STRATIFICATION AND VARIATION OF THE HUMAN GUT MICROBIOTA

Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences,

Julius-Maximilians-Universität Würzburg,

Faculty of Biology

submitted by

Paul Igor Costea

from

Brasov, Romania

Würzburg 2016

Submitted on:

Members of the *Promotionskomitee*:

Chairperson:	Prof. Dr. Thomas Rudel
Primary Supervisor:	Prof. Dr. Peer Bork
Supervisor (Second):	Prof. Dr. Thomas Dandekar

Date of Public Defence:

Date of Receipt of Certificates:

a billion friends

living, eating, dividing

just feculent

ABSTRACT

The microbial communities that live inside the human gastrointestinal tract -the human gut microbiome- are important for host health and wellbeing. Characterizing this new "organ", made up of as many cells as the human body itself, has recently become possible through technological advances. Metagenomics, the high-throughput sequencing of DNA directly from microbial communities, enables us to take genomic snapshots of thousands of microbes living together in this complex ecosystem, without the need for isolating and growing them. Quantifying the composition of the human gut microbiome allows us to investigate its properties and connect it to host physiology and disease. The wealth of such connections was unexpected and is probably still underestimated. Due to the fact that most of our dietary as well as medicinal intake affects the microbiome and that the microbiome itself interacts with our immune system through a multitude of pathways, many mechanisms have been proposed to explain the observed correlations, though most have yet to be understood in depth.

An obvious prerequisite to characterizing the microbiome and its interactions with the host is the accurate quantification of its composition, i.e. determining which microbes are present and in what numbers they occur. Historically, standard practices have existed for sample handling, DNA extraction and data analysis for many years. However, these were generally developed for single microbe cultures and it is not always feasible to implement them in large scale metagenomic studies. Partly because of this and partly because of the excitement that new technology brings about, the first metagenomic studies each took the liberty to define their own approach and protocols. From early meta-analysis of these studies it became clear that the differences in sample handling, as well as differences in computational approaches, made comparisons across studies very difficult. This restricts our ability to cross-validate findings of individual studies and to pool samples from larger cohorts. To address the pressing need for standardization, we undertook an extensive comparison of 21 different DNA extraction methods as well as a series of other sample manipulations that affect quantification. We developed a number of criteria for determining the measurement quality in the absence of a mock community and used these to propose best practices for sampling, DNA extraction and library preparation. If these were to be accepted as standards in the field, it would greatly improve comparability across studies, which would dramatically increase the power of our inferences and our ability to draw general conclusions about the microbiome.

Most metagenomics studies involve comparisons between microbial communities, for example between fecal samples from cases and controls. A multitude of approaches have been proposed to calculate community dissimilarities (beta diversity) and they are often combined with various preprocessing techniques. Direct metagenomics quantification usually counts sequencing reads mapped to specific taxonomic units, which can be species, genera, etc. Due to technology-inherent differences in sampling depth, normalizing counts is necessary, for instance by dividing each count by the sum of all counts in a sample (i.e. total sum scaling), or by subsampling. To derive a single value for community (dis-)similarity, multiple distance measures have been proposed. Although it is theoretically difficult to benchmark these approaches, we developed a biologically motivated framework in which distance measures can be evaluated. This highlights the importance of data transformations and their impact on the measured distances.

Building on our experience with accurate abundance estimation and data preprocessing techniques, we can now try and understand some of the basic properties of microbial communities. In 2011, it was proposed that the space of genus level variation of the human gut microbial community is structured into three basic types, termed enterotypes. These were described in a multi-country cohort, so as to be independent of geography, age and other host

properties. Operationally defined through a clustering approach, they are “densely populated areas in a multidimensional space of community composition”(source) and were proposed as a general stratifier for the human population. Later studies that applied this concept to other datasets raised concerns about the optimum number of clusters and robustness of the clustering approach. This heralded a long standing debate about the existence of structure and the best ways to determine and capture it. Here, we reconsider the concept of enterotypes, in the context of the vastly increased amounts of available data. We propose a refined framework in which the different types should be thought of as weak attractors in compositional space and we try to implement an approach to determining which attractor a sample is closest to. To this end, we train a classifier on a reference dataset to assign membership to new samples. This way, enterotypes assignment is no longer dataset dependent and effects due to biased sampling are minimized. Using a model in which we assume the existence of three enterotypes characterized by the same driver genera, as originally postulated, we show the relevance of this stratification and propose it to be used in a clinical setting as a potential marker for disease development. Moreover, we believe that these attractors underline different rules of community assembly and we recommend they be accounted for when analyzing gut microbiome samples.

While enterotypes describe structure in the community at genus level, metagenomic sequencing can in principle achieve single-nucleotide resolution, allowing us to identify single nucleotide polymorphisms (SNPs) and other genomic variants in the gut microbiome. Analysis methodology for this level of resolution has only recently been developed and little exploration has been done to date. Assessing SNPs in a large, multinational cohort, we discovered that the landscape of genomic variation seems highly structured even beyond species resolution, indicating that clearly distinguishable subspecies are prevalent among gut microbes. In several cases, these subspecies exhibit geo-stratification, with some subspecies only found in the Chinese population. Generally however, they present only minor dispersion limitations and are seen across most of our study populations. Within one individual, one subspecies is commonly found to dominate and only rarely are several subspecies observed to co-occur in the same ecosystem. Analysis of longitudinal data indicates that the dominant subspecies remains stable over periods of more than three years. When interrogating their functional properties we find many differences, with specific ones appearing relevant to the host. For example, we identify a subspecies of *E. rectale* that is lacking the flagellum operon and find its presence to be significantly associated with lower body mass index and lower insulin resistance of their hosts; it also correlates with higher microbial community diversity. These associations could not be seen at the species level (where multiple subspecies are convoluted), which illustrates the importance of this increased resolution for a more comprehensive understanding of microbial interactions within the microbiome and with the host.

Taken together, our results provide a rigorous basis for performing comparative metagenomics of the human gut, encompassing recommendations for both experimental sample processing and computational analysis. We furthermore refine the concept of community stratification into enterotypes, develop a reference-based approach for enterotype assignment and provide compelling evidence for their relevance. Lastly, by harnessing the full resolution of metagenomics, we discover a highly structured genomic variation landscape below the microbial species level and identify common subspecies of the human gut microbiome. By developing these high-precision metagenomics analysis tools, we thus hope to contribute to a greatly improved understanding of the properties and dynamics of the human gut microbiome.

ZUSAMMENFASSUNG

Die mikrobiellen Gemeinschaften innerhalb des menschlichen Darmtrakts – das menschliche Darm-Mikrobiom - sind wichtig für das Wohlbefinden und die Gesundheit des Wirts. Die Charakterisierung dieses neuen "Organs", welches aus ähnlich vielen Zellen besteht wie der menschliche Körper, ist in jüngster Zeit durch technologische Fortschritte möglich geworden. Die Metagenomik, die direkte Hochdurchsatz-Sequenzierung mikrobieller DNA, ermöglicht die Aufnahme "genomischer Schnappschüsse" tausender verschiedener, in einem komplexen Ökosystem zusammenlebender Bakterien, ohne dafür auf deren Isolierung und Wachstum angewiesen zu sein. Die Quantifizierung des menschlichen Mikrobioms erlaubt es uns, seine Eigenschaften zu untersuchen und Verbindungen zu Wirtsphysiologie und -krankheiten zu knüpfen. Der Reichtum dieser Informationen ist unerwartet hoch und wahrscheinlich noch immer unterbewertet. Aufgrund der Tatsache, dass der Großteil unserer Ernährung und unseres Medikamentenkonsums unser Mikrobiom, welches wiederum selbst über verschiedene Arten mit unserem Immunsystem interagiert, beeinflusst, wurden viele Mechanismen vorgeschlagen, um die beobachteten Korrelationen zu erklären. Die meisten davon sind jedoch noch nicht vollständig verstanden.

Eine offensichtliche Komponente zur Charakterisierung des Mikrobioms und dessen Interaktionen mit dem Wirt ist eine akkurate Quantifizierung seiner genauen Zusammensetzung, womit sowohl die Anwesenheit von bestimmten Bakterien als auch deren Anzahl gemeint ist. Obwohl etablierte Standardprozeduren zur Probenbehandlung, DNA-Extrahierung und Datenanalyse existieren, sind sie nicht immer für metagenomische Studien anwendbar, da sie für isolierte Bakterienkulturen entwickelt worden. Deswegen und auch wegen der Begeisterung, die neuartige Technologien mit sich bringen, nahmen sich die ersten metagenomischen Studien jeweils die Freiheit, ihre eigenen Protokolle und Herangehensweisen zu definieren. Die Metaanalyse dieser Studien zeigte, dass Unterschiede sowohl in der Probenbehandlung als auch in der statistischen Auswertung den Vergleich zwischen Studien sehr schwierig machen. Das wiederum beschneidet unsere Fähigkeit, Entdeckungen zu bestätigen und Daten über Studien hinweg zu kombinieren. Um die zwingend notwendige Standardisierung voranzutreiben haben wir einen umfassenden Vergleich von 21 verschiedenen DNA-Extraktionsmethoden sowie verschiedener weiterer Probenbehandlungen, welche Quantifizierungen beeinflussen, vorgenommen. Wir haben eine Reihe von Kriterien entwickelt, um die Messqualität in Abwesenheit von Mock-Kontrollen zu bestimmen und schlagen anhand dieser Methoden für Probenbeschaffung, DNA-Extraktion und Library-Generierung optimale Verfahren vor. Wenn diese als Standard akzeptiert werden, würde das eine stark verbesserte Vergleichbarkeit zwischen Studien ermöglichen und damit sowohl einen extremen Zuwachs an statistischer Power als auch unserer Fähigkeit, generelle Schlüsse über das Mikrobiom zu ziehen, zur Folge haben.

Die meisten metagenomischen Studien teilen ihre Datensätze auf um Vergleiche anzustellen, z.B. zwischen Stuhlproben gesunder und erkrankter Menschen. Eine Vielzahl verschiedener Ansätze, welche wiederum oft mit verschiedenen Datenvorbehandlungen kombiniert werden, wurden vorgeschlagen, um Dissimilarität zwischen Gemeinschaften (Beta-Diversität) zu berechnen. Um metagenomische Daten auf Spezies-, Genus- und höheren Ebenen zu quantifizieren werden üblicherweise reads auf Referenzgenome bestimmter taxonomischer Einheiten aligniert und gezählt. Aufgrund technologieabhängiger Unterschiede in Sequenziertiefe müssen reads normalisiert werden, z.B. indem man alle counts durch die Gesamtanzahl der counts einer Sequenzierung teilt (total sum scaling), oder durch subsampling. Für die Messung der Gemeinschafts(dis)similarität wurden viele Distanzmaße vorgeschlagen. Da es schwierig ist diese Ansätze theoretisch zu vergleichen, haben wir ein biologisch

motiviertes Konzept entwickelt, mit dem man Distanzmaße evaluieren kann. Dies unterstreicht die Wichtigkeit der Datentransformation und dessen Einwirkung auf Distanzmaße.

Aufbauend auf unserer Erfahrung mit Häufigkeitsabschätzungen und Techniken zur Datenvorbehandlung können wir nun versuchen, grundlegende Eigenschaften mikrobieller Gemeinschaften zu verstehen. 2011 wurde vorgeschlagen, dass sich die Variation auf Genusebene im menschlichen Darm auf drei grundlegende Typen beschränkt, welche Enterotypen getauft wurden. Diese wurden in Datensätzen verschiedener Länder als unabhängig von Herkunft, Alter und anderer Wirtseigenschaften beschrieben. Die Enterotypen sind durch einen Cluster-Ansatz als „dicht besiedelte Bereiche in einem multidimensionalen Raum der Gemeinschaftszusammensetzung“ definiert und wurden als grundlegende Stratifikatoren für die menschlichen Population vorgeschlagen. Spätere Studien, welche dieses Konzept auf andere Datensätze anwandten, erhoben Zweifel bezüglich der optimalen Anzahl an Clustern und an der generellen Robustheit des Ansatzes. Dies leitete erneut eine langanhaltende Debate über die Existenz von Strukturen und die besten Wege, diese zu bestimmen und einzufangen, ein. Hier überdenken wir, in Anbetracht der stark gestiegenen Anzahl an verfügbaren Daten, das Enterotypen-Konzept. Wir schlagen ein überarbeitetes Konzept vor, in welchem die verschiedenen Enterotypen als schwache Attraktoren im multidimensionalen Raum verstanden werden und implementieren einen Ansatz zur Berechnung des Attraktors, der dem Datensatz am ähnlichsten ist. Dafür trainieren wir einen Klassifizierer auf einen Referenz-Datensatz, um neue Datensätze zuzuordnen. Damit ist Enterotypisierung nicht mehr datensatzabhängig und der Effekt von sampling bias ist minimiert. Indem wir ein Modell nutzen für das wir die Existenz dreier Enterotypen (definiert durch die selben Genera wie ursprünglich postuliert) annehmen, zeigen wir die Relevanz dieser Stratifikation und schlagen es in einem klinischen Zusammenhang als potentiellen Marker für Krankheitsfortschritt vor. Außerdem glauben wir, dass diese Attraktoren verschiedene Regeln mikrobieller Zusammensetzung widerspiegeln und schlagen vor, sie bei der Analyse von mikrobiellen Daten zu berücksichtigen.

Während Enterotypen Struktur in der Gemeinschaft auf Genusebene beschreiben, kann metagenomische Sequenzierung prinzipiell Auflösung auf Nukleotidebene erreichen, womit single nucleotide polymorphisms (SNPs) und andere genomische Variationen im Darm-Mikrobiom identifiziert werden können. Analysemethoden für dieses Auflösungsniveau wurden erst kürzlich entwickelt und bis heute wurden diese erst wenig erforscht. Wir zeigen, dass die Landschaft an genomischer Variation von SNPs in einer großen, multinationalen Kohorte sogar über die Speziesebene hinaus geht und hochgradig strukturiert ist, was das Vorkommen klar abgrenzbarer Subspezies unter Darmmikroben suggeriert. In mehreren Fällen zeigen diese Subspezies geographische Stratifikation, wobei einige Subspezies nur in chinesischen Populationen vorkommen. Im Allgemein zeigen Sie jedoch nur eine geringfügige Beschränkung der Dispersion und sind in der Mehrzahl der Populationen vorhanden. Innerhalb eines Individuums dominiert häufig eine bestimmte Subspezies, nur selten dominieren verschieden gemeinsam im gleichen Ökosystem. Eine Analyse von Zeitreihenexperimenten deutet darauf hin, dass die dominante Subspezies über Zeiträume von mehr als drei Jahren stabil bleibt. Wenn man ihre funktionalen Eigenschaften untersucht findet man viele Unterschiede, von denen bestimmte relevant für den Wirt erscheinen. Zum Beispiel identifizieren wir eine Subspezies von *E. rectale,* welcher das Flagellum-Operon fehlt, die signifikant assoziiert ist mit geringerem BMI und geringerer Insulinresistenz ihres Wirts; sie korreliert zudem mit höherer mikrobieller Diversität. Diese Assoziationen konnten auf Speziesebene nicht gesehen werden (auf der mehrere Subspezies überlagert sind), was die Wichtigkeit dieser erhöhten Auflösung für ein umfassenderes Verständnis mikrobieller Interaktionen innerhalb des Mikrobioms und mit dem Wirt illustriert.

Zusammenfassend bieten unsere Ergebnisse eine präzise Grundlage für vergleichende Metagenomik des menschlichen Darms, einschließlich Empfehlungen über experimentelles Sampling und statistische Analysen. Weiterhin verfeinern wir das Konzept der Enterotypen-Stratifikation in Gemeinschaften, entwickeln referenzbasierte Ansätze für Enterotypen-Zuordnung und bieten überzeugende Beweise für ihre Relevanz. Indem wir die volle Auflösung metagenomischer Sequenzierungen nutzen entdecken wir eine Landschaft hochgradig strukturierter genomischer Variation unterhalb der Speziesebene und identifizieren gemeinsame Subspezies des menschlichen Darm-Mikrobioms. Durch die Entwicklung dieser hochpräzisen metagenomischen Untersuchungsansätze tragen wir zu einem verbesserten Verständnis der Eigenschaften und Dynamiken des menschlichen Darm-Mikrobioms bei.

TABLE OF CONTENTS

1. INTRODUCTION

Microbes and the communities they form have captivated people's imagination ever since the first observations of Antonie Van Leeuwenhoek and Robert Hooke back in the 17th century[1]. From pond water and soil (the first communities to be studied) to the ocean and then human related habitats such as the oral cavity, skin and stool, microbes are our ubiquitous companion. They include bacteria, archaea, fungi, protists, algae and viruses, of which innumerable different types have been catalogued and classified to date, with many more remaining unknown[2]. They are estimated to make up approximately 60% of all living matter on earth. Modern microbiology is focused on the systematic cataloguing and characterization of this vast microbial diversity, with the ultimate goal of understanding not only the varied functions that each unit performs, but also how these units come together to form stable assemblages, i.e. ecosystems.

This cataloguing, or classification, of various micro-organisms (starting with bacteria) is a relatively recent development, due to the fact that most of human history was spent in complete ignorance of the existence of microbes. After the observations of Leeuwenhoek and Hooke, microscopists of the 17th and 18th century thought of bacteria ('infusion animalcules') as all belonging to the same shape-shifting species (pleomorphic species) and kept busy with detailed drawing and descriptions of them. The first phenotypic classification of bacteria was put forth by Otto Mueller in 1773[3], more than 100 years after the first descriptions of bacteria. He split these 'animalcules' into two form genera: punctiform (Monas) and elongated (Vibrio). More were added in consequent years, though all still based around microscopic morphology and simply motivated by curiosity. In comparison, the classification of the macro world had already been systematized in the early 18th century by Carolus Linnaeus and his introduction of a hierarchical classification system, which he used to partition the over 40,000 specimen he had collected throughout his life. Thus, the now familiar categories of kingdom, phylum, class, order, family, genus, species and strain were introduced and used to categorize specimen based on their relative similarity.

The drive for classification in the microbial world changed radically with Pasteur and Koch. The first postulated the germ theory[4] and the latter proved its truth value and concluded that it was necessary to view the different morphologies of pathogenic bacteria as belonging to distinct units[5] (i.e. a bacterial species). Now, systematization was driven by a need to identify and name pathogenic bacteria, with very real implications to human health.

In the process of proving Pasteur's postulate, Koch introduced a cultivation technique, allowing the growing of one single microbe on solidified gelatin (later agar), and thus revolutionized the way microbiology studies were performed[6]. This way, scientists could isolate a specific strain and investigate its properties, instead of just observing a collection of microbes through the microscope. Thus, classification could be based on phenotypic descriptions, obtained through specific tests performed on single units. This brought about an explosion in the amount of bacteria described, with multiple compendia being published, listing an ever increasing amount of types. This culminated in the first modern taxonomy, "Bergey's Manual of Determinative Bacteriology", in 1923, which became the reference work for further additions and modifications[7]. The manual's relevance consisted in the fact that it provided unifying criteria for describing bacterial phenotypes and morphology, as well as introducing a consistent nomenclature which greatly facilitated communication and exchange of results in the emerging microbiology community. Moreover, this first systematic description of microbial diversity was elaborated in the same terms as Linnaeus's macro taxonomy, preserving all the different levels and seamlessly expanding the net cast over the diversity of living things to include a swath of new members.

In the various environments where microbes live, they perform a varied number of functions, without which life on the planet would not have arisen as we know it and its continuation would certainly not be possible. For example, photosynthetic bacteria (cyanobacteria) living in the oceans provide a large amount of oxygen in the atmosphere[8] and are responsible for the original oxygenation of the earth's atmosphere in the proterozoic era[9]. Of similar importance are various rhizobia (soil bacteria), which usually form nodules on the roots of legumes and fix nitrogen[10,11]. These are relevant because nitrogen is the most common nutrient deficiency of soils around the world[12] and thus the most commonly supplied one. This supplementation, usually done through fertilizers, has been shown to have negative environmental impact and thus is not desirable[13]. Our understanding and ability to manipulate microbial environments would have multiple beneficial implications on such issues.

Generally, due to their sheer numbers, microbes represent a cornucopia of enzymatic function, most of which we have barely begun to understand, though we are increasingly aware of their impact and importance for the environment. A recent study has shown that there are up to 40 million different genes in the ocean alone[14], most of which we know nothing about. It is hard to overstate the potential importance of all these different functionalities, from new avenues for tackling antibiotic resistance to enzymatic machinery that can be employed at will by synthetic biology.

Focusing on human associated microbes, we find a large historical bias towards pathogenic bacteria, with most of microbiology work in the past century dealing with the relatively small number of microbes out there whose lifestyle seems to involve a huge discomfort for the host, sometimes even culminating in death. Such bias is easily explicable, especially when one considers that some of the greatest loss of life in human history was caused by bacteria (*Yersinia pestis* is estimated to have caused 200 million deaths in-between the years 1346 and 1353). Recently, however, the focus has shifted towards a more holistic and unbiased view of the human associated flora. So much so, that some have gone as far as equating the totality of microbes living in or on the human body (known as the human microbiome) to a separate organ. This may indeed be warranted, as the number of cells making up the human microbiome is comparable to the total number of host cells. Most of these bacteria, it should be noted, are referred to as "commensal", indicating that they are unlikely (though not unable) to be harmful to the host. These consortia of microbes constitute the object of study that this thesis is built around.

1.1 TAXONOMIC QUANTIFICATION OF MICROBIAL COMMUNITIES

The key to the vastly increased knowledge about microbial communities has come from the discovery of systematic ways of assessing diverse assemblies of microbes and an ever increasing resolution of classification. From the development of gram staining by Hans Christian Gram in 1884, through to the work of Woese and Fox in 1977 in using the 16S rRNA gene as a marker to describe a high resolution phylogeny of prokaryotes[15], we have become increasingly adept at classifying (i.e. naming) the units of a very complex system. It is important to note at this stage that this classification is not necessarily natural[16]. Namely, there exist philosophical objections to what microbiology calls a species and its existence in the real environment. Competing models have been put forth, none of which actually necessitate the existence of a microbial species[17–19], given what we currently know about genomic diversification, microbial evolution, lateral gene transfer[20] and conjugation[21]. While these caveats do not invalidate the structure imposed by pragmatic classification schemes, they serve to highlight the fact that they are only an approximation of reality.

Unabated by these issues, the advent of polymerase chain reaction (PCR) and sequencing technologies allowed us to move to culture-independent techniques, which pushed the development of a systematic description of the microbial taxonomy in a less biased way. After using Sanger sequencing to decode the first bacterial genomes (Haemophilus influenzae[22] and Mycoplasma genitalium[23]), characterization of microbes continued at breakneck speed with the advent of next generation sequencing technologies. This made the number of available references grow quickly, as it became easy to get a good quality bacterial genome assembly at a low cost[24]. This in turn meant that taxonomy had to keep up and the systematization needed to quickly generalize.

Currently, there are two main avenues to estimating the relative abundance of prokaryotic taxa in a given microbial community: 16S rRNA gene amplicons ("16S rRNA gene sequencing") and shotgun sequencing. Sequencing one of the hypervariable regions of the 16S rRNA gene (or the entire gene in some cases) can be used to assign reads to a taxonomic unit (i.e. species). 16S rRNA gene sequencing is made affordable by the fact that it only sequences a very small subset of the available DNA and thus can be heavily multiplexed. It was shown that as few as 10,000 reads will yield a good approximation of the sample composition at genus level. The low cost and high throughput do however come with some disadvantages. The method introduces an amplification bias and the result is influenced by the choice of primers. This problem has recently been minimized by the generalized use of only one set of primers, making studies comparable, even if the resulting abundances are biased compared to reality. More importantly though, the phylogenetic resolution of the 16S rRNA gene is far from perfect and assignments at high levels of resolution (such as species or below) are often impossible. Lastly, even with an assignment at species level, this approach misses the functional variation often observed even between very closely related bacteria.

