

Metagenomic analysis of genetic variation in human gut microbial species

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



Vorgelegt von
Ana Cheng Zhu
aus
Lissabon, Portugal
Würzburg 2015

Eingereicht am:

Bürostempel

Mitglieder des Promotionskomitees:

Vorsitzende/r:

1. Betreuer: Dr. Habil. Peer Bork

2. Betreuer: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums:

Doktorurkunden ausgehändigt am:

Erklärung

I hereby declare that my thesis entitled “Metagenomic analysis of genetic variation in human gut microbial species” 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.

Ana Zhu

I would like to dedicate this thesis
to my family.



Acknowledgements

I would like to thank the following people for helping me throughout my PhD. Your contribution has helped me grow as a scientist and as a person. The completion of this dissertation has a special meaning for me as I went through a personal struggle. The hope and knowledge I have received during my PhD are invaluable and has helped me and my family to handle this period and therefore to all of you I am truly grateful.

First, my supervisor **Peer Bork**, for giving me the opportunity to do a PhD in his group. I would like to thank him for providing means of funding, and for the period I struggled the most I am thankful for your patience. He provided invaluable breath of scientific knowledge, critical thinking, guidance and shaped my view of being a scientist.

Thomas Dandekar, Lars Steinmetz and Kiran Patil for accompanying my progress through my PhD and helpful feedback. Also would like to thank **Thomas Dandekar** for helping me with all the bureaucratic issues related with the university enrollment, and for always promptly coming to Heidelberg for my TAC meetings.

Helke Hillebrand for providing invaluable moral support and for acknowledging my special situation and arranging funding for the last 4 month of the PhD. And to the **EMBL international PhD programme** for creating such a welcoming environment to work in.

Manimozhiyan Arumugam for guiding and supporting through the first part of my PhD. He provided in depth understanding of metagenomics and population genetics.

Shinichi Sunagawa, for guiding and supporting me through the second part of my PhD and showing me the significance of my work in times when I could

not understand it. Among other areas his guidance was crucial for forming my current understanding of prokaryotic species genomes and phenotypic impact.

Daniel Mende for all the scientific discussion especially about pan-genomes and data analysis. **Siegfried Schloissnig** for generating the SNP catalogue and discussion about population genetics. **Takuji Yamada** for all the knowledge about metabolic pathway and figure design.

Members of the Bork group, especially to the members involved in the “Genomic variation of the gut microbiome” which mainly involved **Siegfried Schloissnig, Manimozhayan Arumugam, Shinichi Sunagawa, Julien Tap** and **Alison Waller**. I learned a lot throughout the group discussions. **Yan Yuan** for technical support in handling and maintaining the servers. **Pablo Minguez** for all the help with programming issues.

Simone Li, Marja Driessen, Falk Hildebrand and **Jens Kultima** for helping me proofreading the thesis and **Falk Hildebrand** and **Martina Kluenemann** for translating the summary to German. **Amanda Di Giulio** for being my writing buddy.

Usual members of the coffee crew **Pablo Minguez, Kalliopi Trachana, Jaime Cepas, Luca Parca, Luz Garcia** and **Marc Sitges** for all our random conversations, the utility of the word “whatever” and the many words ending with “-ero/-era”. A special thanks also to **Anita Voigt**.

Nicole Prior and **Gary Male** for turning Heidelberg a place I can call home, with wonderful English Christmas and British humor that I now find so endearing. **Lionel Newton** for showing me father Ted. **Nicole family** to somehow always manage to be with me on my birthday and bring champagne with strawberries. **Nicole Prior, Simone Li, Gary Male, Joana Pinto, Sara Ferreira, Ana Catarina, Lionel Newton** thank you for being there for me without fear, I am truly grateful for having you guys in my life.

My grandfather **Zhu Zhi Hua** and my dad **Zhu Chang Long** for incentivating me to pursue a career in science. To all of my family, specially my brothers **Antonio Zhu, Mario Zhu**, dad, mom **Cheng A Lan**, grandma **Li Xiu De**, uncle **Zhu Chang Bin** and aunt **Xia Xiao Xiao** to teach me the meaning of courage.

Table of Contents

Erklärung	v
Abbreviations.....	18
Summary	20
Zusammenfassung	22
INTRODUCTION	26
1.1 Prokaryotic species in the human gut	26
1.2 Characterization of gut prokaryotic species variation to complement human population genetics	28
1.3 Genetic variation of gut prokaryotic species and its unexplored impact on human phenotype	30
1.4 Mechanistic understanding of genetic variation of gut prokaryote	32
1.5 Metagenomics for characterization of genetic variation in prokaryotic species	33
1.6 Outline.....	35
METHODS.....	38
2.1 Resource building	38
2.1.1 Source of faecal metagenomic samples.....	38
2.1.2 Generation of a non-redundant reference genome	39

2.1.3	Mapping of Illumina reads to non-redundant reference genome catalogue.....	39
2.1.4	Detection of species presence in a metagenome.....	40
2.1.5	Functional annotation of genes	41
2.2	Single nucleotide variation analysis in metagenomes	41
2.2.1	Prevalent and dominant species in our cohort.....	41
2.2.2	SNP calling.....	42
2.2.3	Measurement of species evolution (based on pN/pS ratio).....	43
2.2.4	Measurement of gene evolution (based on pN/pS ratio).....	43
2.2.5	Comparison between the evolution of two species (<i>Roseburia intestinalis</i> and <i>Eubacterium eligens</i>).....	44
2.3	Gene content based methods	45
2.3.1	Detection of species presence in a metagenome for gene content analysis.....	45
2.3.2	Determination of core and accessory genes	46
2.3.3	Estimation of accessory genes fraction	47
2.3.4	Comparison of gene content differences between pairs of individuals and pairs of reference genomes	49
2.3.5	Determination of accessory gene deletion blocks	50

2.3.6	Determination of paralog within and between reference genome	50
2.3.7	Detection of genomic islands.....	51
RESULTS AND DISCUSSION		52
3.1	SNP variation in gut prokaryotic species with phenotypic implications	52
3.1.1	Introduction.....	52
3.1.2	Faster SNP evolution of prokaryotic species within individuals rather than between individuals	53
3.1.3	Type IV secretion system is slow evolving and bile acid hydrolase is fast evolving in the human gut prokaryotic species	58
3.1.4	<i>galK</i> gene evolution is uncoupled from prokaryotic species evolution.....	59
3.1.5	Discussion.....	60
3.2	Inter-individual variation in gene content of human gut bacterial species	63
3.2.1	Introduction.....	63
3.2.2	Fraction of accessory genes increases with genome size.....	66
3.2.3	Gut strains of the same species have large inter-individual variation in gene content	70

3.2.4	Accessory genes are enriched in mobile elements and functions associated with cell wall and membrane	75
3.2.5	Single gene deletions are highly abundant and associated with mobile elements.....	76
3.2.6	Accessory genes have functions that imply phenotypic differences of an individual.....	77
3.2.7	Discussion.....	81
	CONCLUSIONS	84
	SUPPLEMENTARY TABLES	86
	SUPPLEMENTARY FIGURES.....	93
	BIBLIOGRAPHY.....	98

List of Tables

Table	Page
Table 1. Definitions used in the scope of sub-chapter 3.2	66
Table 2: Information regarding the 11 species references genomes and the metagenomes used in the current study.....	67

List of Figures

Figure	Page
Figure 1: Procedure used for selection of metagenomes and species for SNP-based methods	42
Figure 2: Procedure used for selection of metagenomes and species used in gene content analysis	46
Figure 3: Diagram illustrating gene coverage of core and accessory genes of one species (<i>Dialister invisus</i>) for 10 individual metagenomes.	47
Figure 4: Statistics of genomic variation across 101 gut microbial species prevalent in 252 metagenomes from 207 individuals.....	55
Figure 5: pN/pS ratio distribution across 66 dominant species and across 207 individuals.....	58
Figure 6: Comparison between <i>Eubacterium eligens</i> and <i>Roseburia intestinalis</i>	59
Figure 7: Estimation of the fraction of accessory genes (%) based on exponential model	68
Figure 8: Display of the percentage of accessory genes for 11 gut bacterial species.	70
Figure 9: Variability between pairs of metagenomes	72
Figure 10: Variability between (1) sequenced reference genomes, (2) Bacteroidetes and Firmicutes reference genomes and (3) metagenomes.....	73

Figure 11: Variability between sequenced reference genomes and metagenomes for <i>Parabacteroides distasonis</i>	74
Figure 12: Difference plot of orthologous group's functional categories between gene number of core and accessory genes	76
Figure 13: Gene deletion block size distribution	78
Figure 14: Cumulative number of genes in deletion blocks of a given size.	79

Abbreviations

COG = Cluster of Orthologous Group;

IBD = Inflammatory Bowel Disease;

NOG = Non-supervised Orthologous Group;

OG = orthologous groups

HMP = Human Microbiome Project;

HGT = Horizontal gene transfer;

MetaHIT = Metagenomics of the Human Intestinal Tract;

PUL = Polysaccharide utilization loci;

CPS = Capsular polysaccharide synthesis loci;

Summary

Microbial species (bacteria and archaea) in the gut are important for human health in various ways. Not only does the species composition vary considerably within the human population, but each individual also appears to have its own strains of a given species. While it is known from studies of bacterial pan-genomes, that genetic variation between strains can differ considerably, such as in *Escherichia coli*, the extent of genetic variation of strains for abundant gut species has not been surveyed in a natural habitat. This is mainly due to the fact that most of these species cannot be cultured in the laboratory. Genetic variation can range from microscale genomic rearrangements such as small nucleotide polymorphism (SNP) to macroscale large genomic rearrangements like structural variations. Metagenomics offers an alternative solution to study genetic variation in prokaryotes, as it involves DNA sequencing of the whole community directly from the environment. However, most metagenomic studies to date only focus on variation in gene abundance and hence are not able to characterize genetic variation (in terms of presence or absence of SNPs and genes) of gut microbial strains of individuals.

The aim of my doctorate studies was therefore to study the extent of genetic variation in the genomic sequence of gut prokaryotic species and its phenotypic effects based on: (1) the impact of SNP variation in gut bacterial species, by focusing on genes under selective pressure and (2) the gene content variation (as a proxy for structural variation) and their effect on microbial species and the phenotypic traits of their human host.

In the first part of my doctorate studies, I was involved in a project in which we created a catalogue of 10.3 million SNPs in gut prokaryotic species, based on metagenomes. I used this to perform the first SNP-based comparative study of prokaryotic species evolution in a natural habitat. Here, I found that strains of gut microbial species in different individuals evolve at more similar rates than the

strains within an individual. In addition, I found that gene evolution can be uncoupled from the evolution of its originating species, and that this could be related to selective pressure such as diet, exemplified by galactokinase gene (*galK*). Despite the individuality (i.e. uniqueness of each individual within the studied metagenomic dataset) in the SNP profile of the gut microbiota that we found, for most cases it is not possible to link SNPs with phenotypic differences. For this reason I also used gene content as a proxy to study structural variation in metagenomes.

In the second part of my doctorate studies, I developed a methodology to characterize the variability of gene content in gut bacterial species, using metagenomes. My approach is based on gene deletions, and was applied to abundant species (demonstrated using a set of 11 species). The method is sufficiently robust as it captures a similar range of gene content variability as has been detected in completely sequenced genomes. Using this procedure I found individuals differ by an average of 13% in their gene content of gut bacterial strains within the same species. Interestingly no two individuals shared the same gene content across bacterial species. However, this variation corresponds to a lower limit, as it only accounts for gene deletion and not insertions. This large variation in the gene content of gut strain was found to affect important functions, such as polysaccharide utilization loci (PULs) and capsular polysaccharide synthesis (CPS), which are related with digestion of dietary fibers.

In summary, I have shown that metagenomics based approaches can be robust in characterizing genetic variation in gut bacterial species. I also illustrated, using examples both for SNPs and gene content (*galK*, PULs and CPS), that this genetic variation can be used to predict the phenotypic characteristics of the microbial species, as well as predicting the phenotype of their human host (for example, their capacity to digest different food components). Overall, the results of my thesis highlight the importance of characterizing the strains in the gut microbiome analogous to the emerging variability and importance of human genomics.

Zusammenfassung

Mikrobielle Arten (Bakterien und Archaeen) im menschlichen Darm sind wichtige Begleiter für unsere Gesundheit. Jedoch gibt es nicht nur starke Unterschiede zwischen individuellen Wirten in der Artenzusammensetzung des Darmmikrobioms, sondern es scheint sogar Individuen-spezifische Bakterienstämme zu geben. Analysen von Bakterien wie z.B. *Escherichia coli* haben schon früh gezeigt, dass die Genome von Bakterienstämmen derselben Art große Unterschiede aufzeigen können; jedoch wurden diese Unterschiede bisher noch nicht in einer natürlichen Umgebung gezeigt. Genetische Variation kann viele Ausprägungen haben und reicht von kleinen Veränderungen wie „small nucleotide polymorphism“ (SNP) zu makroskopischen Veränderung, wie z.B. chromosomalen Restrukturierungen. All diese genetischen Variationen wurden bis jetzt nicht in der natürlichen Umgebung der Bakterien studiert, vorallem bedingt durch fehlende Methoden um die meisten dieser Bakterien im Labor zu kultivieren. Metagenomische Studien können hier helfen, da sie unabhängig von Kultivierungen jegliche DNS aus einer natürlichen Bakteriengemeinschaft untersuchen. Jedoch wurde dies in den meisten bisher veröffentlichten metagenomischen Studien nicht ausgenutzt da diese hauptsächlich auf die Anzahl der gefunden Gene ausgerichtet waren.

Das Ziel meiner Doktorarbeit war es, die genetische Variation in Darmbakterien zu beschreiben und phenotypische Veränderungen zu untersuchen. Dies habe ich umgesetzt durch die Erforschung (1) der SNP-Varianz in Darmbakterien, mit besonderem Augenmerk auf Gene, die unter einem selektivem Druck stehen und (2) der Variationen in der Genzusammensetzung eines Genomes (als eine Annäherung an strukturelle Variationen) und welchen Effekt dies auf Mikrobenarten und Wirtsphenotypen hat.

Im ersten Kapitel meiner Doktorarbeit beschreibe ich meine Arbeit in einem Projekt unserer Gruppe, in dem wir basierend auf metagenomischen Daten 10

Millionen SNPs in menschlichen Darmbakterien beschrieben haben. Diesen Datensatz habe ich verwendet um die erste SNP-basierte, vergleichende Studie der Bakterienevolution in einem natürlichen Habitat zu realisieren. Ich entdeckte, dass Bakterienstämme unabhängig vom Wirt ähnliche evolutionäre Raten haben. Genauer gesagt, die evolutionäre Rate für eine Art ist stabiler zwischen Wirten, als die von verschiedenen Spezies innerhalb eines Wirtes. Ausserdem fand ich heraus, dass die Evolution von einzelnen Genen unabhängig vom restlichen Genom einer Spezies ist. Dies könnte durch einen Selektionsdruck wie z.B. die Ernährung des Wirtes ausgelöst werden, was ich am Beispiel des Galactokinasegenes (*galK*) gezeigt habe. Obwohl wir zeigen konnten, dass das SNP-Profil der Darmbakterien spezifisch für den jeweiligen Wirt ist, konnten wir keine Assoziation zwischen SNPs und Wirtsphänotypen finden. Auch aus diesem Grund habe ich mich in meiner weiteren Arbeit verstärkt auf makroskopische Genomvariationen konzentriert.

Im zweiten Teil meiner Doktorarbeit entwickelte ich eine neue Methode, um Variationen in der genomische Zusammensetzung von einzelnen Bakterienarten zu beschreiben, wieder basierend auf metagenomischen Daten. Hierbei fokussiere ich mich insbesondere auf Gene, die in unseren metagenomischen Daten im Vergleich zum Referengenom fehlen und wende dies auf die 11 dominantesten Bakterienspezies an. Diese neue Methode ist robust, da die gefundene Genomvarianz in unseren metagenomischen Daten übereinstimmt mit Daten aus komplett sequenzierten Genomen. So konnte ich herausfinden, dass im Durchschnitt 13% der Gene einer Bakterienart zwischen einzelnen Wirten variieren. Besonders interessant ist hier, dass wir keine zwei Wirte gefunden haben, die für eine Bakterienart genau diesselben Gene haben. Jedoch ist die erwartete Varianz aller Wahrscheinlichkeit nach noch größer, da ich mit dieser Methode nur fehlende Gene beschreiben kann, aber nicht neu hinzugekommene. Diese Varianz kann auch wichtige bakterielle Funktionen betreffen, z.B. Gene für „polysaccharide utilization loci“ (PULs) und „capsular polysaccharide synthesis“ (CPS), welche wichtig sind um Ballaststoffe in der Nahrung zu verwerten.

Zusammenfassend konnte ich in dieser Arbeit zeigen, dass metagenomische Methoden robust genug sind um die genetische Varianz von Darmbakterien zu beschreiben. Ausserdem konnte ich zeigen, dass die beschriebene Varianz benutzt werden kann, um phenotypische Veränderungen von Bakterien vorherzusagen (demonstriert für die *galK*, PULs and CPS-Gene). Dies wiederum könnte benutzt werden um Vorhersagen für den Wirt über z.B. seine Ernährung zu machen. Meine Doktorarbeit zeigt wie wichtig es ist, einzelne Bakterienstämme zu charakterisieren, ganz analog zu der Bedeutsamkeit der genetischen Varianz des menschlichen Genomes.

CHAPTER 1

INTRODUCTION

1.1 Prokaryotic species in the human gut

The human gastrointestinal tract houses a complex community of microbial species (bacteria and archaea) that are important for human health, termed the human microbiota. The complexity of this community is one of the major challenges when characterizing the genetic variation within gut microbial species and therefore this paragraph is dedicated to illustrate its complexity. In terms of dimension the gut microbiota weighs up to 1.5kg [1] and is composed by 3-fold [2] to 10-fold [3] more cells than the number of human cells. The gut community is mostly composed of species from two phyla, Bacteroidetes and Firmicutes, constituting more than 90% [4–7]. Besides these two phyla, prokaryotic species belonging to Actinobacteria, Proteobacteria and Verrucomicrobia phyla are also found in minority [4, 8]. In contrast to phyla, both gut microbial species composition and abundance have a large variability among individuals, ranging between 10- and 10.000-fold [9]. Recently it was found that human individuals mainly stratify into three clusters that were named “enterotypes” [8]. These enterotypes were mostly driven by species composition, with *Prevotella*, *Bacteroides* and *Ruminococcus* genus being predominant in each enterotype. Two of these enterotypes have been related with long-term dietary habits [10] suggesting that taxonomical composition might be linked with individuals diet. Independently of the variation in taxonomic composition, the number of species found per human individual is rather large, highlighting the complexity of this community, estimated with different technologies and varied between 101 ± 21 species, which is 16S rRNA based, and 160 species, which is metagenomic based (Faith et al. 2013; Qin et al. 2010).

One of the key roles of the gut microbiota is to help our gut to obtain nutritional value from our food [12], without the help of the microbiota the gut would not be able to assimilate nutritional value from a substantial fraction of dietary components as demonstrated in a germ-free mice study (mice without a microbiota). The study found that germ-free mice required 30% more caloric intake to reach the same weight as normal mice [12]. The composition of the gut microbiota is also relevant [13–15]. In addition, the relation between gut microbiota and diet is dynamic that is dietary changes have been associated with changes in the composition of the gut microbiota [16, 17]. To understand the importance of the gut microbiota, one can view these communities as a natural and stable bioreactor where they break down several indigestible food components (indigestible by human enzymes), and the components that they do not digest are excreted in the faeces [18]. Food components, which only the microbiota can degrade include the majority of complex carbohydrates and plant polysaccharides [19], present in vegetables, fruits, cereals and leguminous seeds [18]. Examples include, plant cell wall polysaccharides, lignin, resistant starch and inulin [20]. Human genome do not encode these enzymes because of the wide structural diversity of these dietary fibres, which would require a large repertoire of catalytic enzymes with different specificities [21]. Instead the prokaryotic gene reservoir (which is 20-fold higher than the number of human genes, and also larger in number of carbohydrate enzymes [22–24]), can provide these functions. For example, *Bacteroides cellulosilyticus*, has 56 carbohydrate active enzymes that are not seen in the human genome, highlighting the increase in metabolic capacity given by only a single species [25].

The vastness of the prokaryotic gene reservoir (in terms of the total number of prokaryotic genes, which is a proxy for prokaryote richness) has also been linked to differences in individuals both dietary habits and health conditions. Higher fibre consumption, in the form of fruit and vegetables, appear to be associated with a higher number of prokaryotic genes [26]. Whereas, low prokaryotic gene number and thus decrease prokaryotic richness was found to be associated with more

pronounced disease and inflammatory phenotype [27]. Decreased prokaryotic richness has also been associated with inflammatory bowel disease [9, 28, 29] and elderly populations [30]. Also recently it has been shown that the structure of the gut microbiota is related with long-term dietary habits [10, 16, 17]. Overall these studies highlight the importance of understanding the relation between long-term diet and microbial gene reservoir variation.

