625	N utilization substance protein B	0.0179	COG0781
	K03043	DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	0.0203	COG0085
	K03168	DNA topoisomerase I [EC:5.99.1.2]	0.0243	COG0550 COG0551 COG1754
	K03466	DNA segregation ATPase FtsK/SpoIIIE, S-DNA-T family	0.0257	COG1674
	K00288	methylenetetrahydrofolate dehydrogenase (NADP+) [EC:1.5.1.5]	0.033	-
	K00600	glycine hydroxymethyltransferase [EC:2.1.2.1]	0.0403	COG0112
	K03657	DNA helicase II / ATP-dependent DNA helicase PcrA [EC:3.6.1.-]	0.0423	COG0210
D	K03497	chromosome partitioning protein, ParB family	7.59e-05	COG1475
	K03496	chromosome partitioning protein	0.0299	COG1192
	K03292	glycoside/pentoside/hexuronide:cation symporter, GPH family	0.0312	COG2211
	K00642	glutamate N-acetyltransferase [EC:2.3.1.35]	0.0346	-
	K03892	ArsR family transcriptional regulator	0.0426	COG0640
	K00789	S-adenosylmethionine synthetase [EC:2.5.1.6]	0.0446	COG0192 COG1812

Supplementary Table C-8. KEGG modules overrepresented in enterotypes.

KEGG modules overrepresented in enterotypes and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text.

Entero-type	**Module**	**Description**	**P-value**
Acore	M00155	Keratan sulfate degradation	2.05E-24
	M00006	Pentose phosphate pathway, oxidative phase	2.30E-17
	M00248	Biotin biosynthesis, pimeloyl-CoA => biotin	3.46E-11
	M00247	Cobalamin biosynthesis, cobinamide => cobalamin	2.13E-08
	M00159	Fatty acid biosynthesis, initiation	2.73E-08
	M00008	Entner-Doudoroff pathway	3.15E-08

	M00102	UDP-Xylose biosynthesis, UDP-Glc => UDP-Xyl => Xyl	2.72E-07
	M00083	beta-Alanine biosynthesis, L-aspartate => beta-alanine	6.02E-07
	M00001	Glycolysis, fructose-6P => pyruvate	5.88E-06
	M00160	Fatty acid biosynthesis, elongation	1.40E-05
	M00017	Glutamate biosynthesis, oxoglutarate => glutamate (glutamate dehydrogenase)	4.57E-05
	M00046	Asparagine degradation, asparagine => aspartate +NH3	0.0028
	M00037	Histidine biosynthesis, PRPP => histidine	0.0057
	M00255	Ascorbate biosynthesis, animals	0.006
	M00012	Glyoxylate cycle	0.0104
	M00104	CMP-N-Acetylneuraminate biosynthesis (mammals), ManNAc => Neu5Ac-9P => CMP-Neu5Ac	0.013
	M00297	CAM, dark	0.0137
	M00608	PTS system, beta-glucosides-specific II component	0.0137
	M00249	Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P	0.0201
Bcore	M00105	dTDP-Glucose, dTDP-galactose and dTDP-rhamnose biosynthesis	3.47E-30
	M00273	Complex I (NADH dehydrogenase), NADH dehydrogenase I	1.63E-23
	M00252	Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P	3.34E-22
	M00099	GDP-Mannose biosynthesis, fructose-6P => GDP-Man	1.59E-20
	M00239	Ascorbate biosynthesis, plants	1.72E-20
	M00034	Tryptophan biosynthesis, chorismate => tryptophan	3.88E-20
	M00156	Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen	2.70E-17
	M00323	Putative spermidine/putrescine transport system	2.15E-16
	M00117	Uronic acid metabolism	3.49E-12
	M00278	Complex II (succinate dehydrogenase / fumarate reductase), succinate dehydrogenase	1.29E-08
	M00032	Cysteine biosynthesis, serine => cysteine	5.50E-07
	M00240	NAD biosynthesis, aspartate => NAD	2.75E-05
	M00192	C5 isoprenoid biosynthesis, non-mevalonate pathway	3.37E-05
	M00302	Reductive carboxylate cycle	7.77E-05
	M00299	C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type	0.0001
	M00042	Urea cycle	0.0001
	M00062	Cysteine metabolism, cysteine => 3-sulfino-L-alanine => pyruvate	0.0003
	M00009	Citrate cycle	0.0003
	M00063	Cysteine metabolism, cysteine => 3-mercaptopyruvate => pyruvate	0.0004
	M00660	RuvABC complex	0.0005