The second approach to measuring taxonomic and functional composition of microbial communities is the metagenomics shotgun sequencing approach, which does not involve the amplification of the starting material and thus subjects the entire DNA pool to sequencing. This way, taxonomic abundances as well as gene and functional abundances can be estimated. There are a series of approaches for obtaining community composition estimates from shotgun metagenomic data, which will be introduced and discussed in the following subsections. It is important to understand the caveats and strengths of each one of these approaches, as they are the basis of all future inferences made from metagenomic data.

1.1.1 METAPHLAN

This method involves a heavy pre-computing step, but results in a small background to map to for abundance estimation[25]. Briefly, metaphlan searches the reference genome collection (a set of all available genome assemblies) for parts of genomes that can be used as a unique proxy identifier for a given taxonomic unit. For example, taking the NCBI taxonomy as a ground truth, the method searches for genes that are unique to most members of a predefined species. This then identifies units that act as a proxy for the presence and abundances of that species. This approach can be applied at all levels, thus resulting in pools of genes that are markers for different taxonomic units at different resolutions. Note that if the NCBI taxonomy has falsely classified a species, this will impact the outcome of this precomputing step. Finally, reads from a metagenomic sample are aligned against these proxy sequences and taxa abundances are estimated.

The resulting gene pool is considerably smaller than the entire NCBI genome collection, which makes aligning reads to it very fast. Once mapped, the software estimates the relative abundance of a specified taxonomic level. The main drawback of this approach is its inability to estimate the

fraction of the measured community that it cannot quantify. This problem will disappear when the NCBI genome collection will have a representative for each extant bacterial species, but this does not seem to be a likely occurrence in the near future.

To illustrate this issue, consider a community that is only composed of three members. Two are bacterial species for which we have a reference genome (call them bugA and bugB) and the third one is a type of bacterium we have not encountered before (call it bugX). Let us assume that all three taxa are in equal abundance (33.3% each). As we have marker genes for bugA and bugB, we will be able to map some of our metagenomic reads to them. Reads from bugX will not map to our reference set. Ultimately, our result will show a community composed of bugA and bugB, both representing roughly 50% of the community. Both these estimates would be wrong, because we are unable to quantify the unknown fraction. Fundamentally, this is because a non-mapping read can be so for two undistinguishable reasons: because it comes from a region of bugA or bugB that is not unique to them, or because it comes from bugX. The inability to assess the unknown fraction is thus a property of the method.

In an environment that has been extensively studied and for which a lot of reference genomes are available, this method should perform very well. Moreover, as the number of genomes increases, given that the pre-computing step is still feasible, this approach will only improve.

1.1.2 SPECI CLUSTERS

Unlike metaphlan, the specI method[26] does not take the NCBI taxonomy as a necessary ground truth. Specifically, it defines and globally applies a molecular cutoff that is a better proxy for a microbial species. Given this definition, it clusters genomes into "species pools" and uses only one genome to represent the species. This results in a collection of representative genomes to which metagenomic samples can be aligned and species abundances quantified.

As mentioned earlier, a considerable amount of research has focused on defining bacterial units (i.e. species) on the basis of their 16S rRNA gene. Recognizing that this definition is not perfect and a single similarity cutoff is a poor approximation of the bacterial phylogeny, Mende et al.[26] expanded the base on which a similarity cutoff may be used. As such, they focused on 40 universal marker genes that are present in a single copy in most bacteria. Using these, it is possible to calibrate a per gene cutoff that allows the identification of bacterial species consistently across the phylogeny. Thus, any two bacterial genomes that are more similar than the similarity cutoff are considered to belong to the same bacterial species. This approach recovers a surprising overlap with the existing NCBI taxonomy, though, for clarity, the genomic clusters it defines are termed specI clusters. We generally think of these specI clusters as being a more systematic definition of what a bacterial species is.

Before outlining the abundance estimation step, we need to also introduce the concept of a representative genome. This is a genome from a specI cluster, used to represent the species. It is usually chosen as the most complete genome in the collection. To test that this genome is indeed a representative of the cluster, we simulate reads[27] from all the reference genomes and map them back to a collection of all representative genomes (referred to as the representative set). We observe that reads sampled from genomes of a specI cluster map exclusively to the representative of that cluster. We note however, that not all reads map uniquely to this representative set. Such multiple mappers (reads that have multiple equally good mappings) are discarded.

Finally, using this representative set, we are able to quantify the relative abundances of microbes in a given sample. By mapping all reads to this set and normalizing for genome length

and other issues, we are able to estimate relative abundances of different microbes for which a specI cluster is available. Moreover, this method is able to estimate the unknown fraction of microbes, because reads that do not map to any of the representative are highly likely to have been samples from an unknown specI.

1.1.3 MOTUS

Building on the specI clusters, Sunagawa et al.[28] proposed an approach that would not only quantify the known specI abundances in metagenomic data, but also the unknown ones. To achieve this, specI clusters are defined on marker genes assembled from metagenomic samples in addition to reference genomes. First, genes are assembled from every sample and a hidden markov-model (HMM) is used to pick out those genes that are orthologues of the 10 universal marker genes. This results in a collection of unconnected marker genes, representing the full diversity of the habitat being studied. All marker genes from reference genomes are added to this collection.

To reduce the number of features in use, these genes are then clustered together at the similarity cutoff determined in the specI approach. Thus, for each marker gene, there will be clusters of genes that all belong to one species. At this point, there are pools of genes that represent a species, but they are not connected to each other. For this, samples are mapped back to a representative gene for each gene cluster and the abundance of these is measured. Then, all pairwise correlations are computed, to determine genes that covary. These are then linked together to form what is termed mOTU linkage groups. These groups link most of the 10 marker genes that a specI representative genome should have and are thus equivalent. They do however also identify linkage groups for which there is no reference genome. Due to this, the method does not suffer from the ”unknown” problem, making the quantification accurate. In the worst case scenario, species that are lowly abundant in all samples will be missed. This is because their marker genes will never be assembled. However, these would be of little consequence to the estimation error, as they are, by definition, lowly abundant across all samples.

1.1.4 STRAIN LEVEL RESOLUTION

The subsections above introduced the most widely used level of high resolution characterization for microbial communities, namely species level. However, strain level resolution has been a goal of multiple research groups, trying to disentangle the phenotypic difference observed even within a species.

Previous work in the Bork group has highlighted the fact that most individuals harbor multiple strains of any given species, with a surprising degree of individuality[29]. Other approaches have taken the resolution even further, considering each variant position to determine a new strain[30,31], and further highlighting the individuality of each sample. This level of resolution however is not very practical, as it makes it almost impossible to compare between samples, which is probably one of the reasons why most of the strain-level variation papers are focused on bacterial evolution and diversity rather than specific associations with the host or comparison between individuals. Of note, these investigations have also shown considerable functional differences at strain-level[31].

1.2 FUNCTIONAL QUANTIFICATION OF MICROBIAL COMMUNITIES

There are two aspects to consider when sampling microbial populations. To put it simply, these are "who is there" and "what are they doing". In an ideal case, the latter could be directly and easily inferred from the former. However, when it comes to microbes, this is not the case. Fundamentally, this is because our ability to read out "who is there" is limited by our understanding of microbial genomics. Here, we have, through the systematization schemes introduced above, approximated the real variation and discretized it into countable types in order to be able to make sense of the complex system. As mentioned earlier, it is even possible that we have imposed a non-existing structure onto some of these units, as some have argued that there is no reason to presuppose the existence of prokaryotic species[17].

The natural inference would follow the path of generalization from one genome within the type to the entire set. So, we should be able to take the representative specI genome and infer from there that all species within that specI cluster perform the same function. While this has been used as an approximation and may indeed be a helpful one, there are clear examples in which this generalization does not hold. For example, *E. coli* is a prevalent species of the human gut microbiome, being found in most individual's colons and thus can be considered a commensal bacteria. However, many pathogenic strains of *E.coli* also exist, which cause severe disease in humans. Thus, the label of *E.coli* is not a very informative one, as the outcome of having a lot of these bacteria in your gut can vary from non-problematic to death. Generally the differences will probably be at a much finer level, but they will certainly exist. As has been demonstrated by pan-genomic studies, most prokaryotic species have a core genome (i.e. genes shared by all strain within a type) that is considerably smaller than the size of the genome, allowing for considerable functional potential differences between strains[32].

Due to all of these considerations, it is perhaps better to circumvent the species generalization and quantify the functional potential of a community directly. While a 16S rRNA gene based approach does not allow for this, it can be done with a shotgun approach. Generally, these methods are predicated on the generation of a gene catalog encompassing the totality of genetic variation. If we have a complete survey of all potential genes in a habitat, we are then able to compare all samples from the habitat on the background of that collection of genes.

1.2.1 GENE CATALOGUES

A gene catalogue is constructed by assembling all reads within all samples and determining the complete set of genes that these samples contain. First, we assemble contigs within each sample of interest. Then, within each sample, genes are determined on these contigs and separated. This results in a pool of genes for each sample. Next, all genes are put together and clustered at a 95% identity cutoff. That is, all genes that have an identity above 95% are clustered and a representative is chosen. The resulting set of representative genes forms the catalog. For the ocean for example, this catalog contains over 40 million genes[14]. The most recent human gut catalog contains roughly 10 million genes[33]. These genes are then annotated using functional databased such as KEGG[34] and eggNOG[35], to allow for functional inference. It should be noted here that the majority of genes in these catalogues do not actually get annotated with a function, highlighting the future need to better characterize these[36].

When a gene catalog is available for a habitat, samples can be mapped against it and their functional potential can be compared to others. This allows us to look for functional enrichments, independent of the taxonomic universe. Furthermore, using these, we can try and determine de-novo units of genes that co-correlate across multiple samples. These have been termed metagenomic species (MGS)[37] and are yet another systematization of the genomic

variation observed in the environment. Thus, even with circumventing the taxonomic space, we have gone full circle and come back to a catalog based approach to determining core genomes of bacterial species. A comprehensive comparison of these two approaches to estimating the core genome has not been performed, though initial analysis suggests they are mostly consistent, lending weight to the hypothesis of the natural emergence and existence of microbial species.

1.3 THE HUMAN GUT MICROBIOME

All of the methodology introduced above allows us to determine the microbial composition of any metagenomic sample. In the context of this thesis, we are however narrowly interested in the human gut microbiome, that is, the collection of bacteria and archaea that populate the lower gastrointestinal tract (GIT). Viruses and fungi are also found across the GIT, but are commonly disregarded, due mostly to methodological issues. When performing 16S studies, viruses and fungi are ignored by definition, as they do not possess such a gene. In metagenomic studies, it is possible to recover both[38,39], though focus has historically been on the bacterial and archaeal fraction.

The colon is the part of the GIT with the highest diversity and number of microbes (comparable to the total number of cells in the host[40]). It represents a collection of all other organisms present throughout the tract and was thus chosen as the focal point of study. It is also the easiest to sample, as humans excrete the content of their colon on a daily basis.

With stool in our hands and sophisticated measuring techniques, we can start to understand the complex community of prokaryotes that have made a home of our digestive tract.

1.3.1 EARLY LIFE COLONIZATION

A natural starting point for interrogating the gut associated microbiome is at birth, where humans should emerge with a blank slate and a pristine colon. Blank slate aside, it has recently become obvious that we may not have such a pristine starting point after all.

Firstly, the question of the microbial load of the placenta has gotten renewed attention, as evidence has surfaced that it may not actually be sterile after all[41], even under normal conditions. It seems to contain a specific microbiome, distinguishable from other human body sites, composed of a diverse collection of commensal microbes from the Firmicutes, Tenericutes, Proteobacteria, Bacteroidetes, and Fusobacteria phyla and can be linked to antenatal infections and even pre-term birth. More work is necessary to determine the robustness of these findings, as they are dealing with extremely low amounts of microbes and are thus highly prone to contaminations. For example, it has been shown that each extraction kit used in these studies has its own distinct microbiome[42,43], requiring caution when drawing conclusions from small amounts of material. In the status quo, pregnant women are regularly screened against bacterial infections and are generally told to abstain from consuming raw fish or unpasteurized products as they may contain pathogenic bacteria which are highly problematic for the pregnancy. The existence of commensal placental microbes and their potential implications for fetus development add a new dimension to these considerations.

Independently of the placental microbial population, the fetus gets into contact with a plethora of microbes once it gets born. And here, the way in which it emerges from the mother may also play a role in the development and acquisition of microbes. Two options exist, that of natural birth or that of caesarian section. In the former, the fetus is exposed to the vaginal and stool microbiome as it emerges, while in the latter it is directly exposed to the mother's skin and thus

is more likely to acquire a different set of microbes[44–46]. In the initial weeks of life, the delivery mode strongly imprints the newborn's composition. However, this fades over time, though diversity differences can be observed even later in development. This has led researchers to hypothesize that the original exposure may have a long lasting influence on the development of the immune system, in part because caesarian section births have long been associated with a slight increase in asthma, allergies[47] or type I diabetes[48].

The considerations above have led some to suggest that caesarian section born babies should be inoculated with vaginal bacteria through artificial means[49] and such studies have shown that such a transfer would indeed bring the composition of c-section born babies to a more similar state to that of vaginally born ones. There is, however, no long term evidence that such an inoculation would be advantageous, so - given the higher risk of infection associated with vaginal births[50] - it may be too soon to ask this of your doctor.

After the initial weeks of life, the microbiome is highly unstable, up to age 6[51], though it is consistently characterized by high levels of Bifidobacterium, which is comparatively low in adults. It has been indicated that the composition can be influenced by breast feeding and long-lasting effects are observed after antibiotic treatment, with implications for the later development of obesity and type I diabetes[52–55]. Moreover, given the interconnection between the microbiome and the immune system, extensive research has associated compositional shifts in early life to the development of the immune system and inflammatory disorders such as allergies and asthma[56,57].

1.3.2 THE ADULT MICROBIOME

Following the original colonization, the gut microbiome of children up to the age of 6 is dramatically distinct from that of adults[51]. Enriched in *Bifidobacterium spp.*, it is highly variable over time. After childhood, the composition of the gut microbiome becomes fixed, through mechanisms not yet understood. While there is still temporal variation within an individual[58], the magnitude of this variation is smaller than the differences observed between people. For example, in a cohort of nearly 2000, samples taken from the same individual are clearly distinguishable from all other ones, when considering taxonomic composition. The functional space is a lot less variable, with no distinguishable individual specific features[59]. Of note, this may be due to our poor understanding of the functional diversity and our inability to annotate most of the genes in the human gut to a known function. Thus, the low variation functional space is made up of very basic functions, relating to amino acid synthesis, DNA repair and replication and other clearly necessary machinery for the life and replication of microbes. It is however also possible that a lot of the differences observed on a taxonomic level do not reflect in the functions of the microbiome, as different microbes can perform the same functions and are thus, from this perspective at least, interchangeable. Pan-genomic and phenotypic studies have suggested that a bacterial genus is considerably homogenous in term of functional potential[60], making the choice of species within that genus irrelevant to the broad functional space. Again, though, none of these statements should be generalized too much, as clear violations of these rules can easily be found in the literature, where even at strain level there are considerable and relevant differences in functional potential[32].

The variation observed in both the taxonomic and functional spaces represent highly relevant considerations when thinking about the human associated gut microbiome and the type and size of experiments needed. For example, in the taxonomic space, any association with other variables will need to overcome the large background variation. Consider *Prevotella corpi*, which is one of the most prevalent bacterial species of the gut (i.e. is preset in most samples). It ranges from 0 to almost 40% in some samples, which, given our detection limit of approximately

0.001%, is a span of five orders of magnitude. Conversely, large populations will be needed to distinguish very small effect sizes at the functional level.

1.3.3 THE MICROBIOME IN HEALTH AND DISEASE

The consortia of gut microbes are believed to have evolved together with their human host and formed a working symbiosis[61,62]. While the gut provides protection and a rich environment for the bacteria to thrive, they in turn are important to the wellbeing and health of the host. Most importantly, they provide an additional barrier against outside pathogenic microbes, by occupying the space through which these would have to pass in order to infect the host. This barrier aids the colon mucosal layer, which is the physical barrier that the host interposes between the epithelial layer and the gut lumen. This layer is rich in glycoproteins and water, forming a slippery film on top of the epithelium[63]. Apart from this apparent out-numbering game, there are specific mechanisms through which certain members of the commensal microbiome offer protection from pathogens. For example, commensal *E. coli* has been shown to protect against *Salmonella* infection in mice, by engagement of the NLRC4 inflammasome and IGF-1 signaling[64]. This suggests the possibility of immunizing humans from various pathogens by cultivating the right microbiome, which would potentially circumvent the emerging antibiotic resistance problem.

Apart from pathogen protection, the microbiome participates in important metabolic functions that humans cannot perform, acting as a fermenter to degrade various complex molecules and transform them into valuable compounds that can be taken up and used by the host. Among these are a range of vitamins that can be produced by members of the human gut community[65]. For example, bacteria like *Lactobacillus reuteri* and others are capable of producing cobalamin (vitamin B12), which is crucial for the functioning of the brain and nervous system, though the host does not possess the necessary enzymes to produce it. Moreover, some of the gut bacteria are able to break down otherwise indigestible dietary fibers into short chin fatty acids (SCFAs) (butyrate, propionate and acetate), which can then be taken up by the host[66]. For example, colonocytes (epithelial cells that line the colon) uptake butyrate and use it as their main energy source. Moreover, this small molecule has been shown to influence expression patterns in the colon lining by a direct activity on DNA acetylation[67]. Studies following the compound outcome of different levels of butyrate in stool have however been contradictory. There is evidence of a small overall effect of butyrate enemas on infection outcomes[68] and a beneficial effect on colitis[69], though others have shown there was little effect on expression of mucins, which had been thought to be the main mode of action[70].

Some species of bacteria, such as *Akkermansia muciniphila* have been shown to feed directly on the mucus[71], suggesting the interaction between the host and microbiome is complex and two-sided. This view of interacting entities gave birth to the main assumptions behind the human microbiome field, namely that this colossal gut ecosystem interacts with, and thus must have an effect on the host.

While work has been done to catalogue the variation in the healthy population, a lot of research has focused on associations between the gut microbiome and disease. These associations range from the expected to the unexpected and all the way to the unlikely. One potential reason for this wide variation resides in all the confounding factors that we have yet to identify and thus control for[72]. Moreover, it is important to note that for most of these correlations, causation directionality remains to be proven. It should also be noted that a biased expectation is acceptable in these conditions. By this we mean that the expectation should be that for most of these cases the causality flows from host changes to changes in the microbiome, simply because this is a more parsimonious explanation. One other issue to consider is that of confounding

factors, of which we know very little at the moment. Interactions between drugs and the microbiome as well as effects driven by host inflammation can rarely be accounted for in our correlations, but they can represent a significant effect. Moreover, the expectation of the involvement of the microbiome should be contextualized by the already mentioned vast variation observed between healthy individuals, suggesting little restrictions when it comes to which microbes inhabit the host. The case for disease though is different with multiple examples of specific changes associated with host phenotypes. For example, a robust association between the microbiome and host has been observed in the case of colorectal cancer (CRC)[73,74]. Here, we were able to build a classifier based on microbial abundance data which was able to confidently classify CRC cases. We further identified marker species which were informative in this classification. Most of these were known oral pathogens, with the most informative being *Fusobacterium nucleatum,* which showed a dramatic enrichment in samples from CRC patients. Such classifiers have since been built on other cohorts and similar marker species recovered, suggesting a robust signal for CRC, independently of the study. The question remains whether these marker species thrive because of the existence of cancer or they are implicated in its etiology. *F. nucleatum* for example has been shown to play a role in the carcinogenesis of proximal colon cancer[75] and to be able to invade the epithelium. Of the other marker species it is known that they are capable of adhering to non-mucosal layers, which could explain their growth around the tumor site and thus why they are more abundant in cancer samples.

Other applications where the microbiome can be used as a good predictive measure include liver cirrhosis[76] and obesity[77], though the predictive power in the latter is more limited. A case where it was thought that the microbiome was a good predictor was that of type II diabetes, though it was later discovered that these studies were confounded by a specific drug that the diabetic patients were prescribed, namely metformin[72]. Controlling for that confounder showed that the microbiome of type II diabetics cannot be distinguished from that of controls. This should be a cautionary tale when considering other reports, as the microbiome seems sensitive to many factors we are currently not aware of and which may ultimately explain the effects we are observing.

In the direction of the more unlikely associations, it has been proposed that the microbiome is involved in controlling satiation by direct signaling to the brain, through the so-called gut-brain-axis. Other effects on the brain are specific to the development of the immune system, with some recent implications for Alzheimer and even autism. More work is needed before any of these associations can be considered robust or before we can understand their causal directionality.

1.3.4 STRUCTURE OF THE HUMAN GUT MICROBIOME

It is becoming clear that the presence or absence of specific species is unlikely to explain complex disease states. Much more likely, such states are caused by an imbalance (dysbiosis[78]) in the community structure which in turn causes the "bioreactor"[79] to alter its function[80,81]. The reasons and consequences of this imbalance constitute another major focus of research. This underlines the need to view the microbiome as a complex interconnected system and asses its properties accordingly. Stability over time, robustness and stress response become crucial areas of investigation.

The compositional space of the western adult human gut microbiome is not uniformly populated. That is to say, that there appear to be preferred states that the taxonomic units assemble into. Trying to get a handle on the properties of these "attractors", Arumugam et al proposed clustering the space into what they termed enterotypes[82]. These appeared at the time to be independent of geography, age and other measure variables and were thus proposed as a general feature of human associated gut communities. In the meantime, it turned out that

geography does play a role, at least insofar as highly distinct populations, such as those from Malawi, clearly harbor a different microbiome[51]. This consideration aside, multiple other datasets from western populations, analysed using a range of methods[83–86], have been found to cluster into a similar structure. Furthermore, such states may exist in a similar structure across other primates[61], are influenced by long-term diet[83] and can be associated with health status[86]. Further refinements in enterotypes will bring about a better understanding of community structure. Some methods have yielded as many as 4 enterotypes[87], while others have argued for only two distinguishable clusters[88]. Some have even gone as far as suggesting there is no structure in the human gut and everything should be treated as a gradient, though an exact mathematical formulation for what that implies is lacking. The number of sub-communities is not as important as their predictive power, from different metabolic behavior to interactions with the host. Once stable communities have been identified, some of the above considerations can be applied to them.

The level of resolution needed to make inferences about the bacterial community may have to be increased if the right dynamics are to be captured. Variation at genus level has been a long-standing approach. However, with the ability to robustly detect bacteria down to species level[28], unhampered by the lack of reference genomes, it has become clear that considerations on the genus level have limited power. Going down to species and beyond will greatly increase the complexity of the system, but it will allow us to capture more fine-grained features which can be highly relevant for understanding both the microbial ecosystem and its interaction with the host.

2. RESULTS

In the following chapters we shall present a summary of multiple studies, each geared towards a better understanding of microbial communities. These studies are at times narrowly technical, focusing on seemingly small issues such as data normalization or sample storage. Other times they are overtly general, aiming for a grand scheme understanding of these complex systems. They are not presented in chronological order, but in the order in which they are logically dependent on each other. Thus, the discussion will start from the small and technical aspects, in the hope of building a firm backbone on which a more global view can come to rest.