Besides extracting nutritional value, the microbiota plays other important roles, such as the synthesis of vitamin K and B [31, 32], production of short-chain fatty acids (SCFAs) such as acetate, butyrate and propionate, which are the main end-product of bacterial fermentation. Also 60-70% of the energy content of indigested carbohydrates are stored in the form of SCFA after digestion [33]. SCFA are important in stimulating the intestinal blood flow [34, 35], affect epithelial proliferation and differentiation [36–38] and have anti-inflammatory properties [39–41]. Furthermore, the gut community forms a layer around the mucosa layer, creating a barrier against the colonization of foreign pathogenic strains through competitive exclusion [42]. The gut microbiota has also a role in the development of an individual immune system, for example by helping in the development of lymphoid structures and epithelial functions [43, 44]. For a more detail view of gut microbiota effect on human health please have a look into Hooper *et al.* 2003 and 2012 [45, 46]. Due to its importance the human gut microbiota has been loosely termed “another human organ”, and therefore it is important to study the impact of microbiota on human health.

1.2 Characterization of gut prokaryotic species variation to complement human population genetics

Genetic variations, such as SNPs and structural variations, across a genome have been extensively characterized in large human populations. The publication of the first human genome in 2004 was a milestone in human genetics [47, 48]. Created using Sanger sequencing technology, the project cost 3 billion dollars and took 15 years to complete. The sequencing rate was the main limiting

step and was associated with both monetary and time constraints. Subsequently these limitations motivated the development of high-throughput sequencing technologies, such as sequencers produced by Illumina (Solexa) company. The project provided a reference genome that would serve as a guide for future sequencing projects, paving the way towards the usage of whole genomes in human population genetic studies. For example, two large-scale international landmark projects were the HapMap [49–51] and the 1000 Genome Project [52, 53] with the goals of characterizing the genetic variation of human population and contribution of genetic variation to human health.

The HapMap project and the 1000 Genome Project aimed at cataloguing genetic variation in terms of SNPs (both common and rare) and structural variations. The HapMap project was mainly focused on SNP discovery, as it aimed at identifying haplotype blocks of common SNPs (that are present in 5% or greater allele frequency). The outcome of the project was the identification of more than 8 million common SNPs, constituting an important catalogue of variation across different ethnic backgrounds [49–51]. Soon we realised that individuals differences were beyond single nucleotide differences and included a panoply of small and large variations, such as insertions and deletions (indels) and large structural variations [54, 55]. Therefore, the 1000 Genome Project was created with the objective of characterizing both common and rare SNPs and structural variations. They catalogued the haplotype map of 38 million SNPs, 1.4 million short insertions and deletions and 14.000 large deletions across 1.092 individuals from 14 populations (drawn from European, East Asia, sub-Saharan Africa and the Americas) [52, 53]. Together both projects revealed the extent of and formulated conceptual aspects of genomic variation observed across the human population. They also lead to some clinical applications such as diagnostic testing of hereditary disorders (as reviewed in Rehm *et al.* 2013 [56]). As gut prokaryotic species are important for human health, characterizing the genomic variability of these species in a similar fashion as has been done for the human genome may largely improve our understanding of their effect on human phenotypic traits.

In contrary to variations in the human genome, such variations had not yet been characterized prior to the commencement of my PhD studies. Most human associated microbiota studies focus on either the study of changes in the relative abundance of either microbial species or gene repertoire, whereas the genetic variation of each species is not yet characterized. Similar to human population genetics, characterization of prokaryotic species genetic variation can also be based on reference genomes. Reference genomes for a large number of gut prokaryotic species were previously unavailable in public domains prior to 2010, this reality changed with the first catalogue of reference genomes associated with the human microbiome which was later expanded to 800 reference genomes [57, 58], among other resources that became available [9]. These constituted the necessary resources for us to move towards characterization of the genetic variation (SNP's and structural variation which gene content is used as a proxy) of species within the gut microbiome.

1.3 Genetic variation of gut prokaryotic species and its unexplored impact on human phenotype

Genetic variation of gut prokaryote community members can change not only the species' phenotypes but also lead to changes in their human host phenotypes and health conditions. Examples of gut associated human phenotypes include the capacity to diggest different food sources and drugs. Determination of the exact phenotypic impact on each gut microbiota member is still impossible due to the infinitesimal conditions that would need to be tested to explore the complete phenotypic landscape of a given genome. A small number of studies for a few species have linked genetic variation to certain aspects of the bacterial phenotype (reviewed by Read *et al.* 2014 [59]). Initial screening of genetic variation can provide an estimation of potential prokaryote and human phenotypes.

Mechanistic demonstration of links between key gut community members and human phenotypic variability is hard to establish. This is because the degree of variability between individual's strain is unknown and therefore it is hard to

establish what constitute the key members. While the prokaryotic species composition vary greatly (in the thousands range) between human individuals [8, 23], the number of genes of the prokaryote community can vary even more (in the millions range) [23]. From the total of 10 million microbial genes identified in faecal samples from 1.267 individuals only 300.000 are shared by at least 50% of the individuals [23]. Note that these number of gut microbial genes per individual are calculated for the whole gut microbiota independently of the species and hence the actual number of genes in each individual gut strains are not known. The diversification of the gene pool can be a result of three different scenarios: 1) by difference in the composition of the gut community members, 2) by differences in the relative abundance of the gut community members or 3) by changes to the genomes of gut community members through mutation or horizontal gene transfer (HGT) [60]. Whereas (1) and (2) have been studied, (3) is still unknown. Therefore, it is important to investigate whether the variability in gene content and other genomic traits, such as SNP' variation, is mainly due to changes in taxonomical composition, or reflects differences in strain composition for the same species, if it is the latter strain variation needs to be characterized.

Recent pan-genome studies [61–65] have revealed that bacterial strains within a species varied greatly in their gene content, such as the gut bacterial species *Escherichia coli*, indicating that their genomes are highly dynamic. These studies focused on the entire gene repertoire of a given species, where bacterial strains of the species are compared. Such pan-genome studies typically involves the categorization of genes into core genes that are shared among strains and accessory genes that are present in only some strains. Accessory genes can even be specific to only a single strain, for example in *Haemophilus influenza*, 19% of their genes were unique genes, exemplifying the extent of strain-specific variability. Accessory genes can encode antibiotic resistance, virulence factors or other loci that contribute to the adaptation of the organism to the environment [62]. In terms of the human gut, variation of specific genes in the microbiota has also been associated with differences in individuals' capacity to degrade fibres (for example

cellulose and algae) [21, 66–68]. The algae example (porphyran enzyme) was only found in Japanese but not in American individuals. This enzyme is an interesting case as it was likely acquired by HGT from marine bacterial species associated with the algae. This example suggests that dietary changes might introduce novel genes to the resident gut prokaryote strains from non-sterile food related bacteria and hence might change the original genetic composition of resident strains.

Besides gene content variation, SNPs are another type of genetic variation that can also have a phenotypic impact, even when the variation is due to a few mutations. For example, a single nucleotide exchange in *E. coli* is sufficient to cause differences in lipopolysaccharide phenotype and serum sensitivity [69] also three point mutations in two genes are sufficient to confer clinically relevant antibiotic resistance [70]. Variation in terms of SNPs can explain phenotypic variability among strains for any species [71] but it is especially important in distinguishing strains in clonal species where recombination rate is low, such as in *Mycobacterium tuberculosis* [72]. In 2013, the first bacterial GWAS based on SNP profiling investigated host adaptation in 192 *Campylobacter jejuni* and *C. coli* [73]. In the same year GWAS study for *Mycobacterium tuberculosis* found mutations in genes involved in resistance to anti-tuberculosis drugs which were detected in resistant lineages of *M. tuberculosis* [71]. These whole-genome characterization of SNP variation among strains (referred here as SNP profiling) have revealed that SNPs variation was associated with certain phenotypic traits in genome-wide association studies (GWAS), such as host adaptation, virulence and antibiotic resistance [71, 73–76].

1.4 Mechanistic understanding of genetic variation of gut prokaryote

Understanding the mechanism that lead to genetic variation are important as they determine the frequency of changes in genetic variation and the architecture of prokaryote genomes. Genetic variation in prokaryotic species originate from different driving forces than in human where they are mainly a result of gene duplication and mutation. In prokaryotes, genetic variation are mainly a

result of HGT and gene deletion except for highly clonal species, where mutation is the main driving force [59]. HGT can occur through transformation, conjugation or transduction [77]. Transformation is a process where extracellular DNA is stably uptaken and integrated into the genome. For natural transformation to happen bacterial cells need to develop a regulated physiological state that involves between 20 and 50 proteins and is time-limited in response to an environmental stimulus called natural competence [77]. Conjugation is a transfer process that involves the direct contact between two cells, process mediated by cell-to-cell junctions and a pore through which the DNA is transferred [78]. Transduction is a gene transfer mechanism that is mediated by certain types of bacteriophages. A HGT event is usually divided into three steps excision (if the genes are present in the host chromosome), transfer through conjugation, transduction or transformation, and integration through homologous recombination, although the first and second step are not mandatory [77]. Large part of the gene content variation is associated with mobile genetic elements (MGE) [77, 79]. Mobile genetic elements are segments of DNA that encode proteins used to build the machinery that mediates the transfer of DNA within genomes (intracellular mobility) or between bacterial cells (intercellular mobility) [80]. The classical MGE include plasmids, bacteriophages and transposons. The list now includes more ancient MGE that have underwent certain erosion throughout the history of the strain [81, 82]. These includes various genomic islands (syntenic blocks of accessory genes acquired by HGT detected through strain genome comparison), large megaplasmids and other elements with less defined structure and function (e.g. even restriction-modification systems have been referred as part of MGEs).

1.5 Metagenomics for characterization of genetic variation in prokaryotic species

Population genomic studies of a complex community such as the human gut have been partly impossible due to technological limitations mainly related with isolation. Isolation is a key step in the traditional whole genome sequencing based approach. First, 56%±4% of the gut species are still not cultivable and the value

can even be lower depending on the phyla the species belongs to, for Firmicutes 79% are not cultivable [83]. Also, culture conditions do not represent well the situation bacteria live in their natural environment, constituting a simplified version of the environment (where environmental stresses such as antibiotic therapy and immune system effect are not emulated) [64]. Furthermore, culture conditions can introduce changes to the bacterial genome [84]. Finally, both isolating and whole-genome sequencing a complex community such as the human gut have time and monetary cost limitations that do not enable to scale the method across a large human population. Instead, there is a pre-selection of the strains to be studied. For example, in a recent study of *Staphylococcus epidermidis* only 9% out of 800 strains were chosen based on morphology and sequenced [85].

As an alternative, metagenomics offers a culture-independent technique that enables the study of complex communities such as the human gut *in nature* conditions, meaning set in natural environment where the community is under natural selection. Metagenomics involves extracting the DNA directly from environmental samples without requiring isolation. The field of gut metagenomics has progressed dramatically in the last few years, and several metagenomic studies have associated changes in the gut microbiota with several diseases, both related with the GI tract [26, 27, 86–89], but also not directly related with the GI tract, such as atherosclerosis [90]. However, most of these metagenomic studies either focus on taxonomical differences, or in terms of genetic content they only study variation in relative abundances [91]. In either cases the link between strain genotypes and their respective phenotype is hard to establish, as discussed in **sub-chapter 1.3**. Instead characterization of gene content in terms of gene presence or absence and SNP's profiling can provide a better estimate of phenotype.

Initial population genetic studies using metagenomics have been done in simpler systems (in terms of number of dominant strain's) such as acid mine drainage [92, 93] and infant gut microbiota [94, 95], but not for complex communities, such as the adult human gut. The acid mine drainage system

described is mainly dominated by one dominant bacterium (*Leptospirillum* group II), and sequence variation between strains were reconstructed across a time-scale of 9 years. The infant gut microbiota, which was tracked over short time-frame (one month) was mostly composed of 6 genomes comprising 96% of the sequenced reads.

At the start of my PhD studies, two large gut metagenomic datasets had recently been published from two consortiums the European Metagenomics of the Human Intestinal Tract consortium (MetaHIT) [9] and the NIH Human Microbiome Project (HMP) [58], which included 139 and 124 faecal metagenomes, respectively. The MetaHIT project was a large and collaborative consortium effort with the central objective to establish associations between the genes of the human gut microbiota and the health of the sampled individuals. Participants were either healthy, or had either obesity and/or inflammatory bowel disease. The the HMP project aimed at characterizing the gut microbiota from the human gut, but also from other body sites, in healthy individuals [96]. These dataset had an unprecedented coverage (in total the 252 samples amounted to 1.56 terabases) and were ideal datasets to characterize the genetic variation of gut prokaryotic species in terms of SNPs and gene content.

1.6 Outline

During my PhD I participated in a project to characterize the genomic variation of gut microbiota, in the project we characterized the SNP catalogue of 101 gut prokaryotic species, the catalogue contained 10.3 million SNPs across 252 faecal metagenomes from 207 individuals. The size of the catalogue is close to the SNP catalogue created for human genomes which was based on 179 individuals (14.4 million) [52]. My role in the project was to find an approach to estimate the potential phenotypic impact of strains SNP variation, as the phenotype of individual's bacterial strains is not known. For this purpose I measured the bacterial gene and genome evolution based on SNPs using a measure akin to the

classical dN/dS ratio but applicable to haplotype independent SNP data, pN/pS ratio, the results of this project can be found in **sub-chapter 3.1**.

Estimation of phenotypic outcome is still impossible for the majority of SNPs, instead estimation of bacterial strain's gene content provides a direct prediction. For this purpose I developed a methodology applicable to abundant bacterial species for determination of genes presence or absence in a metagenome. The method provides a clear estimation of phenotypic outcome in comparison to typical functional studies in metagenomes which are based on relative abundance and where this is not possible.

CHAPTER 2

METHODS

2.1 Resource building

The generation of a reference genome set and the mapping of the reads to the reference genomes for both SNP profiling and gene content projects were done by several members of the Bork group (Jens Kultima, Shinichi Sunagawa and Siegfried Schloissnig). The generation of a set of representative genomes for a given species was critical to provide a consistent taxonomical definition of species across the tree-of-life, further details about the motivation for this classification can be found in mOTU or specl paper [97].

2.1.1 Source of faecal metagenomic samples

A set of illumina reads from 252 out of the initial 266 published faecal metagenomes (124 from the European MetaHIT Project [9] and 142 from the US HMP [57]) were used in my doctoral studies, **Appendix 1**. 14 metagenomes were removed after quality control, for more details see Schloissnig *et al.* [98].

The metagenomes from the MetaHIT Project originated from danish individuals and spanish individuals. The danish individuals (part of a obesity study) were composed of healthy controls and patients with obese/diabetes, whereas the spanish (part of an inflammatory bowel disease or IBD study) were composed of healthy controls and patients with ulcerative colitis or Crohn's disease in clinical remission. Both controls and patients collected the faecal samples and were asked to freeze the samples in their home freezer after collection. The samples were delivered to the Hospital using insulating polystyrene foam containers, and stored at -80% until analysis. The DNA was isolated as previously described [28].

The faecal samples from HMP studies originated from healthy individuals and were collected by the individual and stored in a Styrofoam box surrounded by

frozen gel packs for a maximum periods of 24h until the samples were delivered to the clinical laboratory and DNA was isolated using the commercial kit MoBio PowerSoil TM [57]. From the total of 51 individuals, 41 of those were sampled twice and 2 individuals sampled three times.

2.1.2 Generation of a non-redundant reference genome

The number of sequenced genomes for a prokaryotic species vary considerably, and this variation introduce biases in the quantity of genetic variation that can be detected in metagenomic samples. For example, reads belonging to a given species are more likely to find a perfectly matched sequence in species with more genomes available and hence would capture less polymorphic sites (SNPs). To take into consideration these biases, we generated a set of non-redundant reference genomes as follows.

A set of 1.511 prokaryote genomes were downloaded from GenBank and the MetaHIT Consortium (<http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/>). A set of 40 universal single copy marker genes was identified for each genome using 40 HMM profiles (one HMM profile for the orthologous group in eggNOG for each marker gene). For each marker gene, pairwise comparison between all genomes were calculated and the median identity between all marker genes was used as an approximation for average nucleotide identity (ANI) between two genomes. An operational cut-off of more than 95% ANI was used for species identification as recommended [99], and genomes were clustered into 929 clusters using complete linkage, constituting the non-redundant reference genome set used in the following analysis. Mapping of Illumina reads to reference genomes, **Appendix 2**.

2.1.3 Mapping of Illumina reads to non-redundant reference genome catalogue

Illumina reads from 266 fecal metagenomes (124 from the European MetaHIT study [9], 139 from the US Human Microbiome Project [57], 3 obtained

from Washington University [100]) were quality controlled by applying a customized trimming and filtering pipeline (described in [8] in section 5.2 of Schloissnig *et al.* 2013) with minor modifications, **Appendix 3**. In short, (1) bases of reads 5' end were trimmed when the number of base calls for any base (A, T, G and C) was within the average across all cycles plus/minus two standard deviations, (2) bases of reads 3' end were trimmed if the quality score was below 20 and (3) both reads shorter than 45 bp and reads with a median quality score below 20 were removed from further analysis. 252 samples passed this quality control step and were used throughout the study. With the intent of choosing a reference genome that can represent each cluster (defined above), high quality reads obtained from a subset of metagenomes (TS1, TS4, TS25, MH0006 and MH0012) were mapped to the 1,511 genomes using an alignment identity cutoff of 85% in the Mosaik program (version 1.1.0021; <http://bioinformatics.bc.edu/marthlab/Mosaik>) with the options “-a all -m all -hs 15 -mmp 0.85 -mmal -minp 0.9 -mhp 100 -act 20”. Next, for each genome included in the cluster based on the ANI criteria (see above), the genome possessing the highest read coverage was selected resulting in a set of 929 reference genomes, each likely representing a unique species. Finally, all metagenomic dataset (252 samples) were mapped to these 929 reference genomes with an alignment identity cutoff of 95% applying the same options as above, except for using “-mmp 0.95” instead of “-mmp 0.85” using Mosaik. The relative abundance of each genome within a sample was calculated by counting the number of reads mapped to a reference genome, and this number is normalized by the genome size.

2.1.4 Detection of species presence in a metagenome

Reads can map to a species without the species actually existing within the sample, for example if: (1) the reads map to genes that originally belonged to this species but have been acquired by another species through HGT or (2) reads map to highly conserved genes. When two species are closely related highly conserved genes can be more than 95% identical.

To ensure that the species is actually present within the sample we required that reads mapped to at least 40% of the nucleotide positions of the representative reference genome. The 40% was chosen as a proxy for the lower bound of shared gene content for an average bacterial species based on an *E. coli* study, where they used a wide range of strains and measure the minimum fraction of genes shared between two strains.

2.1.5 Functional annotation of genes

Genes from each species were mapped to orthologous groups (OGs), which were composed of both Cluster of Orthologous Groups (COGs) and Non-supervised Orthologous Groups (NOGs) from the eggNOG v3.0 pipeline [105] by using Blastp [106] and a bit-score higher than 60. The orthologous groups were further clustered in their COG functional categories [107]. Enrichment of COG functional categories were done using fisher test and multiple testing was adjusted with FDR. Genes in **sub-chapter 3.2** were also annotated with KEGG v62 [108] and MEROPs [109] by using Blastp [106] and a bit-score also higher than 60.

2.2 Single nucleotide variation analysis in metagenomes

In this **sub-chapter 2.2** I describe a metagenomic-based methodology to characterize the evolution of bacterial species and genes that is based on SNP profile. The method used pN/pS ratio since this method does not require the reconstruction of haplotypes.

2.2.1 Prevalent and dominant species in our cohort

A species was selected if it (1) was detected in at least one sample (according to the detection criteria in **section 2.1.3**), (2) accumulated a depth of genome coverage of at least 10x when pooled over all samples. The second filter is used to avoid rare and transient species that are present in a small number of individuals. In total there were 101 species out of the original 929 species (non-redundant genomes) that fulfill the criteria (1) and (2), and these species were

named “prevalent”. From the 101 prevalent species selected we found that 99% of the 7.4 billion reads mapped to only 66 species (**Figure 1**), suggesting that these are the “dominant” species in our cohort **Appendix 2**.

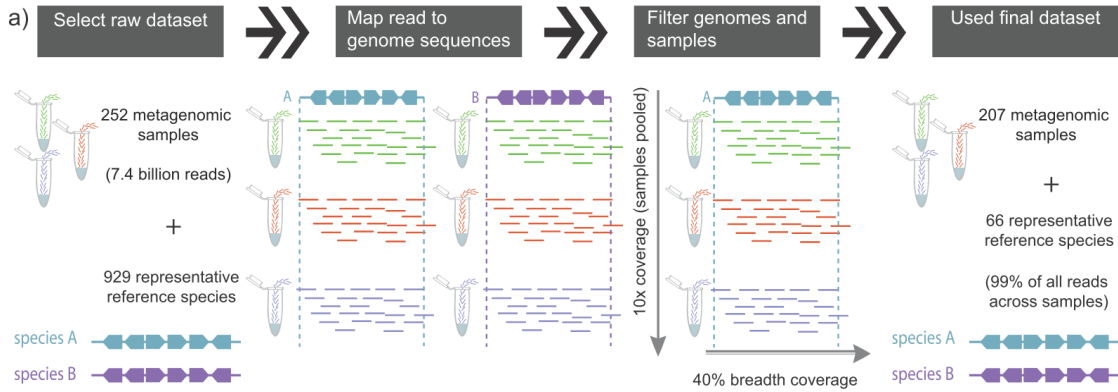


Figure 1: Procedure used for selection of metagenomes and species for SNP-based methods

The initial dataset consisted of 252 metagenomic samples and a non-redundant set of reference genomes representative of 929 species based on 40 universal, single-copy marker genes. Metagenomic reads from each sample were aligned to each species and was followed by a multi-step filtering procedure used in sample and genome selection. The final dataset corresponded to 207 samples that mapped to 66 species. For individuals where there existed more than one time-point the first time point was used.