	M00028	Leucine biosynthesis, pyruvate => leucine	0.0005
	M00053	Methionine biosynthesis, intermediates, homoserine => O-acetylhomoserine => L-homocysteine	0.0011
	M00029	Isoleucine biosynthesis, pyruvate => isoleucine	0.0013
	M00282	Cytochrome bd complex	0.0018
	M00010	Citrate cycle, first carbon oxidation	0.0033
	M00090	Inosine monophosphate biosynthesis, PRPP + glutamine => IMP	0.0068
	M00300	C4-dicarboxylic acid cycle, NAD+ -malic enzyme type	0.0069
	M00269	Tetrahydrofolate biosynthesis	0.0071
	M00020	Proline biosynthesis, glutamate => proline	0.0126
	M00298	CAM, light	0.0198
	M00011	Citrate cycle, second carbon oxidation	0.0244
	M00301	C4-dicarboxylic acid cycle, NADP+ -malic enzyme type	0.0424
	M00058	Leucine degradation, leucine => acetoacetate + acetyl-CoA	0.0494
C	M00308	Ribosome, bacteria	6.27E-35
	M00309	Ribosome, archaea	1.18E-27
	M00351	Simple sugar transport system	2.77E-23
	M00318	Sulfonate/nitrate/taurine transport system	2.62E-10
	M00320	Iron(III) transport system	2.58E-07
	M00367	Branched-chain amino acid transport system	5.41E-07
	M00095	Pyrimidine deoxyribonuleotide biosynthesis, CDP/CTP => dCDP/dCTP,dTDP/dTTP	0.0002
	M00372	Zinc transport system	0.0004
	M00270	C1-unit interconversion	0.0007
	M00597	DNA polymerase III complex	0.0015
	M00641	MutHLS complex	0.0016
	M00092	Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP	0.0018
	M00262	Ptrescine metabolism, N-acetylation, putrescine => 4-aminobutanoate	0.0029
	M00659	RecFOR complex	0.0037
	M00370	Iron complex transport system	0.0038
	M00313	RNA polymerase, bacteria	0.0039
	M00648	uvrA2B2 complex	0.0087
	M00103	UDP-N-Acetylmuramate biosynthesis, UDP-GlcNAc => UDP-MurNAc	0.0284
	M00091	Adenine nucleotide biosynthesis, IMP => ADP/dADP,ATP/dATP	0.0369
	M00242	Ubiquinone biosynthesis, chorismate => ubiquinone, prokaryotes	0.042

	M00031	Glycine biosynthesis, serine => glycine	0.0429
	M00658	RecBCD complex	0.0445
D	M00378	Antibiotic ABC transport system	6.88E-10
	M00610	PTS system, fructose-specific II component	2.78E-05
	M00324	Maltose/maltodextrin transport system	0.001
	M00326	Multiple sugar transport system	0.0011
	M00344	Methyl-galactoside transport system	0.0024
	M00380	Lipopolysaccharide transport system	0.003
	M00377	Putative ABC transport system	0.0032
	M00260	Polyamine biosynthesis, arginine => putrescine => spermidine	0.0054
	M00007	Pentose phosphate pathway, non-oxidative phase	0.0088
	M00118	Pentose interconversion, arabinose/ribulose/xylulose/xylose	0.0091
	M00005	Pentose phosphate pathway and PRPP biosynthesis	0.0165
	M00293	ATP synthase	0.0177
	M00004	Pentose phosphate pathway	0.0276
	M00055	Methionine degradation	0.0277
	M00056	S-Adenosylmethionine biosynthesis, methionine => S-adenosylmethionine => methionine	0.0304
	M00671	Sermidine/putrescine transport system	0.0381
	M00211	Diphosphatidylglycerol biosynthesis, CDP-glycerol => cardiolipin	0.0491
	M00054	Methionine salvage pathway	0.0496

Supplementary Table C-9. Orthologous groups enriched in groups of individuals.

Orthologous groups overrepresented in correlation with host properties and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text.

Feature	**Value**	**OG**	**Description**	**P-value**
gender	female	COG0673	Predicted dehydrogenases and related proteins	0.001502
	male	COG1192	ATPases involved in chromosome partitioning	0.04023
		COG0463	Glycosyltransferases involved in cell wall biogenesis	0.045692
nationality	Danish	COG2801	Transposase and inactivated derivatives	9.98E-54
		COG0249	Mismatch repair ATPase (MutS family)	6.92E-05
		COG0826	Collagenase and related proteases	0.000495
		COG0564	Pseudouridylate synthases, 23S RNA-specific	0.000912
	French	COG1472	Beta-glucosidase-related glycosidases	0.000132
		COG0058	Glucan phosphorylase	0.008734
		COG0458	Carbamoylphosphate synthase large subunit	0.047006

			(split gene in MJ)	
	Italian	COG1609	Transcriptional regulators	0.000764
		COG1070	Sugar (pentulose and hexulose) kinases	0.003002
		COG0436	Aspartate/tyrosine/aromatic aminotransferase	0.006779
		COG1940	Transcriptional regulator/sugar kinase	0.018231
	Japanese	COG0583	Transcriptional regulator	9.07E-17
		COG0454	Histone acetyltransferase HPA2 and related acetyltransferases	3.46E-08
		COG0500	SAM-dependent methyltransferases	4.21E-08
		COG0480	Translation elongation factors (GTPases)	2.29E-07
		COG0438	Glycosyltransferase	2.31E-05
		NOG75023	Regulator protein	8.83E-05
		COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	0.003332
		COG1253	Hemolysins and related proteins containing CBS domains	0.004759
		COG1032	Fe-S oxidoreductase	0.007059
		COG2244	Membrane protein involved in the export of O-antigen and teichoic acid	0.019799
	Spanish	COG1506	Dipeptidyl aminopeptidases/acylaminoacyl-peptidases	1.59E-31
		COG0526	Thiol-disulfide isomerase and thioredoxins	5.62E-12
		COG0110	Acetyltransferase (isoleucine patch superfamily)	8.59E-06
		COG4974	Site-specific recombinase XerD	5.34E-05
		COG0514	Superfamily II DNA helicase	0.000222
		COG5545	Predicted P-loop ATPase and inactivated derivatives	0.000705

Supplementary Table C-10. KEGG modules enriched in groups of individuals.

Functional modules overrepresented in correlation with host properties and the P-values for their enrichment. Functions mentioned in the main text are emphasized in bold text.