Metagenomics studies crucially depend on sample collection, storage and DNA extraction. These steps have been shown to greatly influence the measured community composition and thus are prone to confounding results of different studies[89–94]. These combined factors add up to such a great effect that most studies to date are not comparable[95], making US collected samples in one study appear to have a completely different microbial composition compared to European or Chinese samples collected within other studies. While there is an expectation that geography plays a role in the assembly of the gut community, this effect cannot at the moment be distinguished from the batch effect introduced by the different sample handling steps performed by the different laboratories. This is currently the greatest factor limiting comparison between studies and thus hampering our ability to acquire a global picture of the human gut associated microbiome. When striving for an understanding of a complex system, one needs to be able to measure its relevant variables. If direct measurement is not possible, a proxy measurement must be devised and its error needs to be quantified. This is exactly the case for the gut microbiome too. Firstly, most studies to date focus on excreted stool as a proxy measurement for the community composition in the gut. While there have been multiple studies showing that the makeup of feces along the gastro-intestinal tract (GIT) is variable[63,96], stool composition has been selected as the proxy measurement point, for two main reasons: it is proximal to the colon, which is the most dense and diverse community along the GIT and it is easy to sample, which is a very important consideration when planning large cohort studies.

Multiple problems then arise, related to measuring the community composition of stool, two of which we have studied and assessed in depth. First is the issue of storage: after a stool sample has been collected, it needs to be shipped to a lab for analysis and then stored until such analysis can take place. The initial storage can be for hours or even days and weeks, depending on the setup of the study. Ideally, the donor would freeze the specimen at -20 degrees Celsius and simply ship it to the laboratory that will further process it. In reality, the logistics of such freezing are not feasible as they would impose very high additional costs. Furthermore, such an approach is actually impossible in some circumstances, where the donor does not have access to means of achieving such low temperatures (high powered freezers, dry ice, etc.). We show that using RNALater as a fixing agent for the collected specimen allows it to be handled and shipped at room temperature without significantly distorting the final measurement. Next, we focused on DNA extraction. Once a stool sample arrives in the lab, bacterial cells within it need to be broken and their DNA recovered, in such a way that the measured community composition matches that of the sample. To test the effect of DNA extraction methods on observed bacterial composition, we performed an extensive comparison across 21 different protocols used around the world. These protocols are representative of the main alternatives available for extraction and cover methodologies used by the most extensive metagenomic studies to date. We show that the biggest component of the measurement error is indeed the DNA extraction method, and propose a standard, automatable approach to be used by everyone performing DNA extractions from human stool samples.

The next step in the logical chain relates to data transformations. Given the nature of our measurement, we are currently unable to obtain absolute counts of our features within a sample and can only measure relative composition. That is to say, we cannot know that an individual has 10^6 *Prevotella copri* (a common gut bacterium), we can only know that *Prevotella corpi* represents 20% of the total amount. Such measures are called "compositional" and have some specific properties. Most importantly, after normalization, they are defined on a unit simplex and thus the increase of one necessitates the decrease of all others (i.e. it is a 1 sum game). Such compositional issues are discussed in the methods section. One other issue that is presented is that of variance stabilization and additional data transformation to allow for comparison between samples. These problems arise from the way abundances are distributed, with more than six orders of magnitude between the most abundant taxon and the least abundant one we can measure. Furthermore, once comparable, deciding on a distance measure between community compositions remains a vexing question. While we do not claim to have solved these issues, we strive to present the reader with a digestion of the main challenges and a set of simple guiding rules for choosing a distance measure.

Accumulated expertise in dealing with sample collection and data transformations allowed me to be involved in multiple other projects where I aided in data analysis. These include work on using the gut microbiome to predict colorectal cancer[73], a description of the confounding effects of medication (specifically metformin) on type II diabetes microbiome studies[72], as well as work on the taxonomic and functional diversity of bacteria in the world's oceans[14]. Involvement in all these studies allowed me to get familiar with microbial data from a wide range of environments and understand the wide applicability of acquired technical notions.

Returning to the main thread of my work, having satisfactorily established that we can store and extract DNA from stool bacterial communities and that we can consistently compare across samples, we proceed to investigate the properties of these communities in relation to each other and the host they inhabit. To this end, we followed up and refined a previously described clustering of individual microbiomes, termed enterotypes[82].

2.1 STANDARDS FOR METAGENOMIC SAMPLE COLLECTION AND DNA EXTRACTION

Any inference, and indeed any attempt to understand the microbial ecosystem of the human gut is bound to be impacted by measurement error. While this error can be split into multiple components, there are two important ones that we have considered.

The first consideration is that of sample storage prior to arrival and DNA extraction in the laboratory. In Paper 1, we show that RNALater storage is comparable to frozen storage at -20 degrees Celsius, in that it introduces only minor (non-significant) changes in the observed microbial composition at species level, while eliminating the need for immediate freezing or sample processing. RNALater is an aqueous solution used to preserve biological specimen, generally used for protecting RNA. According to the manufacturer, samples fixed in this buffer are stable at room temperature for up to one week, at +4°C for one month and at −20°C and −80°C indefinitely.

Briefly, we used seven subjects, with samples collected over a period of up to two years to investigate temporal variability and assess preservation-induced variation. We assess community composition using the mOTU approach[28] to determine bacterial species relative abundances. We collect multiple samples from the same stool specimen (without homogenization) and subject them to different storage methods: freezing at -20 degrees, chilling at between 4 and 10 degrees and finally room temperature storage, with and without RNALater

as a buffer. We show that RNALater storage does not introduce a significant effect for any of the observed species when compared to freezing. Furthermore, samples from the same individual cluster together, even in the context of 888 additional samples. This suggests that the storage effect is comparable to the natural variation in a stool specimen and thus that alternative storage methods are desirable.

The second consideration is that of DNA extraction bias, by which we mean the error resulting from using different methodologies to break open bacterial cells in order to access and sequence their DNA. For quantifying this error and determining an optimal extraction approach, we have, in Paper2, undertaken a comparison 21 of the most widely used extraction protocols. In the first phase, we quantify measurement errors introduced by various factors and present a comprehensive comparison between them, which allows us to rank these factors according to effect size. We show that the extraction method has the highest effect and can dramatically distort observed community composition. In phase two of the study we show that the natural biological variation within a stool specimen is greater than the one induced by applying a standard protocol across different laboratories, supporting a standardization effort.

Generally, methods relying on mechanical breaking of cells walls, using shaking and bead beating, perform well in terms of recovering a highly diverse bacterial community and lysing gram-positive cell walls. Furthermore, we note that formaldehyde based extractions also perform well, though we recommend against them, as they are potentially dangerous. Finally, we propose a standardized extraction method, the use of which would allow for better comparison across future studies.

Taken together, these two studies pave the way to large scale studies which one may compare across, allowing new work to compare with and build on old one.

2.2 DATA NORMALIZATION AND TECHNICAL CONSIDERATIONS

Because of the nature of the measurement performed in metagenomics studies, we cannot at the moment obtain absolute counts for the different taxonomic units. All measurements are compositional, i.e. defined on a simplex. This results in an intrinsic negative correlation between the considered features; this is easy to model if we were to imagine that we are measuring only two variables. As they have to sum up to 100%, if one increases, that would cause the other one to decrease, thus leading to a correlation of -1. However, we cannot assume that the total number of bacteria in the gut community will always be the same, so it is entirely possible that the growth of one bacterium does not influence the number of cells of the others. The inability to disentangle such dynamics is termed the “compositional” issue.

Complications further arise when trying to compare between samples, as the total number of bacteria may be different, or each may have been sampled to a higher depth (by sequencing more DNA molecules). We hold that a total sum scaling (i.e. dividing by the sum of counts per each sample) is a sufficient normalization and will allow comparison between samples. We illustrate this point in the following publication, highlighting also the necessity of a variance normalizing transform after the normalization.

2.3 STRUCTURE OF GUT COMMUNITY COMPOSITION

Of primary interest in the papers covered in this section are the properties of the overall community composition, such as they can be observed through different resolutions.

As a general framework of what structure may conceivably be, the reader should consider community compositions of different human associated body-sites. These consist of different microbes, in different abundances, making them easy to distinguish from each other. This however does not mean that compositions are homogenous within a body-site. Indeed, much variation remains within each of these groups, though this also varies by site. For example, the urogenital tract is a low diversity community, consisting mostly of Lactobacilii, which makes this body-site more homogenous than, for instance, the skin, which can itself be further split into multiple sub-sites, each with its specific habitat.

2.3.1 ENTEROTYPES IN THE SPACE OF MICROBIAL COMPOSITIONAL LANDSCAPE

One of the first observations made on the compositional variation of the human gut was that different people have very different communities that are stable over time. Even in a large cohort, one can easily identify samples from an individual for which a previous time-point exists. So, the closest sample in composition space will be another sample from the same individual and all other samples will be relatively distant.

When further investigating the general composition landscape, it became clear that this space is not evenly populated. So, while each individual is unique, there are some states that are more likely than others, causing the multi-dimensional space resulting from all abundance measures to appear somewhat structured.

The identification of this structure is the topic of paper 4. It mainly focuses on the issues that have been encountered when trying to identify this structure de-novo in new datasets and deals with concerns that such structure is artificial. It goes to great lengths to discuss the properties that these enterotypes have, both from a species composition perspective and from a functional one. Finally, it proposes a new, classifier based, method of enterotype assignment. It shows that this method is robust at recovering known structure across datasets and outlines the advantages of the new approach, specifically for small datasets where recovery of de-novo structure is cumbersome. It furthermore provides a filtering step in which the composition of the new samples is compared to a reference set and a determination is made regarding the likelihood that these samples are similar to those. This is important as we would like to limit the classification into enterotypes to samples for which such an exercise is meaningful. For example, a determination of enterotype for an ocean sample should not be made. This filter step also ensures that new types can be found. With this improved concept of enterotypes it should be possible to consistently identify and compare between different datasets.

2.3.2 SUB-SPECIES STRUCTURE

There is constant tension between the desire for a simple categorization and the ability to make confident inferences based on this simplification. This tension can be seen in prokaryotic taxonomy between the “splitter” and the “lumper” camps. The former want more high resolution determination of types, with constant additions and refinements to current taxonomy, while the latter prefer condensing the nomenclature to a restricted number of types. In many cases, there are no objective measures by which to decide on a level of resolution, because in these cases we are dealing with units that we know little about. In the few cases where more is known, it seems that the splitters have an advantage. Such cases are those of human pathogens, like *Escherichia coli* or *Samonella enterica*. Both are bacterial species which contain strains that can be dangerous to the human host and thus studies split them into very fine-grained types, which are specifically informative of the phenotypic outcome the human host will experience when exposed to them[32].

In Paper 5, we tried to determine the existence and properties of structure below the currently accepted level of microbial species. We were able to investigate the genomic variation landscape of 73 highly abundant species of the human gut, using more than 2000 samples. These include the most abundant archaeon in the gut, namely *Methanobreivbacter smithii*, as well as the most commonly studied gut bacteria such as *Prevotella copri*, *Eubacterium rectale* and many more. Of these 73 species, 39 showed strong clustering structure in the variation landscape and these clusters were termed subspecies.

The subspecies we have identified appear to be mutually exclusive within an individual, stable over time and show specific geography distributions. Mostly in the Firmicutes phylum, we see strong geographic enrichments of subspecies based on geography. Of note, *E. rectale*, *E. eligens* and others show subspecies that are specific to Chinese samples. Moreover, general dispersal limitations are observed for most species, whose subspecies are not randomly distributed across the globe. In individuals for whom we have time-series sampling of up to 3 years, we rarely see the dominant subspecies changing. In some of the few cases where this happens, we can explain this change by the fact that the individual in question went through a course of antibiotic treatment.

Investigating host associations and functional differences between subspecies, we find flagellated and non-flagellated subspecies of *E. rectale* associated with host inflammation, insulin resistance and body mass index. We propose that the presence in high abundance of the flagellum carrying subspecies causes low-grade inflammation in the gut, which has been repeatedly associated with higher insulin resistance and obesity.

Thus, this new level of taxonomic resolution is a naturally emerging, host significant one, which should in the future be considered when trying to associate the human gut microbiome to the host. Furthermore, attempts at understanding the interactions between microbes and the networks and ecosystems they form should also account for this newly described structure.

3. CONCLUDING REMARKS

Complete solutions to complex issues rarely exist. Thus, it is unsurprising that the work presented in this thesis cannot be taken to have completely solved the issues it set out to tackle. However, it is our contention that each publication presented here is a well formulated attempt at getting closer to such a solution.

The sample storage method proposed in Paper 1, together with the DNA extraction method proposed in Paper 2, form the basis for comparable estimations of bacterial composition in stool samples across multiple studies. This tackles one of the most pressing issues in the field of human metagenomics, namely the impossibility of comparison across studies. Large dataset specific effects have been reported for all major cohorts[95], making their comparison cumbersome. With everyone in the field adhering to the same protocol, which we have shown is easily implementable and reproducible across locations, study batch effects should stop being an issue. Moreover, our proposed DNA extraction method is easy to automate and thus can be applied in large, high-throughput situations.

One drawback of the proposed standard is that it does not necessarily provide the exact sample composition and is thus subject to change and improvement. As Paper 2 does not provide a sample in which the ground truth is known, it is impossible to assess what the actual extraction biases are. The reason for which such a mock community does not exist has to do with the technical challenges of building one, highlighted by the repeated failure of multiple groups to definitively quantify extraction biases. For these reasons, we use a series of other proxy measures for assessing the quality of the extraction, but leave open the possibility that measurement error still exists.

Our studies, as well as others on the same topic, ask the question of statistically significant changes in composition between a control set and a treated set. For example, the storage method is shown to not significantly bias the estimate of any of the measured species. This, however, does not mean that the storage method does not introduce an effect, but just that this variation is undetectable given the number of replicates used. Same considerations hold true of Paper 2, in which the question is similarly phrased, only this time regarding the effect of DNA extraction methods. Here, a more thorough analysis focuses not only on the significance of the differences, but on their size as well. The size of the effect then becomes the important consideration. Namely, we can now tradeoff accuracy for practicability. This is not to say that we should not aim for better extraction methods or better storage alternatives, just that we should consider the added benefit of these in the context of how they influence the overall outcome of the measurement and how relevant this is for the questions that the data is then used for.

The standard DNA extraction protocol we propose is not guaranteed to recover the exact bacterial composition of the sample. However, the size of the measurement error that it may introduce will be smaller than all biological observation quantified to date. For example, it will be smaller than the differences observed between samples taken from different parts of the same specimen, or differences between samples taken from the same individual at different time points. Thus, as long as the effect of interest is of a greater magnitude than this background biological variation, the implementation of our standard sample storage and DNA extraction methodology will allow for its measurement, as well as facilitate future comparisons to other cohorts that have used the same approach.

Having tackled some of the major challenges of metagenomic sampling, we revisited the discussion related to structure of the human gut microbiome. We show, using additional data, that the originally proposed concept of enterotypes can be recovered across three extensive data

sets. Moreover, we develop a classification approach that allows the assignment of enterotypes independently of the study, which was a major hurdle to comparability. With this new approach, there is no requirement to the study size that is necessary for considering such structure and, importantly, enterotypes can be controlled for even in small studies.

Zooming in, below the genus level at which enterotypes are found, we have defined a new, subspecies resolution taxonomic level that we show is relevant to ecological considerations as well as to the host. We propose using this new resolution in addressing questions about the interactions between the microbiome and host, as functional differences between subspecies can be dramatic and thus will have an impact on our observations.

Taken together, the findings and proposals herein are a considerable advancement of the field of human metagenomics that will strengthen our ability to understand important and complex interactions and ensure a more rigorous inference framework, in which findings can be compared and tested across multiple studies.

4. REFERENCES

1. Lawson, I. CRAFTING THE MICROWORLD: HOW ROBERT HOOKE CONSTRUCTED KNOWLEDGE ABOUT SMALL THINGS. *Notes Rec. R. Soc. Lond.* **70,** 23–44 (2016).

2. Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proc. Natl. Acad. Sci.* **113,** 6585–6587 (2016).

3. Müller, O. *Vermium terrestrium et fluviatilium, seu animalium infusoriorum, helminthicorum et testaceorum non marinorum succincta historia*. (1773).

4. Schwartz, M. The life and works of Louis Pasteur. *J. Appl. Microbiol.* **91,** 597–601 (2001).

5. Brock, T. *Robert Koch*. (1999).

6. Koch, R. Methods for the study of pathogenic organisms 1881. *Mitth. aus dem Kais. Gesundheitsamte* **1,** 1 – 48 (1881).

7. Bergey, D. BERGEY'S MANUAL OF DETERMINATIVE BACTERIOLOGY. *Am. J. Med. Sci.* (1934).

8. Mazard, S., Penesyan, A., Ostrowski, M., Paulsen, I. T. & Egan, S. Tiny Microbes with a Big Impact: The Role of Cyanobacteria and Their Metabolites in Shaping Our Future. *Mar. Drugs* **14,** (2016).

9. Hamilton, T. L., Bryant, D. A. & Macalady, J. L. The role of biology in planetary evolution: cyanobacterial primary production in low-oxygen Proterozoic oceans. *Environ. Microbiol.* **18,** 325–340 (2016).

10. Barrett, L. G., Zee, P. C., Bever, J. D., Miller, J. T. & Thrall, P. H. Evolutionary history shapes patterns of mutualistic benefit in Acacia-rhizobial interactions. *Evolution* **70,** 1473–85 (2016).

11. Yuan, S. *et al.* RNA-Seq Analysis of Differential Gene Expression Responding to Different Rhizobium Strains in Soybean (Glycine max) Roots. *Front. Plant Sci.* **7,** 721 (2016).

12. Fields, S. Global nitrogen: cycling out of control. *Environ. Health Perspect.* **112,** A556–63 (2004).

13. Hoeft, R. G. in 235–243 (2003). doi:10.1021/bk-2004-0872.ch017

14. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).

15. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74,** 5088–5090 (1977).

16. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2567–72 (2005).

17. Gevers, D. *et al.* Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3,** 733–739 (2005).

18. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6,** 431 (2008).

19. Rosselló-Mora, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25,** 39–67 (2001).

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

21. Lanka, E. & Wilkins, B. M. DNA Processing Reactions in Bacterial Conjugation. *http://dx.doi.org/10.1146/annurev.bi.64.070195.001041* (2003).

22. Fleischmann, R., Adams, M., White, O. & Clayton, R. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* (1995).

23. Fraser, C., Gocayne, J., White, O. & Adams, M. The minimal gene complement of Mycoplasma genitalium. *Science (80-.).* (1995).

24. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10,** 599–606 (2012).

25. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–4 (2012).

26. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10,** 881–4 (2013).

27. Mende, D. R. *et al.* Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One* **7,** e31386 (2012).

28. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–9 (2013).

29. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature*

493, 45–50 (2013).

30. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33,** 1045–52 (2015).

31. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13,** 435–438 (2016).

32. Gordienko, E. N., Kazanov, M. D. & Gelfand, M. S. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J. Bacteriol.* **195,** 2786–92 (2013).

33. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).

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

35. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284–9 (2012).

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

37. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).

38. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci.* 1305923110– (2013). doi:10.1073/pnas.1305923110

39. Duerkop, B. A., Clements, C. V, Rollins, D., Rodrigues, J. L. M. & Hooper, L. V. A composite bacteriophage alters colonization by an intestinal commensal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17621–6 (2012).

40. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164,** 337–340 (2016).

41. Aagaard, K. *et al.* The Placenta Harbors a Unique Microbiome. *Sci. Transl. Med.* **6,** 207–214 (2014).

42. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).

43. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12,** 118–123 (2010).

44. La Rosa, P. S. *et al.* Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci.* **111,** 12522–7 (2014).

45. Biasucci, G., Benenati, B., Morelli, L., Bessi, E. & Boehm, G. Cesarean delivery may affect the early biodiversity of intestinal bacteria. *J. Nutr.* **138,** 1796S–1800S (2008).

46. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 11971–5 (2010).

47. Thavagnanam, S., Fleming, J., Bromley, A., Shields, M. D. & Cardwell, C. R. A meta-analysis of the association between Caesarean section and childhood asthma. *Clin. Exp. Allergy* **38,** 629–633 (2008).

48. Cardwell, C. R. *et al.* Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. *Diabetologia* **51,** 726–735 (2008).

49. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat. Med.* **22,** 250–253 (2016).

50. Signore, C., Hemachandra, A. & Klebanoff, M. Neonatal Mortality and Morbidity After Elective Cesarean Delivery Versus Routine Expectant Management: A Decision Analysis. *Semin. Perinatol.* **30,** 288–295 (2006).

51. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–7 (2012).

52. Ardeshir, A. *et al.* Breast-fed and bottle-fed infant rhesus macaques develop distinct gut microbiotas and immune systems. *Sci. Transl. Med.* **6,** 252ra120–252ra120 (2014).

53. Mach, N. *et al.* Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environ. Microbiol. Rep.* (2015). doi:10.1111/1758-2229.12285

54. Paolella, G. & Vajro, P. Childhood Obesity, Breastfeeding, Intestinal Microbiota, and Early Exposure to Antibiotics: What Is the Link? *JAMA Pediatr.* (2016). doi:10.1001/jamapediatrics.2016.0964

55. Laursen, M. F. *et al.* Infant Gut Microbiota Development Is Driven by Transition to Family Foods Independent of Maternal Obesity. *mSphere* **1,**

56. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352,** 539–44 (2016).

57. Claassen-Weitz, S. *et al.* Current Knowledge and Future Research Directions on Fecal Bacterial Patterns and Their Association with Asthma. *Front. Microbiol.* **7,** 838 (2016).

58. David, L. a *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).

59. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).

60. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7,** 830–8 (2013).

61. Moeller, A. H. *et al.* Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat. Commun.* **3,** 1179 (2012).

62. Moeller, A. H. *et al.* Cospeciation of gut microbiota with hominids. *Science (80-.).* **353,** (2016).

63. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G799–807 (2011).

64. Schieber, A. M. P. *et al.* Disease tolerance mediated by microbiome E. coli involves inflammasome and IGF-1 signaling. *Science* **350,** 558–63 (2015).

65. LeBlanc, J. G. *et al.* Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* **24,** 160–168 (2013).

66. Encarnação, J. C., Abrantes, A. M., Pires, A. S. & Botelho, M. F. Revisit dietary fiber on colorectal cancer: butyrate and its role on prevention and treatment. *Cancer Metastasis Rev.* **34,** 465–78 (2015).

67. Rada-Iglesias, A. *et al.* Butyrate mediates decrease of histone acetylation centered on transcription start sites and down-regulation of associated genes. *Genome Res.* **17,** 708–19 (2007).

68. Raqib, R. *et al.* Efficacy of sodium butyrate adjunct therapy in shigellosis: a randomized, double-blind, placebo-controlled clinical trial. *BMC Infect. Dis.* **12,** 111 (2012).

69. Pacheco, R. G. *et al.* Use of butyrate or glutamine in enema solution reduces inflammation and fibrosis in experimental diversion colitis. *World J. Gastroenterol.* **18,** 4278–87 (2012).

70. Hamer, H. M. *et al.* Butyrate enemas do not affect human colonic MUC2 and TFF3 expression. *Eur. J. Gastroenterol. Hepatol.* **22,** 1134–40 (2010).

71. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1219451110

72. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).

73. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).

74. Sobhani, I. *et al.* Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One* **6,** e16393 (2011).

75. Yu, J. *et al.* Invasive Fusobacterium nucleatum may play a role in the carcinogenesis of proximal colon cancer through the serrated neoplasia pathway. *Int. J. cancer* **139,** 1318–26 (2016).

76. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513,** 59–64 (2014).

77. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).

78. Tamboli, C. P., Neut, C., Desreumaux, P. & Colombel, J. F. Dysbiosis in inflammatory bowel disease. *Gut* **53,** 1–4 (2004).

79. Festi, D. *et al.* Gut microbiota and its pathophysiology in disease paradigms. *Dig. Dis.* **29,** 518–24 (2011).

80. Chang, J. Y. *et al.* Decreased diversity of the fecal Microbiome in recurrent Clostridium difficile-associated diarrhea. *J. Infect. Dis.* **197,** 435–8 (2008).