2.2.2 SNP calling

A multi-sample SNP calling was done based on the pooled samples to identify a catalogue of 10.3 million SNPs. Only bases with a quality of at least 15 were considered. The following criteria were used to identify SNPs: (1) the single nucleotide variant had a minimum allele frequency of 1%, (2) was supported by a minimum of 4 reads. The first criterion is the classical definition of polymorphism and is used to exclude random sequencing errors that accumulate in the same position when depth of coverage is very high. The second criterion eliminates

sequencing errors that are randomly distributed across the genome. For further details please read Schloissnig *et al.* 2013 [98].

2.2.3 Measurement of species evolution (based on pN/pS ratio)

To measure the evolutionary pressure I used pN/pS ratio a measure akin to the classical dN/dS ratio. To calculate pN/pS ratio we needed to calculate the expected and observed ratio of non-synonymous and synonymous substitutions for the detectable parts of genomes and genes.

For the calculation of expected ratio first all the codon of the genome/gene were taken and determined the effect of all possible mutational events on the codon. The outcome could be either a synonymous or a non-synonymous mutation, and both outcomes were counted for each genome or gene in a given sample. This expected ratio was calculated assuming a uniform model for the occurrence of mutations across the genomic sequence.

To calculate the observed ratio between non-synonymous and synonymous substitutions all codons that contained polymorphic sites were extracted and the alleles were categorized into non-synonymous and synonymous. This observed ratio between non-synonymous and synonymous substitutions was then compared to the expected ratio resulting in the pN/pS ratio.

For all dominant species with at least 10x genome coverage found in a given metagenome, the pN/pS ratio was calculated for all the genes in each species and the average across the genes was used to estimate the pN/pS ratio of a species, **Figure 5**.

2.2.4 Measurement of gene evolution (based on pN/pS ratio)

To reliably estimate pN/pS ratio for a gene I followed the methodology in **section 2.2.3** and further required that average gene base pair coverage of 3 reads and that non-protein coding genes were discarded. Next I estimated the

average pN/pS ratio of a gene across all samples. Also, to ensure that the genes examined are ubiquitous I only considered genes that were found in at least half of the samples (≥ 126).

Orthology was assumed when the bit score was higher than 60 bits. The pN/pS ratio of a given OG was calculated based on the average across all genes mapped to an OG pooled from all species (8,122 OGs). The OGs with the lowest and highest score were then analyzed.

2.2.5 Comparison between the evolution of two species (*Roseburia intestinalis* and *Eubacterium eligens*)

To evaluate how different species respond under the gut environment I chose two species with similar conditions, which differed considerably in their pN/pS ratio (and average genome pN/pS ratio of 0.236 for *R. intestinalis* and 0.141 for *E. eligens*) that is:

(1) had similar genome coverage (*R. intestinalis* had an average coverage: 5.05x and a sum coverage overall samples of 1,046x; while *E. eligens* had an average coverage: 5.65x and a sum coverage overall samples of 1,169x) and

(2) were observed in similar number of samples (106 and 147 respectively).

Gene pN/pS ratio were determined and mapped to OG according to **subchapter 2.2**. Genes either of the species, *R. intestinalis* or *E. eligens*, were pooled separately and the median was used to calculate the pN/pS ratio of a given OG. For the 611 OGs that were detected in both species the log₂ ratio between the OGs of the *E. eligens* (low genome pN/pS ratio) and *R. intestinalis* (high pN/pS ratio) were calculated. Finally, the *galK* gene was selected in both species to illustrate (1) the difference in the mutation profile between genes with high and low pN/pS ratio and (2) to show that the pN/pS ratios are not affected by SNP density.

2.3 Gene content based methods

In **sub-chapter 2.3** I describe a metagenomic-based methodology to characterize the gene content variation of microbial species within an environment that can be applied to any complex microbial community. The framework, which is based on gene deletions [101], avoids known biases in pan-genome studies (such as cultivation requirement and genome pre-selection) since it captures the species in their natural habitats.

2.3.1 Detection of species presence in a metagenome for gene content analysis

A species was detected in a metagenomic sample according to **section 2.1.4** with the following additional filtering criteria, and a random set of 10 samples were picked (**Figure 2**).

(1) a minimum of 30x genome coverage (considered the *de facto* standard for high coverage [102]) was necessary to guarantee that the determination of gene presence is not influenced by sequencing depth and

(2) the 10 universal single-copy marker genes [103] had to be present. Criteria (1) and (2) are used to ensure that the genome detected corresponds to the species in study and not a close relative species with a similar genome composition.

(3) species seen in at least 10 samples were selected to increase statistical power.

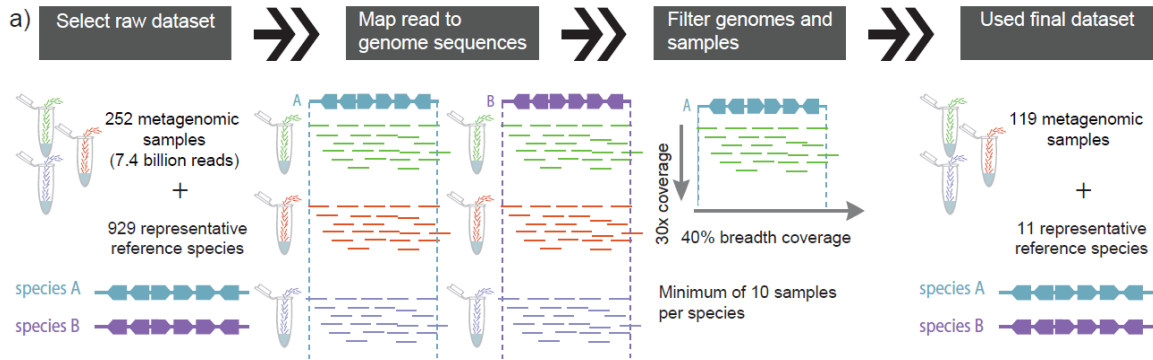


Figure 2: Procedure used for selection of metagenomes and species used in gene content analysis

The initial dataset consisted of 252 metagenomic samples (HMP and MetaHIT) and a non-redundant set of reference genomes representative of 929 species based on 40 universal, single-copy marker genes. Metagenomic reads from each sample were mapped to each representative genome of a given species and was followed by a multi-step filtering procedure used in metagenomic samples and species selection. The final dataset corresponded to 119 samples that mapped to 11 species.

2.3.2 Determination of core and accessory genes

For each species, a gene was categorized as core if detected in all 10 samples (that is at least 40% of the basepairs in a gene are covered by reads) otherwise was categorized as accessory, see **Figure 3**. The 40% gene length coverage filter was used to ensure that the gene is not called present due to spuriously assigned reads or reads that have origins in an ortholog from a close relative species. To chose the optimal gene length coverage filter, the gene content between pairs of biological replicates (time-series) was calculated for different cut-offs ranging between 0% and 100% in intervals of 10% (**Appendix 15**) and affected 3% of the genes (**Appendix 16**).



Figure 3: Diagram illustrating gene coverage of core and accessory genes of one species (*Dialister invisus*) for 10 individual metagenomes.

Dialister invisus is used to exemplify the typical variability in the coverage of core and accessory genes and their location across the genome in different individual metagenomes. Green represents core genes; red represents accessory genes and white represents missing genes, or genes below the 40% gene length coverage filter cut-off. The bottom bar shows the consensus between the 10 individual metagenomes and illustrates the definition of core and accessory genes but also illustrating the presence of regions where no gene in the reference genome was found in the 10 metagenomes.

Since the species had different abundances and coverage within metagenomes, either of the factors could influence the core and accessory gene categorization. However, neither abundance or coverage have an effect on the fraction of accessory genes, since none of the two variables correlated with the fraction of accessory genes (correlation with coverage has a $R=0.08$, $p\text{-val}=0.82$ and correlation with abundance has a $R=0.07$ and a $p\text{-val}=0.84$, **Appendix 17**).

2.3.3 Estimation of accessory genes fraction

The fraction of accessory genes was estimated by applying a subsampling procedure followed by model building. The subsampling procedure was based on random subsample sets that had a defined sample size. Several sample sizes

were used from 2 to 10 and for each sample size all combinations of random subsamples were used, except when the number of combinations was too high, in which cases I limited the sample size to only 500 combinations due to time constraints. For every subsample set I calculated the fraction of genes that were missing in at least one of the samples and was termed “subsample-based fraction”. The mean subsample-based fraction was calculated for each sample size independently and used to build the model. Two main models have been applied in pan-genome studies, exponential regression model [58] and power law regression model [6]. Additionally, negative exponential model and the spline function were also tested.

To evaluate the models it is necessary to determine the expected value in order to compare with the estimated values generated by the models. An appropriate candidate for the expected value is the calculation of subsample-based fraction for sample sizes larger than 10. For the species where I had these many samples, I calculated the subsample-based fraction for sample sizes ranging between 11 and the total sample size of a species. These subsample-based fractions were named the “expected fraction” to differentiate them from the observed subsample-based fraction. Subsequently, I compared the curve that was extrapolated from the model with the curve observed using the expected fraction, as an example see **Appendix 14**.

I observe that the exponential model predicts the curve that is closest to the expected fraction. While the exponential model deviates from the expected fraction by 8%, the other models performed worse with the closest the power law regression model deviating by 12%. Consequently, the exponential model was chosen to predict the fraction of accessory genes. The exponential model tended to underestimate the fraction, and hence the values obtained are a lower bound to the fraction of accessory genes.

2.3.4 Comparison of gene content differences between pairs of individuals and pairs of reference genomes

For a given bacterial species pair-wise comparison were done between all pairs of individuals A and B in the following way. The set of genes found in individual A and B are respectively defined as:

$$A = \{a|a \text{ is a gene found in species } X \text{ in individual } A \text{ and}$$

$$B = \{b|b \text{ is a gene found in species } X \text{ in individual } B\}$$

Then, I calculated the number of genes in the symmetric difference $|A\Delta B|$ and the number of genes in the union $|A \cup B|$ of individual pairs. The symmetric difference defines the set of genes that are present in either of the individuals but not found in the intersection of the two individuals, whereas the union defines the set of all distinct genes that are found in both individuals and in the reference genome. The gene content difference between two individuals is calculated according to the following formula and corresponds to the number of genes in the symmetric difference after normalization by the number of genes in the union.

$$f = \frac{|A\Delta B|}{|A \cup B|} \times 100$$

The pair-wise individual comparison was dependent on the reference genome selected to map the sequencing reads of samples. To emulate this reference genome dependency in comparison of gene content between sequenced genomes, a randomly selected genome was used for each species as “reference” genome. This “reference” genome was used as a third genome and only genes in this “reference” genome were considered in the pair-wise comparison. Hence, the genes that were counted in the symmetric difference, corresponds to genes found in the “reference” genome and in either of the two compared genomes but not in both, whereas the union corresponds to genes that were found in the “reference” genome and in either or both the other two genomes.

2.3.5 Determination of accessory gene deletion blocks

To determine the accessory gene deletion blocks, the representative reference genome was used as a reference of the gene order, and the location of the genes that were missing in each sample was determined. Contiguous accessory genes that are absent were clustered and called gene deletion blocks. In situations where an accessory gene is absent and neither of its neighbour genes are present, the absent gene is called single-gene deletion block. For 7 of our species, the representative reference genome was not completely assembled and the genome was composed of several contigs (Table 1). For these species, gene deletion block were counted in each contig. Therefore, there is the possibility that a gene deletion block is split into two contigs. Also, single-gene deletion blocks that occur in the start or end of a contig were not counted, so as to not inflate the number of single-gene deletion blocks. For each species gene deletion blocks were determined for 10 samples separately and the number of genes and number of gene deletion blocks were counted for each block size.

2.3.6 Determination of paralog within and between reference genome

To find paralog genes within and between reference genomes the methodology described in Alonso-Saéz *et al.* 2012 [104] was used. In summary, 95% ANI of the 40 universal single copy marker genes was applied to find the sequenced genomes that belonged to the *Bacteroides thetaiotaomicron* species. Three more genomes were found besides the genome used in our study (*Bacteroides sp.* 1.1.6), the type strain *Bacteroides thetaiotaomicron* VPI-5482 (ATCC, NCBI TaxID 226186), *Bacteroides thetaiotaomicron* dnLKV9 (NCBI TaxID 1235785) and *Bacteroides sp.* 1.1.14 (NCBI TaxID 469585). These genomes were used to construct the *Bacteroides thetaiotaomicron*-specific orthologous groups (NOG) using the eggNOG pipeline. The genes were assigned to a given *B. thetaiotaomicron* NOG by using Blastp with a bit score cut-off of 60 and 95% identity. In addition, to ensure that genes detected in the large-deletion block also

do not have less conserved paralogs within *Bacteroides sp.* 1.1.6 a less stringent cut-off was used (40% identity and 80% protein length).

2.3.7 Detection of genomic islands

Genomic islands were detected using Islandviewer methods IslandPath-DIMOB and SIGI-HMM using default options [105].

CHAPTER 3

RESULTS AND DISCUSSION

SCHLOISSNIG, S., ARUMUGAM, M., SUNAGAWA, S., MITREVA, M., TAP, J., ZHU, A., ET AL., NATURE 2013

“Genomic variation landscape of the human gut microbiome”

3.1 SNP variation in gut prokaryotic species with phenotypic implications

This **sub-chapter 3.1** is integrated in a larger group project intended to characterize the human gut microbiome based on SNP profiling and has been published in Nature 2013 [98]. In this project my aim was to create a methodology to estimate the phenotypic outcome of SNP variability from metagenomic dataset using a measurement of evolution based on SNP's (pN/pS ratio). The calculation of SNP density, nucleotide diversity (π) and the downsampling used in **Appendix 13** were done by Siegfried Schloissnig.

3.1.1 Introduction

As detailed in **sub-chapter 1.3** SNP variation can have a phenotypic effect in both the commensal bacteria where the SNP is found but also in its human host. Since most species genomes were only recently sequenced [57], evaluation of the phenotypic impact of each SNP is impossible, especially for such a large SNP catalogue composed of 10.3 million SNPs (**sub-chapter 1.6**). Another approach to study the phenotypic impact is to focus on genes undergoing selection. Natural selection acts by increasing the survival of the fittest microbial strains, and genes that contribute to strains fitness. Over time, microbial strains can maintain inherited traits (conservation via purifying selection), novel beneficial traits can emerge (adaptation via positive selection) or traits can be lost by a relaxation of purifying

selection. In terms of genetic variation, long-term purifying and positive selection results in genes which are slow or fast evolving. Whereas for some genes the evolutionary outcome is similar across the majority of species, for others it can differ among species [106]. To measure the evolutionary effect of SNPs for both gut microbial strains and their respective microbial genes, I measure the ratio of non-synonymous to synonymous polymorphisms (pN/pS ratio) a measurement akin to classical dN/dS ratio [92, 107]. pN/pS ratio has been previously applied to the study of genome evolution for isolation based genome and also metagenomes of simple communities [92, 108, 109]. This method has the advantage compared to other population genetic measurements that it does not require the establishment of genomes haplotypes, as this is still currently not possible for complex metagenomic samples. pN/pS ratio tends to capture the evolutionary effect of relatively recent events in the history of the population [109] and pN/pS ratio is relatively insensitive to confounding factors related with divergence data such as generation time (contrary to dN/dS ratio) [109]. Hence pN/pS ratio constitutes a good measurement of the effect of natural selection on SNPs and can be used to predict the phenotypic outcome. In this project I measured pN/pS ratios for 229,692 genes from 66 dominant species measured in metagenomic samples of 207 individuals.

3.1.2 Large variation in SNP evolution of prokaryotic species within individuals rather than between individuals

The genome coverage varies greatly across gut prokaryotic species in the metagenomes, therefore we evaluated whether pN/pS ratio is affected by genome coverage, as other SNP based measurements such as SNP density (SNP kb⁻¹) and nucleotide diversity are affected by coverage biases, **Figure 4**. Nucleotide diversity is a measurement for the degree of polymorphism in a population. I found that pN/pS ratio is not affected by genome coverage, as shown in **Figure 4**. Independency of pN/pS ratio from coverage is also supported by the finding that pN/pS ratio remained considerably stable when the genome coverage was downsampled to different coverages using genomes with more than 10x coverage.

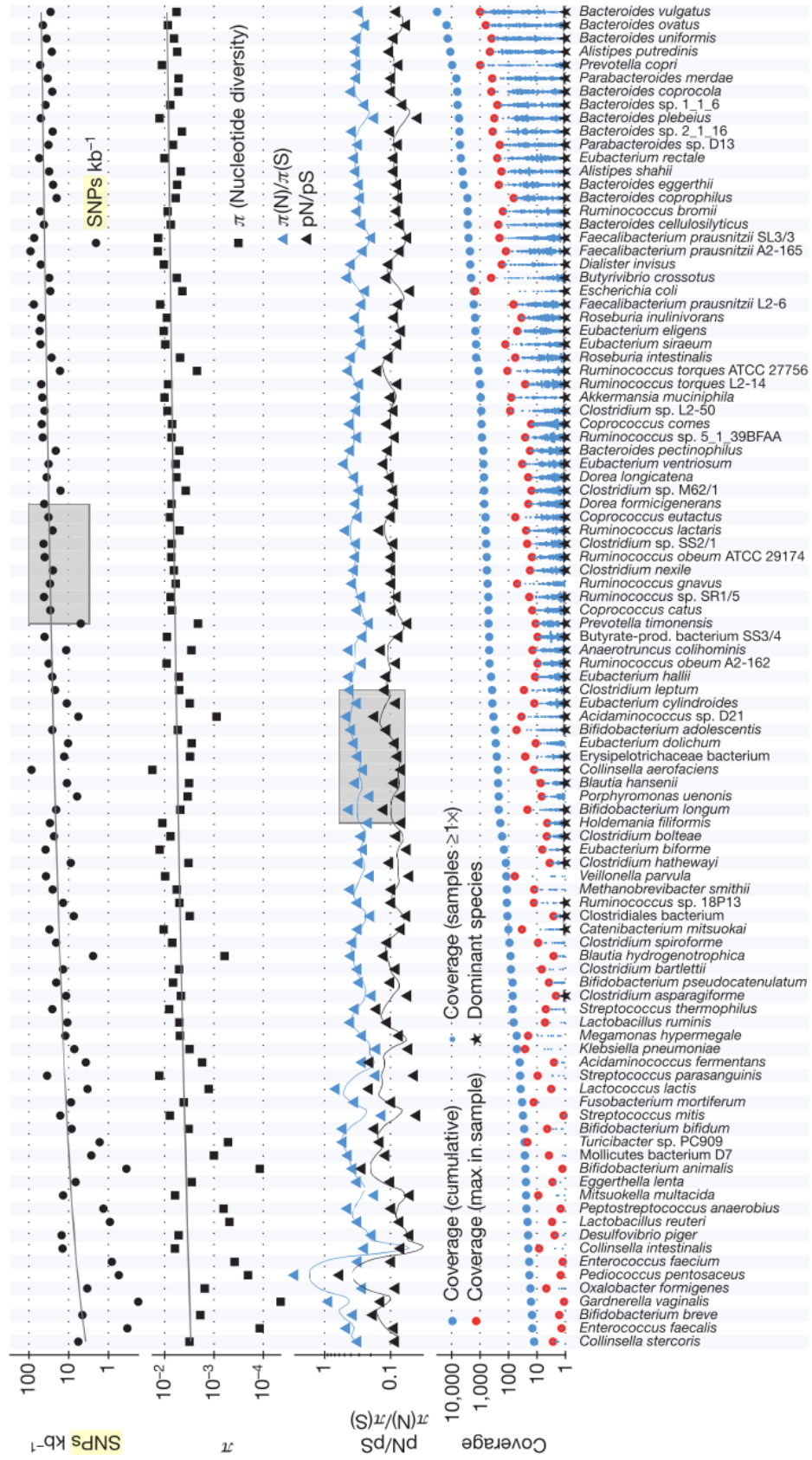
The reason for this independency is related with pN/pS ratio corresponding to a ratio and therefore the coverage effect is nullified between nominator and denominator. Moreover, we found that pN/pS ratio is in accordance with $\pi(N)/\pi(S)$ ratio and $\pi(\text{non-degenerate sites})/\pi(\text{4-fold degenerate sites})$ ratio, with the latter being less affected by properties of the mutation spectra such as the likelihood of transversion and transitions. These agreements further support pN/pS ratio as reliable and appropriate to estimate the degree of polymorphism in the population and to infer evolution.

To ensure that sufficient samples were available with high sequencing depth pN/pS ratios were only calculated for dominant species, which we defined as species that have at least 10x coverage across samples and 99% of the reads in metagenomes where these genomes were found, in **Figure 1**. Since the SNP profile is relatively constant across time-points, usage of time-points from the same individual would introduce biases towards the individuals where more time-points are available, hence only the first time-point for a given individual was used. The pN/pS ratio for each pair of species-individual was calculated for 66 dominant species across 207 individuals. I found that pN/pS ratios were relatively stable across individuals but varied considerably in range among species with an average pN/pS ratio of 0.11, and ranging between 0.03 (*Bacteroides plebeius*) and 0.17 (*Acidaminococcus* sp. D21) across species in **Figure 5** and **Appendix 4**. The pN/pS ratio calculated for gut dominant prokaryotic species is in agreement with previous reports of pN/pS ratio and dN/dS ratio [92, 108–111]. The relatively low pN/pS ratios were constant across different individuals, despite the variation in individuals phenotype (gender, age, disease status, etc), except for two genomes *Parabacteroides merdae* and *Bacteroides uniformis*. The reason for these two exception remains currently unknown. These low pN/pS ratios might indicate that similar constraints exist across individuals, hence the evolution of gut microbial species is likely to be dominated by long-term purifying selection and drift instead of rapid adaptations to a given host environment. I found that the mean pN/pS ratio between the two continents were statistically significant), however the signal could

also be due to lower sequencing depth of European samples which could lead to less dominant species (such as the top species with high pN/pS ratios) not to be detected, **Figure 5**. Instead the wide variety of pN/pS ratios across species, in comparison to across individuals suggest that the evolution constraint acts differently in different gut species, and could be related to niche specialization of each bacteria.