Feature	Value	Module	Description	P-value
gender	female	M00370	Iron complex transport system	0.001796
		M00279	Complex II (succinate dehydrogenase / fumarate reductase), fumarate reductase	0.002759
		M00097	UDP-glucose and UDP-galactose biosynthesis	0.026468
		M00242	Ubiquinone biosynthesis, chorismate => ubiquinone, prokaryotes	0.048277
	male	M00021	Aspartate biosynthesis, oxaloacetate => aspartate	0.021322
nationality	Danish	M00318	Sulfonate/nitrate/taurine transport system	2.13E-08

		M00658	RecBCD complex	5.26E-05
		M00649	uvrBC complex	0.003858
		M00372	Zinc transport system	0.005882
		M00092	Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP	0.00678
		M00659	RecFOR complex	0.019367
		M00104	CMP-N-Acetylneuraminate biosynthesis (mammals), ManNAc => Neu5Ac-9P => CMP-Neu5Ac	0.023721
		M00203	Glyceroglycolipid biosynthesis	0.028296
		M00154	Heparan sulfate degradation	0.029536
		M00101	UDP-N-Acetylgalactosamine and UDP-N-acetylmannosamine biosynthesis	0.044988
		M00153	Chondroitin sulfate degradation	0.049869
	French	M00005	Pentose phosphate pathway and PRPP biosynthesis	0.00108
		M00004	Pentose phosphate pathway	0.001968
	Japanese	M00366	Polar amino acid transport system	7.80E-22
		M00247	Cobalamin biosynthesis, cobinamide => cobalamin	8.24E-06
		M00608	PTS system, beta-glucosides-specific II component	5.54E-05
		M00076	Dopamine / noradrenaline / adrenaline metabolism	8.77E-05
		M00023	Lysine biosynthesis, aspartate => lysine	0.000131
		M00159	Fatty acid biosynthesis, initiation	0.003303
		M00386	Cell division transport system	0.005068
		M00024	Methionine biosynthesis, apartate => homoserine => methionine	0.005229
		M00030	Serine biosynthesis, glycerate-3P => serine	0.00805
		M00098	UDP-N-Acetylglucosamine biosynthesis, fructose-6P => UDP-GlcNAc	0.032239
		M00025	Threonine biosynthesis, apartate => homoserine => threonine	0.035052
	Spanish	M00105	dTDP-Glucose, dTDP-galactose and dTDP-rhamnose biosynthesis	2.87E-19
		M00273	Complex I (NADH dehydrogenase), NADH dehydrogenase I	1.50E-17
		M00253	Phylloquinone biosynthesis, chorismate => phylloquinone	1.12E-15
		M00241	Menaquinone biosynthesis, chorismate => menaquinone	4.23E-14
		M00034	Tryptophan biosynthesis, chorismate => tryptophan	5.68E-11
		M00323	Putative spermidine/putrescine transport system	2.96E-10

		M00158	Pectin degradation	9.35E-10
		M00252	Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P	1.95E-08
		M00036	Tyrosine biosynthesis, chorismate => tyrosine	6.79E-08
		M00156	Lipopolysaccharide biosynthesis, inner core => outer core => O-antigen	3.02E-07
		M00340	Putative ABC transport system	8.55E-07
		M00119	CMP-Kdo biosynthesis	3.08E-06
		M00278	Complex II (succinate dehydrogenase / fumarate reductase), succinate dehydrogenase	1.87E-05
		M00032	Cysteine biosynthesis, serine => cysteine	0.000191
		M00248	Biotin biosynthesis, pimeloyl-CoA => biotin	0.000198
		M00249	Pyridoxal biosynthesis, erythrose-4P => pyridoxal-5P	0.000445
		M00012	Glyoxylate cycle	0.000546
		M00282	Cytochrome bd complex	0.000643
		M00016	Glucuronate pathway (uronate pathway)	0.001058
		M00083	beta-Alanine biosynthesis, L-aspartate => beta-alanine	0.002434
		M00160	Fatty acid biosynthesis, elongation	0.003515
		M00046	Asparagine degradation, asparagine => aspartate +NH3	0.007557
		M00300	C4-dicarboxylic acid cycle, NAD+ -malic enzyme type	0.010344
		M00255	Ascorbate biosynthesis, animals	0.011157
		M00018	Glutamine biosynthesis, glutamate => glutamine	0.017718
		M00244	Pantothenate biosynthesis, valine => pantothenate	0.021502
		M00010	Citrate cycle, first carbon oxidation	0.042009
		M00003	Gluconeogenesis, oxaloacetate => fructose-6P	0.042725
clinical status	obese	M00211	Diphosphatidylglycerol biosynthesis, CDP-glycerol => cardiolipin	0.003436
		M00319	Molybdate transport system	0.007701
		M00118	Pentose interconversion, arabinose/ribulose/xylulose/xylose	0.008401

Supplementary Table C-11. Orthologous groups significantly correlating with age when combined into a linear model.

Orthologous group	Description
COG0205	6-phosphofructokinase
COG0366	Glycosidases
COG0438	Glycosyltransferase
COG4646	DNA methylase

C.3 Supplementary Notes

C.3.1 Functions identified in gut metagenomes

Histidine kinases (COG0642) formed the most frequent orthologous group (OG) in the gut metagenomes. Using the phylogenetic origins of genes forming this OG (See Section 2.2.6 for details), we observed that *Bacteroides* and *Prevotella* contribute a very high fraction of this function in our metagenomes. Thus the observed expansion of signaling genes such as this OG in *Bacteroides thetaiotaomicron* must extend also to the *Prevotella*, which is also from the phylum Bacteroidetes.

The most variable OG in our gut metagenome samples is an ATPase (COG1132) component of ABC-type transporters. ABC type transport system is one of the most conserved molecular machines, which contributes not only for efflux but also for influx of compounds. These transporters participate in the persistence of bacteria in their ecological environment[152]. Their tremendous variety is also observed in the STRING database (Supplementary Figure C-12), suggesting the contribution to the diversity of bacterial ability, such as drug resistance.