81. Willing, B. *et al.* Twin studies reveal specific imbalances in the mucosa-associated

microbiota of patients with ileal Crohn's disease. *Inflamm. Bowel Dis.* **15,** 653–60 (2009).

82. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–80 (2011).

83. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–8 (2011).

84. Hildebrand, F. *et al.* Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* **14,** R4 (2013).

85. Roager, H. M., Licht, T. R., Poulsen, S. K., Larsen, T. M. & Bahl, M. I. Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet. *Appl. Environ. Microbiol.* **80,** 1142–9 (2014).

86. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–84 (2012).

87. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One* **7,** e30126 (2012).

88. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9,** e1002863 (2013).

89. Cardona, S. *et al.* Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.* **12,** 158 (2012).

90. Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R. & Fierer, N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol. Lett.* **307,** 80–6 (2010).

91. Nechvatal, J. M. *et al.* Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. *J. Microbiol. Methods* **72,** 124–32 (2008).

92. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2,** 19 (2014).

93. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7,** e33865 (2012).

94. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutr. J.* **9,** 23 (2010).

95. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23,** 1704–14 (2013).

96. Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* **31,** 107–33 (1977).

5. ACKNOWLEDGMENTS

I have had the fortune of meeting and working with many extraordinary people during my PhD. Their knowledge, insights and cruel jadedness have made my studies highly enjoyable at times. Working with them has influenced the way I see the world and interact with it. Whether this change was for the better remains to be seen.

I would like to thank the following people for their continued support:

Peer Bork for your supervision and infinite patience in guiding me through the tedious development necessary to achieve this stage; for putting up with my confused emails and my sometimes unfoundedly strong opinions. For introducing me to the academic environment and giving me the tools necessary to navigate it.

My TAC members, **Thomas Dandekar**, **Wolfgang Huber** and **Kiran Patil**, for your valuable input and support during my PhD.

Georg Zeller for the constantly impressive scientist that you are. It has been an honour and a great pleasure to work with you, fight with you and get continuously annoyed at your unproductively high standards.

Shinichi Sunagawa for taking the time to show me the ropes and offering a great role model. I will never be able to be as civil, calculated or relaxed as you, while getting so much work done. **Sean Powell** for always reminding me of the big picture and making me look nice by comparison. **Luis Coelho** for all the really fun arguing that kept my brain from rotting. **Deepikaa Menon** for always making everything fun. Could have lived with less hair in the shower drain though. **Ignacio Ibarra Del Rio** for being a beacon of cheer in an otherwise mostly gloomy existence. **Kristoffer Forslund** for always being so frustratingly nice and encouraging. **Gustaf Lundgren** for just being. **Roman Valls** for the continuous reminder that I can always do just fine in industry. My sister, **Irina Costea** for invaluable feedback and copy-editing of the thesis.

All the members of the **Bork group** for making my stay at EMBL fun and productive. And EMBL itself, for generally managing to keep me from trouble, irrespectively of how much of it I was constantly trying to bring onto myself.

Everyone that had to put up with me during the past four years (though I can be persuaded to extend this to include a longer time span). I know it wasn't easy, but here is something for your trouble: you now all have 1000 worthless internet points! (don't spend them all in one place)

PAPER 1

Voigt et al. Genome Biology (2015) 16:73
DOI 10.1186/s13059-015-0639-8

Genome Biology

RESEARCH Open Access

Temporal and technical variability of human gut metagenomes

Anita Y Voigt[1,2,3], Paul I Costea[1], Jens Roat Kultima[1], Simone S Li[1,4], Georg Zeller[1], Shinichi Sunagawa[1] and Peer Bork[1,3,5*]

Abstract

Background: Metagenomics has become a prominent approach for exploring the role of the gut microbiota in human health. However, the temporal variability of the healthy gut microbiome has not yet been studied in depth using metagenomics and little is known about the effects of different sampling and preservation approaches. We performed metagenomic analysis on fecal samples from seven subjects collected over a period of up to two years to investigate temporal variability and assess preservation-induced variation, specifically, fresh frozen compared to RNALater. We also monitored short-term disturbances caused by antibiotic treatment and bowel cleansing in one subject.

Results: We find that the human gut microbiome is temporally stable and highly personalized at both taxonomic and functional levels. Over multiple time points, samples from the same subject clustered together, even in the context of a large dataset of 888 European and American fecal metagenomes. One exception was observed in an antibiotic intervention case where, more than one year after the treatment, samples did not resemble the pre-treatment state. Clustering was not affected by the preservation method. No species differed significantly in abundance, and only 0.36% of gene families were differentially abundant between preservation methods.

Conclusions: Technical variability is small compared to the temporal variability of an unperturbed gut microbiome, which in turn is much smaller than the observed between-subject variability. Thus, short-term preservation of fecal samples in RNALater is an appropriate and cost-effective alternative to freezing of fecal samples for metagenomic studies.

Background

Microbial communities that inhabit the human gut are essential to human health. To better understand the role of gut microbes in health, major efforts have been undertaken including large-scale studies such as the European Metagenomics of the Human Intestinal Tract (MetaHIT) project and the US American Human Microbiome Project (HMP) [1,2]. These studies have provided insights into the gut microbial community composition in healthy human individuals. Changes in the microbial community composition have been associated with diet [3,4] as well as with multiple diseases, such as atherosclerosis, inflammatory bowel diseases and obesity [5-7].

In addition to these cross-sectional studies that compared healthy and diseased cohorts, longitudinal studies have helped shed light not only on the community compositional variability but also on the temporal variability, providing a more complete picture of the factors that shape the gut microbiome in health and disease. Several studies have demonstrated considerable between-subject variability of the gut microbial composition. However, the gut microbiome has been described to be constrained around a highly personal and stable composition within each healthy subject over time [8-12].

Perturbation of the human gut microbiome is known to occur as a result of antibiotics treatment, a frequently prescribed medication. Antibiotic intervention leads to a rapid decrease of diversity and post-treatment recovery is slow and incomplete, even up to 4 years after the treatment [13-17]. Resistant bacterial species, as a result

* Correspondence: bork@embl.de
[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[3]Molecular Medicine Partnership Unit (MMPU), University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany
Full list of author information is available at the end of the article

of antibiotics treatment, can persist over years [18-20] and the resistance potential of gut microbiota displays regional differences [21,22]. Similarly, there are indications of other long-term community shifts caused by endogenous (for example, disease) or environmental perturbations (for example, diet and lifestyle change [3,4,23]) that have not yet been studied in depth.

Studies on the temporal variability of the gut microbiome have mostly been performed over short periods (weeks to one year; for example, [8,12,23,24]) and only rarely over long periods (5 and 12 years [9,11]). The methods of deriving the taxonomic community composition were primarily based on PCR-denaturing gradient gel electrophoresis (PCR-DGGE; for example, [25]), 16S rRNA gene sequencing (for example, [26,27]), and the HITChip microarray (for example, [11,28]). Only two studies [12,29] have so far analyzed longitudinal non-amplified metagenomic shotgun sequencing data that were collected from 43 subjects in the context of the HMP [1]. However, the majority (41 out of 43) were only sampled twice, making it difficult to assess temporal stability.

Despite their common aim to better understand microbial community shifts over time, the aforementioned studies do not attempt to quantify different sources of variability, from technical to biological ones. In particular, technical aspects have been shown to be important for the comparison between data sets. Limited comparability in human microbiome data sets often results from differences in sample preservation and DNA isolation protocols as well as readout methods (for example, sequencing of different 16S rRNA gene regions or application of different sequencing technologies). A meta-analysis [30] assessing the effect size of technical differences on data comparability showed that samples rather cluster by study or the methods applied (for example, for DNA isolation) than by the parameter of interest (for example, disease state). To counteract these batch effects, the International Human Microbiome Standards (IHMS) project was launched to suggest standards for sample processing (mainly DNA isolation) with the goal to maximize future data comparability. However, different storage conditions of a fecal sample can also impact the compositional readout, as different microbes respond differently to environmental exposure [31]. Research in this direction has been conducted previously to compare different storage and preservation conditions (for example, different temperatures or preservatives such as RNALater) [32,33]. RNALater, a quaternary ammonium salts-based solution, is commonly used as a logistically convenient solution to preserve RNA from biological samples at room temperature when freezing is not possible, and was recently also considered for omics technologies [34]. It was shown to have a minor effect on the recovered composition and thus represents a potential alternative to immediate freezing [35-37]. To date, the technical variability on a taxonomic and functional level has not been put in the context of temporal and within-sample variability (meaning within the stool from a single bowel movement).

We collected fecal samples over up to two years from seven subjects to investigate the temporal variability and individuality of the human gut microbiome using metagenomic shotgun sequencing. To disentangle technical, temporal and between-subject variability we contrasted the variability of microbial community composition within a fecal sample [38,39] with the variability introduced by different preservation methods, RNALater or freezing after two different time intervals. By comparing the fecal metagenomes of the seven subjects over time and in the context of 888 published metagenomes, we generally found between-subject variability to be much larger than within-subject variability. This high degree of individuality can, however, be disrupted by antibiotic treatment, which in one subject triggered a large and long-lasting community shift. Bowel cleanse was also investigated but did not appear to cause a major disturbance. Technical variability (within-sample and preservation-induced variability) was smaller than temporal within-subject variability and therefore we propose RNALater as an alternative to fresh freezing fecal samples.

Results and discussion

Study design

Fecal samples were self-collected from seven adults at short (few days) and longer (weeks to months) time intervals (Additional file 1). All subjects were considered healthy at the time of sampling, unless stated otherwise (see Material and methods). The study was split into five sub-studies as shown in Figure 1. Out of the seven subjects, five subjects performed sampling for more than one year while three subjects collected over more than two years (sub-study 3). At two time points, seven days (d_7; sub-study 1) and 392 days (d_{392}; sub-study 2) after the first sampling event, feces from three and five subjects, respectively, were collected and replicates either frozen or preserved in RNALater. One subject (*Alien*) collected additional fecal samples after antibiotics treatment ($d_{376–380}$, sub-study 4) and bowel cleanse ($d_{630–637}$, sub-study 5).

All fecal samples were subjected to whole genome shotgun sequencing and the data analyzed at species-level using mOTUs (metagenomic operational taxonomic units based on single-copy phylogenetic marker genes [29]), and at a number of functional levels: clusters of orthologous groups (COGs) [40], KEGG (Kyoto Encyclopedia of Genes and Genomes) groups of orthologous genes (KOs), modules and pathways [41].

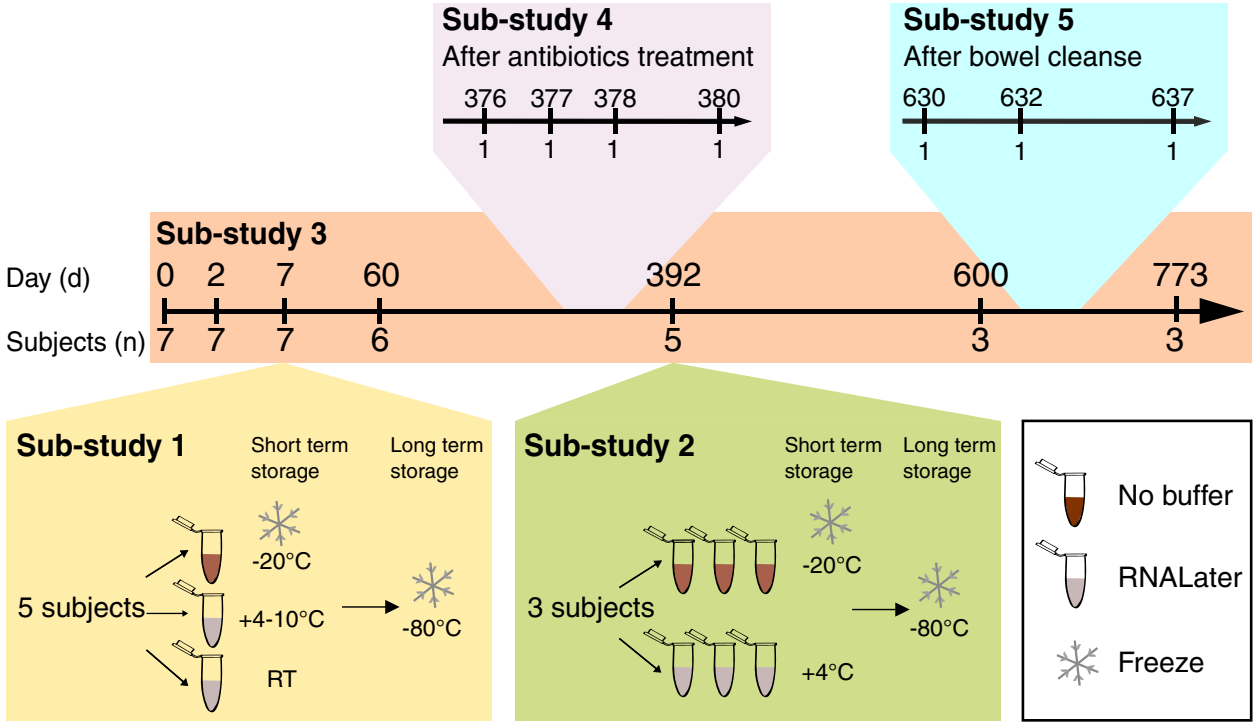


Figure 1 Overview of the study design. This variability assessment was subdivided into five sub-studies. The comparison of technical variability due to fecal preservation methods and stool homogeneity was addressed in sub-studies 1 and 2, respectively. The conditions for the different preservation methods and the numbers of replicates taken are described for these sub-studies. Sub-study 3 was about the temporal variability within and between subjects. One subject underwent antibiotics treatment (sub-study 4) and bowel cleanse (sub-study 5) in the time course of the study. The sampling time points and number of subjects that collected fecal samples are indicated in the timeline.

Preservation-induced variability of the fecal species community

Biological samples are generally stored frozen or processed immediately to maintain their integrity. However, this is often logistically inconvenient, especially in remote areas. In contrast, preservation in RNALater eliminates the need for immediate freezing or sample processing. RNALater is an aqueous solution that preserves biological samples by protecting especially RNA from degradation (for example, [42]). The solution penetrates and stabilizes the sample for later analysis. According to the manufacturer's instruction, these samples are stable at room temperature (RT) for up to one week, at +4°C for one month and at −20°C and −80°C indefinitely. Thus, the usage of RNALater for sample collection would facilitate sample preservation and shipping prior to metagenomic analysis.

We collected fresh samples from which aliquots were frozen immediately (to be used as reference samples) or preserved in RNALater. At d_7, RNALater-preserved samples from five subjects were kept at both +4 to 10°C and RT for one week. However, at d_{392} RNALater-preserved samples were kept at +4°C for 24 h from three subjects before storing at −80°C (sub-studies 1 and 2; Figure 1).

To analyze the taxonomic variability of frozen and RNALater-preserved replicates, we performed hierarchical clustering of the mOTU abundances based on Euclidean distance. This analysis revealed that samples from the same subject clustered together irrespective of the preservation method. This similarity held true for both d_7 and d_{392} with the exception of subject *Alien*, who underwent an antibiotics treatment in between these time points (Figures 2A,B and 3A). Within the cluster of each subject (at d_{392}), the replicates did not cluster by the preservation protocol (Figure 2B), suggesting that biological within-sample variability was larger than preservation-induced effects.

To extend this observation, we clustered all collected samples from all subjects in the context of 888 published metagenomes from MetaHIT and HMP (Figure 4; details in Material and methods). We found that the samples from d_7 and d_{392} had other samples from the same subject as nearest neighbors. All d_7 samples had the other two replicates from d_7 as nearest neighbors (Figure 4). For d_{392}, the first three *Peacemaker* and four *Bugkiller*

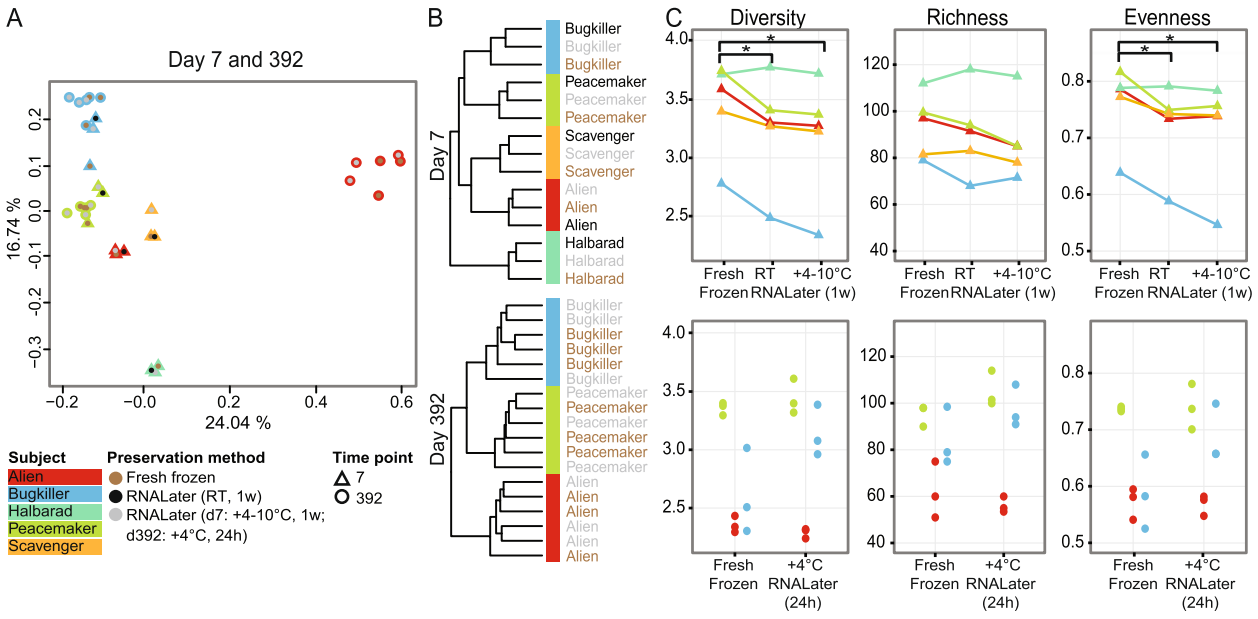


Figure 2 Comparison of methods for fecal sampling. **(A,B)** The samples collected in sub-studies 1 (d_7) and 2 (d_{392}) clustered by subject **(A)** and time point **(B)** but not according to the preservation method (frozen, RNALater (+4 to 10°C or RT for 1 week (1w) on d_7; frozen and RNALater +4°C, 24 h on d_{392}) applied (complete linkage clustering based on Euclidean distance). **(C)** Shannon diversity index, richness and evenness are shown for d_7 (upper panels) and d_{392} (lower panels) and statistically significant differences are indicated by asterisks with *P*-value ≤0.05 (unpaired Wilcoxon test).

neighbors were their corresponding replicates from d_{392}. For subject *Alien*, all samples from d_{392} clustered together but the nearest neighbor was not necessarily the samples preserved under the same condition. Thus, the samples did not cluster by preservation method. Taken together, these results show that RNALater does not introduce a bias in the overall microbiome composition and its effect is smaller than within-subject variability.

We repeated the taxonomic analyses using gene abundances summarized at different functional levels. Relative abundance of orthologous groups, that is, COG and KO profiles, (see Material and methods) of all collected samples were clustered in the context of 888 published metagenomes from MetaHIT and HMP (Additional files 2 and 3). For both COG and KO profiles, the nearest neighbor of samples from d_7 and d_{392} were very similar to those seen in taxonomic clustering (Figure 4). Using COG abundances, with the exception of one *Peacemaker* sample, all d_7 replicates clustered together, and for *Peacemaker* and *Alien* four and five of the d_{392} sample replicates, respectively, clustered together. Using KO profiles, the clustering of samples was similar.

To get a deeper insight into potential preservation-induced changes of the microbiome, we compared indices for species diversity and community evenness. At d_7, RNALater-preserved samples (storage at RT or at +4 to 10°C) compared to the immediately frozen samples showed a significant decrease in their Shannon diversity index ($P = 0.016$ and $P = 0.0008$, unpaired Wilcoxon-test) and species evenness ($P = 0.016$ and $P = 0.016$, unpaired Wilcoxon-test) but not richness ($P = 0.056$ and $P = 0.056$, unpaired Wilcoxon-test). In contrast, at d_{392}, RNALater preservation did not have the same effect on these ecological indices (Figure 2C).

To determine preservation-induced and temporal within-subject and between-subject differences, we correlated mOTU, COGs, KOs, KEGG modules and pathways (Spearman correlation) between different preservation techniques, sampling time points and subjects (Figure 3A,B; Figure S3A-C in Additional file 4). We found that the similarity between protocols is consistently high for both species and COGs (minimum Spearman's r = 0.82 and 0.95, respectively), similar to previous findings [36]. Due to our longitudinal study design we could extend the analysis performed by Franzosa *et al.* [36], and verify that the correlation between time points was lower for both species and COGs (maximum Spearman's r = 0.75 and 0.93, respectively) than between preservation methods. Between-subject correlations were even lower than between-time point correlations.