Figure 4: Statistics of genomic variation across 101 gut microbial species prevalent in 252 metagenomes from 207 individuals.

The statistics regarding genomic variation were calculated for 101 prevalent species. Prevalent species are defined as having a cumulative coverage bigger than 10 and at least 40% breadth of the genome covered. The 66 dominant species that are indicated by an asterisk account for 99% of the mapped reads and were further selected for pN/pS ratio analysis. Species names are given without specifying the strain details. In the last subplot the blue point cloud plots shows the coverage (more than 1x) across all metagenomes, with the blue dot above each line corresponding to the cumulative coverage and red dot the maximum coverage observed across all metagenomes where the species is found. Nucleotide diversity (π) follows the trend in SNPs kb⁻¹ and pN/pS ratio is in agreement with $\pi(N)/\pi(S)$ ratio.



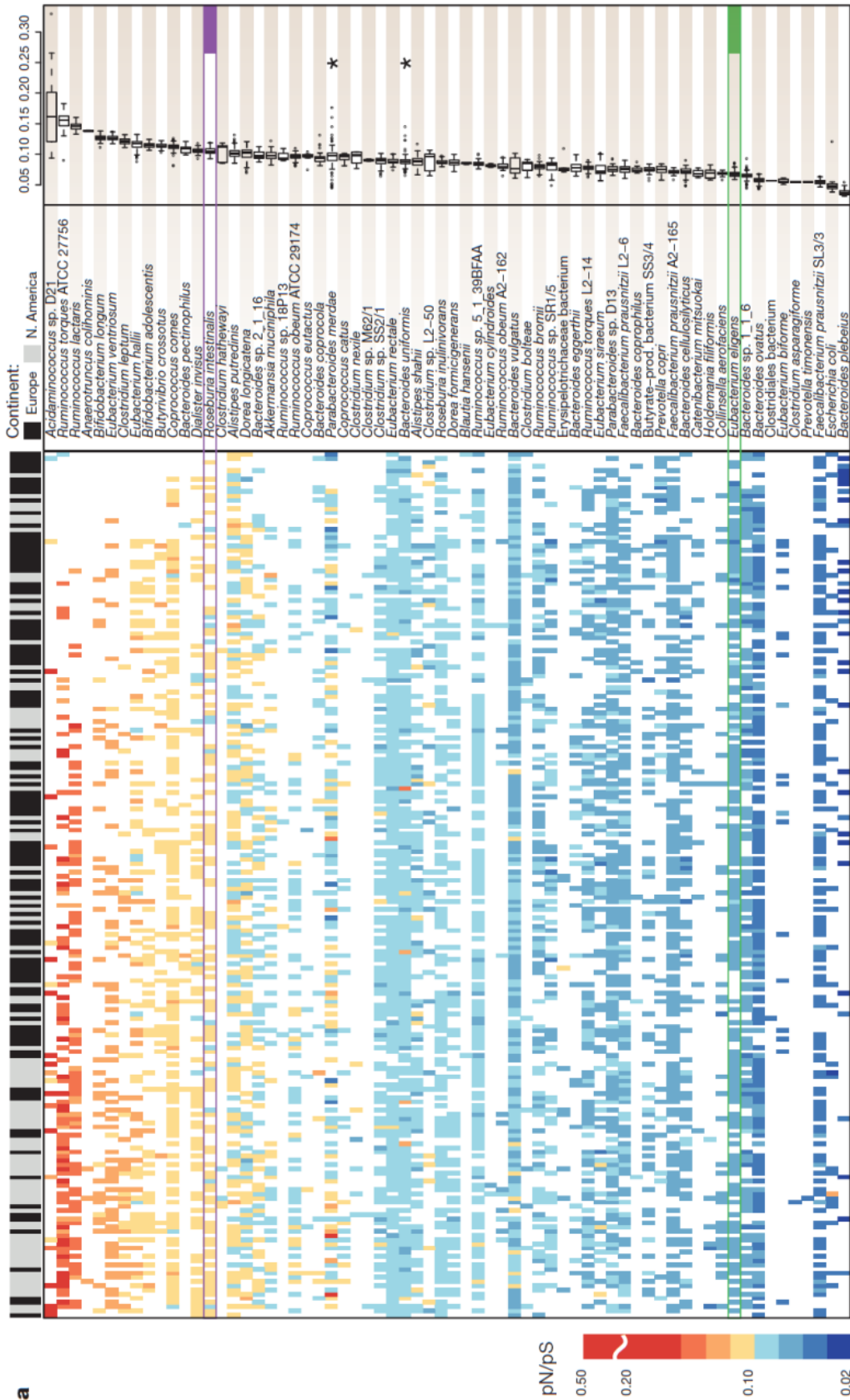


Figure 5: pN/pS ratio distribution across 66 dominant species and across 207 individuals

*Heat map of pN/pS ratios for 66 dominant species (rows) and 207 individuals (columns, only the first time-point was used per individual) is displayed and summarized by species (boxplot on the right). Rows and columns are sorted by their mean pN/pS ratio, which varied greatly across species compared to across individuals. The panel above the heat map annotated the continent of residence for each individual, North American or Europe. The purple box and green box highlight the two genomes (*Roseburia intestinalis* and *Eubacterium eligens*) which are used in further analysis in Figure 6. The mean pN/pS ratio was significantly different between the two continents, with European individuals tending to have lower pN/pS ratios. This continental difference is likely to be due to an effect of lower sequencing depth in samples from European individuals and therefore these samples are not able to capture the less abundant species, as for example shown in species displayed in the top-right corner.*

3.1.3 Type IV secretion system is slow evolving and bile acid hydrolase is fast evolving in the human gut prokaryotic species

For each metagenome I calculated the pN/pS ratio for each gene among all genes found across the 66 dominant species (229,692). The genes were further grouped into their orthologous groups (8,000) using the eggNOG pipeline [112]. Here I focused on the genes with lowest and highest pN/pS ratios. As expected, housekeeping genes were frequently associated with genes that had the lowest pN/pS ratios (**Appendix 5**). For example, tRNA synthetases, DNA polymerases, RNA polymerases, Transcription elongation factors, chaperonin among others were found in this list. Less obvious were the finding of genes related with type IV secretion system. Type IV secretion system are used for gene transfer among prokaryotic species [113] and are involved in the interaction of the host with pathogenic [114] and commensal bacteria [115], especially in immune modulation and anti-inflammatory responses [116]. Several genes with conserved but currently unknown functions also have low pN/pS ratio, suggesting that their functions could be targeted for further phenotypic characterization **Appendix 6**.

On the other side of the spectrum, several transposases and antimicrobial resistance genes are among the genes with high pN/pS ratios. Interestingly I also find among the highest pN/pS ratios gut-specific genes such as bile salt hydrolase (BSH) [117] in **Appendix 5**. BSH encodes the gene involved in the “gateway reaction” for a wide variety of pathways involved in production of secondary bile acids from conjugated bile acids (CBAs) initially produced by the human host.

3.1.4 *galK* gene evolution is uncoupled from prokaryotic species evolution

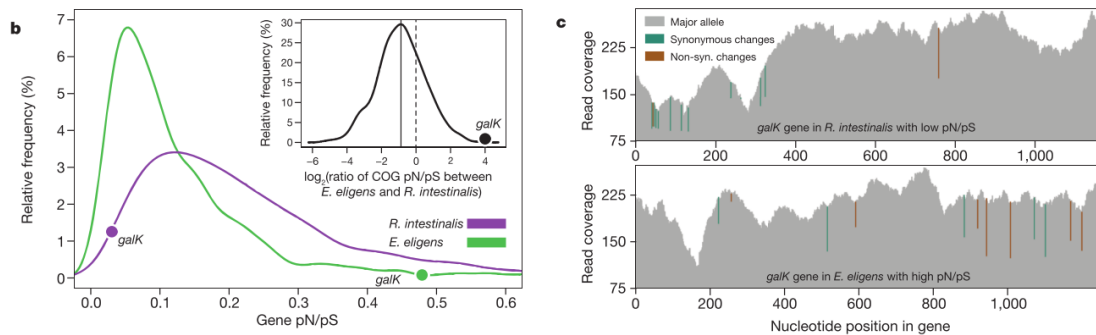


Figure 6: Comparison between *Eubacterium eligens* and *Roseburia intestinalis*

(b) Distribution of pN/pS ratio for each gene in two genomes, *R. intestinalis* and *E. eligens*, based on the average pN/pS ratio across all metagenomes where each species is detected. Despite having a similar genome coverage pN/pS ratios for most genes are higher in *R. intestinalis*. Inset of **(b)** shows the log₂ ratio between orthologues from *E. eligens* and orthologues from *R. intestinalis* with the average log₂ ratio showed by solid vertical line and random expectation by a dashed vertical line. Outliers can be shown in this way, like the galactokinase gene (*galK*), whose pN/pS ratio is among the lowest in *R. intestinalis* and highest in *E. eligens*. **(c)** Illustrates the distribution of synonymous (green) and non-synonymous (brown) SNPs across the *galK* genes from *R. intestinalis* (top panel) and *E. eligens* (bottom panel), both plots are based on cumulative read coverage. *E. eligens* displays a higher number of polymorphisms and is enriched in synonymous SNPs, whereas the opposite is observed in *R. intestinalis*.

In order to investigate the congruence between pN/pS ratio of a given species and pN/pS ratio of its respective genes, I compared the pN/pS ratios of all

genes between two species *Roseburia intestinalis* and *Eubacterium eligens*. These species differed considerably in their species pN/pS ratios across similar metagenomes and similar genome coverage range. Whereas the average pN/pS ratio of *R. intestinalis* (across 106 metagenomes) is 0.236 for *E. eligens* (across 147 metagenomes) the ratio is only 0.131, **Figure 5**. The majority of genes (75%) in *R. intestinalis* had systematically higher pN/pS ratios compared to their orthologous genes in *E. eligens*, with only a few exceptions **Figure 6b** and **Appendix 7**. The exceptions whose pN/pS ratios deviate considerably point to different evolutionary constraints for these genes. For instance, *galK* gene which encodes galactokinase had a low pN/pS ratio for *R. intestinalis* but a high pN/pS ratio for *E. eligens* (0.03 and 0.48 respectively **Figure 6b** and **Figure 6c**). *galK* is an essential enzyme in the Leloir pathway for galactose metabolism for the majority of organisms. The *galK* example reveals that the same gene can undergo different evolutionary outcomes that are independent of the species evolutionary pressures (one species can undergo tight negative selection while relaxed negative selection for another).

3.1.5 Discussion

The goal of this project was to estimate whether pN/pS ratio can be used to estimate the phenotypic impact of strains variation in their SNPs. pN/pS ratio was found to be reliable and applicable for metagenomic datasets. The stable pN/pS ratio observed for gut microbial species (based on 66 dominant species) across 207 individuals suggest that host conditions (which include variation in diet, host genetic composition and immune tolerance) have minor effects on the evolution of species compared to constraints that are constant across the human population (such as gut physiology, anaerobic conditions and pH). Instead most of the variation is species specific suggesting the niche that a given species occupies might have an effect upon genome evolution, although there is also the possibility that the species have naturally different speeds of evolution.

Next, I analyzed the evolution of 229 000 genes from 8 000 orthologous groups, and found as expected several housekeeping genes to be slow evolving and transposases to be fast evolving. Unexpectedly I found that among the slowest evolving there were genes related with type IV secretion system (associated with conjugation) which is associated typically with strain variability [62]. This suggests that genes in type IV secretion system undergo purifying selection because bacteria need to maintain genome plasticity by gene transfer through conjugation in a changing environment such as the human gut [113] and/or due to their important role in the cross-talk with the host immune system [116]. The contrast between type IV secretion system and transposases evolution indicates that mobile machineries do not necessarily undergo the same evolutionary trajectory. Among the fast evolving genes, I found a gut-specific BSH, which catalyzes deconjugation of CBAs to release free primary bile acids and amino acids [118]. CBAs are cholesterol derivatives synthesized in the liver that can inhibit bacterial growth and upregulate the defense system of the host mucosa [117]. Deconjugation and modification of bile acids have been associated with colorectal cancer and gallstone formation [119, 120], and more directly BSH has been linked to dysbiosis in IBD patients [121]. The high pN/pS ratio may indicate genome plasticity required to metabolize and respond to a variety of different bile acids that exist in the gut and also be related with dysbiosis in different diseases.

Finally, I measured whether gene evolution can be uncoupled from species evolution. Here I compared genes from *E. eligens* with *R. intestinalis* and found that most genes (75%) follow the genome trend, with only a small proportion of outliers. Among the outliers I found a carbohydrate degradation enzyme *galK* (galactokinase). Galactokinase is the first enzyme used in the Leloir pathway and is the only reaction that can convert galactose to glucose [122]. This pathway is important for usage of galactose or other more complex carbohydrates that have galactose in their constitution, such as melibiose. Although *galK* is present in *E. eligens*, the gene may not perform its main function (see also [123]), since *E. eligens* is not able to ferment galactose or galactose-containing disaccharides

lactose and melibiose [124], which are substrates of *galK*. The statement is supported by findings in *Bacillus subtilis* in which the pathway is inactivated and *galK* must be non-functional in order to not accumulate galactose derivatives that are toxic for the cell [125]. In comparison, *R. intestinalis* is able to ferment melibiose [126], validating that *galK* is functional in this species.

Overall, the results show that certain phenotypic inferences can be already estimated based on evaluation of gene evolution, however the method is limited to genes in the extreme spectrum of evolution, whereas for the remaining further studies are needed.

“Inter-individual differences in the gene content of human gut bacterial species”

3.2 Inter-individual variation in gene content of human gut bacterial species

In **subchapter 3.2** I describe the findings obtained from using a metagenomic-based approach to determine gene content of bacterial species from complex environments. Taken as an example, I analyzed 11 abundant gut species across 10 human individuals for each species, and find a large inter-individual variation in gene content between gut bacterial species. Moreover, I find that for the same species the gene content differences between individual’s strains are associated with important functional traits such as polysaccharide degradation and capsule polysaccharide synthesis, both of which have an effect on the digestive capacity of the human host. These findings imply that functional variation cannot be explained by species composition alone.

The results of this project are currently in revision with Genome Biology where I am the leading author. Within the project, Daniel Mende contributed with analysis to benchmark my approach to completely sequenced genomes.

3.2.1 Introduction

We found when analysing the SNP catalogue of the gut microbiota that individuals have a temporally stable SNP profile that is individual-specific (individuality) [98]. This suggests that the pool of strains that inhabit the human population is rather large and not limited to a set of strains. The interpretation of the phenotypic outcome due to variation in SNP profile is however rather

complicated. Instead variation in bacterial gene content is easier to interpret and provides a more direct estimation of a bacterium phenotypic traits, therefore the aim of **sub-chapter 3.2** was to create a methodology to characterize bacterial strain gene content based on metagenomic data.

In the current study I categorized genes into core (seen in all metagenomes) and accessory genes (seen in a subset of metagenomes) for a given species, (see **Table 1** for core and accessory definition and **Table 2** for details of species used in the study). This nomenclature derives from pan-genome studies, please note that other nomenclatures have been used to describe the same or similar concepts. The accessory genes [62] have also been named dispensable [63] and variable genes [127]. Genes have also been categorized by their frequency in HGT, with genes being separated into hard core (genes that seldom or never undergo exchange), soft core (genes where it is hard to establish if they have undergone exchange) and shell (freely exchangeable genes) [128]. Moreover, the species accessory genes can further be separated into unique genes (genes that only exist in one strain) and non-unique genes (genes that exist in two or more, but not all strains) [129] and can also be named volatile genes (present in less than 20% of the strains) or persistent genes (present in more than 90% of the strains) [64]. Since for a few species that I analyzed only 10 metagenomes were available, I have simplified the nomenclature and focused on core, accessory and unique genes. As accessory genes are what differentiate strains from each other in terms of gene content, they provide the genetic basis for strain specificity. Accessory genes encode supplementary biochemical pathways and functions that can provide a selective advantage to the bacteria [130]. Note that the core and accessory concept is not restricted to strains of the same species as used in the context of this **sub-chapter 3.2** but can also be extended to other taxon. Baptiste *et al.* 2004 [131] named taxon core to define the set of genes that are shared by all members of a given taxonomic rank, hence any taxonomic rank can be used as an taxon core. Taxon core have been explored among species, genus, phyla, bacterial domain and even the universal core (reviewed in [130]). For example the

universal core represent the genes that are present in all organisms and is composed of 40 marker genes. These marker genes are present as a single copy in all domains and have constraints against HGT [132, 133]. The bacteria core, which is present in all bacterial species is estimated to be composed of approximately 250 genes [134].

Gene content affects the genome architecture, two forces interplay in accessory genome architecture, one in maintaining the gene synteny (operon, genomic islands) after gene transfer, and another in destroying it. Examples of both cases have been previously reported in the literature. Operons, mosaic operons [135] and genomic islands [82] have been shown to maintain their original synteny after HGT. While another study has shown that megaplasmids suffered selective deletion removing up to 10% of the plasmid genome after insertion into the bacterial cell [84]. On the other hand, Price *et al.* 2005 has shown that operon structure can be destroyed in order to form new operon structures, either through multiple gene addition or through deletion of DNA between functionally unrelated genes that are close to each other [136]. I intend to investigate how accessory genes distribution affects genome architecture in the abundant gut bacterial species.

Previous studies of accessory genes and pan-genomes have either been conducted on single species (reviewed in [62]), or via comparisons of closely related species (such as species within the same genus) [137, 138]. Since in a metagenomic sample several species exist in the same habitat, my approach enables a comparative study of bacterial species within the same environment. Here I performed a comparative study of the 11 most abundant gut bacterial species that passed my filtering procedure (described in **section 2.3.1**). This comparative study enabled me to calculate the fraction of accessory genes in these species and characterize details of the genomic architecture and functions of these genes. Taken together, our study provides the first metagenomic insight into gene content variability of abundant gut microbial species across individuals.

In addition, I show capsular polysaccharide synthesis (CPS) and polysaccharide utilization loci (PUL) as examples for associated functional implications.

Term	Definition
core gene	species specific gene seen in all samples
accessory gene	species specific gene seen in some samples
single-gene deletion block	single gene missing in a sample when compared to the reference genome
gene deletion block	block of one or consecutive neighbour genes missing in a sample compared to reference genome
large-gene deletion block	deletion of 50 or more genes when compared to the reference genome in a sample
consecutive-gene block	consecutive genes that are present in a given sample
Individual	refers to a individual gut sample for a given species

Table 1: Definitions used in the scope of **sub-chapter 3.2**

*Description of definitions used throughout **sub-chapter 3.2**, the terminology was adapted from pan-genome studies to suit metagenomic studies.*

3.2.2 Fraction of accessory genes increases with genome size

A description of 11 species (**Figure 2**) used in this chapter can be viewed in **Table 2** and for each species 10 random metagenomes were used, **Appendix 8**. For all the species studied the accessory genes were found to have a patchy distribution [139], which means that they are not evenly distributed across the genome with certain regions such as genomic islands being highly concentrated in accessory genes. The patchy distribution of accessory genes are likely a result of HGT and gene deletion [139] and are common observations in pan-genome studies of other species [62]. In total 60 genomic islands were detected across the 11 species with sizes up to 57kb, **Appendix 9**. These genomic islands are likely derived from mobile elements such as prophages, integrative plasmids or ICEs [82].

NCBI taxID	Representative strain name	N° of contigs	N° genes in reference	N° of individuals	N° genes in metagenomes
592028	<i>Dialister invisus</i> DSM 15470	1	2015	16	1905
657321	<i>Ruminococcus bromii</i> L2-63	1	1852	22	1807
511680	<i>Butyrivibrio crossotus</i> DSM 2876	31	2576	13	2493
657322	<i>Faecalibacterium prausnitzii</i> SL3/3	1	2816	11	2670
445970	<i>Alistipes putredinis</i> DSM 17216	11	2795	58	2790
537012	<i>Bacteroides cellulosilyticus</i> DSM 14838	66	5771	15	5542
483216	<i>Bacteroides eggerthii</i> DSM 20697	20	3769	10	3714
717959	<i>Alistipes shahii</i> WAL 8301	1	2616	29	2584
563193	<i>Parabacteroides</i> sp. D13	22	4558	32	4473
469586	<i>Bacteroides</i> sp. 1_1_6	71	5648	41	5639
537011	<i>Prevotella copri</i> DSM 18205	28	3413	32	3195

Table 2: Information regarding the 11 species references genomes and the metagenomes used in the current study.

Description of the 11 species representative genomes that were selected after the filtering procedure and to which our metagenomes were mapped. General information regarding the reference genomes such as their corresponding NCBI TaxID, strain name, number of contigs and number of genes are described. The total number of individuals (from which metagenomes were obtained) where a given species was found is shown together with the number of the genes from each species that were detected in any of the metagenomes.