Contributions

Transcriptome Complexity in a Genome-Reduced Bacterium

Marc Güell, Vera van Noort, Eva Yus, Wei-Hua Chen, Justine Leigh-Bell, Konstantinos Michalodimitrakis, Takuji Yamada, Manimozhiyan Arumugam, Tobias Doerks, Sebastian Kühner, Michaela Rode, Mikita Suyama, Sabine Schmidt, Anne-Claude Gavin, Peer Bork, Luis Serrano

Science **326**(5957):1268-1271. doi: 10.1126/science.1176951

This paper shows an unsuspected complexity of transcriptional regulation through alternative transcripts, antisense RNAs and multiple regulatory sites per gene in *Mycoplasma pneumoniae* that has one of the smallest known genomes with 689 protein-encoding genes.

I performed assembly validation of the resequenced genome of *Mycoplasma pneumoniae*, compared the new genome to an existing reference genome, and identified the real genomic changes in the resequenced genome. This laid the groundwork for designing the tiling microarray used in the study.

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, et al.

Nature **464**(7285):59-65. doi:10.1038/nature08821

This is a fundamental paper regarding the human gut metagenome that appeared in the top journal Nature (London). It establishes a gene catalogue of 3.3 million genes in the human gut microbiome. There are 53 coauthors, and I am the fourth coauthor as I gave the following essential contributions to the paper:

Techniques and tasks: I participated in designing the analyses, generated vector and quality trimmed Sanger reads from human gut metagenomes of 13 Japanese individuals (used for Fig. 1), performed validation of Solexa metagenomic assembly by comparing to 454 assembly and Sanger reads, estimated the Solexa assembly error rate, estimated the completeness of the gene catalogue by comparing to the gene set

from 89 frequent gut microbial genomes (Fig. 2b), contributed to the functional rarefaction analysis (Fig. 2c).

Manuscript: I proofread and edited all versions of the manuscript for its structure as well as biological implications, which include a deeper understanding of the different microbial species in the human gut and their frequencies as well as their functional (enzymatic) repertoire.

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, et al.

Submitted to *Nature*. Manuscript # 2010-03-03138

In this paper, submitted to Nature (London), we identify and study enterotypes of the human gut microbiome. My supervisor Prof. Bork led the study. I am the first author (shared authorship with one other author; there are 39 other coauthors) as I provided the following key contributions to the paper:

Techniques and tasks: I participated in designing the analyses, trimmed the sequence reads from gut metagenome from 39 individuals, assembled the reads, predicted genes, performed BLASTP searches of predicted proteins against STRING and KEGG databases, downloaded sequences and added missing annotations for the 1152 microbial reference genomes, estimated the sequence similarity thresholds for accurately mapping reads at phylum/genus levels, developed the phylogenetic mapping procedure using reference genomes, developed the quantitative functional characterization procedure using gene abundance, estimated quantitative functional and phylogenetic profiles of samples, showed high correlations of gene and genus abundance distributions between Sanger and 454 technology-derived sequences from same samples, estimated the eukaryotic fraction of the metagenomes, identified highly abundant functions from low-abundance microbes, generated clusters of 39 samples with bootstrap support that were used to establish the enterotypes, performed the jack-knife tests to test robustness of enterotypes.

Manuscript: I wrote a significant portion of the manuscript, proofread and edited all versions of the manuscript, discussing all biological implications of different microbial species and their functional (enzymatic) repertoire.

Curriculum Vitae

Personal Data

Name	Manimozhiyan Arumugam
Nationality	Indian
Date of Birth	06.01.1978
Place of Birth	Madurai, India

Education

Since Sep. 2006	PhD, EMBL – Heidelberg
	Comparative metagenomic analysis of the human intestinal microbiota.
	Supervisor: Dr. habil. Peer Bork
	Doktorvater: Prof. Thomas Dandekar
Aug. 2000 – Aug. 2003	M.S. Computer Science
	University of Nebraska-Lincoln
	EMPRR: A High dimensional EM-based piecewise regression algorithm.
	Advisor: Prof. Stephen D. Scott
Aug. 1995 – Jun. 1999	B. Tech. Biotechnology
	Anna University, India
July 1993 – Mar. 1995	Higher Secondary School
	I.T.O. Higher Secondary, Ayakudi, India
June 1983 – Apr. 1993	Secondary School
	St. Mary's Higher Secondary, Madurai, India

Experience

Apr. 2003 – Aug. 2006	Research Associate
	Washington University in St. Louis
Aug. 2002 – Dec. 2002	Teaching Assistant
	University of Nebraska-Lincoln
Aug. 2000 – Dec. 2002	Research Assistant
	University of Nebraska-Lincoln
July 1999 – July 2000	Software Engineer
	Infosys Technologies Ltd., India

Lebenslauf

Persönliche Daten

Name	Manimozhiyan Arumugam
Nationalität	Indisch
Geburtsdatum	06.01.1978
Geburtsort	Madurai, Indien

Studium und Ausbildung

Seit Sep. 2006	PhD, EMBL – Heidelberg
	Comparative metagenomic analysis of the human intestinal microbiota.
	Betreuer: Dr. habil. Peer Bork
	Doktorvater: Prof. Thomas Dandekar
Aug. 2000 – Aug. 2003	M.S. Computer Science
	University of Nebraska-Lincoln
	EMPRR: A High dimensional EM-based piecewise regression algorithm.
	Betreuer: Prof. Stephen D. Scott
Aug. 1995 – Juni 1999	B. Tech. Biotechnology
	Anna University, India
Juli 1993 – März 1995	Higher Secondary School
	I.T.O. Higher Secondary, Ayakudi, India
Juni 1983 – Apr. 1993	Secondary School
	St. Mary's Higher Secondary, Madurai, India