To estimate differences in taxonomic (species) and functional (eggNOG COGs, and KEGG KOs, modules and pathways) composition between frozen and RNALater-preserved samples from d_7 and d_{392}, we performed two-way ANOVA testing on both the taxonomic and functional relative abundances (see Material and methods). We

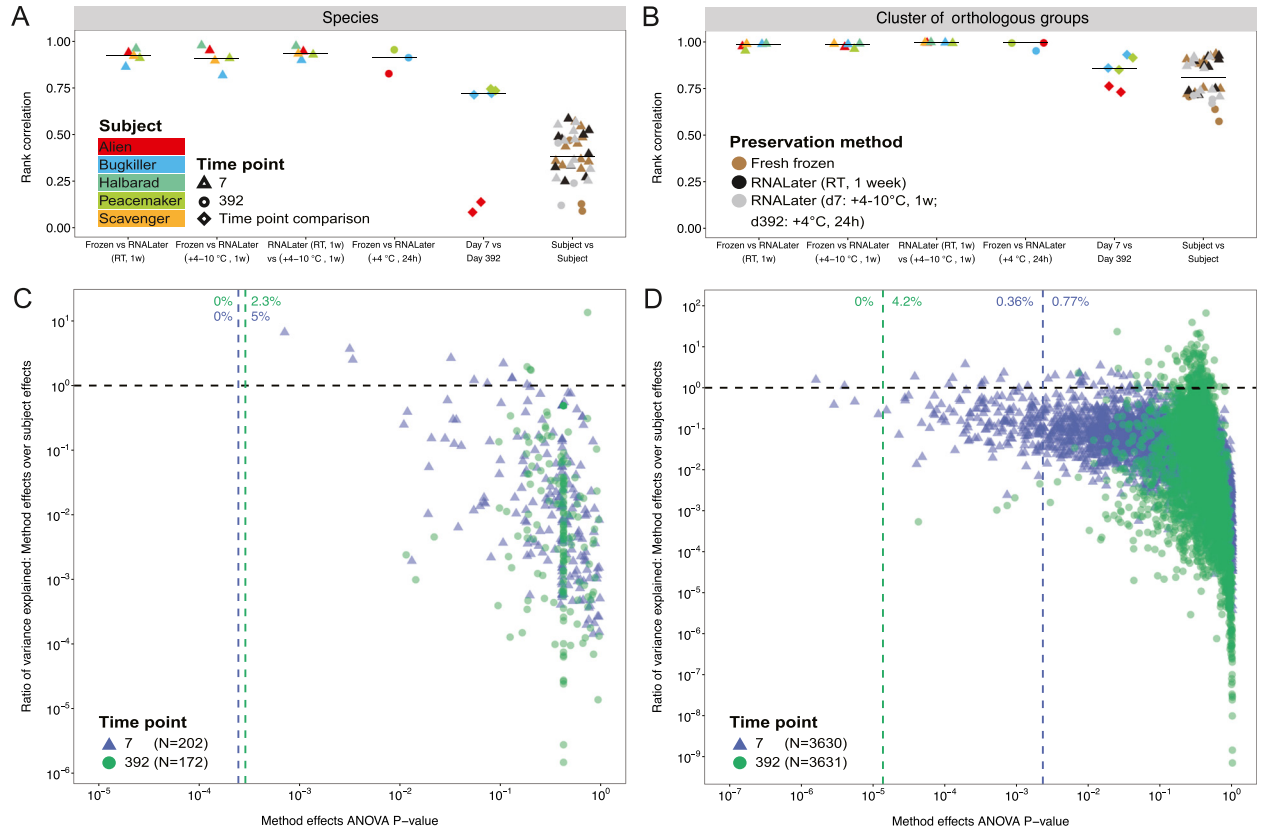


Figure 3 Comparison of technical, temporal within-subject and between-subject variability at taxonomic and functional levels. **(A,B)** Spearman's rank correlations of species profiles **(A)** and cluster of orthologous groups (COG) profiles **(B)** were highly correlated between different preservation methods (frozen versus RNALater (+4 to 10°C or RT for 1 week (1w) each), RNALater at +4 to 10°C versus RNALater at RT and frozen versus RNALater at +4°C, 24 h), less correlated between sampling time points (d_7 versus d_{392}), with lower correlation seen between subjects. Black bars represent group-wise medians. Relatively lower correlation between time points was apparent for the *Alien* samples (red squares), which were taken before and after antibiotics treatment (see main text). **(C,D)** Preservation-induced changes relative to between-subject variability was quantified by two-way ANOVA for species **(C)** and COGs **(D)**, including only features with a relative abundance of at least 0.01% in three or more samples. In total, 7.3% and 5.33% of species and COGs, respectively, showed greater between-method variation than between-subject variation (features above the horizontal black line), but this was statistically significant for only 0.36% of COGs and none of the species tested (Benjamini-Hochberg false discovery rate, $\alpha = 0.05$, Additional file 5). Vertical blue and green lines represent the threshold for statistical significance for d_7 and d_{392}, respectively. Percentages left and right of these lines identify the fractions of statistically significant and insignificant features with larger between-protocol variation for each time point, respectively.

found that 5.0% (d_7) and 2.3% (d_{392}) of the species with a relative abundance exceeding 0.01% in at least 3 of the 33 tested samples varied more between preservation methods than between subjects. However, none of these were statistically significant after correction for multiple hypothesis testing (Benjamin-Hochberg $\alpha = 0.05$; Figure 3C; Additional file 5). For d_7 and d_{392}, 0.77% and 4.2% of the COGs, respectively, varied more between the preservation methods than between subjects, but only 0.36% of COGs were statistically significant (Figure 3D; Additional file 5), which is in the range of previous findings [36]. We found that 0.72%, 0% and 0% of the KOs, modules and pathways, respectively, varied more between preservation methods than between subjects (Figure S3D-F in Additional file 4).

In summary, RNALater appears, in line with a previous publication [36], to be a suitable alternative to immediate freezing at least for short-term storage of a few days, as the variability between protocol replicates is lower than that between time points of the same subject and between subjects.

Within-sample variability of the fecal species community

It was previously shown that there is considerable spatial within-sample variation of parasites in human feces [38] and low abundant bacteria were only sporadically detected in all replicates of the same sample [39]. To address within-sample and technical reproducibility in our study, triplicates at distinct sites of the same fecal

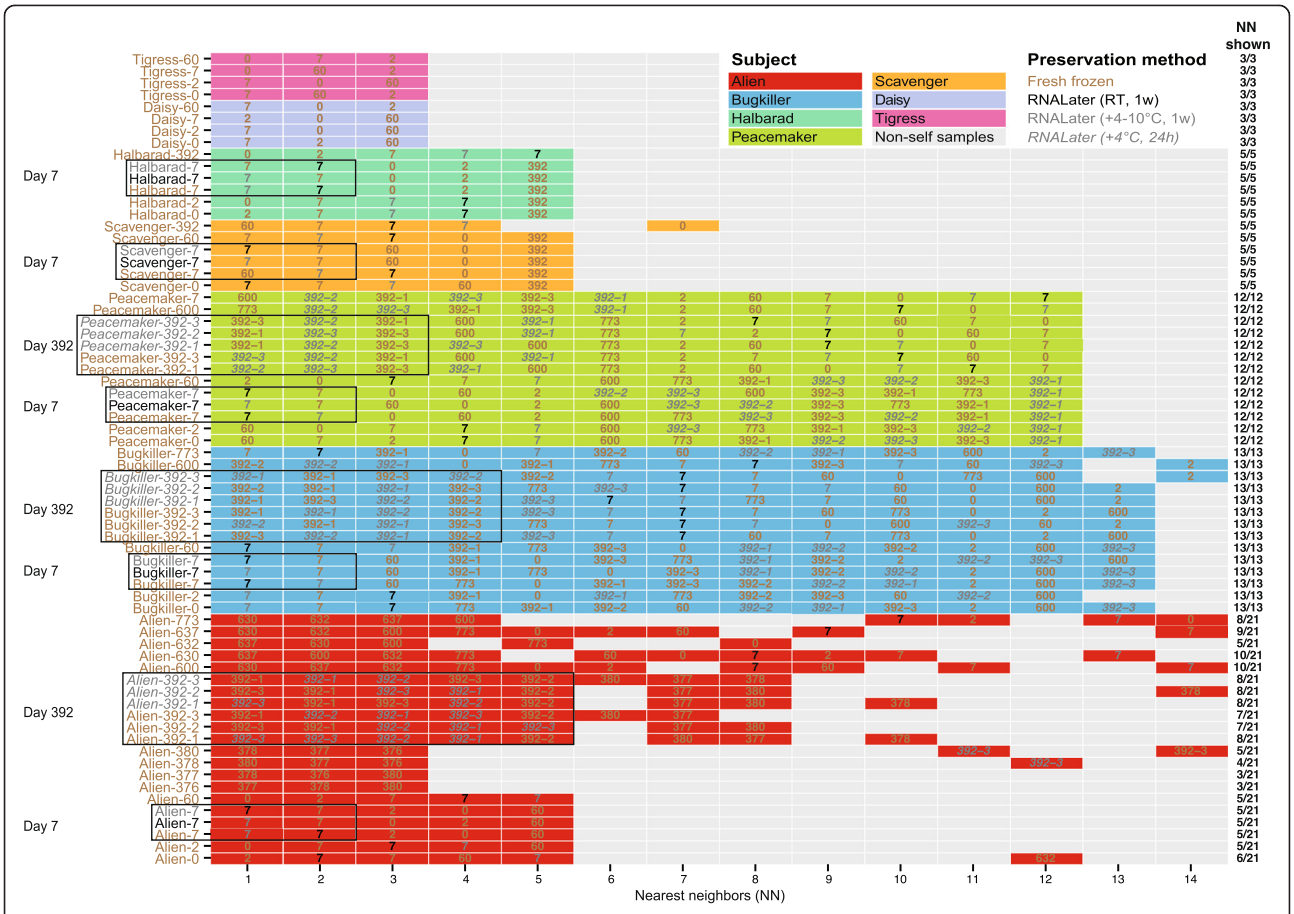


Figure 4 Nearest neighbor plot. The mOTU abundances of the fecal metagenomes of the time series and replicates were clustered in the context of 888 published metagenomes. Only the 14 nearest neighbors (NN) are shown for visual clarity. The colored boxes indicate the respective subject. Non-self samples (samples from another subject, including HMP and MetaHIT) are shaded in grey. Subjects are color-coded, sampling time points are indicated and text color corresponds to the preservation condition of each sample (see key). The column on the right shows how many NNs of each respective sample are depicted, indicating the subject-specificity of the clustering (complete linkage clustering based on Euclidean distances). The figure shows that, with very few exceptions, all time series samples and all fecal replicates (from d_7 and d_{392}) from one subject were closer to each other than to any other sample from another subject. Pre-treatment samples from subject *Alien* were nearest neighbors to each other while the samples right after the treatment ($d_{376–380}$) had highest similarity to each other but not to the pre-treatment samples. The samples collected long after the treatment ($d_{600–773}$) were most similar to each other but a slow recovery to the pre-antibiotics state was visible since pre-treatment samples are among the 14 neighbors shown.

sample were collected from three subjects at d_{392} using two preservation protocols (RNALater and freezing; Figure 1, sub-study 2).

For two subjects the replicates showed only minor variation in ecological indices (Shannon diversity index, species richness and community evenness). Larger fluctuations were detected for diversity and evenness of fresh frozen samples from subject *Bugkiller* only (Figure 2C, lower panel). Nevertheless, all replicates clustered by subject (including *Bugkiller*) in the context of the samples collected on d_{392} (Figure 2B). To set within-sample variation in the context of all time series samples and the MetaHIT and HMP samples (N = 888), we clustered all samples together (Figure 4). The replicates from all three subjects had the other replicates from the same subject as nearest neighbors. All replicates from subject *Alien* clustered by d_{392} but not with the pre-/post-treatment samples, highlighting the drastic change introduced by the treatment. These results for *Alien* remained the same when clustering based on abundances of functional categories (Additional files 2 and 3). This implies that subject-specificity and community similarity is high for all replicates of a fecal sample with only minor fluctuations in diversity and evenness. Together with the fact that replicates preserved under different conditions did not cluster by preservation method (Figures 2B and 4) this supports our study design which was based on samples that were deliberately not homogenized before aliquoting since

(i) samples included in our study were self-collected by lay participants and usually not homogenized in large metagenomic studies and (ii) we aimed to assess within-sample variability.

Temporal variability of fecal microbial communities

In order to assess how technical variability compares to temporal variability, all samples collected by the seven subjects were clustered. It showed that the temporal variability was small and the samples clustered by subject except for *Alien* (Figure 5A,B). Omitting all samples taken from *Alien* after antibiotics treatment resulted in consistent clustering by subject (Figure 5C), showing high subject-specificity and individuality of the gut microbiome. In order to test whether the individuality of the gut microbiome persists on the background of 888 published metagenomes from MetaHIT and HMP, we clustered them together and show the nearest neighbors in Figure 4. The time series samples from the seven subjects were closest to other samples from the same subject rather than to another subject. This was also seen for the 43 subjects in the HMP study, which have multiple time-points. Only few samples from our dataset had the sample of another subject as closer neighbor than a time series sample when comparing the relative taxonomic abundances. For example, the d_{392} sample from *Scavenger* has a sample from another subject as fifth neighbor instead of the *Scavenger* d_0 sample. The number of samples having another subject as closer neighbor rather than a time series sample from the same subject increased when clustering was performed using relative COG and KO abundances (Additional files 2 and 3).

To characterize the temporal variability of the community structure, we calculated the ecological indices, such as the Shannon diversity index, and found that they varied little over time for all subjects (Figure 5A, lower panel for diversity) except *Alien*, who underwent antibiotics treatment.

Our results support previous studies reporting that the temporal variability of the species composition within a subject is smaller than between-subject variability and that in the absence of larger perturbation each individual's microbiota remains relatively stable over time [6,8,9,11,12,14,24,26-29]. However, here we show that even in the context of a large cohort of fecal metagenomes the subjects can be resolved based on the taxonomic composition of their fecal metagenomes with very few exceptions. Thus, the gut microbiome, if unperturbed, is highly subject-specific and the variability is small compared to the between-subject variability.

The effect of perturbations on fecal microbial communities

During the time period of the study, one subject (*Alien*) suffered from an infection that was treated with antibiotics

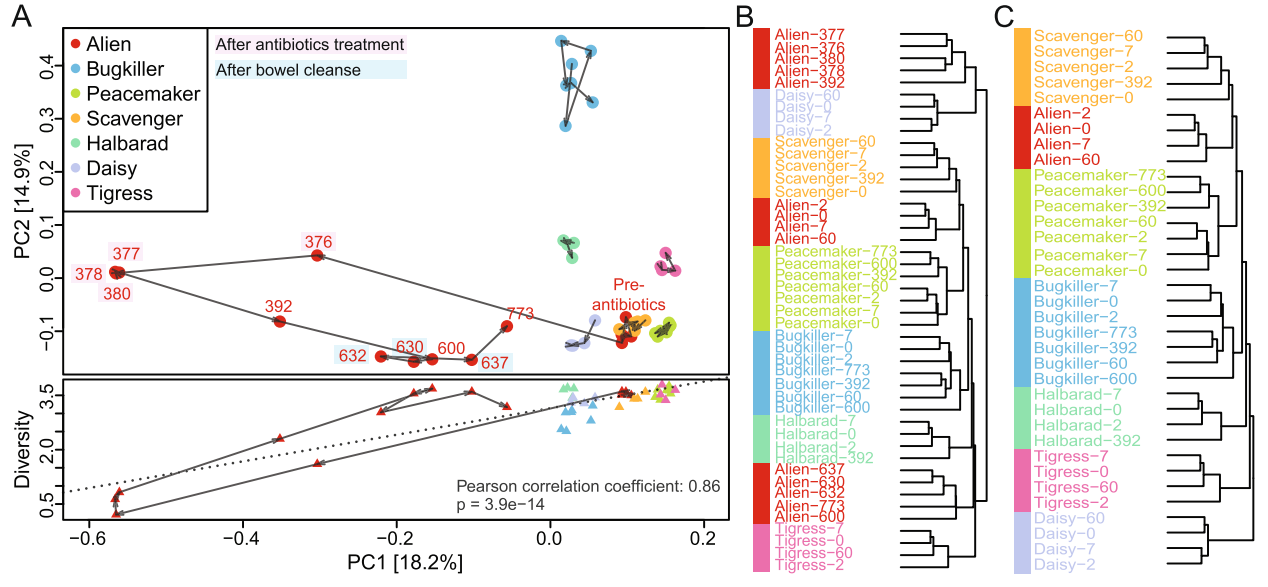


Figure 5 Clustering of the complete time series data set. **(A)** The unperturbed microbiomes of the subjects were stable except for the *Alien* samples, which showed a decline of the Shannon diversity index upon antibiotics treatment ($d_{376–392}$) while the bowel cleanse ($d_{630–637}$) had no detectable effect. The separation of the post-antibiotics from the pre-antibiotics samples along the first principal coordinate (PC1) based on Jensen-Shannon divergence distances was significantly correlated with the decline of the Shannon diversity index (dotted line, *P*-value = $3.9e^{-14}$) (**A**, lower panel) and explained the separate clustering of pre-and post-treatment samples (**A**, upper panel and **(B)**). **(C)** The unperturbed gut microbiome was highly personal and when omitting the post-antibiotics samples, all subjects that collected time series samples over up to two years can be resolved.

and underwent colonoscopy screening, which required bowel cleanse. The antibiotics treatment comprised four days with ceftriaxone, a third-generation cephalosporin antibiotic with broad-spectrum activity against Gram-positive and Gram-negative bacteria. To investigate the consequences of these medical treatments, additional samples were collected for this subject after antibiotics intake (d_{376}, d_{377} and d_{380}) and bowel cleanse (d_{630}, d_{632}, and d_{637}) (Figure 1, sub-studies 4 and 5; for further details see Material and methods).

To observe the response of the fecal microbiome to these two perturbations, we performed hierarchical clustering and found that the post-antibiotics samples separated from the pre-antibiotics samples, but were still distinct from other subjects (Figure 5A,B). The samples taken 226 to 399 days after antibiotics treatment ($d_{600–773}$) clustered closer to the pre-treatment samples (Figure 5A,B), suggesting a (partial) recovery. Determining the nearest neighbor samples in the context of the 888 HMP and MetaHIT samples, using Euclidean distance on taxonomic and functional profiles, confirmed the aforementioned observation suggesting that the gut community composition was still distinct even 399 days (d_{773}) post-antibiotics treatment but gained similarity with the pre-antibiotics community composition (Figure 4). A similar pattern was observed for nearest-neighbor analysis of COG abundances, but individual specificity was less clear for KO abundances (Additional files 2 and 3).

The immediate post-treatment samples ($d_{376–380}$) showed a drastic reduction in Shannon diversity index, species richness and evenness, indicating that fewer and less evenly abundant microbial species were detected. The Shannon diversity index, species richness and evenness of the post-treatment samples dropped from 3.5 to 0.2, 100 to 37 and 0.75 to 0.05, respectively, and were still reduced at d_{392} (18 days after the treatment), compared to the pre-treatment state. At d_{600}, the diversity had returned to its initial level (Figure 5A, lower panel), yet the samples still clustered separately from the pre-treatment samples, indicating that the recovery is not complete (Figures 4 and 5B). Separation of community profiles from the initial state along the first principal coordinate in an ordination analysis (using Jensen-Shannon divergence) correlated with the decline of the Shannon diversity index (Pearson correlation rho = 0.86, $P = 3.9e^{-14}$; Figure 5A).

The bowel cleanse on the day before d_{630} did not have a considerable effect on the community composition: the samples cluster closely with d_{600}, which was before colonoscopy and the fluctuation of the Shannon diversity index was similar to the other subjects and notably smaller than the impact of the antibiotics treatment (Figure 5A,B). Although this case study comprises only one subject, our result that bowel cleanse has little effect on gut microbiome composition is in line with the finding by O'Brien *et al*. [43].

It has been reported that antibiotics have a strong impact on the gut microbial community composition for an extended period of time, although the community was sometimes found to be similar to its pretreatment state within weeks. The return was subject-dependent and often incomplete, at least for some species monitored for time periods of two to six months [14,17] and up to two years [19,44]. Even though we studied the effect of antibiotics in only one subject, we can show that, at least in this subject, despite species diversity recovery, the gut microbial composition was still distinct from the pre-treatment state, even 399 days after the antibiotics treatment. It would be worthwhile exploring in the future how antibiotics effects vary between subjects and depends on factors such as dosage, duration of the treatment and type of antibiotics.

Conclusion

Several studies have addressed the temporal variability or the technical variability (for example, induced by different DNA isolation methods or preservation techniques) of the gut microbiome separately but none set them in a broader context. Hence, these studies have so far not disentangled the biological temporal variability (like, for example, community shifts due to disease or medication) from technical variability (for example, induced by preservation conditions or insufficient stool homogeneity).

In our study we provide the to date largest metagenomic data set of fecal samples collected over more than two years. We addressed the aspects of comparing technical and temporal variability, finding that temporal variability within each subject's gut microbiome was smaller than that between subjects. Even in the context of 888 metagenomes, all time series samples could be recovered using taxonomic abundances, as long as antibiotics did not perturb the gut microbiome. The technical variability introduced by RNALater was small compared to freezing, for both taxonomic and functional features, and does not disrupt subject-specificity nor time point-specificity of the gut microbiome. Thus, we suggest RNALater as an alternative to freezing for the preservation of the fecal microbiome for metagenomic studies.

Material and methods

Sample collection

Fecal sample collection for time series

Informed consent to obtain time series samples of fecal samples was obtained from seven healthy subjects in Germany through the my.microbes project [45]. The study protocol was approved by the EMBL Bioethics Internal Advisory Board, and is in agreement with the

WMA Declaration of Helsinki. All subjects were living in Heidelberg, Germany at the beginning of the study and the mean age of the subjects upon enrollment was 34 ± 6 years. Among these subjects were five males (*Alien*, *Bugkiller*, *Peacemaker*, *Halbarad* and *Scavenger*) and two females (*Daisy* and *Tigress*). Subjects reported themselves as healthy, if they did not undergo prescribed medical treatment or showed any indication of disease symptoms. Fecal samples were collected and conserved under anaerobic conditions in a sealed bag, kept at −20°C for short-term storage and stored at −80°C upon arrival in the laboratory. The fecal samples were collected at days 0, 2, 7, 60, 392, 600 and 773 (sub-study 3) and are referred to here as d_0, d_2, d_7 and so on. One male subject (*Alien*) contracted a bacterial infection and collected further samples after being hospitalized and receiving 2 g of ceftriaxone. Ceftriaxone is an antibiotic with broad-spectrum activity against Gram-positive and Gram-negative bacteria that was administered parenterally over 4 days. The last injection was two days before the first sampling time point (d_{376}) and further samples were then collected on the subsequent days (d_{377}, d_{378} and d_{380}; sub-study 4). Additional samples were taken starting one day after undergoing bowel cleanse for routine colonoscopy (d_{630}, d_{632} and d_{637}; sub-study 5). Figure 1 shows the study design in detail and metadata and sequencing information are given in the Additional file 1.

Fecal sample collection for method comparison

In parallel to the fresh frozen fecal samples, additional samples (1 g each) were collected from five subjects at time point d_7 and from three subjects at d_{392} (without homogenization) and were stored in 10 ml RNALater® Stabilization Solution (Life Technologies GmbH, Darmstadt, Germany). Short-term storage was either at +4 to 10°C or at RT for one week (d_7, sub-study 1) or at +4°C (d_{392}, sub-study 2) for 24 h and frozen at −80°C upon arrival in the laboratory. At d_{392}, each subject collected samples in triplicate, preserved in both RNALater and freshly frozen (Figure 1).

Inclusion of published fecal metagenomes

Published metagenomes from MetaHIT [2,46,47] and HMP [1] were included in our study to set our time series in context of a large collection of metagenomes.

Sample processing and sequencing

DNA isolation from fecal samples

One milliliter of defrosted samples immersed in RNALater was taken and diluted with sterile phosphate-buffered saline and pelleted by centrifugation. Genomic DNA was extracted from frozen or RNALater-preserved fecal samples as previously described [48] using the G'NOMEs kit (MP Biomedicals, Illkirch, France). The following minor modifications were made to the protocol: cell lysis/denaturation was performed (30 minutes, 55°C) before protease digestion was carried out overnight (55°C). Mechanical lysis was followed by RNAse digestion (50 μl, 30 minutes, 55°C). The purified DNA was resuspended in TE buffer after final precipitation for storage at −20°C.

Library preparation and metagenomic sequencing

Library generation and whole genome shotgun sequencing of the fecal samples was carried out on the Illumina HiSeq 2000/2500 (Illumina, San Diego, CA, USA) platform as described in Zeller *et al*. [49]. All samples were paired-end sequenced with 100 bp read lengths at the Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg, to a sequencing depth of approximately 5 Gbp (see Additional file 1 for sequencing results).

Data processing

Taxonomic profiling of fecal samples

Using MOCAT [50], a software package used to process raw Illumina reads to generate taxonomic and functional profiles (option screen with alignment length cutoff 45 and minimum 97% sequence identity), taxonomic relative abundance profiles were generated by mapping screened HQ reads from each metagenome to a database consisting of 10 universal single-copy marker genes extracted from 3,496 NCBI reference genomes and 263 human gut metagenomes that had previously been clustered and linked by co-variance into mOTUs [29,51]. Quantification of mOTU linkage groups was performed using MOCAT, but is also available as a standalone tool at [29].

Functional profiling of fecal samples

Using MOCAT [50] (option screen with alignment length cutoff 45 and minimum 95% sequence identity) functional relative abundance profiles were generated by first calculating gene abundance profiles by mapping screened HQ (high quality) reads from each metagenome to an functionally annotated database consisting of predicted genes from 263 human gut metagenomes [29,49], and estimating each gene's abundance as gene length-normalized nucleotide counts of all reads that matched the protein-coding region of the gene. And second, for each functional feature, its abundance in the metagenomic gene pool was estimated as the sum of the relative abundances of all genes belonging to this family. The genes were summarized into COGs [40], and KEGG KOs, modules and pathways [41]. The metagenomic gene catalog had already been functionally annotated to the KEGG database [48], and was additionally annotated to different COGs by aligning the translated

amino acid sequence of each gene to the eggNOG (version 3) [40] database using BLAST (version 2.2.24) [52] (maximum e-value 0.01) and then annotating the genes using SmashCommunity (version 1.6) [53].