The number of accessory genes can differ greatly among bacterial species [62, 130]. To estimate the fraction of accessory genes across the 11 species I used

the methodology described in **section 2.3.3** consisting of a subsampling procedure followed by a model fitting. The exponential model and power law were tested (typically used in pan-genome studies, see **Appendix 14** [63, 140]). In addition the negative exponential and spline were also tested. The exponential model provided the best fitting to the observed data and was therefore chosen. The asymptotic number obtained from the exponential model was used to extrapolate the fraction of accessory genes for the 11 species as shown in **Figure 7**.

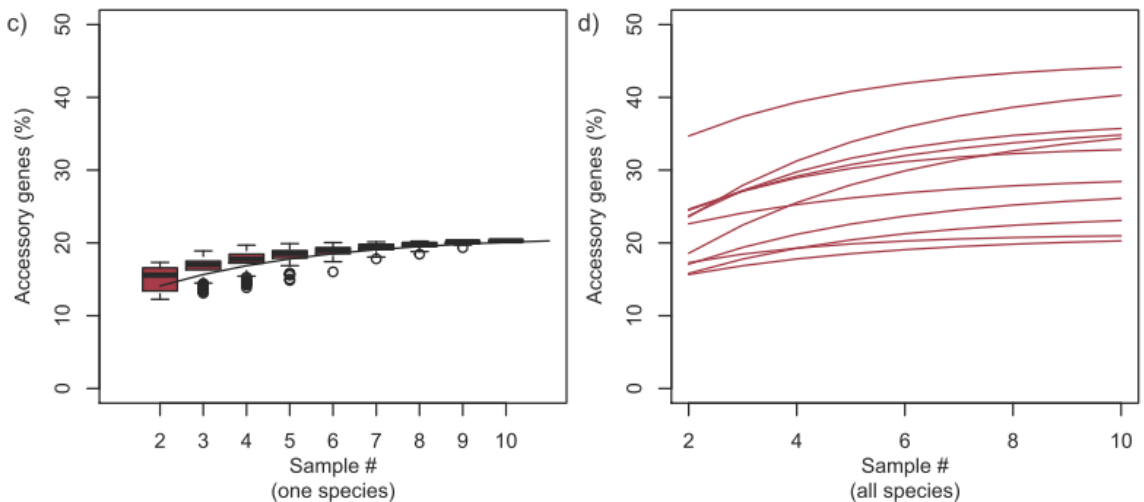


Figure 7: Estimation of the fraction of accessory genes (%) based on exponential model

Boxplot shows the estimated fraction of accessory genes based on the exponential regression model fitting for the 11 gut bacterial species.

The fraction of accessory genes ranges between 20.94% (*Dialister invisus*) and 45.16% (*Prevotella copri*, which is one the main drives of the Prevotella-enterotype) with an average of 32.28% (**Figure 7** and **Figure 8**). Note that only gene deletions are accounted for in the estimates, that is genes that are missing in the metagenomes in comparison to the species respective reference genome. Since I only account for gene deletions, there will be individual unique genes (i.e. genes specific to an individual's strains, and hence not present in the reference genome). Individual unique genes can significantly increase the percentage of

accessory genes, such as observed in *Haemophilus influenza*, where unique genes constitute 19% of a species gene repertoire [65]. Therefore, these estimates should be considered as the lower limit of gene content variability. To have a proxy for individual unique genes, the fraction of unique genes in the 11 reference genomes (i.e. genes in reference genome not detected in any of the individuals) was calculated, to circumvent the impossibility of estimating the number of individual unique genes. On average only 3% (up to 5%) of the genes in reference genomes were found to be unique, and can be used as a minimum estimate of percentage of genes that are missed by my reference dependent approach. In conclusion, the high fraction of accessory genes per species that is estimated in metagenomes is in similar range as estimates found in pan-genome studies [61–65]; and the estimates are likely to increase as only gene deletions are counted and the number of individuals is limited (as more individuals are sequenced the likelihood of finding an individual where the gene is missing also might increase, as shown in **Appendix 14**).

The 11 species belonged to two phyla, Firmicutes and Bacteroidetes. Interestingly, Bacteroidetes had a larger fraction of accessory genes compared to Firmicutes ($p\text{-val} < 0.01$) and that the fraction of accessory genes correlated with genome size ($r=0.72$), see **Figure 8**. This correlation is not simply a result of an increase of accessory genes in larger genomes, since the number of core genes also correlate with genome size, instead it reflects a pattern where the number of accessory genes increases faster than those for core genes in larger genomes (**Figure 8**). Further investigation using species from more phyla is needed to elucidate if this between-phylum difference is a general pattern or solely an effect of differences in genome size.

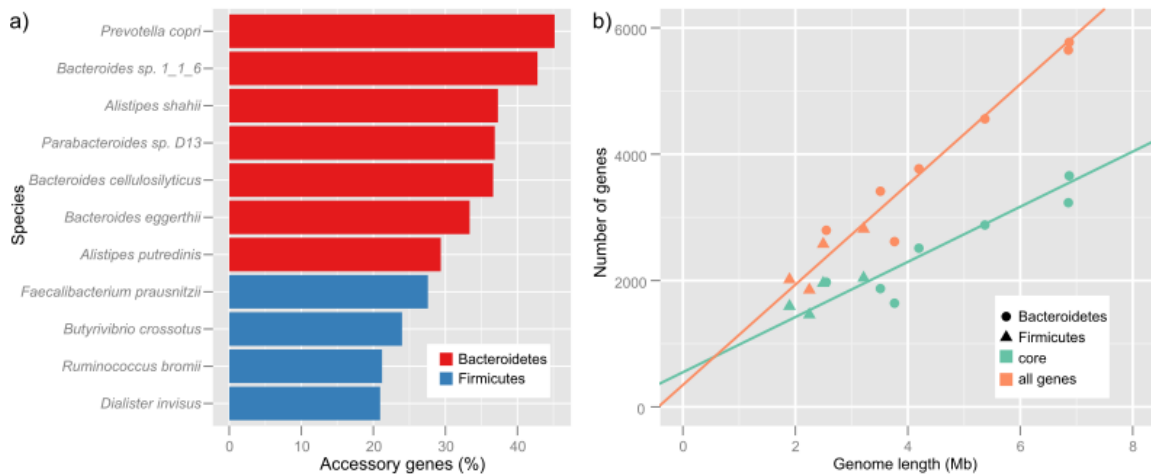


Figure 8: Display of the percentage of accessory genes for 11 gut bacterial species.

(a) The bars correspond to the percentage of accessory genes which were calculated based on the asymptotic number originated from the exponential regression model. The values were estimated for the 11 gut bacterial species which are grouped according to their phyla. (b) Dot plot displays the relation between number of core genes or total number of genes and genome size. The graph shows that number of core genes also correlates with genome size; however the total number of genes grows faster with genome size than the number of core genes.

3.2.3 Gut strains of the same species have large inter-individual variation in gene content

To measure the gene content differences of strain's that exist between individual's gut, pairwise comparison between individual's strains was done for each of the 11 species that is shared independently of the other species present in the two individuals. The genes that are present in one individual but not in the other were counted according to **section 2.3.4** and the average across the 11 species corresponded to $13\% \pm 4.5\%$ (mean \pm SD), as shown in **Figure 9**. This inter-individual difference was considerably larger than differences observed between biological replicates (same individual, samples from different time points) and between technical replicates (same individual, same sample, different

sequencing reactions), which were on average 0.81% and 0.51%, respectively, and statistically not significantly different among each other (p-val=0.71, **Figure 9**).

Among the 11 species, *Bacteroides thetaiotaomicron* (represented by reference genome *Bacteroides* sp. 1_1_6 [98]) has the highest average inter-individual difference in gut bacterial gene content (16%), whereas *Dialister invisus* (represented by reference genome *Dialister invisus* DSM 15470 [98]) has the lowest (6%). As mentioned before these estimates correspond to lower limits due to the dependency on reference genomes. Yet, for all 11 species, no two individuals shared the same gene content, even when the analysis was extended to all 103 individuals.

To measure how the gene content differed between strains in their natural habitat and to compare with those found classically in pan-genome studies, we created a database of complete genome sequences for Firmicutes and Bacteroidetes species (1,077 genomes belonging to 35 species). This dataset was used for comparison as 10 out of 11 species investigated in this **sub-chapter 3.2** do not have enough completely sequenced genomes of other strains available in public databases. The metagenomic dataset studied here had a significantly higher gene content variation compared to published pan-genomes (a mean of 12.97% \pm 4.51% and 10.69% \pm 5.13% respectively, p-val < 10^{-16} **Figure 10**). The difference is even higher if we consider the pan-genomes from all available species (a mean of 9.19% \pm 6.22% for completely sequenced genomes, p-val < 10^{-16} , **Figure 10**). For only one of the species, *Parabacteroides* D.13, sufficient data was also available for completely sequenced genomes (8 different strains). The results show that pairwise differences of metagenomes (4.32%- 22.64%) are in similar ranges as those observed for completely sequenced genomes (6.68%-20.61%, **Figure 11**). Overall, these comparisons show no large systematic differences between gene content estimations obtained from metagenomes and the ones obtained from genomes from isolated strains, demonstrating the validity of the proposed method.

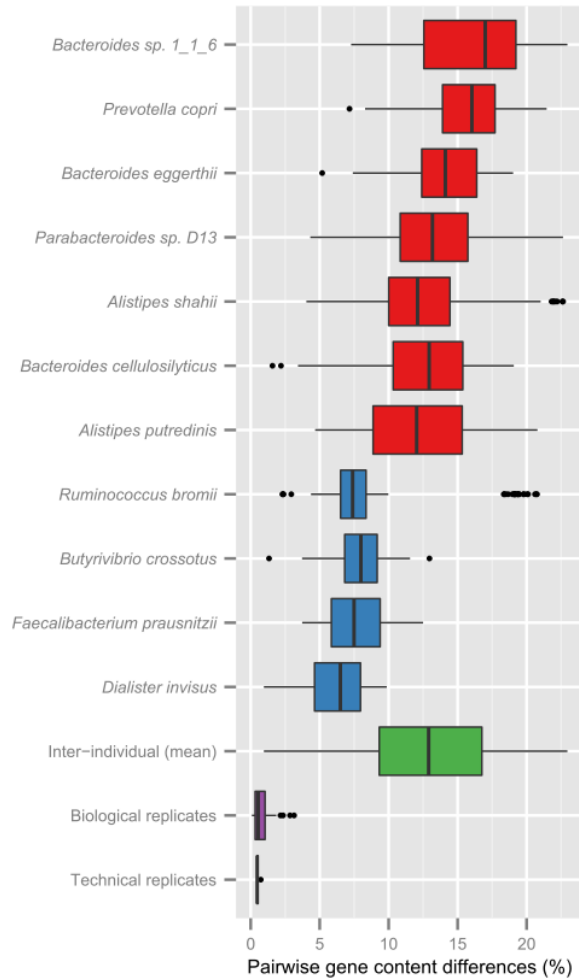


Figure 9: Variability between pairs of metagenomes

For each of the 11 species, the gene content differences between pairs of individual were calculated. Boxplots are colored according to phyla (red for Bacteroidetes and blue for Firmicutes), and the species are ordered according to their mean. The inter-individual boxplot (green) represent pairwise comparison of the same species between different individuals for 11 species. Biological replicates boxplot (purple) show pairwise comparisons of the same species in samples from the same individual at different time-points (11 species). Technical replicates represent gene content differences of *Prevotella copri* of the same sample in four sequencing replicates (no other technical replicates were available).

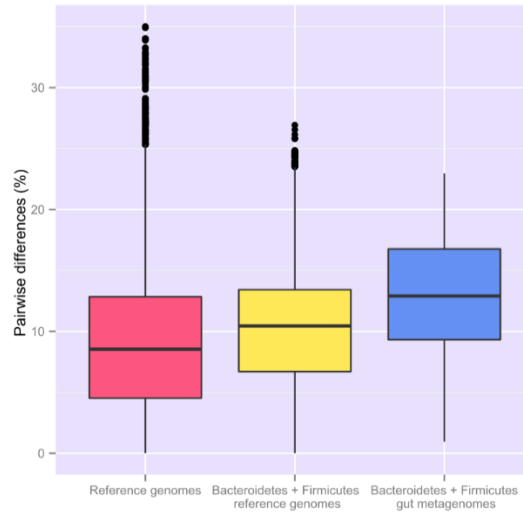


Figure 10: Variability between (1) sequenced reference genomes, (2) Bacteroidetes and Firmicutes reference genomes and (3) metagenomes.

Boxplot show difference in number of genes (%) between pairs of: (1) sequenced reference genomes across bacterial species from all available phyla, (2) sequenced reference genomes across species from Bacteroidetes and Firmicutes phyla (3) metagenomes across the 11 gut bacterial species used in this study (belonging to Bacteroidetes and Firmicutes). Only species with at least 10 sequenced reference genomes or metagenomes are included. Each boxplot corresponds to a pooling of all available species (based on pairwise comparisons between two samples of the same species). The differences observed in metagenomes were significantly higher than the ones observed in completely sequenced genomes, even when considering reference genomes from the same phyla.

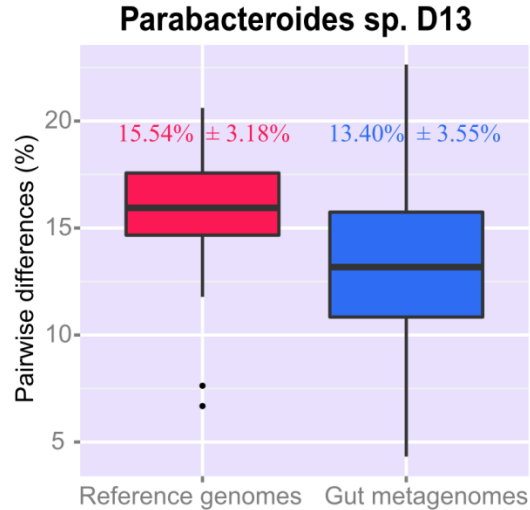


Figure 11: Variability between sequenced reference genomes and metagenomes for *Parabacteroides distasonis*

Boxplots show differences in number of genes (%) between pairs of sequenced reference genomes and pairs of metagenomes of P. distasonis. The differences observed in metagenomes were in similar ranges as found for sequenced reference genomes.

The large inter-individual variation in gut bacteria's strain's gene content implies considerable structural variation that needs to be factored into interpretation of metagenomic studies (note that gene content variation covers a large proportion of structural variation in prokaryotes due to high coding density). In addition, the structural variability of gut bacterial strains across individuals (based on gene content variation as a proxy) is considerably larger compared to that of human genomes, since less than 1% of base pairs in structurally variable regions are different between two individuals human genome [52]. This large inter-individual variation in gut bacteria's strain gene content could be because of a particularly high frequency of HGT events in the gut compared to any other human body site or non-human habitats [141], which has been linked to antibiotic usage (e.g. tetracycline) and inflammation [142, 143]. Independently of the underlying mechanisms, I found that the concept of individuality based on SNP variations [98] also holds true at the level of gene content variations, at least within this limited data set.

3.2.4 Accessory genes are enriched in mobile elements and functions associated with cell wall and membrane

To determine the functions that contribute the most to individual variability in bacterial gene content, the core and accessory genes were mapped to orthologous groups (OGs) and each OG to their respective functional categories [26]. Expectedly, variation in bacterial gene content was associated with functions related to mobile elements, like recombination (this functional category includes several transposases and viral proteins), and defence mechanisms (such as modification-restriction systems [144], ABC-type antimicrobial and multidrug transporters), in **Figure 12**. Accessory genes were also enriched in functions pertaining to cell wall and cell membranes, which is dominated by genes encoding glycosyltransferases (33%). Glycosyltransferases are important for modification of surface epitopes like capsular polysaccharides, O-antigens and exopolysaccharides. The large diversity of glycosyltransferases may help the bacteria in colonizing the gut environment [145, 146]. In agreement with my observations, glycosyltransferases have been associated with HGT in Bacteroidetes living in the gut [145]. Finally, accessory genes were also enriched in unknown genes, suggesting that there is a large range of unexplored functions that could potentially have an impact on an individual's phenotype. On the other hand, the core genes were mainly associated with genes involved in essential functions such as translational and ribosomal related genes, as well as amino-acid transport and metabolism, **Figure 12**.

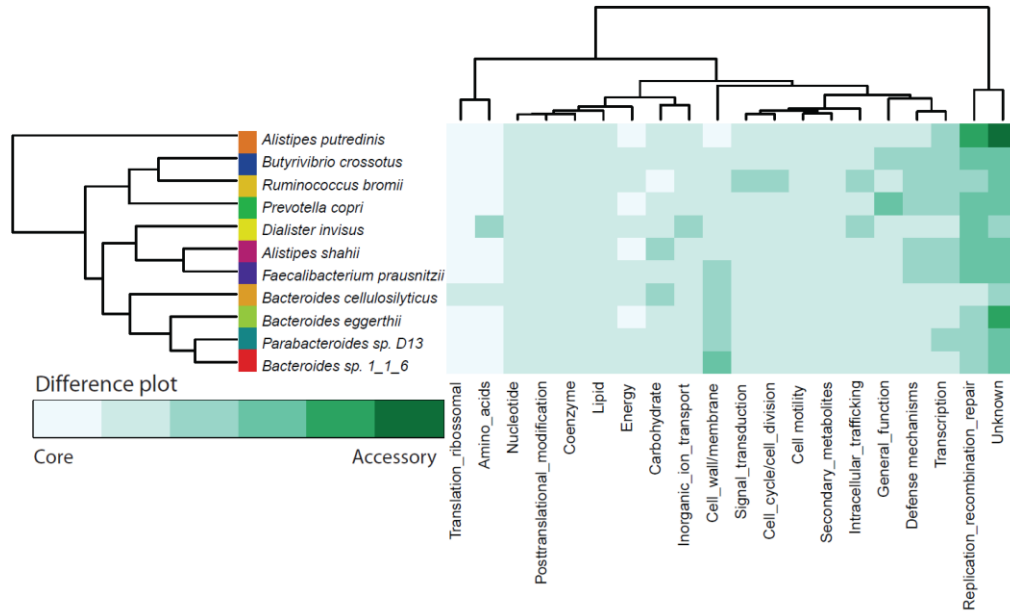


Figure 12: Difference plot of orthologous group’s functional categories between gene number of core and accessory genes

The difference plot shows the difference between the numbers of core genes and the number of accessory genes which belong to a certain functional category. Darker green shades corresponds to functional categories with higher ratio between accessory genes than core genes.

3.2.5 Single gene deletions are highly abundant and associated with mobile elements

To study the effect of gene content variability on bacterial genome architecture I took an approach based on gene deletion blocks [30, 31]. I define a gene deletion block as a group of contiguous accessory genes that are absent in one individual when compared to the reference and a single-gene deletion block as a single gene that is absent, but whose neighbouring genes are certainly present in order to have a very strict criterion (**Table 1**). The following analysis, contrary to the previous sections are focused on accessory genes that are absent

instead of present. I focus on absent genes because this allows the study of genome architecture without requiring the determination of gene origin, age and ancestry within the genome [30]. Gene deletion blocks were found in each metagenomic sample by comparison with the reference genome. Please note that the gene deletion blocks detected can arise either by gene deletion within the strains of an individual or by gene insertion(s) in the reference genome. I determined the number of gene deletion blocks and the number of genes contained in each block for a given species in a metagenomic sample (**Figure 13a** and **Figure 14**).

I found that the most frequent number of genes in a gene deletion block was single-gene deletion block, corresponding to a mean of 33.74% of all blocks and 25% of all deleted genes (**Figure 13** and **Figure 14**). Across the 11 species, several ATPases, transcription and recombination related proteins (e.g. retron-type reverse transcriptase, transcriptional regulators and recombinases) were at the top of functions found in single-gene deletion blocks (**Appendix 10**). These functional categories clearly associate single gene deletion blocks to mobile elements and the functional nature of the genes involved supports hypothesis claiming that previously integrated elements underwent erosion through deletion of their mobilization and integration machinery [147], even though I cannot exclude the possibility that some of these are gene insertions occurring in the reference genome.

3.2.6 Accessory genes have functions that imply phenotypic differences of an individual

Apart from the single-deletion blocks, I also detected large-gene deletion blocks with 50 or more genes in several species, with the longest containing 172 genes. These blocks have 50 to 172 genes, and their sizes ranges between 37Kb and 135Kb (**Figure 13a**, **Figure 13b** and **Appendix 11**). On average per species 21% of all deleted genes were found in these large-gene deletion blocks (**Figure 14**), which contain substantial number of operons that are likely to be integrated

into active mobile elements. Indeed, these large-gene deletion blocks have large integrons containing functions that may confer functional difference between strains (such as a likely queuosine biosynthetic pathway, several peptidases, that can for example be involved in the lysis of cell wall peptidoglycan, and a toxin-antitoxin system), **Appendix 12**. Since differences in large-gene deletion are likely to have phenotypic consequences in the respective individuals, they were studied in more details. In total 21 large-gene deletion blocks were found in 8 species, and each species had between 1 and 7 large-gene deletion blocks. Not unexpectedly, these deletion blocks were often associated with prophages of both Bacteroidetes and Firmicutes or conjugative transposable elements for Bacteroidetes (**Appendix 11**) implying a mechanism for the transfer of functionality.