Beruflicher Werdegang

Apr. 2003 – Aug. 2006	Research Associate
	Washington University in St. Louis
Aug. 2002 – Dec. 2002	Teaching Assistant
	University of Nebraska-Lincoln
Aug. 2000 – Dec. 2002	Research Assistant
	University of Nebraska-Lincoln
Juli 1999 – Juli 2000	Software Engineer
	Infosys Technologies Ltd., India

List of publications

Publications associated with this thesis

1. Guell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S *et al*: **Transcriptome Complexity in a Genome-Reduced Bacterium**. *Science* 2009, **326**(5957):1268-1271.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.
3. Arumugam M, Raes J, Pelletier E, Paslier DL, Yamada T, Mende DR, Fernandes GR, Bruls T, Batto J-M, Bertalan M *et al*: **Enterotypes of the human gut microbiome**. Submitted to *Nature*.

Publications resulting from work before my PhD

1. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE *et al*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 2004, **428**(6982):493-521.
2. Wu JQ, Shteynberg D, Arumugam M, Gibbs RA, Brent MR: **Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing**. *Genome Res* 2004, **14**(4):665-671.
3. Wei C, Lamesch P, Arumugam M, Rosenberg J, Hu P, Vidal M, Brent MR: **Closing in on the C. elegans ORFeome by cloning TWINSCAN predictions**. *Genome Res* 2005, **15**(4):577-582.
4. Arumugam M, Wei C, Brown RH, Brent MR: **Pairagon+N-SCAN_EST: a model-based gene annotation pipeline**. *Genome Biol* 2006, **7 Suppl 1**:S5 1-10.
5. Giannakis M, Stappenbeck TS, Mills JC, Leip DG, Lovett M, Clifton SW, Ippolito JE, Glasscock JI, Arumugam M, Brent MR *et al*: **Molecular properties of adult mouse gastric and intestinal epithelial progenitors in their niches**. *J Biol Chem* 2006, **281**(16):11292-11300.

6. Keibler E, Arumugam M, Brent MR: **The Treeterbi and Parallel Treeterbi algorithms: efficient, optimal decoding for ordinary, generalized and pair HMMs**. *Bioinformatics* 2007, **23**(5):545-554.
7. Lu DV, Brown RH, Arumugam M, Brent MR: **Pairagon: a highly accurate, HMM-based cDNA-to-genome aligner**. *Bioinformatics* 2009, **25**(13):1587-1593.

Bibliography

1 Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* **31**, 107-133 (1977).

2 Lederberg, J. & McCray, A. 'Ome Sweet 'Omics-- A Genealogical Treasury of Words. *The Scientist* **17** (2001).

3 Xu, J. & Gordon, J. I. Honor thy symbionts. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 10452-10459 (2003).

4 Guarner, F. & Malagelada, J.-R. Gut flora in health and disease. *The Lancet* **361**, 512-519 (2003).

5 Duncan, S. H. *et al.* Oxalobacter formigenes and Its Potential Role in Human Health. *Appl. Environ. Microbiol.* **68**, 3841-3847 (2002).

6 Hooper, L. V., Midtvedt, T. & Gordon, J. I. How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu Rev Nutr* **22**, 283-307 (2002).

7 Gronlund, M. M., Arvilommi, H., Kero, P., Lehtonen, O. P. & Isolauri, E. Importance of intestinal colonisation in the maturation of humoral immunity in early infancy: a prospective follow up study of healthy infants aged 0-6 months. *Arch Dis Child Fetal Neonatal Ed* **83**, F186-192 (2000).

8 Fox, G. E. *et al.* The phylogeny of prokaryotes. *Science* **209**, 457-463 (1980).

9 Eckburg, P. B. *et al.* Diversity of the Human Intestinal Microbial Flora. *Science* **308**, 1635-1638 (2005).

10 Hayashi, H., Sakamoto, M. & Benno, Y. Phylogenetic Analysis of the Human Gut Microbiota Using 16S rDNA Clone Libraries and Strictly Anaerobic Culture-Based Methods. *MICROBIOLOGY and IMMUNOLOGY* **46**, 535 (2002).

11 Lay, C. *et al.* Colonic Microbiota Signatures across Five Northern European Countries. *Appl. Environ. Microbiol.* **71**, 4153-4155 (2005).

12 Seksik, P. *et al.* Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut* **52**, 237-242 (2003).

13 Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences* **104**, 13780-13785 (2007).

14 Zoetendal, E. G., Rajilic-Stojanovic, M. & de Vos, W. M. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57**, 1605-1615 (2008).

15 Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022-1023 (2006).
16 Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proceedings of the National Academy of Sciences* **105**, 2117-2122 (2008).
17 Schwiertz, A. *et al.* Microbiota and SCFA in Lean and Overweight Healthy Subjects. *Obesity* **18**, 190 (2009).
18 Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031 (2006).
19 Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205-211 (2006).
20 Sokol, H. *et al.* Specificities of the fecal microbiota in inflammatory bowel disease. *Inflamm Bowel Dis* **12**, 106-111 (2006).
21 Parracho, H. M., Bingham, M. O., Gibson, G. R. & McCartney, A. L. Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children. *J Med Microbiol* **54**, 987-991 (2005).
22 Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucl. Acids Res.* **37**, D412-416 (2009).
23 *List of 1152 reference genomes retrieved from the public domain* <http://www.bork.embl.de/Docu/metagenomics/reference_genomes/20090501/Refgenomes_20090501.pdf> (Accessed: 25 Mar 2010).
24 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43 (2004).
25 Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).
26 Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
27 Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**, 1374-1379 (2009).
28 Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355-1359 (2006).
29 Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**, 169-181 (2007).
30 Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).