Data analysis

For the statistical data analysis at the species level, mOTU abundances were used [29] and samples were included in the data analysis if they had more than 3,800 insert counts.

Ecological indices

For the comparison of RNALater with fresh frozen feces and time series samples with each other, mOTU abundances [29] were used to calculate Shannon diversity index, evenness and species richness. To standardize sampling depth, richness, Shannon diversity index and evenness were assessed after rarefaction of the insert count tables to 3,800 insert counts per sample. Differences were assessed using the Wilcoxon test (unpaired) on the deviation from the mean of each subject. *P*-values ≤0.05 were considered statistically significant.

Clustering

Principal coordinate analyses and complete linkage clustering of Euclidean distances (Figure 2A) and Jensen-Shannon divergence distances (Figure 5A) were performed using the *ape* and *ade4* R packages. The dendrograms shown are based on Euclidean distance measurements on the logged abundances (Figures 2B and 3B,C). Nearest neighbors were determined to be the samples with the smallest Euclidean distance (Figure 4; Additional files 2 and 3). Due to large differences in sequencing depth, the metagenomes collected in this study and the HMP [1] and MetaHIT [2,46,47] taxonomic data were only analyzed after rarefaction to an insert count of 5,000 per sample. All samples that passed these criteria were included in the functional analysis.

Two-way ANOVA

To observe taxonomic and functional-specific biases introduced by RNALater preservation across all subjects, a two-way ANOVA was performed for d_7 and d_{392} separately. Our setup was analogous to a previous analysis [36]. Relative species, COG, KO, module and pathway abundances were arcsine square root transformed (for variance stabilization) and only features with a relative abundance of more than 0.01% in at least three samples were included. For d_{392}, the median value of the three replicates for each feature was used.

Data availability

The shotgun metagenomic sequencing data from this study are available from the European Nucleotide Archive (ENA) database [54], accession number ERP009422.

Description of additional data files

The following additional data are available with the online version of this paper. Additional file 1 is a table listing the metadata and sequencing information of the analyzed samples. Additional file 5 is a table listing statistically significant taxonomic and functional features resulting from the method comparison.

Additional files

Additional file 1: Table S1. Overview of the metadata of the subjects and sequencing information of the samples included.

Additional file 2: Figure S1. Nearest neighbor plot based on COGs.

Additional file 3: Figure S2. Nearest neighbor plot based on KOs.

Additional file 4: Figure S3. Comparison of technical, temporal and between-subject variability based on functional profiles.

Additional file 5: Table S2. Overview of taxonomic and functional features. The features include those above the horizontal black line in Figure 3 and Additional file 4.

Abbreviations

COG: cluster of orthologous groups; HMP: Human Microbiome Project; KEGG: Kyoto Encyclopedia of Genes and Genomes; KO: KEGG group of orthologous genes; MetaHIT: Metagenomics of the Human Intestinal Tract; mOTU: metagenomic operational taxonomic unit; PCR: polymerase chain reaction; RT: room temperature.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PB, SS, AYV conceived and managed the project. AYV, SS, PIC, JRK, GZ, SSL and PB designed and performed data analysis. AYV wrote the manuscript with contributions from all other authors. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to the members of the Bork group for inspiring discussions. We thank the volunteers for collecting the samples and acknowledge Yan Ping Yuan and the EMBL Information Technology Core Facility for support with high-performance computing as well as the EMBL Genomics Core Facility for sequencing support. We are grateful for statistical advice from Bernd Klaus from the Centre for Statistical Data Analysis at EMBL. We are also grateful to the European MetaHIT consortium and the NIH Common Fund Human Microbiome Project Consortium for making available the data sets that were used in this study. This work has received funding through the CancerBiome project (European Research Council project reference 268985), the METACARDIS project (FP7-HEALTH-2012-INNOVATION-I-305312), the International Human Microbiome Standards project (HEALTH-FP7-2010-261376), Syste*Mtb* grant (HEALTH-FP7-2010-241587) and funding from EMBL. SSL is the recipient of an Australian Postgraduate Award, and EMBL Australia International PhD Fellowship.

Author details

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [2]Department of Applied Tumor Biology, Institute of Pathology, University Hospital Heidelberg, 69120 Heidelberg, Germany. [3]Molecular Medicine Partnership Unit (MMPU), University of Heidelberg and European Molecular Biology Laboratory, 69120

Heidelberg, Germany. [4]School of Biotechnology and Biomolecular Sciences, University of New South Wales, 2052 Sydney, Australia. [5]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.

Received: 11 February 2015 Accepted: 19 March 2015
Published online: 08 April 2015

References

1. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486:215–21.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
3. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334:105–8.
4. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63.
5. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, et al. Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med. 2013;19:576–85.
6. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457:480–4.
7. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55:205–11.
8. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. Science. 2009;326:1694–7.
9. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science. 2013;341:1237439.
10. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
11. Rajilic-Stojanovic M, Heilig HG, Tims S, Zoetendal EG, de Vos WM. Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol. 2012;15:1146–59.
12. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50.
13. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 2008;6:e280.
14. Dethlefsen L, Relman DA. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc Natl Acad Sci U S A. 2011;108:4554–61.
15. Jakobsson HE, Jernberg C, Andersson AF, Sjolund-Karlsson M, Jansson JK, Engstrand L. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PLoS One. 2010;5:e9836.
16. Perez-Cobas AE, Gosalbes MJ, Friedrichs A, Knecht H, Artacho A, Eismann K, et al. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut. 2012;62:1591–601.
17. De LaCochetiere MF, Durand T, Lepage P, Bourreille A, Galmiche JP, Dore J. Resilience of the dominant human fecal microbiota upon short-course antibiotic challenge. J Clin Microbiol. 2005;43:5588–92.
18. Andersson DI, Hughes D. Persistence of antibiotic resistance in bacterial populations. FEMS Microbiol Rev. 2011;35:901–11.
19. Jernberg C, Lofmark S, Edlund C, Jansson JK. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. ISME J. 2007;1:56–66.
20. Jernberg C, Lofmark S, Edlund C, Jansson JK. Long-term impacts of antibiotic exposure on the human intestinal microbiota. Microbiology. 2010;156:3216–23.
21. Forslund K, Sunagawa S, Coelho LP, Bork P. Metagenomic insights into the human gut resistome and the forces that shape it. Bioessays. 2014;36:316–29.
22. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, et al. Country-specific antibiotic use practices impact the human gut resistome. Genome Res. 2013;23:1163–9.
23. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014;15:R89.
24. Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, et al. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc Natl Acad Sci U S A. 2011;108:4586–91.
25. Zoetendal EG, Akkermans AD, De Vos WM. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl Environ Microbiol. 1998;64:3854–9.
26. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome Biol. 2011;12:R50.
27. Martinez I, Muller CE, Walter J. Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. PLoS One. 2013;8, e69621.
28. Jalanka-Tuovinen J, Salonen A, Nikkila J, Immonen O, Kekkonen R, Lahti L, et al. Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. PLoS One. 2011;6:e23035.
29. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10:1196–9. http://www.bork.embl.de/software/mOTU.
30. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23:1704–14.
31. Swan BK, Ehrhardt CJ, Reifel KM, Moreno LI, Valentine DL. Archaeal and bacterial communities respond differently to environmental gradients in anoxic sediments of a California hypersaline lake, the Salton Sea. Appl Environ Microbiol. 2010;76:757–68.
32. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett. 2010;307:80–6.
33. Flores R, Shi J, Gail MH, Gajer P, Ravel J, Goedert JJ. Assessment of the human faecal microbiota: II. Reproducibility and associations of 16S rRNA pyrosequences. Eur J Clin Invest. 2012;42:855–63.
34. van Eijsden RG, Stassen C, Daenen L, Van Mulders SE, Bapat PM, Siewers V, et al. A universal fixation method based on quaternary ammonium salts (RNAlater) for omics-technologies: Saccharomyces cerevisiae as a case study. Biotechnol Lett. 2013;35:891–900.
35. Vlckova K, Mrazek J, Kopecny J, Petrzelkova KJ. Evaluation of different storage methods to characterize the fecal bacterial communities of captive western lowland gorillas (Gorilla gorilla gorilla). J Microbiol Methods. 2012;91:45–51.
36. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111:E2329–38.
37. Nechvatal JM, Ram JL, Basson MD, Namprachan P, Niec SR, Badsha KZ, et al. Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. J Microbiol Methods. 2008;72:124–32.
38. Krauth SJ, Coulibaly JT, Knopp S, Traore M, N'Goran EK, Utzinger J. An in-depth analysis of a piece of shit: distribution of Schistosoma mansoni and hookworm eggs in human stool. PLoS Negl Trop Dis. 2012;6:e1969.
39. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol. 2010;10:206.
40. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012;40:D284–9.
41. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36:D480–4.
42. Mutter GL, Zahrieh D, Liu C, Neuberg D, Finkelstein D, Baker HE, et al. Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. BMC Genomics. 2004;5:88.
43. O'Brien CL, Allison GE, Grimpen F, Pavli P. Impact of colonoscopy bowel preparation on intestinal microbiota. PLoS One. 2013;8, e62815.
44. Lofmark S, Jernberg C, Jansson JK, Edlund C. Clindamycin-induced enrichment and long-term persistence of resistant Bacteroides spp. and resistance genes. J Antimicrob Chemother. 2006;58:1160–7.

45. Jones N. Social network wants to sequence your gut. Nature. 2011. doi:10.1038/news.2011.523.
46. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500:541–6.
47. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41.
48. Furet JP, Firmesse O, Gourmelon M, Bridonneau C, Tap J, Mondot S, et al. Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. FEMS Microbiol Ecol. 2009;68:351–62.
49. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10:766.
50. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7:e47656.
51. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nat Methods. 2013;10:881–4.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
53. Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. SmashCommunity: a metagenomic annotation and analysis tool. Bioinformatics. 2010;26:2977–8.
54. European Nucleotide Archive. http://www.ebi.ac.uk/ena.

PAPER 2

TOWARDS STANDARDS FOR HUMAN FECAL SAMPLE PROCESSING IN METAGENOMIC STUDIES

Authors: Paul I. Costea[1], Georg Zeller[1], Shinichi Sunagawa[1], Eric Pelletier[2,3,4], Adriana Alberti[2], Florence Levenez[5], Jens Roat Kultima[1], Matthew R. Hayward[1], Emma Allen-Vercoe[6], Laurie Bertrand[2], Michael Blaut[7], Jillian Brown[8], Thomas Carton[9], Stéphanie Cools-Portier[10], Michelle Daigneault[6], Muriel Derrien[10], Anne Druesne[10], Willem M. de Vos[11,12], B. Brett Finlay[13], Harry J. Flint[14], Francisco Guarner[15], Masahira Hattori[16,17], Hans Heilig[11], Johan van Hylckama Vlieg[10], Jana Junick[7], Ingeborg Klymiuk[18], Philippe Langella[5], Emmanuelle Le Chatelier[5], Volker Mai[19], Chaysavanh Manichanh[15], Jennifer C. Martin[14], Clémentine Mery[9], Hidetoshi Morita[20], Paul O'Toole[8], Céline Orvain[2], John Penders[21], Søren Persson[22], Nicolas Pons[5], Milena Popova[9], Anne Salonen[12], Delphine Saulnier[7], Karen P. Scott[14], Bhagirath Singh[23], Kathleen Slezak[7], Patrick Veiga[10], James Versalovic[24], Liping Zhao[25], Erwin G. Zoetendal[11], S. Dusko Ehrlich[5,26,*], Joel Dore[5,*], Peer Bork[1,*]

* - Corresponding authors

[1] European Molecular Biology Laboratory, Germany.

[2] CEA-Institut de Génomique, Genoscope, Centre National de Séquençage, Evry, France.

[3] CNRS UMR8030, Evry France

[4] Université Evry Val d'Essonne, Evry, France

[5] Metagenopolis, Institut National de la Recherche Agronomique, Jouy en Josas, France

[6] The University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

[7] Department of Gastrointestinal Microbiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany

[8] School of Microbiology & APC Microbiome Institute, University College Cork, T12 Y337 Cork, Ireland

[9] Biofortis, Mérieux NutriSciences, France

[10] Life Science, Danone Nutricia Research, Palaiseau, France

[11] Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE, Wageningen, The Netherlands

[12] Department of Bacteriology and Immunology, Immunobiology Research Program, Haartmaninkatu 3 (PO Box 21), FIN-00014 University of Helsinki, Finland

[13] Michael Smith Laboratories, University of British Columbia, Vancouver, B.C., Canada

[14] Rowett Institute of Nutrition and Health, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK

[15] MetaLab, Gastroenterology Dpt, VHIR, Barcelona, Spain

[16] Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan

[17] Graduate School of Advanced Science and Engineering, Waseda University, Tokyo 169-8555, Japan

[18] Center for Medical Research, Medical University of Graz, Graz, Austria

[19] Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, Emerging Pathogens Institute, University of Florida, 2055 Mowry Rd., Gainesville, FL 32610-0009, United States

[20] Graduate School of Environmental and Life Science, Okayama University, Okayama 700-8530, Japan

[21] School of Nutrition and Translational Research in Medicine (NUTRIM) and School for Public Health and Primary Care (Caphri), Department of Medical Microbiology, Maastricht University Medical Center, Maastricht, The Netherlands

[22] Unit of Foodborne Infections, Dept. Microbiology and Infection Control, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen, Denmark

[23] Centre for Human Immunology, Department of Microbiology & Immunology and Robarts Research Institute, University of Western Ontario, London, Ontario N6A 5C1 Canada

[24] Texas Children's Hospital, 1102 Bates Avenue, Feigin Center, Houston, TX 77030, United States

[25] Ministry of Education Key Laboratory for Systems Biomedicine, Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, PR China

[26] King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, UK

ABSTRACT

Metagenomic analysis of fecal samples suffers from challenges in achieving high comparability and reproducibility that need to be addressed in order to better establish microbiota contributions to human health. To test and improve current protocols, we quantified the effect of the DNA extraction process on the observed variation in microbial composition, by comparing 21 representative protocols. Furthermore, we determined the contributions of sequencing, sample storage and biological variability, and show that the DNA extraction process is the strongest technical factor to impact the results. We characterized the biases of different methods, introduced a quality scoring scheme and quantified transferability of the best methods to different labs. Finally, we propose a standardized sample handling and DNA extraction methodology for human fecal samples. Its use will greatly improve the comparability and consistency of different human gut microbiome studies and facilitate future meta-analyses.

INTRODUCTION

Over 3000 publications in the past five years have used DNA- or RNA- based profiling methods to interrogate microbial communities in locations ranging from ice columns in the remote arctic to the human body, resulting in more than 160,000 published metagenomes (both shotgun and 16S)[1]. To date, one of the most studied ecosystems is the human gastrointestinal tract. The gut microbiome is of particular interest due to its large volume, high diversity and potential relevance to human health and disease. Numerous studies have found specific microbial fingerprints that may be useful in distinguishing disease states, for example diabetes[2–4], inflammatory bowel disease[5,6] or colorectal cancer[7]. Others have linked the human gut microbial composition to various factors, such as mode of birth, age, diet and medication[8–11]. Such studies have almost exclusively used their own specific, demographically distinct cohort and methodology. Given the many reports of batch effects[12] and known differences when analyzing data generated using different protocols[13–18], comparisons or meta-analyses will be hampered and, limited in their interpretability. For example, healthy Americans from the HMP study showed lower taxonomic diversity in their stool, than patients with inflammatory bowel disease (IBD) from a European study[19], although it is established that IBD patients worldwide have reduced taxonomic diversity[20]. It is thus currently very difficult to disentangle biological from technical variation when comparing multiple studies[21].

In metagenomic studies, the calculation of compositional profiles and ecological indices is preceded by a complex data generation process, consisting of multiple steps (Figure 1), each of which is subject to technical variability[22]. Usually, a small sample is collected by an individual shortly after passing stool and stored in a domestic freezer, prior to shipment to a laboratory. The location within the specimen that the sample is taken from has been shown to impact the measured composition[23], which is why in some studies[24] larger quantities were homogenized prior to storage in order to generate multiple, identical aliquots. Furthermore, different fixation methods can be used to preserve the sample for shipping and long-term storage. Freezing at below -20°C is the standard, though more practical alternatives exist[23–25]. Eventually, the sample is subjected to DNA extraction, library preparation, sequencing and downstream bioinformatics analysis (Figure 1).

Here we examined to what extend DNA extraction influences the quantification of microbial composition[16,26,27], as compared to other sources of technical and biological variation[23,24,28]. We further compared a wide range of extraction methods, using metagenomic shotgun sequencing, in respect to both taxonomic and functional variability, while keeping all other steps standardized. We investigated the most commonly used extraction kits with varying

modifications and additional protocols, which do not make use of commercially available kits (see Supplementary Table 1 and Supplementary Information). While other studies have previously investigated the differences between extraction methods in a given setting[12,15,16,29], we here systematically tested for reproducibility within and across laboratories on three continents, by applying strict and consistent quality criteria. In addition, we assessed the impact of a more automated library preparation method, for increased throughput of the extraction pipeline. Based on these analyses we recommend a standardized protocol for DNA extraction alongside a library preparation method for application to human stool samples which should serve as a benchmark for new methods and will greatly enhance comparability among metagenomic studies.

RESULTS

STUDY DESIGN

This study consisted of two phases. In the first phase, in order to assess the variability introduced by different extraction methods, we produced multiple aliquots of two stools samples (obtained from two individuals, referred to as sample A and B). Within two hours of emission, the samples were homogenized in an anaerobic cabinet to ensure that the different aliquots have identical microbial compositions, and subsequently aliquoted in 200mg amounts, frozen at -80°C within four hours and shipped frozen on dry-ice, to 21 collaborating laboratories, spanning 11 countries over three continents. These laboratories employed extraction methodologies ranging from the seven most commonly used extraction kits (Invitek's PSPStool, Mobio's PowerSoil, Omega Bio Tek's EZNAstool, Qiagen's QiIAampStoolMinikit, Bio101's G'Nome, MP-Biomedicals's FastDNAspinSoil and Roche's MagNAPureIII) to non-kit-based protocols (Supplementary Table 1 and Supplementary Information). Once extracted, the DNA was shipped to a single sequencing center (GENOSCOPE, France), which tested two different library preparation methods, before performing identical sequencing and analytical methods in an attempt to minimize other possible sources of variation.

After applying a panel of quality criteria, including quantity and integrity of extracted DNA, recovered diversity and ratio of recovered gram-positive bacteria, we selected five protocols (1, 6, 7, 9, and 15) for the second phase. Extractions were then performed in the initial laboratory applying the protocol and in three other laboratories, which had not used the method before, in order to assess reproducibility of these protocols and their transferability between laboratories. For the same samples A and B, three replicates/aliquots were provided per sample per laboratory, as detailed above.

QUALITY CONTROL FOR DNA YIELD AND FRAGMENTATION

Maximizing DNA concentration while also minimizing fragmentation are key aspects to consider when selecting an extraction protocol. This is both because good quality libraries are required for shotgun sequencing and because protocols that consistently recover low yield or highly fragmented DNA are likely to skew the measured composition. We found considerable variation in the quantity of extracted DNA, in line with previous observations[27] (Figure 2). For example, protocol 18 recovered 100 times more DNA than protocols 3 and 12, and 10 times more than protocols 8, 19 and 20 (Figure 2). Furthermore, there was considerable variation in the fragmentation of the recovered DNA, as measured by the percentage of total DNA in fragments below 1.8 kb in length; for example protocols 4, 10 and 12 consistently yielded highly fragmented DNA while for protocol 1 no fragmentation was noticeable. For subsequent analysis,

samples that yielded below 500ng of DNA or were very fragmented (median sample fragmentation above 25%), were not subjected to sequencing. In total, 143 libraries, extracted using 21 different protocols passed the quality requirements imposed above, though as an example only four of 18 samples extracted with protocol 16 (one sample A and three sample B replicates) met the requirements (Supplementary Table 2). For other protocols, a small number of samples were discarded for lack of compliance with quality/quantity criteria.

QUALITY CONTROL FOR VARIABILITY IN TAXONOMIC AND FUNCTIONAL COMPOSITION

All metagenomes were compared with respect to taxonomic and functional compositions to quantify the relative abundances of microbial taxa and their respective gene-encoded functions (Methods). Briefly, based on the extracted DNA, shotgun sequencing libraries were prepared and subjected to sequencing on the Illumina HiSeq2000 platform, yielding a mean of 3.8 Gb (+/- 0.7 Gb) per sample. Raw sequencing data were then processed using the MOCAT[30] pipeline and relative taxonomic and gene functional abundances were computed by mapping high-quality reads to a database of single copy taxonomic marker genes (mOTUs)[19] and annotated human gut microbial reference genes[31], respectively (Methods).

There are, as outlined above (Figure 1), many steps in which sample handling can differ and batch effects can be introduced. The resulting variation in taxonomic and gene functional composition estimates should be considered in terms of both effect size and consistency: if protocol differences lead to an effect larger than the biological variation of interest (e.g. in an intervention study), it will mask that signal. Effects that are systematically consistent (biases) will introduce "batch effects" that can confound any meta-analysis even if their absolute size is comparatively small. It is thus important to minimize these biases in order to facilitate cross-study comparisons.

To contextualize the magnitude of the extraction effect, we compared the technical variation quantified here (caused by extraction protocol) to other technical and biological effects (Figure 3), assessed on data from multiple other studies[23,24,28] (Methods). The biggest biological difference assessed here was that observed between individuals. Next was the within individual variation, as measured between different sampling time points. This effect was much smaller than the between individual variation, resulting in individual-specific microbial composition preserved over time as noted before[19,23,32]. Both these effects are quantified using time-point data from the Human Microbiome Project[28]. The smallest biological effect considered was within specimen variation, resulting from sampling different parts of the stool itself, as quantified in Voigt et al.[23]. In terms of technical sources of variation we have considered measurement errors (assessed through technical replication), library preparation, and effects induced by preservation[23,24] and extraction. It is important to note that these effects have not all been measured independently of each other, resulting in some of the quantified variations being a convolute of multiple effects (Figure3 – checkboxes).

Different distance measures can be used to assess the magnitude of these effects. We focused here on two, which are complementary in terms of the features of the data they consider and thus the dimensions which become relevant. These distance measures were computed on both metagenomics operational taxonomic units (mOTUs[21]) and clusters of orthologous groups (COGs[33]) abundance data, to derive species and functional variation (see Methods). Firstly, we used a Spearman correlation to assess how well species abundance rankings are preserved and found that the variation between most extraction protocols is smaller than the technical within-specimen variation (summarized by the median, Figure 3a). This suggests that, with the exception of protocols 8 and 12, all others recover comparable species rankings. Consequently, if

only the ranks are of interest, most of the available protocols would provide highly comparable results. However, for many applications the abundances of the taxonomic units are important and need to be commensurable. Using a Euclidean distance (which cumulates abundance deviations) we found that many protocols were not comparable and actually introduce large batch effects at the species level, with the median between-protocol distance being higher than the within-specimen variation (Figure 3a), hampering the comparability of samples generated with different extraction methods. To assess similarity between extraction protocol effects, we used principle coordinate analysis (PCoA, see Methods) to visualize these distance spaces (Supplementary Figure 1). These indicated that protocol 12, and to a lesser extent also protocols 3, 8, 11, 16 and 18, had abundance profiles that were different from most of to the other protocols.

Analysis of functional microbiome composition, based on COGs (see Methods, Figure 3b), shows that the majority of protocol effects were greater than biological variation within specimen and across time points within the same individual (Figure 3b), with some of them being greater even than between-subject variability. This may in part be due to the known relatively low variation between individuals in this space[28,34] and would dramatically influence conclusions take from comparative studies.