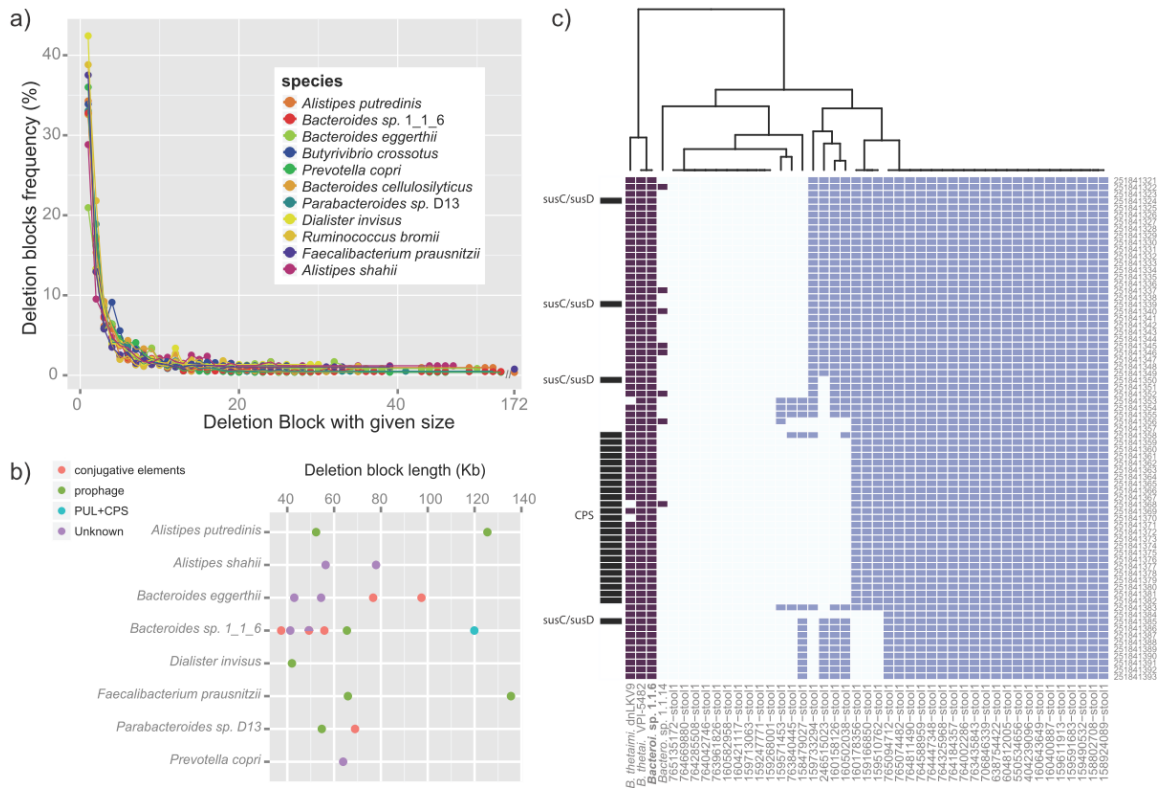


Figure 13: Gene deletion block size distribution

a) Frequency of gene deletion blocks with a given number of contiguous absent genes (%) for 11 gut bacterial species. Each point corresponds to the mean frequency observed across 10 individuals for a given species. **(b)** Length of different large-gene deletion blocks

(50-172 genes) found in 8 gut bacterial species. Several large-gene deletion blocks are associated with prophages and conjugative elements and one of these blocks contains at least PULs loci and a CPS locus. (c) This large-gene deletion block in *Bacteroides* sp. 1_1_6 is shown in more detail across 41 metagenomes and 4 sequenced reference genomes. The sequenced reference genomes are shown in the first four columns in purple and the metagenomes are shown in the remaining columns in blue, whenever the corresponding genes are present. The reference genome used for metagenomes mapping is highlighted in bold. The genes are labelled by its NCBI sequence identifier number (GI). Annotation of *SusC/SusD* and CPS annotation are based on Xu et al. 2003 [148]. Three *SusC/SusD* genes are found upstream of the CPS locus which can be associated with at least one PUL, and one *SusC/SusD* is found downstream of the CPS indicating the existence of another PUL. These PULs have been associated with plant carbohydrate degradation [66]. For the majority of the individuals except one (where one *SusC/SusD* was missing), the CPS and both PULs related sub-regions, in the individuals where the sub-region is present, they show a conserved modularity. The results in metagenomes are also confirmed with sequenced strains, with the CPS and PULs loci present in strains and absent in one.

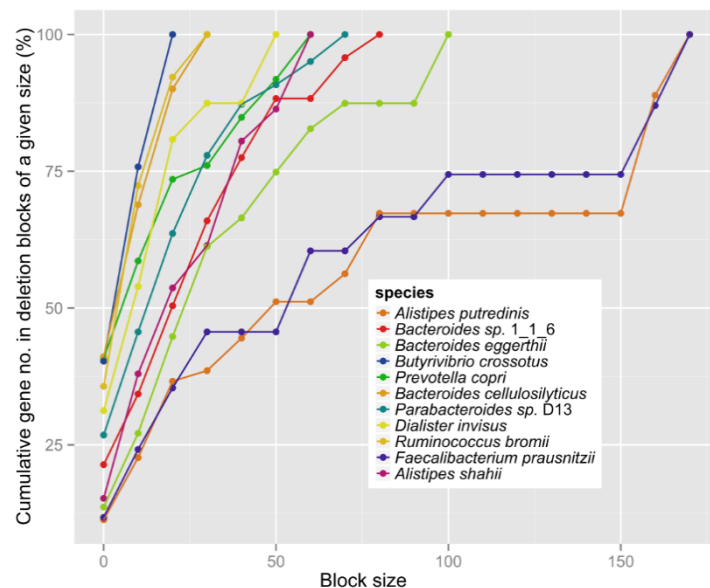


Figure 14: Cumulative number of genes in deletion blocks of a given size.

The total number of genes (%) that are absent in a metagenome and are located in a deletion block with a size smaller or equal to a given block size (x-axis) is plotted for each

of the 11 gut bacterial species. The block size (x-axis) corresponds to the number of consecutive absent genes and the block sizes were binned in bins of sizes multiples of 10. Each data point corresponds to the mean across 10 individuals.

One of these large-gene deletion blocks found in *Bacteroides* sp. 1.1.6 (*B. thetaiomicron*) contains four *susC/susD* genes that are associated with polysaccharide utilization loci (PULs) and one capsular polysaccharide synthesis locus (CPS) containing 25 genes, **Figure 13c**. Bacterial PUL helps the human intestine to forage glycans and polysaccharides [33, 34]. The two PULs detected are associated with plant carbohydrate degradation in the type strain of this species (*Bacteroides thetaiotaomicron* VPI-5482 ATCC) [148, 149]. Type strain denotes the nomenclatural type of a species or subspecies. CPS loci are sensitive to the nutrient availability and are involved in the defense of the bacteria against environmental factors (e.g. host immune system, phage attack and anti-peptide produced by the human body or by other bacteria [150, 151]). Comparison between individuals show that the gene deletion patterns in this large-gene deletion block are further separated into at least three different sub-regions of consecutive-gene blocks **Figure 13c**, with two sub-regions containing PULs and one corresponding to the CPS. Surprisingly, each sub-region is present in some but missing in other individuals independently of the other sub-regions. This pattern is observed not only for the initially randomly chosen 10 individuals, but extends to all individuals where *Bacteroides* sp. 1.1.6 species was detected (41 in total).

The region containing this large-gene deletion block was also tested for presence in other four completely sequenced genomes of *B. thetaiotaomicron* (**Figure 13c**). The whole region was present in the type strain, *Bacteroides thetaiotaomicron* VPI-5482 ATCC, while some genes were missing in the other two strains. In one strain, the region suffered single-gene deletions with the majority of the accessory genes being present, and in another strain, large-gene deletion blocks were found with only a few genes present and not in consecutive-gene blocks. Therefore, the variation in gene content of this large-gene deletion block

observed in metagenomes is in accordance with the variation observed in completely sequenced genomes. Before functional inference of the results, paralogs of the genes in this large-gene deletion blocks had to be checked as they may compensate functionality at an alternative loci. Functional compensation by paralogs have been observed for some glycan-modifying enzymes of *B. thetaiotaomicron* [145], however for the genes in this large-gene deletion block no paralog was found. Hence, in some individuals the two PULs and the CPS seems to be completely absent in their respective *B. thetaiotaomicron*, unless the individuals have the corresponding functionality in insertions that my reference-based methodology is not able to measure. This would limit the *B. thetaiotaomicron* potential for polysaccharide utilization and capsular polysaccharide synthesis and due to the central nature of this functionality for Bacteroides, change the expected phenotype drastically [152, 153].

3.2.7 Discussion

The goal of this project was to create a methodology to robustly measure the gene content of bacterial species within a metagenomic samples based on gene presence or absence. The method, which was applied to 11 abundant species, only accounts for gene deletion and therefore only provides a lower estimate of gene content variability. For each species the core and accessory genes were determined. Using this method I calculated the fraction of accessory genes for each species, and found that fraction of accessory genes correlate with genome size. This results is in line with a number of genome size “scaling laws” which show that gene functional classes scale differently with genome size [154–156], also this concordance shows the robustness of my metagenomic-based method. For example, the number of transcriptional factors, two component system, signal transduction genes increase more than linearly with genome size [155]. It is also in agreement with the observation that larger genomes have higher rates of HGT compared to smaller genomes [157]. Moreover, I found that my metagenomic approach is able to capture a similar range of gene content differences as observed for pan-genomes. This indicates that metagenomic based

comparisons do not show large systematic differences between gene content estimations from the ones obtained in isolated strains, demonstrating the validity of my proposed method.

Another goal of this project was to determine the degree of variability in gene content between gut bacterial strains of the same species. For this purpose, I performed pairwise comparisons between two individuals and found that species differ on average by 13%, and found that this variation was considerably larger than variation detected between biological or technical replicates. Also, no two individuals shared the same gene content, suggesting that individuality might also occur in gene content level as we showed for SNP variability. The proper demonstration of individuality in gene content level will however require a larger sample size and more gut microbial species.

To study the effect of variation in gene content on the architecture of bacterial genome I used an approach based on gene deletion [158]. More specifically, the findings pertaining to the PULs support the idea that carbohydrate degradation is strain-specific as has also been found experimentally in *Bifidobacteria* [159]. In this genus, strains have different key enzymes which are able to diggest different carbohydrate sources. These differences in carbohydrate utilization potential are likely to be a reflection of differential niche adaptation [160–162]. For example, comparison between *B. thetaiomicron* and *B. ovatus* has shown that each species acquires niche-specific PULs [162], which could be an effect of an individual's dietary habits [160, 162]. The latter effect could also be influenced by strain-specific CPS architecture, since expression of *B. thetaiomicron* CPS has also been coupled with dietary changes [149, 151, 163] and since there is indication of coordinated regulation of CPS and PULs [151]. The response of CPS to dietary changes is likely to help the bacteria creating a capsule that mimics the glycan composition found in the individual's gut and also is likely to affect the interaction between the bacteria and the individual's immune system [150, 153]. In summary, CPS and PULs show that important functions such as

carbohydrate degradation can be present or absent in some individuals, indicative that strain information needs to be accounted for phenotypic inferences.

CHAPTER 4

CONCLUSIONS

Species in the gut microbiota play several functions that are important for human health. Metagenomics have enabled the study of gut species in their natural habitat, however most studies focus on either changes in taxonomical composition or in relative abundance. Since the publication of the first pan-genome for *Streptococcus agalactiae* [63], several studies have shown that microbial species can differ greatly in their strains both in small-scale and large-scale rearrangements and also that these variations are phenotypically relevant. However the extent of these variations for gut microbial species and its phenotypic impact in the microbial strains or human host are still unknown.

My main goal of my doctorate studies was to investigate whether using metagenomes it is possible to infer phenotypic impact based on genetic variation (SNPs and gene content as a proxy for structural variations) of gut prokaryotic species. I showed for the first time that both SNPs (**sub-chapter 3.1**) and gene content variations (**sub-chapter 3.2**) can be used for functional inferences of strains in the complex gut community, which was not possible using previous metagenomic approaches based on relative abundances or taxonomical classification.

In **sub-chapter 3.1**, I used a 10.3 million SNP catalogue that we generated to study their phenotypic effect. Due to the size of the SNP catalogue it is currently not possible to infer the effect of single SNP's, instead I evaluated whether evolution could be used for functional inferences. For this purposes I used pN/pS ratio, a SNP based method used in population genetics for tracking recent evolutionary events, and the results have been published in Nature [98]. I found pN/pS ratios to be fairly constant across individuals in contrast to variation within an individual. Also, I exemplified using *galK* that indeed functional potential can be inferred using gene evolution based on pN/pS ratio. Here *galK* was shown to have different evolutionary histories in different species (*Roseburia intestinalis* and

Eubacterium eligens) and this is likely a reflection of these species different capacities in utilizing carbohydrate sources (galactose and galactose derivatives). The different evolutionary outcome of these species, suggest that diet can have a strong effect in the genetic composition of gut species. Although gene evolution (pN/pS ratio) can be use to infer functional inferences, for most genes it is still hard to interpret the functional outcome of SNP variation. Alternative, differences in gene presence or absence are easier for estimation of functional potential.

In **sub-chapter 3.2**, I created a methodology for metagenomes based on gene deletion to detect if genes are present or absent within individual gut strains that was applied to 11 abundant species. The results in this sub-chapter are under minor revision in a publication sent to Genome Biology where I was the leading author. The method is sufficiently robust to find general rules found in prokaryotic species such as the scaling of accessory genes with genome size and is able to capture similar range of gene content variation as classical approaches based on pan-genomes. Hence the method is reliable for application in other complex natural habitats. Despite the limited sample size, I found large inter-individual differences corresponding to an average of 13%. Gene content of gut strains were found to be unique to each individual as we found for SNP pattern [98] although these results needs to be validated with more species and dataset. Importantly, these differences were associated with present or absence of important loci such as PUL and CPS indicative that strain resolution needs to be accounted for functional inferences.

In conclusion, genetic variation both in SNPs and gene content of strains can be used for phenotypic inferences and these affect important functions such as an individual digestive capacities.

APPENDICES

SUPPLEMENTARY TABLES

The supplementary tables that are not displayed in the current document are found in appendix in the CD that accompanies this thesis.

Appendix 1 Information regarding 252 metagenomes used in this thesis.

Details regarding 252 samples are listed, including sample name, data source, subject ID, sampling time point, continent which the sample was originated (NA stands for North America and EU stands for Europe). Statistics about the number of high quality sequencing reads or bases, average read length, and mapping rate to the 929 non-redundant set of reference genomes are shown. The maximum mapping rate for a single sample corresponded to 75.1% and the minimum rate to 11.2%.

Appendix 2 Information regarding 1,497 reference genome used in this thesis.

*Details regarding 1,497 reference genomes are listed, including each genome NCBI taxonomy identifier and species or strain name (column one and two). 14 genomes from the initial set of 1,511 were excluded because no reads in the 252 metagenomes mapped to these genomes. Genomes were clustered into 929 non-redundant species (**section 2.1.2**) and sorted per cluster. For each cluster the representative genome is highlighted in bold. The last column categorizes genomes into either dominant, prevalent or non-prevalent (**section 2.2.1**). In total there are 66 dominant genomes and 101 prevalent genomes (dominant genomes are also considered in the group of prevalent genomes). Genomes that are not selected to be a representative genome are referred as clustered.*

Appendix 3 Statistics regarding number of reads mapped to reference genomes.

*A total of 7.4 billion reads out of the initial 17.85 billion reads, that is 41.6% of the reads (in the 252 metagenomes) could be mapped to the set of 929 reference genomes (**Appendix 2**). Each row represents a reference genome and is identified by NCBI TaxID*

and species or sample name (column 1 and 2). The reference genomes are sorted by decreasing numbers of reads that were mapped across all 252 samples (column 3). The cumulative percentage (column 4) represents the number of reads in all samples that are incrementally summed up and reported as percentage of the total number of mapped reads. The last 252 columns shows the number of reads mapped to each reference genome. The last row corresponds to a count of the total number of unmapped reads in all samples (column 3), total fraction of unmapped reads across samples (column 4), and the remaining cells show the number of unmapped reads in a given samples.

Appendix 4 Table with pN/pS ratios calculated for each species-individual pair.

Each cell corresponds to the average pN/pS ratio calculated for a given species-individual pair across all genes. The pN/pS ratios were calculated for 66 dominant species and 207 individuals (corresponding to the first time-point of an individual, and includes 97 Americans, 39 Spanish and 71 Danish). The pooled column corresponds to pN/pS ratios calculated from SNPs pooled from all samples, whereas the average column corresponds to average pN/pS ratios across samples. The values for pooled column are similar to the ones in the average column. NA corresponds to samples where the genome coverage was not sufficient to call the genome present in a sample, **section 2.2.3**.

Appendix 5 List of top fastest and slowest evolving orthologous groups.

List of the 70 fastest (highest pN/pS ratio) and slowest (lowest pN/pS ratio) OGs in dominant gut species. The pN/pS ratio corresponds to the average of all genes from a given orthologous group across 252 samples and 66 species. Unexpectedly among the lowest pN/pS ratios there are genes COG3451 and COG3505 that are related with type IV secretion systems. Among the highest pN/pS ratios there is a gut specific OG, COG3328 a bile salt hydrolase.

Appendix 6 List of unknown conserved genes among slowest evolving genes.

14 genes whose function is currently unknown were found among the gut microbial genes that have the lowest average pN/pS ratios in at least 126 samples. In ProteinIDs column the identifiers corresponds to the NCBI GI number.

Appendix 7 List of top fastest and slowest evolving genes in *E. eligens* and *R. intestinalis*.

Median pN/pS ratio is displayed for all genes from *E. eligens* and *R. intestinalis* that can be mapped to OGs. For a total of 1.153 genes in *E. eligens* and 1.917 genes in *R. intestinalis* the median pN/pS ratio was calculated for 207 samples (first time point for an individual). 611 OGs were shared between the two species, for each of these OGs the \log_2 ratio between the pN/pS ratios of *E. eligens* and *R. intestinalis* are shown. Among these OGs, COG0153 Galactokinase has the highest ratio between *E. eligens* and *R. intestinalis* and Na^+ /proline symporter is one of the OGs with lowest ratio, both showing cases where the evolution of the gene differs from the species evolution. The proteinIDs corresponds to the NCBI GI numbers.

Appendix 8 List of randomly chosen 10 individuals for each of the 11 species.

Each species can be identified by the representative reference genome NCBI TaxID and strain name. For each species a list of metagenomes names from 10 randomly selected individuals that were used in **sub-chapter 3.2** are shown.

Appendix 9 List of Genomic islands detected by IslandViewer

List of genomic islands detected in 11 species, the location in the contig and length of each genomic island are shown.

Appendix 10 List of OG that are associated with single-gene deletion blocks.

Description of the number of genes in an OG, based on eggNOG, that are found in single-gene deletion blocks. The number of occurrences counts the number of genes of a given OG found across the 11 species. The OGs are sorted by the number of occurrences.

Appendix 11 Mobile elements annotation associated with the 21 large deletion block.

NCBI taxID	start	end	Deletion size	contigID	Annotation
445970	100528	152840	52.312	445970.DS499577	phage proteins
445970	966316	1091786	125.470	445970.DS499577	phage proteins
469586	83	37485	37.402	469586.GG695913	conjugate transposon
469586	89226	130510	41.284	469586.GG695902	phage protein
469586	531042	580320	49.278	469586.GG695902	conjugate transposon
469586	8913	58500	49.587	469586.GG695904	transposase
469586	167866	223715	55.849	469586.GG695904	conjugate transposon, AraC proteins
469586	466497	531978	65.481	469586.GG695899	phage proteins
469586	754106	873999	119.893	469586.GG695900	Lipopoylsaccharide biosynthesis, xylosidase, glycosyl transferase, AracC, Arylsulfatase A
483216	390037	433022	42.985	483216.DS995510	NA
483216	598880	653321	54.441	483216.DS995508	transposase and a phage proteins
483216	511759	588443	76.684	483216.DS995509	conjugate transposon
483216	474391	571712	97.321	483216.DS995508	conjugate transposon, some integrases
537011	26742	90606	63.864	537011.GG703858	glycosultransferase, lipopolysaccharides, integrase and transposase
563193	22957	77703	54.746	563193.GG698743	phage proteins
563193	723512	792457	68.945	563193.GG698739	conjugate transposon, phage proteins

592028	1263683	1305684	42.001	592028.GG698602	phage
657322	2692869	2758766	65.897	657322.FP929046	phage
657322	2248043	2383580	135.537	657322.FP929046	phage proteins
717959	979506	1035890	56.384	717959.FP929032	proteins found in conjugate transposon, integrase
717959	152741	230659	77.918	717959.FP929032	NA

Each large deletion block is identified by its location in the representative reference genome. To identify the location the NCBI TaxID of the representative reference genome, contig, start and end nucleotide positions are listed. The large deletion blocks were annotated based on eggNOG [164], KEGG [165] and MetaCyc [166]. The deletion blocks are annotated as phage or conjugative transposons if the whole machinery is present or defined as such in MetaCyc. Deletion blocks are annotated as phage or conjugative transposon proteins if only some genes but not the whole machinery is annotated. NA is used for large deletion blocks where none of the genes are annotated with functions associated with mobile elements. The size of the deletion blocks are expressed in kilobases. Many large deletion blocks are observed to be related with prophages and conjugative transposons.

Appendix 12: Examples of function encoded in large deletion blocks that are likely to differ between strains.