31 Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* **104**, 13913-13918 (2007).
32 Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* **6**, 693-699 (2008).
33 Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386 (2007).
34 Mitra, S., Klar, B. & Huson, D. H. Visual and statistical comparison of metagenomes. *Bioinformatics* **25**, 1849-1855 (2009).
35 Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
36 Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
37 Wooley, J. C., Godzik, A. & Friedberg, I. A Primer on Metagenomics. *PLoS Comput Biol* **6**, e1000667 (2010).
38 Dandekar, T. *et al.* Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic Acids Res* **28**, 3278-3288 (2000).
39 Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res* **24**, 4420-4449 (1996).
40 Guell, M. *et al.* Transcriptome Complexity in a Genome-Reduced Bacterium. *Science* **326**, 1268-1271 (2009).
41 van Noort, V. *et al.* Genome sequence of a recently evolved thermophilic eukaryote. (in preparation).
42 Handelsman, J. *et al. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* (The National Academies Press, 2007).
43 Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6**, 805-814 (2005).
44 Allen, E. E. & Banfield, J. F. Community genomics in microbial ecology and evolution. *Nat Rev Micro* **3**, 489-498 (2005).
45 Guarner, F. Enteric flora in health and disease. *Digestion* **73 Suppl 1**, 5-12 (2006).
46 MacFarlane, G. T. & Cummings, J. H. in *The Large Intestine: Physiology, Pathophysiology, and Disease* Vol. 116 *Annals of Internal Medicine* eds S.F. Philips, J.H. Pemberton, & R.G. Shorter) 704-704 (Raven Press, 1992).

47 Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**, 776-788 (2008).

48 Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72**, 557-578, Table of Contents (2008).

49 Raes, J., Foerstner, K. U. & Bork, P. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10**, 490-498 (2007).

50 Toft, U. *et al.* The impact of a population-based multi-factorial lifestyle intervention on changes in long-term dietary habits: the Inter99 study. *Prev Med* **47**, 378-383 (2008).

51 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).

52 Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**, 265-272 (2010).

53 Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**, 5623-5630 (2006).

54 Kent, W. J. BLAT--The BLAST-Like Alignment Tool. *Genome Research* **12**, 656-664 (2002).

55 Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**, D250-254 (2008).

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

57 van Dongen, S. M. *Graph Clustering by Flow Simulation*, University of Utrecht, The Netherlands, (2000).

58 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

59 Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ Microbiol* **63**, 2802-2813 (1997).

60 DiGuistini, S. *et al.* De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* **10**, R94 (2009).

61 Arumugam, M., Harrington, E., Raes, J. & Bork, P. SMASH-Community: Simple Metagenomics Analysis Shell for Shotgun Sequences. *Bioinformatics* (submitted).

62 Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-189 (2002).

63 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**, 91-96 (2003).

64 Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucl. Acids Res.* **29**, 2607-2618 (2001).

65 Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115 (1998).

66 National Center for Biotechnology Information. *Complete Microbial Genomes* <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi> (Accessed: 10 Apr 2009).

67 Baylor College of Medicine. *Microbial Genome Projects at BCM HGSC* <http://www.hgsc.bcm.tmc.edu/projects/microbial/microbial-index.xsp> (Accessed: Feb 27 2009).

68 J. Craig Venter Institute. *JCVI: HMP / Human Microbiome Project* <http://hmp.jcvi.org/status.shtml> (Accessed: 01 Mar 2009).

69 MetaHIT Consortium. *MetaHIT draft bacterial genomes at the Sanger Institute* <http://www.sanger.ac.uk/pathogens/metahit/> (Accessed: 01 Apr 2009).

70 Broad Institute of MIT and Harvard. *Nonhuman Genomes & Genetic Variation: Bacteria* <http://www.broadinstitute.org/seq/msc> (Accessed: 01 Apr 2009).

71 Huang, Y., Gilna, P. & Li, W. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**, 1338-1340 (2009).

72 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267 (2007).

73 Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucl. Acids Res.* **36**, e120- (2008).

74 Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucl. Acids Res.* **38**, D190-195 (2010).

75 Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480-484 (2008).

76 Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *Information Theory, IEEE Transactions on* **49**, 1860 (2003).

77 Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences* **106**, 2677-2682 (2009).

78 Felsenstein, J. PHYLIP (Phylogeny Inference Package) v.3.6 (Distributed by the author. Department of Genome Sciences, University of Washington, Seattle., 2005).

79 Sorek, R. *et al.* Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. *Science* **318**, 1449-1452 (2007).

80 Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006 (2002).

81 Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements. *Nucl. Acids Res.* **32**, D45-49 (2004).

82 National Center for Biotechnology Information. *Complete Genomes: Viruses* <http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=5&name=Viruses> (Accessed: 05 Feb 2010).

83 National Center for Biotechnology Information. *Complete Genomes: Phages* <http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=6&name=Phages> (Accessed: 05 Feb 2010).

84 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

85 Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

86 Friedrich, L. & Bettina, G. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software* **28** (2008).

87 Richard, A. B., John, M. C. & Allan, R. W. *The new S language: a programming environment for data analysis and graphics.* (Wadsworth and Brooks/Cole Advanced Books & Software, 1988).