Among the sources of technical variation, the within-protocol variation (i.e. measurement error) was consistently smallest, with the magnitude of the library preparation effect being comparable (Figure 3a,b). The variation introduced by storage method (in RNA Later vs. frozen) was larger than within-protocol variation, and, as previously shown, smaller than within-specimen variation in taxonomic space[23,24].

Taken together, our analysis demonstrates that different DNA extraction protocols induced the largest technical variation, both in taxonomic and in functional space, highlighting that this is a crucial parameter to consider when designing microbiome studies.

QUALITY CONTROL FOR SPECIES-SPECIFIC ABUNDANCE VARIATION

Having quantified and contextualized the different biological and technical sources of variation, we next assessed the quality of different DNA extraction protocols. While a mock community may provide a standard to compare to and thus make a quality assessment more robust, multiple attempts of such approaches have in the past met with problems to recover the expected abundance profiles with either metagenomic or 16S rRNA gene amplicon sequencing[18,27,35]. We thus quantified the variation between protocols independent of a known true composition, first by investigating species-specific effects and secondly by a comprehensive measure of diversity, which we argue provides a good proxy for the overall extraction quality.

We investigated species-specific abundance variation to assess which were most influenced by the extraction protocols. For this, we compared the estimated abundance of a given species in all replicates of a given protocol to the abundances of that species in all replicated of all other protocols, by performing a Kruskal-Wallis test (see Methods). We then applied a false discovery rate (FDR) correction to the obtained p-values. Of the 366 tested species, we found 90 that were significantly affected by extraction protocol (q-value< 0.05). The majority of these were gram-positive, accounting for 37% (+/- 7%) of the sample abundance on average.

These results are in line with previous observations that gram-positive bacteria are more likely to be affected by extraction method[13,35] and are also to be expected based on our extensive knowledge of gram-positive cell walls and their considerably higher mechanical strength. These differences do not reflect the overall performance of any of the protocols, but highlight upper limits of the effect size that may be observed for these species. For a fair comparison, we

contrasted the recovered abundance of some of the significantly affected species, to the mean of the top five highest estimates. This clearly showed that most protocols estimated considerably lower gram-positive bacteria fractions, while the variation in gram-negative abundance estimations is comparatively small (Figure 4).

As the observed biases hint at protocol-dependent incomplete lysis of gram-positive bacteria, we hypothesized that this would correspond to decreased diversity. We thus evaluated whether diversity is a good general indicator of DNA extraction performance. Using the Shannon diversity, which accounts for both richness and evenness, we saw that the recovered relative abundance of gram-positive bacteria correlates with the observed diversity, with a higher fraction of gram-positives resulting in higher diversity (Supplementary Figure 3). Furthermore, we found dramatically reduced diversity in protocols already determined to perform poorly from a DNA quality perspective (i.e. protocols 3, 11 and 12) (Supplementary Figure 2). We conclude that a diversity measure is a good proxy for overall protocol performance and accuracy of the recovered abundance profile.

FACTORS INFLUENCING DNA EXTRACTION OUTCOME

Using diversity as an optimality criterion, we determined protocol parameters that are significantly associated with this indicator (Figure 5). For this purpose we focused on protocols that use the "Qiagen QiIAamp Stool Minikit" [15], namely numbers 5, 6, 8, 9, 11, 13, 15 and 20, which reduces the number of variables that can influence the outcome. We find that "mechanical lysis", "zirconia beads" and "shaking" are positively associated with diversity. We note that there is no association with DNA fragmentation, as all of the samples extracted with these protocols had a low number of fragments below 1.8 kb (Figure2). This was consistent with the notion that mechanical lysis and bead beating are necessary to efficiently extract the DNA of gram-positive bacteria that have cell walls that are harder to break[35] and also in line with our postulation that effective gram-positive recovery will increase the observed diversity. The only significant negative association was with the InhibitEX tablet, which was included in the kit and which the manufacturer recommends for "absorb[ing] substances that can degrade DNA and inhibit downstream enzymatic reactions so that they can easily be removed by a quick centrifugation step"[36], though our assessment suggests an adverse effect on DNA extraction quality.

PROTOCOL REPRODUCIBILITY AND TRANSFERABILITY ACROSS LABORATORIES

Based on the quality of the extracted DNA, species diversity as well as species-specific biases, we selected the five best performing protocols: 15, 7, 6, 9, and 1 (in this order), to be tested for reproducibility across laboratories (phase II). Protocols 15, 6 and 9 use the same Qiagen-based lysis and extraction kit and were combined into a slightly modified protocol, "Q" (Supplementary Information). Protocols 1 and 7 were coded as H and W, respectively.

The laboratories that originally delivered DNA based on the protocol implementations Q, W and H tested once again their variation by replication analysis, ensuring that the variability was comparable to that observed in the first set of extractions (Supplementary Figure 4).

Each extraction method was established and performed in three other laboratories which had no experience with the respective protocol, in order to assess the wider applicability of each as a standard extraction protocol. All three methods were reproducible across locations, though only protocol H had an effect below that of the smallest biological variation (i.e. within-sample). Protocols W and Q introduced a cross-lab effect comparable to within-sample variation.

Although protocol H seemed to be more reproducible across facilities, it underestimated gram-positive bacteria compared to the other two protocols (Figure 5) and so yielded less diverse estimates of microbial composition. Protocol W, while also more reproducible and accurate in terms of extracted relative abundances (Figure 5), is impractical and hard to automate as it involves the use of phenol-chloroform. While no method ranks top on all imposed considerations, protocol Q recovers a highly diverse estimate of the microbial composition which it appears to achieve through lysis of gram-positive bacteria. It is moreover easy to implement and use across facilities. In a tradeoff between practical concerns, reproducibility and accuracy of assessed composition we thus propose protocol Q (see Methods) as a baseline extraction method that future methods should be compared to.

DISCUSSION

We have shown that of all the quantified technical variation considered herein, that introduced by variations in extraction protocol has the largest effects on the observed microbial composition. The outcome of extraction protocols can be influenced by many variables and implementation details, creating a parameter space which is challenging to test exhaustively. Thus, we recognize the limits of our modest recommendations regarding which protocol steps are most crucial to prevent distortions. The 21 laboratories involved in our study have implemented these protocols in order to test reproducibility within and between sites.

In the absence of a universal standard, we stress key criteria comprising DNA integrity, quality and yield, as well as develop a framework of assessing extraction quality independent of a gold-standard. Protocols were validated for transferability (across labs), ensuring reproducible use. Although for particular applications some of the tests are more important than others (e.g. in a multisite consortium reproducibility across labs is more important than in an in depth study in one location) and there is no objective quantification possible, overall protocol Q seems a compromise that should suit most applications.

We anticipate that procedures for DNA extraction will likely further improve in the future. Therefore at this time we recommend that the community adheres to protocol Q as a benchmark for cross-assessment of the diverse protocols used to date. Barring background variation that is difficult to control, any new method that is comparable to these benchmarks within the measurement error assessed in this study should be considered valid. The proposed protocol, together with standard practices for sample collection and library preparation can be found on the IHMS website (http://www.microbiome-standards.org/). Taken together, these recommendations will greatly improve cross-study comparability and with this our ability to make stronger inferences about the properties of the microbiome.

METHODS

LIBRARY PREPARATION AND SEQUENCING

Library preparation started with fragmentation of 250 ng genomic DNA to a 150-700 bp range using the Covaris E210 instrument (Covaris, Inc., USA). The SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter Genomics) were used to perform end repair, A tailing and Illumina compatible adaptors (BiooScientific) ligation. We also performed a 300-600 bp size selection in order to recover most of the fragments.

DNA fragments were then amplified by 12 cycles PCR using Platinum Pfx Taq Polymerase Kit (Life Technologies) and Illumina adapter-specific primers. Libraries were purified with 0.8x

AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and qPCR quantification, the libraries were sequenced using 100 base-length read chemistry in paired-end flow cell on the Illumina HiSeq2000 (Illumina, San Diego, USA).

In the second library preparation protocol, the three enzymatic reactions were performed by a high throughput liquid handler, the Biomek® FX Laboratory Automation Workstation (Beckmann Coulter Genomics) especially conceived for library preparation of 96 samples simultaneously. The size selection was skipped. DNA amplification and sequencing was then performed as in the case of the first approach.

DETERMINING TAXONOMIC AND FUNCTIONAL PROFILES

For determining the taxonomic composition of each sample, shotgun sequencing reads were mapped to a database of selected single copy phylogenetic marker genes[19] and summarized into species-level (mOTU) relative abundances. Functional profiles of clusters of orthologous groups (COGs) were computed using MOCAT[30] by mapping shotgun sequencing reads to an annotated reference gene catalogue as described in Voigt et al.[23]. COG category abundances were calculated by summing the abundance of the respective COGs belonging to each category per sample, excluding NOGs.

COMPARISON TO OTHER TECHNICAL AND BIOLOGICAL VARIATION

To contextualize the size of the effect introduced by different extraction methods, we have assessed different effects caused by either technical or biological factors. These are due to: within protocol variation, library preparation, sample preservation, within specimen variation, between time-points samples from the same individual and between individuals.

For assessing the variation induced by different preservation methods (namely freezing and RNA-later) we use the data from Franzosa et al.[24] and compared the same sample, preserved with the two different methods. For within specimen variation we used data from Voigt et al.[23], where they have sampled the same stool multiple times at different locations along the specimen. As this study also used different storage methods for some samples, we are able to quantify the effect of both within-specimen variation and storage together. For the between time point and individual effect assessment we used the data from the time series data from Voigt et al.[23] as well as a subset of stool samples from the Human Microbiome Project[28].

For assessing library preparation induced variation, we used the same extracted DNA and subjected it to two library preparation methods (Supplementary Information). The first method was the one routinely used for all library preparations presented in the study.

DETERMINING SIGNIFICANTLY DIFFERENT SPECIES

A Kruskal-Wallis test was applied for each species with non-zero abundance in at least two protocols, across both samples. To account for multiple testing we apply a Bonferroni correction to the test p-values and reject the null for any corrected values below 0.05.

PRINCIPAL COORDINATE ANALYSIS

Principal coordinate analysis was performed with the R ade4 package (version 1.6.2), using the *dudi.pco* function.

REFERENCES

1. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
2. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5,** e9085 (2010).
3. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
4. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).
5. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).
6. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G799–807 (2011).
7. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).
8. Dethlefsen, L., McFall-Ngai, M. & Relman, D. a. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* **449,** 811–8 (2007).
9. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 11971–5 (2010).
10. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–7 (2012).
11. Chatelier, E. Le *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).
12. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2,** 19 (2014).
13. McOrist, A. L., Jackson, M. & Bird, A. R. A comparison of five methods for extraction of bacterial DNA from human faecal samples. *J. Microbiol. Methods* **50,** 131–139 (2002).
14. Smith, B., Li, N., Andersen, A. S., Slotved, H. C. & Krogfelt, K. A. Optimising bacterial DNA extraction from faecal samples: comparison of three methods. *Open Microbiol. J.* **5,** 14–7 (2011).
15. Maukonen, J., Simões, C. & Saarela, M. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. *FEMS Microbiol. Ecol.* **79,** 697–708 (2012).
16. Kennedy, N. A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9,** e88982 (2014).
17. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–34 (2010).
18. Ariefdjohan, M. W., Savaiano, D. A. & Nakatsu, C. H. Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens.

Nutr. J. **9,** 23 (2010).

19. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–9 (2013).

20. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9,** 599–608 (2012).

21. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23,** 1704–14 (2013).

22. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.* **6,** 693–9 (2008).

23. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16,** 73 (2015).

24. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2329–38 (2014).

25. Song, S. J. *et al.* Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* **1,** (2016).

26. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* **16,** 276 (2015).

27. Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. & Forney, L. J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7,** e33865 (2012).

28. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–14 (2012).

29. Claassen, S. *et al.* A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples. *J. Microbiol. Methods* **94,** 103–10 (2013).

30. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7,** e47656 (2012).

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

32. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.* **112,** 201423854 (2015).

33. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284–9 (2012).

34. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489,** 220–30 (2012).

35. Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14,** 112 (2014).

36. No Title. Available at: https://www.qiagen.com/fr/shop/lab-basics/buffers-and-reagents/inhibitex-tablets/.

ACKNOWLEDGEMENTS

We would like to acknowledge the help of Sebastian Burz and Kevin Weizer for the editing and web-posting of the SOPs. This study was funded by the European Community's Seventh Framework Programme via International Human Microbiome Standards (HEALTH-F4-2010-261376) grant. We also received support from Scottish Government Rural and Environmental Science and Analytical Services.

AUTHOR CONTRIBUTIONS

PIC, SS, GZ analyzed data, drafter and finalized the manuscript. EP and AA analyzed data, sequenced samples and wrote the manuscript. FL, JRK, MRH and EAV analyzed data and wrote the manuscript. MB, JB, LB, TC, SCP, MD, MD, AD, WMV, BBF, HJF, FG, MH, HH, JHV, JJ, IK, PL, ELC, VM, CM, JCM, CM, HM, CO, POT, JP, SP, NP, MP, AS, DS, KPS, BS, KS, PV, JV, LZ, EGZ extracted samples and wrote the manuscript. SDE, JD and PB designed the study and wrote the manuscript.

FIGURES

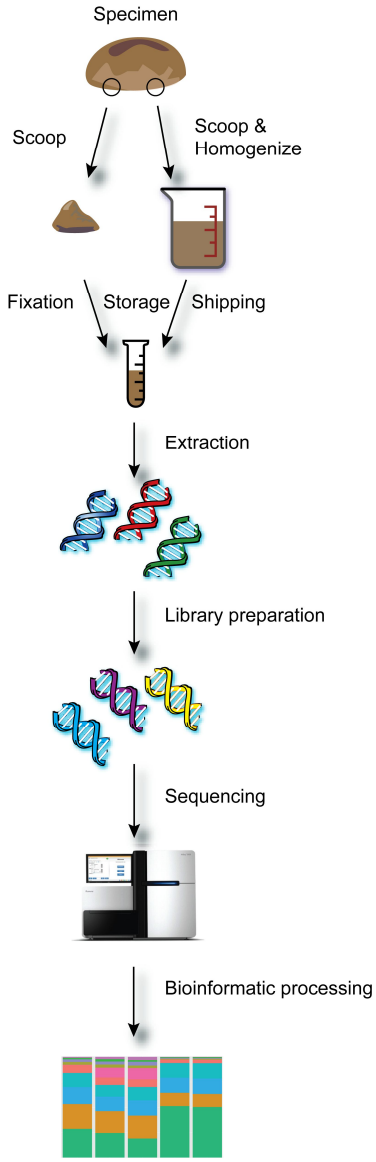


Figure 1: Schematic workflow of human fecal samples processing.

Illustration of the main steps involved in extracting and analyzing DNA sequences from human fecal samples, from collection to bioinformatics analysis. Importantly, none of the outlined steps are standardized, which may introduce strong effects between different studies, making their results hard to compare. For example, differences between freezing and RNA-later fixation have been previously described[23] to bias the measured sample composition.

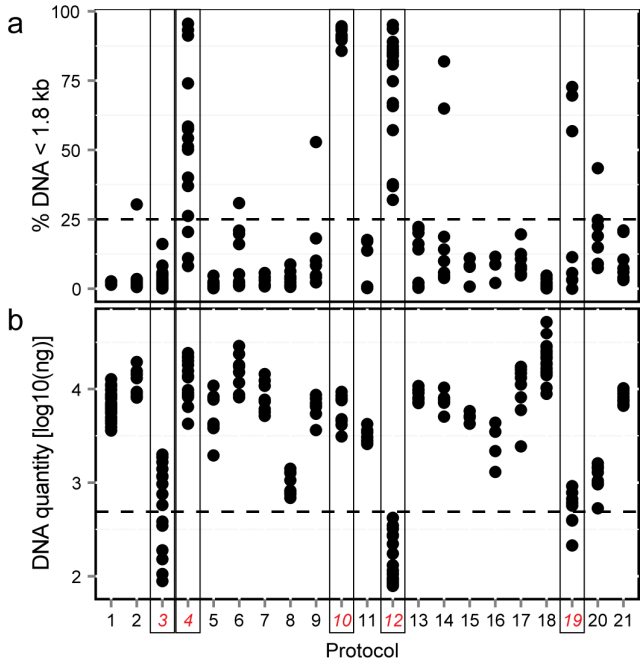


Figure 2: Quality control of extracted DNA

Quality (a) and quantity (b) of extracted DNA from 21 different protocols. a) Percentage of DNA molecules shorter than 1.8 kb, b) quantity of extracted DNA. Protocols failing quality cut-offs (indicated by dashed lines) for either measurement are highlighted in red and boxed.

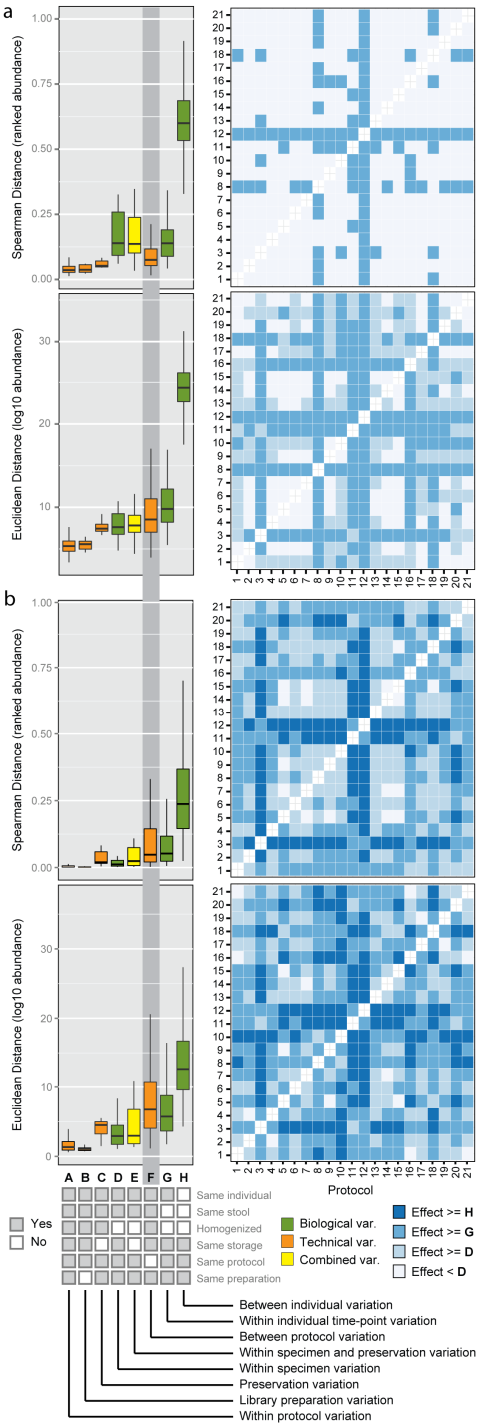


Figure 3: Effect of DNA extraction protocol and library preparation on sample composition

Using both a Euclidean and an Spearman distance measure (see Methods) on species abundances (using mOTU[19]) (a) as well functional abundances (using COGs[33]) (b), shows the relative effect size of different sources of variation. The library preparation and the within-protocol variation are the smallest effects, while the between protocol variation may be greater than some biological effects[23,24]. Heat maps on the right show all pairwise distances between protocols, highlighting which protocol may be considered comparable and which not under different measures of similarity as encoded by letters D,H and G on the bottom-right.

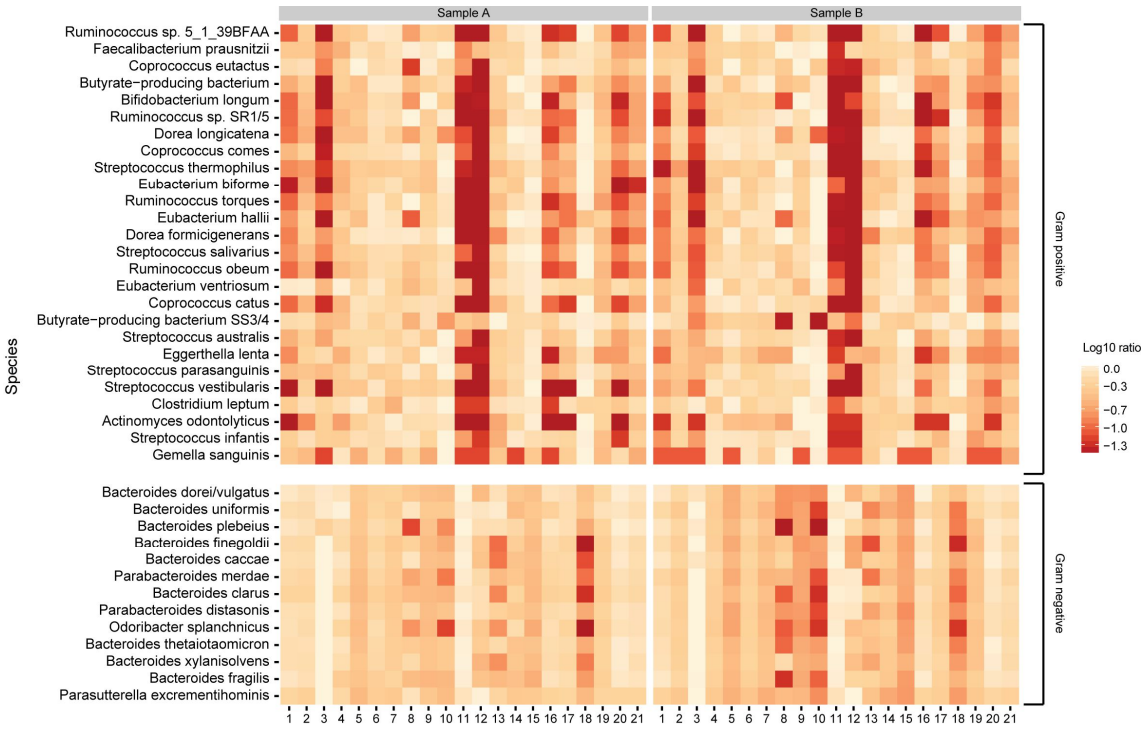


Figure 4: Species specific abundance variation

Assessing variation of species abundances shows that biases are consistent across the two samples. Gram-positive bacteria are heavily under-estimated compared to the mean across the five highest recovered ratios, while gram-negative bacteria are only slightly skewed. Abundances are calculated using mOTUs[19] , with only those having a species level annotation being shown.