Region	NCBI TaxID	NCBI GI	OG	OG description	KEGG ko	Additional Annotation	Additional Annotation source
Region 1	469586	251838631	COG2856	Predicted Zn peptidase	NA	NA	NA
Region 1	469586	251838633	COG0739	Membrane proteins related to metalloendopeptidases	NA	Peptidase M23-involved in lyse of cell wall peptidoglycan	MEROPS
Region 1	469586	251838643	COG0739	Membrane proteins related to metalloendopeptidases	NA	Peptidase M23-involved in lyse of cell wall peptidoglycan	MEROPS
Region 2	445970	167658993	COG0302	GTP cyclohydrolase I	K01495	Folate metabolism- queuosine biosynthetic pathway	KEGG
Region 2	445970	167658995	COG0720	6-pyruvoyl-tetrahydropterin synthase	K01737	Folate metabolism- queuosine biosynthetic pathway	KEGG
Region 2	445970	167658996	COG1738	Uncharacterized conserved protein	K09125	Folate metabolism- queuosine biosynthetic pathway	KEGG
Region 2	445970	167658997	COG0603	Predicted PP-loop superfamily ATPase	K06920	Folate metabolism- queuosine biosynthetic pathway	KEGG
Region 3	657322	locus_tag: FPR_27200	NOG119748		NA	Peptidase M15A - metallopeptidases, mostly specialised carboxypeptidases and dipeptidases	MEROPS
Region 4	483216	217988239	COG0739	Membrane proteins related to metalloendopeptidases	NA	NA	NA
Region 4	483216	217988249	COG0739	Membrane proteins related to metalloendopeptidases	NA	NA	NA

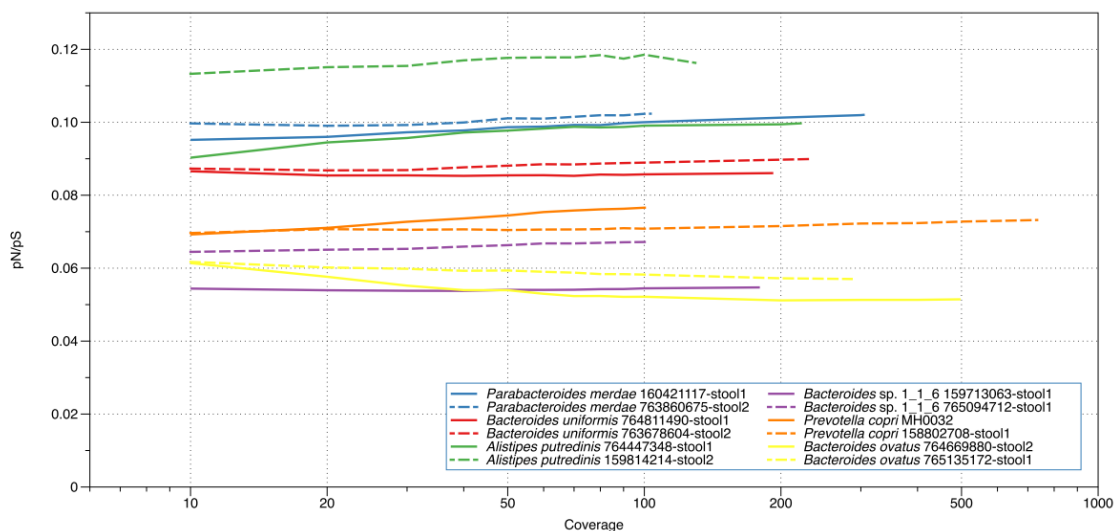
Region 5	483216	217988965	COG0739	Membrane proteins related to metalloendopeptidases	NA	NA	NA
Region 6	592028	260403808	NA	NA	NA	Putative lipoprotein	MetaCyc
Region 6	592028	260403809	COG2856	Predicted Zn peptidase	NA	Putative toxin-antitoxin, toxin component	MetaCyc
Region 6	592028	260403810	COG1396	Predicted transcriptional regulators	NA	Putative toxin-antitoxin, antitoxin component	MetaCyc

Genes found in a given large deletion block are grouped by regions in the deletion block. Each region can contain more genes than the ones listed in the table, however for most of unlisted genes their functions are unknown. Only the relevant genes involved in queuosine biosynthetic pathway (Folate metabolism), toxin-antitoxin system and peptidases are described. Each gene is described by the representative genome of origin (NCBI TaxID), the gene NCBI identifier and their annotation to eggNOG, KEGG [165], MEROPs [167] and MetaCyc [166]. NA is used when no annotation is found in a given database.

APPENDICES

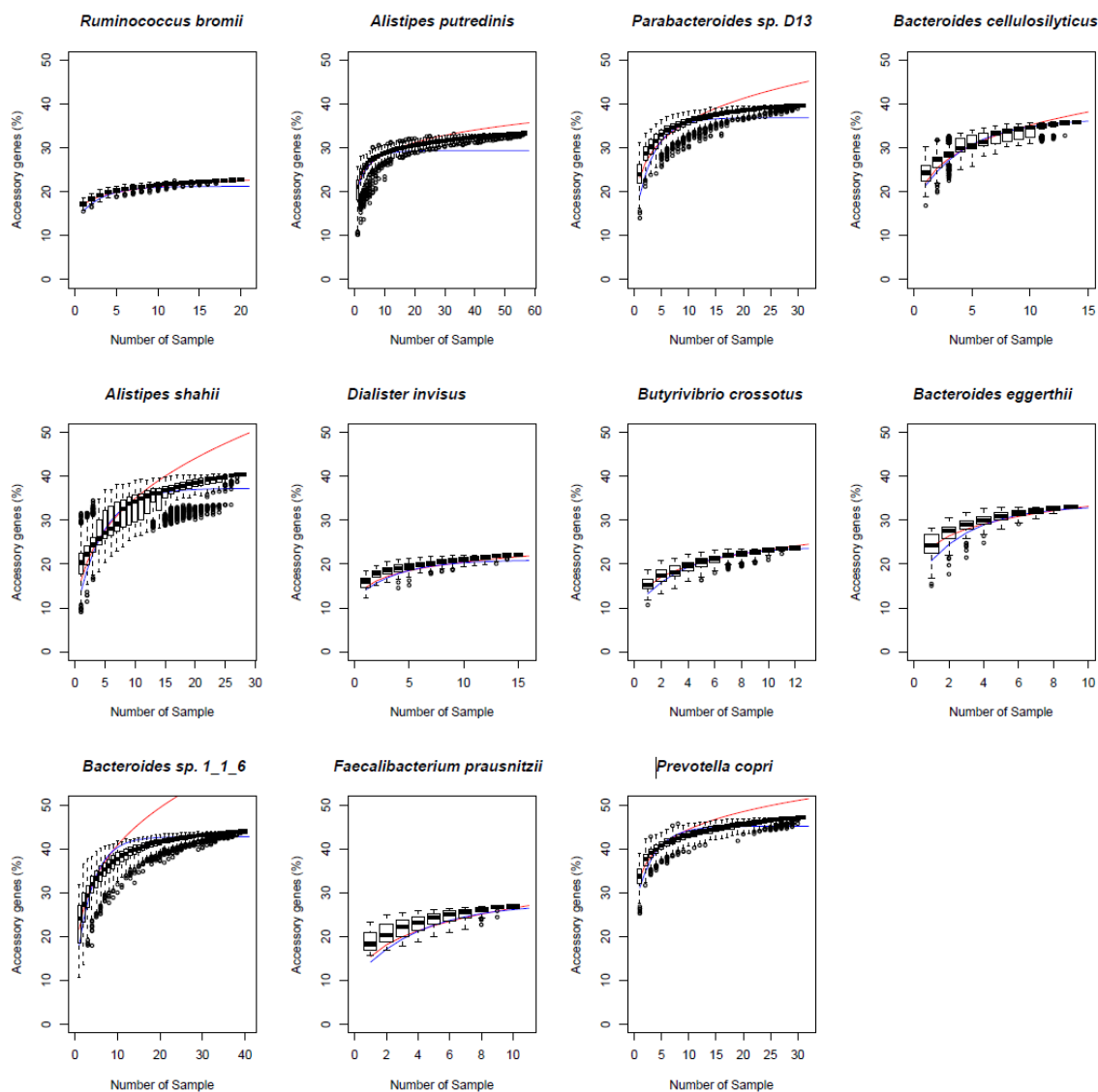
SUPPLEMENTARY FIGURES

Appendix 13 Effect of downsampling on pN/pS ratios.



SNPs of four dominant species (two samples were used per species) were downsampled starting from their native coverage down to 10x. At each downsampling step the SNPs remaining after the downsampling were used for the calculation of the genome pN/pS ratio. The plot illustrates that a strongly stable ratio across the whole coverage range for each of the eight instances. Moreover, for all the genome-sample pairs with a minimum coverage of 50x (635 pairs, in order for the downsampling to have a significant influence on the number of SNPs) we performed downsampling to 10x. Comparison between pN/pS ratio at 10x coverage with native coverage reveal that 87% of these genome-sample pairs have less than 0.01 difference between the two coverages.

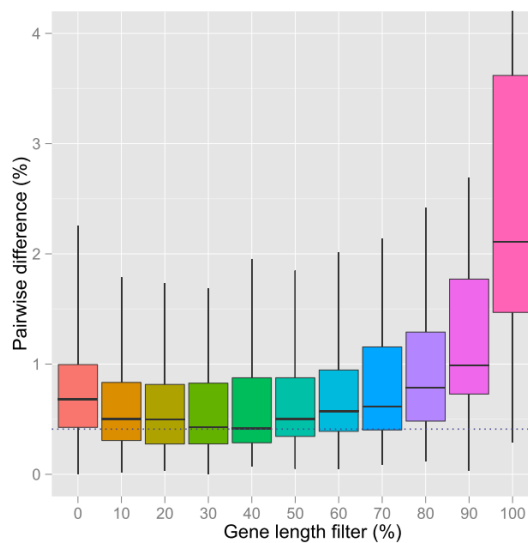
Appendix 14 Subsampling based estimation of percentage of accessory genes curve in the 11 species.



Each graph illustrates the comparison between “expected fraction” and the percentage of accessory genes estimated by the exponential regression and power law regression model. The boxplots show the “expected fraction”, blue curve plots the fitting of the exponential regression model and the red curve plots the fitting of power law regression model. “Expected fractions” were calculated as described in **section 2.3.3** based on a subsampling procedure. The two models were fitted to the median values of “subsampled-based fraction” which were based on 10 randomly chosen individuals. For small sample

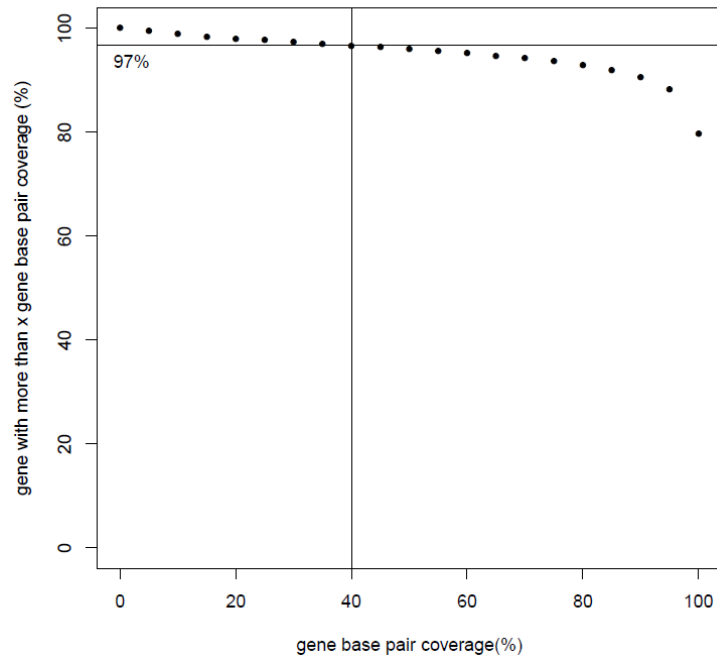
sizes both the two curve fit similarly to the “expected fractions”, as sample size increases exponential regression model curve tend to underestimate the values and power law regression model tend to overestimate. For larger sample sizes the difference between the “expected fractions” and the two curves is smaller for exponential regression model compared to power law regression model.

Appendix 15 Variation in gene content across the 11 species between biological replicates using several gene length coverage filters.



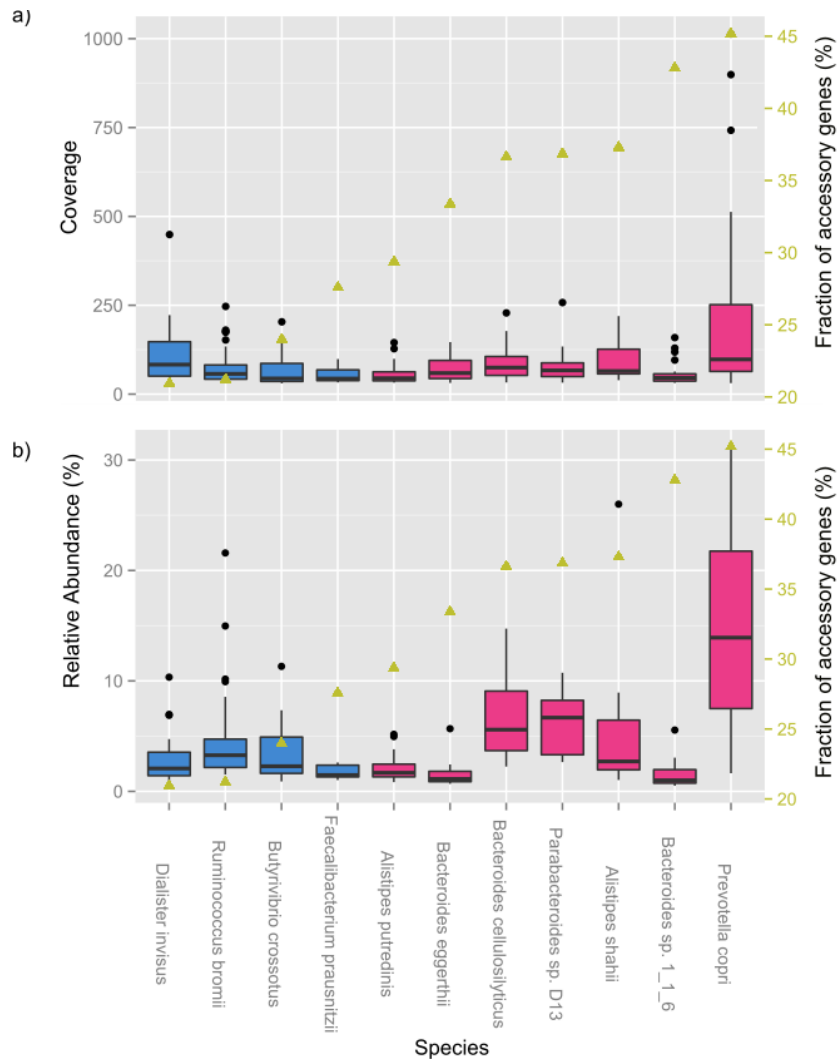
Each boxplot represents the differences in gene content between two biological replicates (that is two time-series for a given individual) after applying a given gene length coverage filter. The gene length coverage filter corresponds to the fraction of a gene length that has been covered by reads. The filters ranged between 0% and 100% (in intervals of 10%). The figure shows that the minimum average variability is found at a gene length coverage filter of 40%.

Appendix 16 Fraction of genes with at least a given gene length coverage.



Dotplot represents the percentage of genes that are called absent in all species-individual pairs (from the 11 species) when the gene length coverage filter is set at a given x value (in x axis). The gene length coverage filter corresponds to the fraction of a gene length that has been covered by reads. With the chosen gene length coverage filter of 40%, 3% of all genes had reads mapped and are considered as a result of spurious read mapping or homology to a closely relative species and therefore are regarded as absent.

Appendix 17 Percentage of accessory genes is not dependent on genome abundance nor genome coverage.



Boxplot shows the (a) depth of genome coverage and (b) relative abundance of each species within an individual. The yellow triangle represents the fraction of accessory genes observed across 10 randomly chosen individuals. Species are sorted by the fraction of accessory genes and boxplots are colored according to the respective phylum a species belong to, pink for Bacteroidetes and blue for Firmicutes.

BIBLIOGRAPHY

1. Xu J, Gordon JI: **Honor thy symbionts.** *Proc Natl Acad Sci* 2003, **100**:10452–9.
2. Jones ML, Ganopoulos JG, Martoni CJ, Labbé A, Prakash S: **Emerging science of the human microbiome.** *Gut Microbes* 2014, **5**.
3. Savage DC: **Microbial ecology of the gastrointestinal tract.** *Annu Rev Microbiol* 1977, **31**:107–33.
4. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308**:1635–8.
5. Lay C, Rigottier-gois L, Holmstrøm K, Rajilic M, Vaughan EE, Vos WM De, Collins MD, Thiel R, Namsolleck P, National I, Recherche D, Cedex J, Al LAYET, Icrobiol APPLNM: **Colonic Microbiota Signatures across Five Northern European Countries.** *Appl Environ Microbiol* 2005, **71**:4153–4155.
6. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR: **Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.** *Proc Natl Acad Sci* 2007, **104**:13780–5.
7. Zoetendal EG, Rajilic-Stojanovic M, de Vos WM: **High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota.** *Gut* 2008, **57**:1605–1615.
8. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borrueal N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M,

Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, et al.: **Enterotypes of the human gut microbiome.** *Nature* 2011, **180**:1–7.

9. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li S, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, et al.: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**:59–65.

10. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh S a, Bewtra M, Knights D, Walters W a, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD: **Linking long-term dietary patterns with gut microbial enterotypes.** *Science* 2011, **334**:105–8.

11. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, Rosenbaum M, Gordon JI: **The long-term stability of the human gut microbiota.** *Science* 2013, **341**:1237439.

12. Gilmore MS, Ferretti JJ: **Microbiology. The thin line between gut commensal and pathogen.** *Science* 2003, **299**:1999–2002.

13. Bäckhed F, Ding H, Wang T, Hooper L V, Koh GY, Nagy A, Semenkovich CF, Gordon JI: **The gut microbiota as an environmental factor that regulates fat storage.** *Proc Natl Acad Sci* 2004, **101**:15718–23.

14. Turnbaugh PJ, Ley RE, Mahowald M a, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006, **444**:1027–31.

15. Bäckhed F, Manchester JK, Semenkovich CF, Gordon JI: **Mechanisms underlying the resistance to diet-induced obesity in germ-free mice.** *Proc Natl Acad Sci* 2007, **104**:979–84.

16. Ley RE, Turnbaugh PJ, Klein S, Gordon JI: **Human gut microbes associated with obesity.** *Nature* 2006, **444**:1022–1023.
17. Duncan SH, Belenguer A, Holtrop G, Johnstone AM, Flint HJ, Lobley GE: **Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces.** *Appl Environ Microbiol* 2007, **73**:1073–8.
18. Grabitske HA, Slavin JL: **Low-Digestible Carbohydrates in Practice.** *J Am Diet Assoc* 2008, **108**:1677–1681.
19. Tremaroli V, Bäckhed F: **Functional interactions between the gut microbiota and host metabolism.** *Nature* 2012, **489**:242–9.
20. Selvendran R: **The plant cell wall as a source of dietary fiber: chemistry and structure.** *Am J Clin Nutr* 1984, **39**:320–337.
21. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL, Coutinho PM, Henrissat B, Leclerc M, Doré J, Monsan P, Remaud-Simeon M, Potocki-Veronese G: **Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes.** *Genome Res* 2010, **20**:1605–12.
22. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.
23. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, et al.: **An integrated catalog of reference genes in the human gut microbiome.** *Nat Biotechnol* 2014, **32**:834–841.

24. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B: **The abundance and variety of carbohydrate-active enzymes in the human gut microbiota.** *Nat Rev Microbiol* 2013, **11**:497–504.
25. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, Pudlo N a, Muegge BD, Henrissat B, Hettich RL, Gordon JI: **Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome.** *PLoS Biol* 2013, **11**:e1001637.
26. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto J-M, Renault P, Doré J, Zucker J-D, Clément K, Ehrlich SD, Blottière H, Leclerc M, Juste C, de Wouters T, Lepage P, Fouqueray C, Basdevant A, Henegar C, Godard C, Fondacci M, Rohia A, Hajduch F, et al.: **Dietary intervention impact on gut microbial gene richness.** *Nature* 2013, **500**:585–8.
27. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, et al.: **Richness of human gut microbiome correlates with metabolic markers.** *Nature* 2013, **500**:541–6.
28. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J: **Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.** *Gut* 2006, **55**:205–11.
29. Lepage P, Häsler R, Spehlmann ME, Rehman A, Zvirbliene A, Begun A, Ott S, Kupcinskas L, Doré J, Raedler A, Schreiber S: **Twin study indicates loss of**

interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology* 2011, **141**:227–36.

30. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O'Sullivan O, Fitzgerald GF, Deane J, O'Connor M, Harnedy N, O'Connor K, O'Mahony D, van Sinderen D, Wallace M, Brennan L, Stanton C, Marchesi JR, Fitzgerald AP, Shanahan F, Hill C, Ross RP, O'Toole PW: **Gut microbiota composition correlates with diet and health in the elderly.** *Nature* 2012, **488**:178–84.

31. Miyakawa M, Kanzaki M, Kotake A: **Vitamin B-6 Deficiency in Germfree Rats¹.** *J Nutr* 1977, **107**:1707–1714.

32. Wostmann BS: **The germfree animal in nutritional studies.** *Annu Rev Nutr* 1981, **1**:257–279.

33. Hooper L V, Midtvedt T, Gordon JI: **How host-microbial interactions shape the nutrient environment of the mammalian intestine.** *Annu Rev Nutr* 2002, **22**:283–307.