88 Rajilic-Stojanovic, M. *et al.* 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* (2009).

89 Raes, J., Harrington, E. D., Singh, A. H. & Bork, P. Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol* **17**, 362-369 (2007).

90 Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Research* **19**, 2317-2323 (2009).

91 Arumugam, M. *SMASH User's Manual* <http://www.bork.embl.de/software/smash/> (Accessed: 25 Mar 2010).

92 Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).

93 Durbin, R., Haussler, D., Stein, L., Lewis, S. & Krogh, A. *GFF (General Feature Format) specifications document* <http://www.sanger.ac.uk/resources/software/gff/spec.html> (Accessed: 25 Mar 2010).

94 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2007).

95 Tjaden, B. *et al.* Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res* **30**, 3732-3738 (2002).

96 Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nat Biotechnol* **18**, 1262-1268 (2000).

97 Reppas, N. B., Wade, J. T., Church, G. M. & Struhl, K. The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. *Mol Cell* **24**, 747-757 (2006).

98 Nelson, C. M. *et al.* Whole genome transcription profiling of Anaplasma phagocytophilum in human and tick host cells by tiling array analysis. *BMC Genomics* **9**, 364 (2008).

99 Akama, T. *et al.* Whole-genome tiling array analysis of Mycobacterium leprae RNA reveals high expression of pseudogenes and noncoding regions. *J Bacteriol* **191**, 3321-3327 (2009).

100 McGrath, P. T. *et al.* High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol* **25**, 584-592 (2007).

101 Toledo-Arana, A. *et al.* The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950-956 (2009).

102 Vogel, J. & Wagner, E. G. Target identification of small noncoding RNAs in bacteria. *Curr Opin Microbiol* **10**, 262-270 (2007).

103 Materials and methods are available as supporting material on Science online. <http://www.sciencemag.org/cgi/content/full/sci;326/5957/1268/DC1> (Accessed: 25 Mar 2010).

104 Weiner, J., 3rd, Herrmann, R. & Browning, G. F. Transcription in Mycoplasma pneumoniae. *Nucleic Acids Res* **28**, 4488-4496 (2000).

105 Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037 (2009).

106 Wang, X. J., Gaasterland, T. & Chua, N. H. Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. *Genome Biol* **6**, R30 (2005).

107 Henz, S. R. *et al.* Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol* **144**, 1247-1255 (2007).

108 Ge, X., Wu, Q., Jung, Y. C., Chen, J. & Wang, S. M. A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics* **22**, 2475-2479 (2006).

109 Lapidot, M. & Pilpel, Y. Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep* **7**, 1216-1222 (2006).

110 Brantl, S. & Wagner, E. G. Antisense RNA-mediated transcriptional attenuation occurs faster than stable antisense/target RNA pairing: an in vitro study of plasmid pIP501. *EMBO J* **13**, 3599-3607 (1994).

111 Andre, G. *et al.* S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of Clostridium acetobutylicum. *Nucleic Acids Res* **36**, 5955-5969 (2008).

112 Brantl, S. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol* **10**, 102-109 (2007).

113 Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566 (2005).

114 Hooper, S. D. *et al.* Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Mol Syst Biol* **3**, 72 (2007).

115 Washio, T., Sasayama, J. & Tomita, M. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res* **26**, 5456-5463 (1998).

116 de Hoon, M. J., Makita, Y., Nakai, K. & Miyano, S. Prediction of transcriptional terminators in Bacillus subtilis and related species. *PLoS Comput Biol* **1**, e25 (2005).

117 Budin-Verneuila, A., Maguin, E., Auffraya, Y., Dusko Ehrlich, S. & Pichereaua, V. An essential role for arginine catabolism in the acid tolerance of Lactococcus lactis MG1363. *le Lait* **84**, 8 (2004).

118 Boue, S., Letunic, I. & Bork, P. Alternative splicing and evolution. *Bioessays* **25**, 1031-1034 (2003).

119 Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* **5**, 285 (2009).

120 Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326**, 1263-1268 (2009).
121 Chang, L. J., Chen, W. H., Minion, F. C. & Shiuan, D. Mycoplasmas regulate the expression of heat-shock protein genes through CIRCE-HrcA interactions. *Biochem Biophys Res Commun* **367**, 213-218 (2008).
122 Kuhner, S. *et al.* Proteome Organization in a Genome-Reduced Bacterium. *Science* **326**, 1235-1240 (2009).
123 Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**, 837-848 (2006).
124 Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307**, 1915-1920 (2005).
125 Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **102**, 11070-11075 (2005).
126 Zhang, H. *et al.* Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci U S A* **106**, 2365-2370 (2009).
127 Zoetendal, E. G., Akkermans, A. D. & De Vos, W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. *Appl Environ Microbiol* **64**, 3854-3859 (1998).
128 Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the Human Infant Intestinal Microbiota. *PLoS Biol* **5**, e177 (2007).
129 Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**, 525-552 (2004).
130 von Mering, C. *et al.* Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science* **315**, 1126-1130 (2007).
131 Suau, A. *et al.* Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65**, 4799-4807 (1999).
132 Supplementary information available on Nature online. <http://www.nature.com/nature/journal/v464/n7285/suppinfo/nature08821.html> (Accessed: 25 Mar 2010).
133 Colwell, R. K. EstimateS: Statistical estimation of species richness and shared species from samples. (2005).
134 Wang, X., Heazlewood, S. P., Krause, D. O. & Florin, T. H. Molecular characterization of the microbial species that colonize human ileal and colonic

mucosa by using 16S rDNA sequence analysis. *J Appl Microbiol* **95**, 508-520 (2003).