34. Mortensen F V, Nielsen H, Aalkjaer C, Mulvany MJ, Hesse I: **Short chain fatty acids relax isolated resistance arteries from the human ileum by a mechanism dependent on anion-exchange.** *Pharmacol Toxicol* 1994, **75**:181–185.

35. Mortensen F V, Nielsen H: **Short chain fatty acids dilate isolated human colonic resistance arteries.** *Gut* 1990:1391–1394.

36. Gamet L, Daviaud D, Denis-Pouxviel C, Remesy C, Murat JC: **Effects of short-chain fatty acids on growth and differentiation of the human colon-cancer cell line HT29.** *Int J Cancer* 1992, **52**:286–289.

37. Mortensen F V., Langkilde NC, Joergensen JCR, Hesse I: **Short-chain fatty acids stimulate mucosal cell proliferation in the closed human rectum after Hartmann's procedure.** *Int J Colorectal Dis* 1999, **14**:150–154.
38. Whitehead RH, Young GP, Bhathal PS: **Effects of short chain fatty acids on a new human colon carcinoma cell line (LIM1215).** *Gut* 1986, **27**:1457–1463.
39. Hamer HM, Jonkers D, Venema K, Vanhoutvin S, Troost FJ, Brummer RJ: **Review article: The role of butyrate on colonic function.** *Aliment Pharmacol Ther* 2008, **27**(October 2007):104–119.
40. Segain JP, Raingeard de la Blétière D, Bourreille a, Leray V, Gervois N, Rosales C, Ferrier L, Bonnet C, Blottière HM, Galmiche JP: **Butyrate inhibits inflammatory responses through NFkappaB inhibition: implications for Crohn's disease.** *Gut* 2000, **47**(Cd):397–403.
41. Hamer HM, Jonkers DMAE, Bast A, Vanhoutvin SALW, Fischer MAJG, Kodde A, Troost FJ, Venema K, Brummer RJM: **Butyrate modulates oxidative stress in the colonic mucosa of healthy humans.** *Clin Nutr* 2009, **28**:88–93.
42. Ashida H, Ogawa M, Kim M, Mimuro H, Sasakawa C: **Bacteria and host interactions in the gut epithelial barrier.** *Nat Chem Biol* 2011, **8**:36–45.
43. Ivanov II, Honda K: **Intestinal commensal microbes as immune modulators.** *Cell Host Microbe* 2012, **12**:496–508.
44. Lee YK, Mazmanian SK: **Has the microbiota played a critical role in the evolution of the adaptive immune system?** *Science* 2011, **330**:1768–1773.
45. Hooper L V., Littman DR, Macpherson a. J: **Interactions Between the Microbiota and the Immune System.** *Science* 2012, **336**:1268–1273.

46. Hooper L V, Stappenbeck TS, Hong C V, Gordon JI: **Angiogenins: a new class of microbicidal proteins involved in innate immunity.** *Nat Immunol* 2003, **4**:269–73.
47. Hood L, Rowen L: **The Human Genome Project: big science transforms biology and medicine.** *Genome Med* 2013, **5**:79.
48. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** 2004:931–945.
49. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299–320.
50. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**:851–61.
51. The International HapMap Consortium: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467**:52–8.
52. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–73.
53. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **135**(V):0–9.
54. Tuzun E, Sharp AJ, Bailey J a, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson M V, Eichler EE: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37**:727–32.
55. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps K a, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K,

Chen L, Malig M, et al.: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56–64.

56. Rehm HL: **Disease-targeted sequencing: a cornerstone in the clinic.** *Nat Rev Genet* 2013, **14**(April):295–300.

57. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward D V, Birren BW, Gibbs R a, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, et al.: **A catalog of reference genomes from the human microbiome.** *Science (80-)* 2010, **328**:994–9.

58. The human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature* 2012, **486**:215–21.

59. Read TD, Massey RC: **Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology.** *Genome Med* 2014, **6**:109.

60. Stecher B, Maier L, Hardt W-D: **“Blooming” in the gut: how dysbiosis might contribute to pathogen evolution.** *Nat Rev Microbiol* 2013, **11**:277–84.

61. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V: **Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species.** *Genome Biol* 2010, **11**:R107.

62. Mira A, Martín-Cuadrado AB, Auria GD, Rodríguez-valera F: **The bacterial pan-genome : a new paradigm in microbiology.** *Int Microbiol* 2010, **13**:45–57.

63. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, et al.: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”**. *Proc Natl Acad Sci* 2005, **102**:13950–5.
64. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui M El, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguéne C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, et al.: **Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths**. *PLoS Genet* 2009, **5**:e1000344.
65. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD: **Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains**. *Genome Biol* 2007, **8**:R103.
66. Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, Klintner S, Pudlo NA, Urs K, Koropatkin NM, Creagh AL, Haynes CA, Kelly AG, Cederholm SN, Davies GJ, Martens EC, Brumer H: **A discrete genetic locus confers xyloglucan metabolism in select human gut *Bacteroidetes***. *Nature* 2014, **506**:498–502.
67. Hehemann J-H, Kelly AG, Pudlo N a, Martens EC, Boraston AB: **Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes**. *Proc Natl Acad Sci* 2012:1–6.

68. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G: **Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota.** *Nature* 2010, **464**:908–12.
69. Grozdanov L, Zähringer U, Blum-Oehler G, Brade L, Henne A, Knirel Y a., Schombel U, Schulze J, Sonnenborn U, Gottschalk G, Hacker J, Rietschel ET, Dobrindt U: **A single nucleotide exchange in the wzy gene is responsible for the semirough O6 lipopolysaccharide phenotype and serum sensitivity of Escherichia coli strain Nissle 1917.** *J Bacteriol* 2002, **184**:5912–5925.
70. Bagel S, Hüllen V, Wiedemann B, Heisig P: **Impact of gyrA and parC Mutations on Quinolone Resistance, Doubling Time, and Supercoiling Degree of Escherichia col.** *Antimicrob Agents Chemother* 1999, **43**:868–875.
71. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M: **Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis.** *Nat Genet* 2013, **45**:1183–9.
72. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S: **Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.** *Nat Genet* 2013, **45**:1176–1182.
73. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley K a, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B 5 biosynthesis as a host specificity factor in Campylobacter.** *Proc Natl Acad Sci* 2013, **110**(May):1–5.

74. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden M, Wilson D, Read TD, van den Elsen J, Priest NK, Feil EJ, Hurst LD, Josefsson E, Massey RC: **Predicting the virulence of MRSA from its genome sequence.** *Genome Res* 2014.
75. Alam MT, Petit R a., Crispell EK, Thornton T a., Conneely KN, Jiang Y, Satola SW, Read TD: **Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association.** *Genome Biol Evol* 2014, **6**:1174–1185.
76. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD: **Dense genomic sampling identifies highways of pneumococcal recombination.** *Nat Genet* 2014, **46**(February):305–9.
77. Frost LS, Leplae R, Summers AO, Toussaint A: **Mobile genetic elements: the agents of open source evolution.** *Nat Rev Microbiol* 2005, **3**:722–32.
78. Thomas CM, Nielsen KM: **Mechanisms of, and barriers to, horizontal gene transfer between bacteria.** *Nat Rev Microbiol* 2005, **3**:711–721.
79. Wozniak R a F, Waldor MK: **Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow.** *Nat Rev Microbiol* 2010, **8**:552–63.
80. Toussaint A, Merlin C: **Mobile elements as a combination of functional modules.** *Plasmid* 2002, **47**:26–35.
81. Burrus V, Pavlovic G, Decaris B, Guédon G: **Conjugative transposons : the tip of the iceberg.** *Mol Microbiol* 2002, **46**:601–610.

82. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW: **Genomic islands: tools of bacterial horizontal gene transfer and evolution.** *FEMS Microbiol Rev* 2009, **33**:376–93.
83. Goodman AL, Kallstrom G, Faith JJ, Reyes A, Moore A, Dantas G, Gordon JL: **Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice.** *Proc Natl Acad Sci* 2011.
84. Lee M-C, Marx CJ: **Repeated, selection-driven genome reduction of accessory genes in experimental populations.** *PLoS Genet* 2012, **8**:e1002651.
85. Conlan S, Mijares LA, Becker J, Blakesley RW, Bouffard GG, Brooks S, Coleman H, Gupta J, Gurson N, Park M, Schmidt B, Thomas PJ, Otto M, Kong HH, Murray PR, Segre J a: **Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates.** *Genome Biol* 2012, **13**:R64.
86. Qin J, Li Y, Cai Z, Li SS, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, et al.: **A metagenome-wide association study of gut microbiota in type 2 diabetes.** *Nature* 2012, **490**:55–60.
87. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F: **Gut metagenome in European women with normal, impaired and diabetic glucose control.** *Nature* 2013, **498**:99–103.
88. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Paul I, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-king P, Tournigand C, Nhieu JT Van, Yamada T, Zimmermann J: **Potential of fecal microbiota for early-stage detection of colorectal cancer.** *Mol Syst Biol* 2014, **10**:1–19.

89. Greenblum S, Turnbaugh PJ, Borenstein E: **Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease.** *Proc Natl Acad Sci* 2012, **109**:594–9.
90. Koeth R a, Wang Z, Levison BS, Buffa J a, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L, Smith JD, DiDonato J a, Chen J, Li H, Wu GD, Lewis JD, Warriar M, Brown JM, Krauss RM, Tang WHW, Bushman FD, Lusis AJ, Hazen SL: **Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis.** *Nat Med* 2013(April):1–12.
91. Greenblum S, Carr R, Greenblum S, Carr R, Borenstein E: **Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species Article.** *Cell* 2015, **160**:1–12.
92. Simmons SL, Dibartolo G, Denev VJ, Goltsman DSA, Thelen MP, Banfield JF: **Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation.** *PLoS Biol* 2008, **6**:e177.
93. Denev VJ, Banfield JF: **In Situ Evolutionary Rate Measurements Show Ecological Success of Recently Emerged Bacterial Hybrids.** *Science* 2012, **336**:462–466.
94. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman D a, Banfield JF: **Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization.** *Genome Res* 2012.
95. Morowitz MJ, Denev VJ, Costello EK, Thomas BC, Poroyko V: **Strain-resolved community genomic analysis of gut microbial colonization in a premature infant.** *Proc Natl Acad Sci* 2010.
96. Figures S, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos E a., Madupu R, Magrini V, Martin JC, Mitreva M,

Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, et al.: **Structure, function and diversity of the healthy human microbiome.** *Nature* 2012, **486**:207–214.

97. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger S a, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P: **Metagenomic species profiling using universal phylogenetic marker genes.** *Nat Methods* 2013.

98. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P: **Genomic variation landscape of the human gut microbiome.** *Nature* 2013, **493**:45–50.

99. Konstantinidis KT, Tiedje JM: **Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead.** *Curr Opin Microbiol* 2007, **10**:504–9.

100. Turnbaugh PJ, Hamady M, Yatsunencko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe B a, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457**:480–4.

101. Ochman H: **Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis.** *Science* 2001, **292**:1096–1099.

102. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nat Rev Genet* 2014, **15**:121–132.

103. Mende DR, Sunagawa S, Zeller G, Bork P: **Accurate and universal delineation of prokaryotic species.** *Nat Methods* 2013, **10**:881–4.
104. Alonso-Sáez L, Waller AS, Mende DR, Bakker K, Farnelid H, Yager PL, Lovejoy C, Tremblay J-É, Potvin M, Heinrich F, Estrada M, Riemann L, Bork P, Pedrós-Alió C, Bertilsson S: **Role for urea in nitrification by polar marine Archaea.** *Proc Natl Acad Sci* 2012, **109**:17989–94.
105. Langille MGI, Brinkman FSL: **IslandViewer: An integrated interface for computational identification and visualization of genomic islands.** *Bioinformatics* 2009, **25**:664–665.
106. Shapiro BJ, Alm EJ: **Comparing patterns of natural selection across species using selective signatures.** *PLoS Genet* 2008, **4**:e23.
107. McDonald, John, Kreitman J: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991.
108. Liu J, Zhang Y, Lei X, Zhang Z: **Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective.** *Genome Biol* 2008, **9**:R69.
109. Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G: **Shifts in the intensity of purifying selection: An analysis of genome-wide polymorphism data from two closely related yeast species.** *Genome Res* 2010, **20**:1558–1573.
110. Novichkov PS, Wolf YI, Dubchak I, Koonin E V.: **Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes.** *J Bacteriol* 2009, **91**:65–73.

111. Friedman R, Drake JW, Hughes AL: **Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles.** *Genetics* 2004, **167**:1507–1512.
112. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P: **eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations.** *Nucleic Acids Res* 2010, **38**(Database issue):D190–5.
113. Alvarez-Martinez CE, Christie PJ: **Biological diversity of prokaryotic type IV secretion systems.** *Microbiol Mol Biol Rev* 2009, **73**:775–808.
114. Alvarez-Martinez CE, Christie PJ: **Biological diversity of prokaryotic type IV secretion systems.** *Microbiol Mol Biol Rev* 2009, **73**:775–808.
115. Nagai H, Roy CR: **Show me the substrates: Modulation of host cell function by type IV secretion systems.** *Cell Microbiol* 2003, **5**:373–383.
116. Kelly D, Conway S, Aminov R: **Commensal gut bacteria: Mechanisms of immune modulation.** *Trends Immunol* 2005, **26**:326–333.
117. Jones B V, Begley M, Hill C, Gahan CGM, Marchesi JR: **Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome.** *Proc Natl Acad Sci* 2008, **105**:13580–5.
118. Ridlon JM, Kang D-J, Hylemon PB: **Bile salt biotransformations by human intestinal bacteria.** *J Lipid Res* 2006, **47**:241–59.
119. Berr F, Kullak-Ublick GA, Paumgartner G, Münzing W, Hylemon PB: **7 alpha-dehydroxylating bacteria enhance deoxycholic acid input and cholesterol saturation of bile in patients with gallstones.** *Gastroenterology* 1996, **111**:1611–1620.

120. Bernstein H, Bernstein C, Payne CM, Dvorakova K, Garewal H: **Bile acids as carcinogens in human gastrointestinal cancers.** *Mutat Res* 2005, **589**:47–65.
121. Ogilvie L a, Jones B V: **Dysbiosis modulates capacity for bile acid modification in the gut microbiomes of patients with inflammatory bowel disease: a mechanism and marker of disease?** *Gut* 2012, **61**:1642–3.
122. Frey P a: **The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose.** *FASEB J* 1996, **10**:461–470.
123. Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castaño-Diez D, Chen W-H, Devos D, Güell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Böttcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin A-C: **Proteome organization in a genome-reduced bacterium.** *Science* 2009, **326**:1235–1240.
124. Holdeman L V., Moore WEC: **New genus, Coprococcus, twelve new species, and emended description of four previously described species of bacteria from human feces.** *Int J Syst Bacteriol* 1974, **24**:260–277.
125. Krispin O, Allmansberger R: **The Bacillus subtilis galE gene is essential in the presence of glucose and galactose.** *J Bacteriol* 1998, **180**:2265–2270.
126. Duncan SH, Hold GL, Barcenilla A, Stewart CS, Flint HJ: **Roseburia intestinalis sp . nov ., a novel saccharolytic , butyrate-producing bacterium from human faeces.** 2002:1615–1620.
127. Lukjancenko O, Wassenaar TM, Ussery DW: **Comparison of 61 sequenced Escherichia coli genomes.** *Microb Ecol* 2010, **60**:708–20.

128. Doolittle WF: **If the Tree-of-Life fell, would we recognize the sound.** 2004.
129. Medini D, Donati C, Tettelin H, Massignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**:589–94.
130. Vernikos G, Medini D, Riley DR, Tettelin H: **Ten years of pan-genome analyses.** *Curr Opin Microbiol* 2014, **23C**:148–154.
131. Bapteste E, Boucher Y, Leigh J, Doolittle WF: **Phylogenetic reconstruction and lateral gene transfer.** *Trends Microbiol* 2004, **12**:406–11.
132. Ciccarelli FD, Doerks T, Mering C von, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311**:1283–7.
133. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM: **Genome-wide experimental determination of barriers to horizontal gene transfer.** *Science* 2007, **318**:1449–52.
134. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome.** *Trends Genet* 2009, **25**:107–10.
135. Omelchenko M V, Makarova KS, Wolf YI, Rogozin IB, Koonin E V: **Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ.** *Genome Biol* 2003, **4**:R55.
136. Price MN, Huang KH, Arkin AP, Alm EJ: **Operon formation is driven by co-regulation and not by horizontal gene transfer.** *Genome Res* 2005, **15**:809–19.
137. Lefébure T, Stanhope MJ: **Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition.** *Genome Biol* 2007, **8**:R71.

138. Gordienko EN, Kazanov MD, Gelfand MS: **Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*.** *J Bacteriol* 2013, **195**:2786–92.
139. Doolittle WF: **If the Tree of Life fell, would we recognize the sound?** In *Microb Phylogeny Evol Concepts Controv.* In J. Sapp. USA: Oxford University Press; 2004:119–133.
140. Tettelin H, Riley D, Cattuto C, Medini D: **Comparative genomics: the bacterial pan-genome.** *Curr Opin Microbiol* 2008, **11**:472–477.
141. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ: **Ecology drives a global network of gene exchange connecting the human microbiome.** *Nature* 2011:2–5.
142. Cheng Q, Paszkiet BJ, Shoemaker NB, Gardner JF, Salyers AA: **Integration and Excision of a *Bacteroides* Integration and Excision of a *Bacteroides* Conjugative Transposon, CTnDOT.** *J Bacteriol* 2000, **182**:4035–4043.
143. Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, Barthel M, Westendorf AM, Krogfelt KA, Walker AW, Ackermann M, Dobrindt U, Thomson NR, Hardt W-D: **Gut inflammation can boost horizontal gene transfer between pathogenic and commensal *Enterobacteriaceae*.** *Proc Natl Acad Sci* 2012, **109**:1269–74.
144. Oliveira PH, Touchon M, Rocha EPC: **The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts.** *Nucleic Acids Res* 2014:1–14.
145. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim K, Fulton RS, Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI: **Evolution of symbiotic bacteria in the distal human intestine.** *PLoS Biol* 2007, **5**:e156.

146. Peterson DA, Frank DN, Pace NR, Gordon JI: **Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases.** *Cell Host Microbe* 2008, **3**:417–27.
147. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and environmental microorganisms.** *Nat Rev Microbiol* 2004, **2**:414–24.
148. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI: **A Genomic View of the Human–*Bacteroides thetaiotaomicron* Symbiosis Title.** *Science* 2003.
149. Bjursell MK, Martens EC, Gordon JI: **Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, *Bacteroides thetaiotaomicron*, to the suckling period.** *J Biol Chem* 2006, **281**:36269–79.
150. Peterson DA, McNulty NP, Guruge JL, Gordon JI: **IgA response to symbiotic bacteria as a mediator of gut homeostasis.** *Cell Host Microbe* 2007, **2**:328–39.
151. Martens EC, Roth R, Heuser JE, Gordon JI: **Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont.** *J Biol Chem* 2009, **284**:18445–57.
152. Comstock LE: **Importance of glycans to the host-bacteroides mutualism in the mammalian intestine.** *Cell Host Microbe* 2009, **5**:522–6.
153. Martens EC, Kelly AG, Tauzin AS, Brumer H: **The Devil Lies in the Details: How Variations in Polysaccharide Fine-Structure Impact the Physiology and Evolution of Gut Microbes.** *J Mol Biol* 2014, **426**:3851–3865.

154. Nimwegen E van: **Scaling laws in the functional content of genomes.** *Trends Genet* 2003, **19**:479–484.
155. Konstantinidis KT, Tiedje JM: **Trends between gene content and genome size in prokaryotic species with larger genomes.** *Proc Natl Acad Sci* 2004, **101**:3160–5.
156. Molina N, Nimwegen E van: **Scaling laws in functional genome content across prokaryotic clades and lifestyles.** *Trends Genet* 2009, **25**:243–7.
157. Cordero OX, Hogeweg P: **The impact of long-distance horizontal gene transfer on prokaryotic genome size.** *Proc Natl Acad Sci* 2009, **106**:21748–53.
158. Ochman H, Jones IB: **Evolutionary dynamics of full genome content in *Escherichia coli*.** *EMBO J* 2000, **19**:6637–43.
159. Pokusaeva K, Fitzgerald GF, van Sinderen D: **Carbohydrate metabolism in *Bifidobacteria*.** *Genes Nutr* 2011, **6**:285–306.
160. Sonnenburg JL, Xu J, Leip DD, Chen C-H, Westover BP, Weatherford J, Buhler JD, Gordon JI: **Glycan foraging in vivo by an intestine-adapted bacterial symbiont.** *Science* 2005, **307**:1955–9.
161. Sonnenburg ED, Zheng H, Joglekar P, Higginbottom SK, Firbank SJ, Bolam DN, Sonnenburg JL: **Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations.** *Cell* 2010, **141**:1241–52.
162. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI: **Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts.** *PLoS Biol* 2011, **9**:e1001221.

163. Sonnenburg ED, Sonnenburg JL, Manchester JK, Hansen EE, Chiang HC, Gordon JI: **A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism.** *Proc Natl Acad Sci* 2006, **103**:8834–9.
164. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P: **eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.** *Nucleic Acids Res* 2012, **40**(Database issue):D284–9.
165. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(Database issue):D109–14.
166. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller L a, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2012, **40**(Database issue):D742–53.
167. Rawlings ND, Waller M, Barrett AJ, Bateman A: **MEROPS: the database of proteolytic enzymes, their substrates and inhibitors.** *Nucleic Acids Res* 2014, **42**(Database issue):D503–9.