135 Kobayashi, K. *et al.* Essential Bacillus subtilis genes. *Proc Natl Acad Sci U S A* **100**, 4678-4683 (2003).

136 Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).

137 Letunic, I., Yamada, T., Kanehisa, M. & Bork, P. iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* **33**, 101-103 (2008).

138 von Mering, C. *et al.* STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35**, D358-362 (2007).

139 Cummings, J. H. & Macfarlane, G. T. The control and consequences of bacterial fermentation in the human colon. *J Appl Bacteriol* **70**, 443-459 (1991).

140 Dongowski, G., Lorenz, A. & Anger, H. Degradation of pectins with different degrees of esterification by Bacteroides thetaiotaomicron isolated from human gut flora. *Appl Environ Microbiol* **66**, 1321-1327 (2000).

141 Wong, J. M., de Souza, R., Kendall, C. W., Emam, A. & Jenkins, D. J. Colonic health: fermentation and short chain fatty acids. *J Clin Gastroenterol* **40**, 235-243 (2006).

142 Hamer, H. M. *et al.* Review article: the role of butyrate on colonic function. *Aliment Pharmacol Ther* **27**, 104-119 (2008).

143 Elango, R., Ball, R. O. & Pencharz, P. B. Amino acid requirements in humans: with a special emphasis on the metabolic availability of amino acids. *Amino Acids* **37**, 19-27 (2009).

144 Metges, C. C. Contribution of microbial amino acids to amino acid homeostasis of the host. *J Nutr* **130**, 1857S-1864S (2000).

145 Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environmental Microbiology* **11**, 2574-2584 (2009).

146 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

147 Saunier, K. & Dore, J. Gastrointestinal tract and the elderly: functional foods, gut microflora and healthy ageing. *Dig Liver Dis* **34 Suppl 2**, S19-24 (2002).

148 Hayashi, H., Sakamoto, M., Kitahara, M. & Benno, Y. Molecular analysis of fecal microbiota in elderly individuals using 16S rDNA library and T-RFLP. *Microbiol Immunol* **47**, 557-570 (2003).

149 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376 (2005).

150 Kleerebezem, M., Quadri, L. E., Kuipers, O. P. & de Vos, W. M. Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. *Mol Microbiol* **24**, 895-904 (1997).

151 Sonnenburg, E. D. *et al.* A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism. *Proceedings of the National Academy of Sciences* **103**, 8834-8839 (2006).

152 Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol Mol Biol Rev* **72**, 317-364, table of contents (2008).

153 Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. *PLoS Biol* **6**, e280 (2008).

154 Walker, A. Say hello to our little friends. *Nat Rev Micro* **5**, 572 (2007).

155 Gerdes, K., Christensen, S. K. & Lobner-Olesen, A. Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* **3**, 371-382 (2005).

156 Kankainen, M. *et al.* Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. *Proceedings of the National Academy of Sciences* **106**, 17193-17198 (2009).

157 Krogfelt, K. A. Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of Escherichia coli. *Rev Infect Dis* **13**, 721-735 (1991).

158 Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).

159 Vanhoutte, T., Huys, G., Brandt, E. d. & Swings, J. Temporal stability analysis of the microbiota in human feces by denaturing gradient gel electrophoresis using universal and group-specific 16S rRNA gene primers. *FEMS Microbiology Ecology* **48**, 437-446 (2004).

160 Tannock, G. W. *et al.* Analysis of the Fecal Microflora of Human Subjects Consuming a Probiotic Product Containing Lactobacillus rhamnosus DR20. *Appl. Environ. Microbiol.* **66**, 2578-2588 (2000).

161 Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326**, 1694-1697 (2009).

162 Salyers, A. A., West, S. E., Vercellotti, J. R. & Wilkins, T. D. Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Appl Environ Microbiol* **34**, 529-533 (1977).

163 Turnbaugh, P. J. & Gordon, J. I. An Invitation to the Marriage of Metagenomics and Metabolomics. **134**, 708 (2008).

164 Duncan, S. H. *et al.* Reduced Dietary Intake of Carbohydrates by Obese Subjects Results in Decreased Concentrations of Butyrate and Butyrate-Producing Bacteria in Feces. *Appl. Environ. Microbiol.* **73**, 1073-1078 (2007).

165 Loliger, J. Function and Importance of Glutamate for Savory Foods. *J. Nutr.* **130**, 915- (2000).

166 Kovacikova, G. & Skorupski, K. The alternative sigma factor sigma(E) plays an important role in intestinal survival and virulence in Vibrio cholerae. *Infect Immun* **70**, 5355-5362 (2002).

167 Fujihashi, K. & Kiyono, H. Mucosal immunosenescence: new developments and vaccines to control infectious diseases. *Trends Immunol* **30**, 334-343 (2009).

168 Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* **284**, 24673-24677 (2009).

169 Shipman, J. A., Berleman, J. E. & Salyers, A. A. Characterization of Four Outer Membrane Proteins Involved in Binding Starch to the Cell Surface of Bacteroides thetaiotaomicron. *J. Bacteriol.* **182**, 5365-5372 (2000).

170 Luton, P. E., Wayne, J. M., Sharp, R. J. & Riley, P. W. The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology* **148**, 3521-3530 (2002).

171 Colombel, J. F. *et al.* [Methanogenesis in man]. *Gastroenterol Clin Biol* **11**, 694-700 (1987).

172 Dubos, R. J. Pasteur's dilemma -- the road not taken. *ASM News* **40**, 703-709 (1974).

173 *IHMC: The International Human Microbiome Consortium* <http://www.human-microbiome.org> (Accessed: 25 Mar 2010).