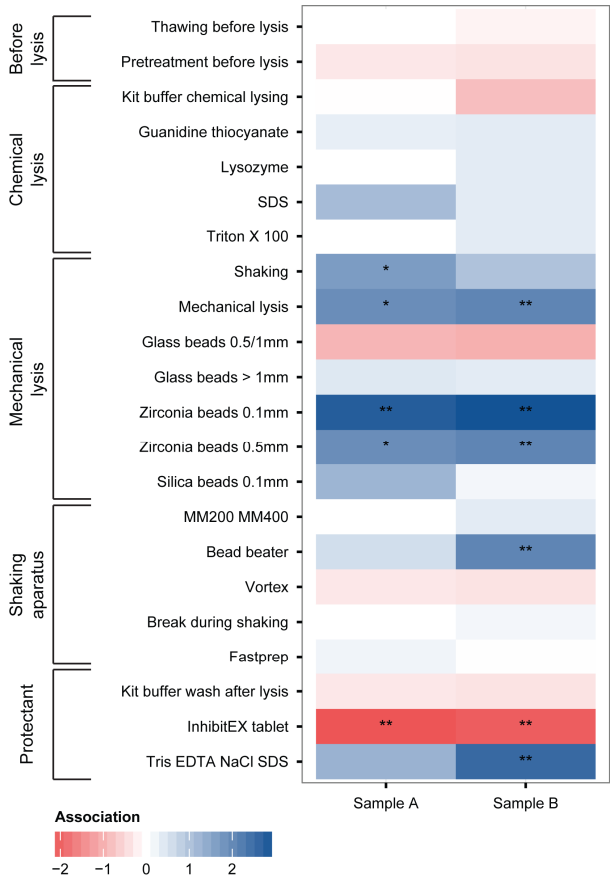
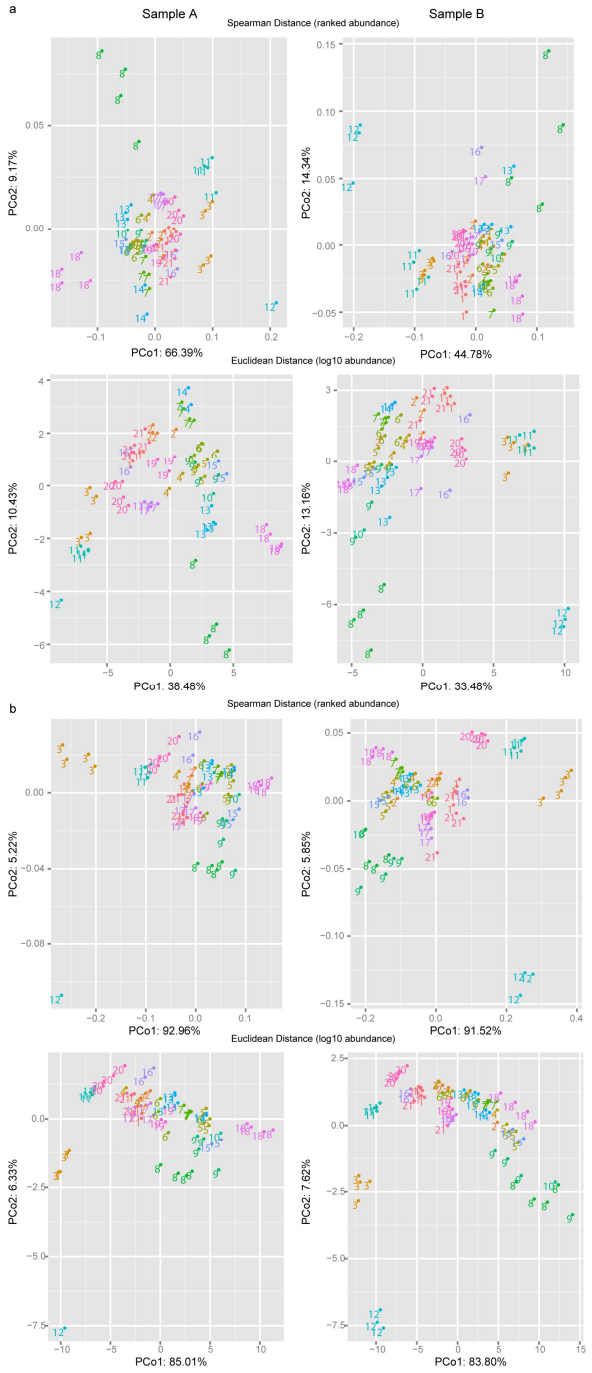


Figure 5: Effects of protocol manipulations on sample composition

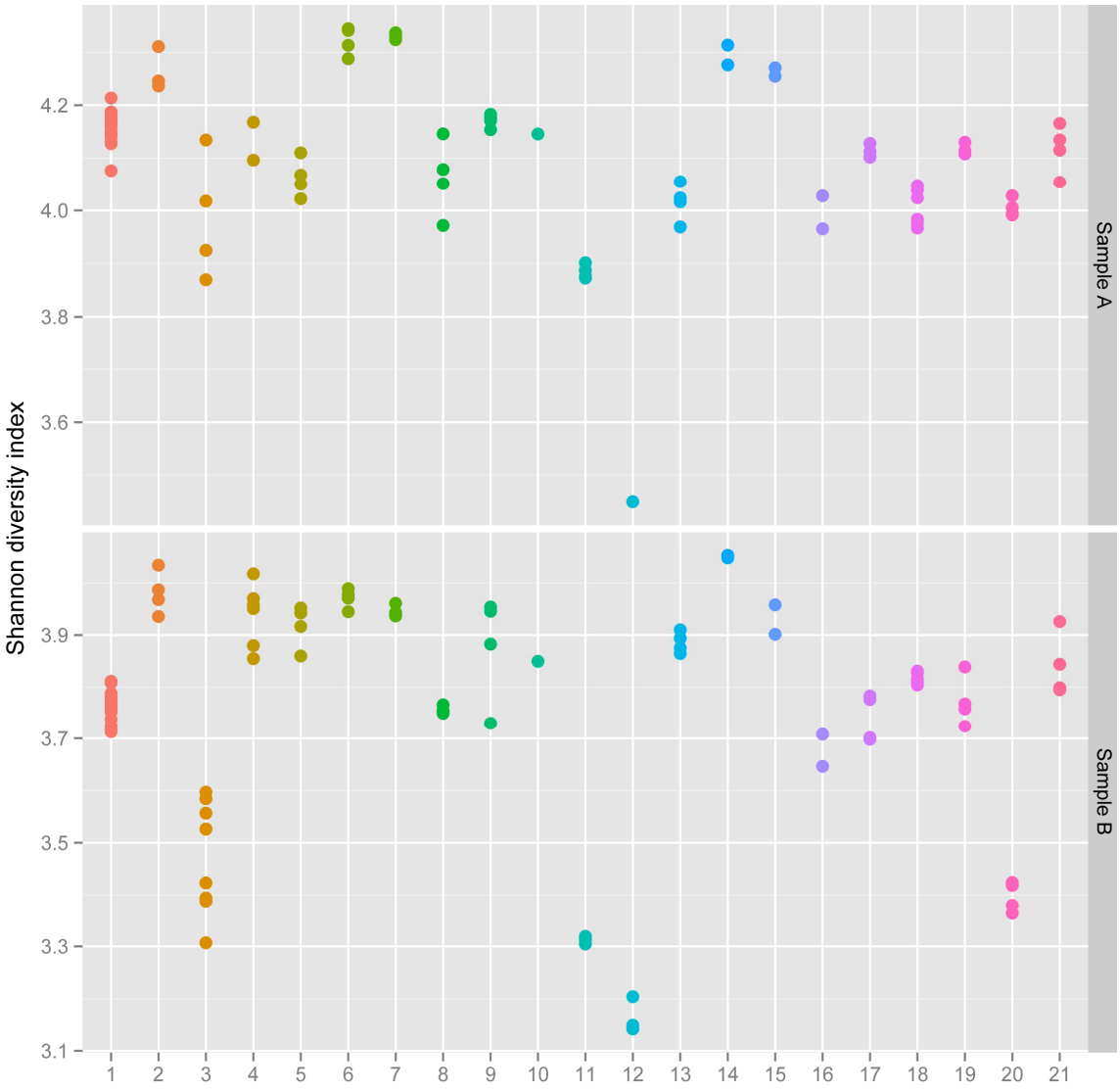
Out of 22 protocol descriptors that vary between the Qiagen based methods, 7 are significantly associated with diversity outcomes. Associations are coded as negative (red) and positive (blue), with significance highlighted by * < 0.05 and ** < 0.01. P-values have been FDR corrected for multiple testing.

SUPPLEMENTARY FIGURES



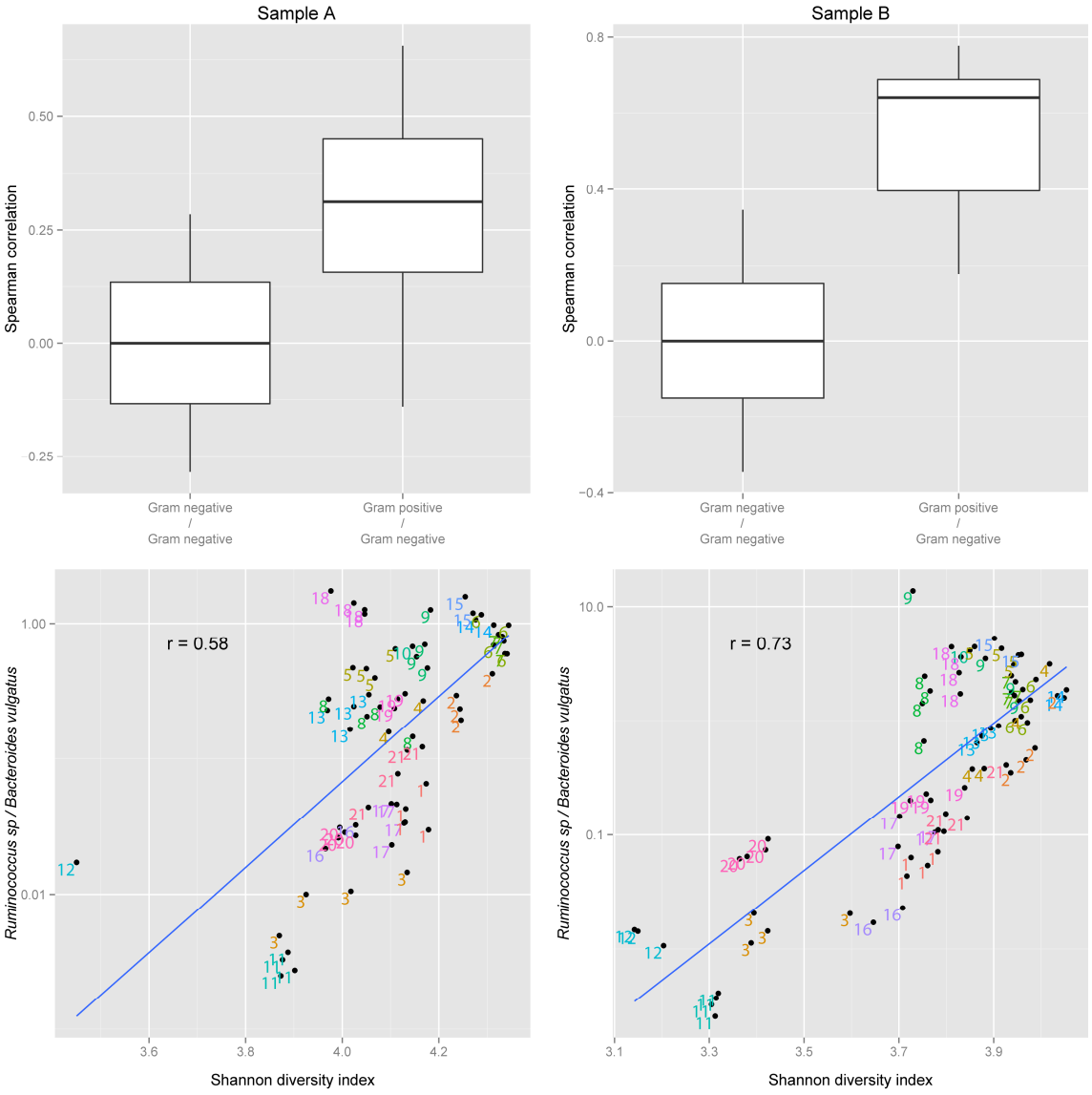
Supplementary Figure 1: Extraction bias across the two samples.

Extraction bias is consistent across the two samples, independent of the distance measure that was used. (a) shows a PCoA projection of the species abundances for each sample, independently, using a Spearman ranked correlation as well as a Euclidean distance. Most of the variation is captured by the first two principal coordinates and the clustering of extraction methods is easily observable. (b) shows a PCoA projection of the functional distance, both Spearman ranked and Euclidean.



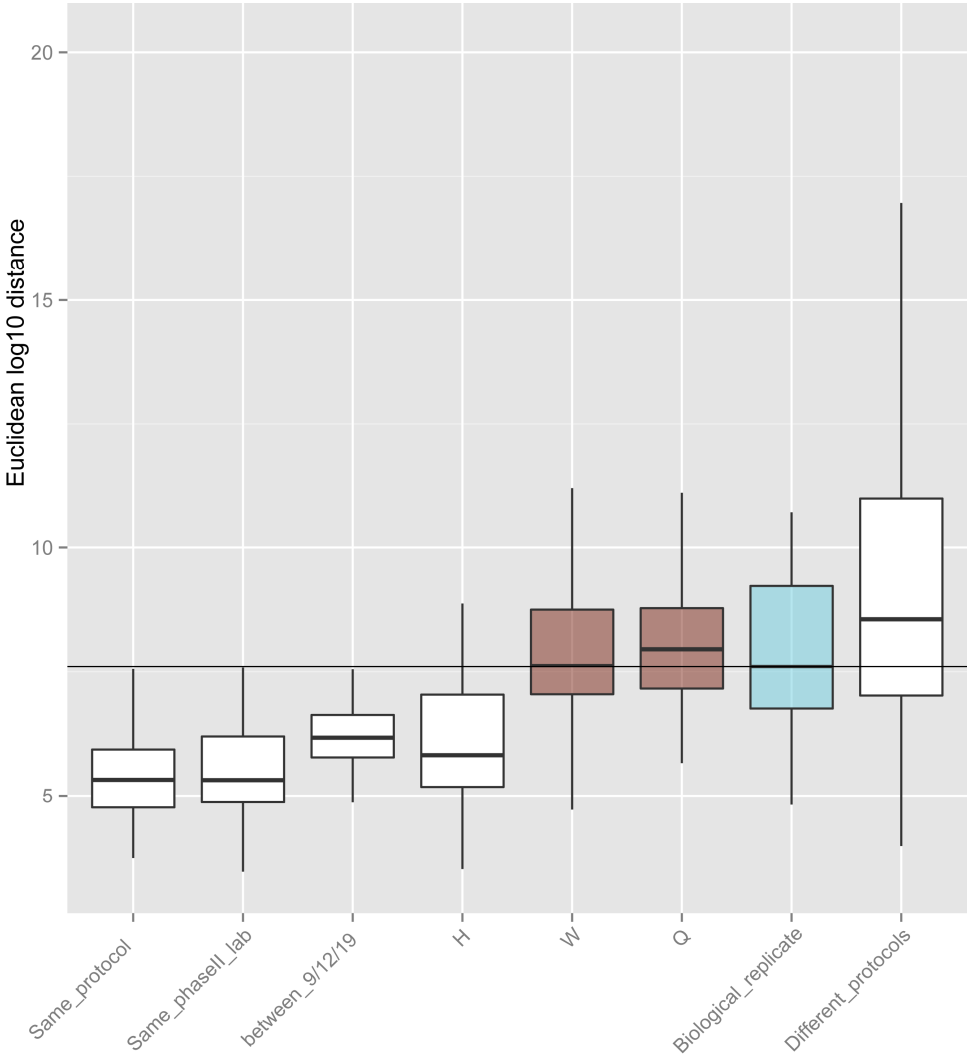
Supplementary Figure 2: Shannon diversity of samples composition

Observed Shannon diversity is consistently influenced by extraction method, as illustrated in both samples. Furthermore, there is a considerable difference in diversity between the two samples, which is not overwritten by extraction bias.



Supplementary Figure 3: Lysis of gram-positive bacteria positively correlates with Shannon diversity.

Recovery of gram-positive bacteria correlates with overall Shannon diversity. Considering only the top 20 most abundant species within each sample, ratios were computed between all gram-positive and gram-negative bacteria as well as gram-negative to gram-negative bacteria. The top panel shows the correlation of these ratios with the Shannon diversity index, while the lower panel exemplifies this correlation on the most abundant gram-positive and gram-negative bacteria that are common to both samples A and B, indicating the strong positive relation between recovery of gram-positive bacteria and observed Shannon diversity.



Supplementary Figure 4: Extraction bias of best performing protocols considered in Phase II

Phase II variation considerations. Extraction variation is the same in Phase II replicates as that of Phase I (bars 1 and 2, respectively). Furthermore, the three protocols that have been merged for Phase II, namely 6, 9 and 15, are similar below the biological replicate variation. The reproducibility of the tree Phase II protocols (H, W and Q) are comparable of below that of biological variation within one specimen.

SUPPLEMENTARY INFORMATION

IHMS DNA EXTRACTION PROTOCOL #1

CELL LYSIS

- Turn on heating block 70°C in advance.
- Take samples out of -80°C and before thawing of stool, add to samples:
    - 250µL Guanidine Thiocyanate *4M*
    - 40µL N-Lauroyl sarcosine 10%,
    - And homogenize with a sterile tooth pick (1 per sample).
- Add 500µL N-Lauroyl sarcosine 5%.
- Vortex vigorously.
- Incubate for 1h at 70°C on heating block.
- Weigh in cupules :
    - 15mg PVPP per sample
    - 200mg for 20 ml of TENP solution per batch of 12 samples.
- After incubation, add 750µL sterile glass beads (0,1mm) to each sample.
- Vortex vigorously.
- Apply bead-beating for 10 minutes in homogenizer set at 25r/s.

REMOVAL OF IMPURITIES

- Add 15mg of PVPP powder to each sample
- Vortex vigorously.
- Centrifuge 5 minutes at 12,700 RPM at 4°C.
- Transfer supernatant to sterile 2mL tube.
- Add to pellet 500µL of TENP solution (Homogenize TENP suspension before pipeting).
- Vortex vigorously until complete dissociation of the pellet. (Use sterile tooth pick if needed)
- Centrifuge 5 minutes at 12,700 RPM at 4°C.
- Collect surnageant and pool with the first one.
- Repeat operation of pellet wash with TENP two more times (3 washes total).
- Centrifuge pooled supernatants for 10 minutes at 12,700 RPM at 4°C.
- Divide supernatant in two equal volumes (≅ 850µl) in two 2mL tubes containing 1 mL isopropanol 2 (2 tubes per sample).

(Attention : take care not to draw the pellet).

PRECIPITATION OF NUCLEIC ACIDS AND PROTEINS

- Gently mix by turning the tubes a few times. (DNA floculation not always visible).
- Let settle for 10 minutes at room temperature.
- Centrifuge 15 to 30 minutes at 12,700 RPM at 4°C.

Attention : pellet not always visible.

- Discard supernatant by inversion of the tube into liquid biohazards bin.
- Let dry.

Attention : for fecal samples in stabilizing suspensions or liquid stool, supernatant must be discarded using a pipet and tips (not by tube inversion).

REMOVAL OF PROTEINS

- Add 450/540µL phosphate buffer (PH8 0,1M) and 50/60µL potassium acetate (5M) in the first tube and flick pellet in suspension.

Attention : adjust volume depending on size of pellet, but use the same volume for all samples of a given project.

- Transfer all liquid of first tube containing resuspended pellet into second tube.
- Resuspend pellets by aspiration using P1000 automatic pipet (adjusted to 350µl).
- Complete resuspension by aspiration using P200 automatic pipet (adjusted to 150µl).
- Leave on ice in the fridge for 1h30 minutes minimum.

PRECIPITATION OF PURIFIED DNA

- Centrifuge 30 minutes at 12,700 RPM at 4°C.
- Turn on heating block 37°C in advance.
- After centrifugation, transfer cupernatant to sterile 1,5mL tube.
- Add 2µL Rnase (10mg/mL) to each DNA preparation (stored at -20°C)
- Incubate for 30 minutes à 37°C in heating block.
- Add 50µL sodium acetate 3M and 1mL absolute ethanol kept at -20°C.
- Mi Gently mix by turning the tubes a few times.
- Leave overnight at -20°C.
- Centrifuge 15 to 30 minutes at 12,700 RPM at 4°C (depending on the amonut of visible DNA flock).
- Discard supernatant in liquid biohazard bin.
- Eliminate droplets by tapping tubes upside down on tissue paper.
- Add 1mL 70% ethanol to wash the pellet.
- Centrifuge 5 minutes at 12,700 RPM at 4°C.
- Discard supernatant in liquid biohazard bin.
- Eliminate droplets by tapping tubes upside down on tissue paper.
- Repeat the wash in 70% ethanol a second time.
- Eliminate droplets of 70% ethanol with P200 automatic pipet.
- Set pellets to dry for 10 minutes in laminar flow hood.
- Resuspend DNA pellet using a pipet in a fixed volume of TE buffer (50 to 300µL depending on pellet size and sample origin).
- Discard all spoiled/exposed material in biohazard bins.

STORAGE

Store long term at -20°C or less.

PREPARATION OF SOLUTIONS

- Phosphate buffer pH 8, 1M
    - 9,32 mL Na2HPO4 : 14,2 g for 100mL H2O (dissolve on heating stirrer)
    - 0,68 mL NaH2PO4 (1M) : 12 g for 100mL H2O
    - 90 mL H2O
    - Check pH = 8 with pH paper
- EDTA, 2 H2O pH 8, 0.5M
    - 9,305g qs for 50 mL H2O (dissolve by heating)
    - Adjust to pH 8 with NaOH pellet (approximately one) using a pH meter
- Tris-HCL (pH 7.5, 1M or pH 8.0, 1M)
    - 6,05 g Trizma base qs for 50mL H2O.
    - Adjust to pH 7.5 or 8.0 with concentrated HCL using a pH meter
- TENP(50 mM Tris pH8, 20 mM EDTA pH8, 100 mM NaCl, 1% of PVPP)

  - 1,5 mL Tris-HCL pH 8, 1M
  - 1,2 mL EDTA pH 8, 0.5M
  - 0,6 mL NaCl, 5M
  - 0,3 g PVPP (attention, will not dissolve)
  - qs for 30ml H2O.
- Guanidine Thiocyanate 4M
  - 12,37g guanidine thiocyanate in a Falcon tube, manipulation under hood (careful with toxicity)
  - 13,5 mL H2O
  - 2,6 mL Tris-HCL 1M pH7.5
  - Shake overnight on a rocking agitator: in closed flacon, protected from light by aluminum foil
  - Completing to 26.1 mL of H2O
  - Heat in Dry bath or in an oven at 60-70 ° C for 10min (if not totally dissolved)
  - Filter through 0.2 microns Millipore filter
  - Store at 4°C protected from light
- NaCl 5M
  - 14.6g qs for 50mL H2O (in a Falcon tube)
- N-Lauroyl Sarcosine 10%
  - 2g for 20mL H2O (in a Falcon tube)
- N-Lauroyl Sarcosine 5%
  - 1g for 20mL phosphate buffer pH8, 0,1M (in a Falcon tube)
- TE pH 8 (10mM Tris.HCl pH8, 1mM EDTA pH8)
  - 200µl Tris HCL 1M pH8
  - 40µl EDTA 0,5M pH8
  - qs for 20mL H2O (in a Falcon tube)
- Potassium acetate (5M for Acetate , 3M for potassium)
  - 29,44g Potassium acetate
  - + 11,5 mL Glacial acetate
  - + 28,5 mL H2O
  - qs for 100 mL with H2O
- *Sodium acetate 3M*
  - 12,304g in 40mL d' H2O (in a Falcon tube)
  - Adjust pH at 5,2 with Glacial acetate
  - qs for 50 mL with H2O
- Rnase (Ribonuclease) 10mg/mL

IHMS DNA EXTRACTION PROTOCOL #2

SYNOPSIS OF THE METHOD

The protocol involves the following steps
- chemical and mechanical lysis of cells
- removal by precipitation and centrifugation of aromatic compounds, cellular debris and proteins
- enzymatic digestion of RNAs
- alcoholic precipitation of purified DNAs

PRODUCTS

Na2HPO4	M : 142 g/mol	*Sigma*		
NaH2PO4	M : 120 g/mol	*Sigma*	*S3139*	
EDTA	M : 372,2 g/mol	*Sigma*	*E5134*	
Trizma base	M : 121,1 g/mol	*Sigma*	*T8524*	
NaOH	M : 40,0 g/mol	*Prolabo*	*28.244.295*	
HCl *concentrated*	M : 36,48 g/mol	*Prolabo*	*20.252.290*	
PVPP(polyvinylpoly-pyrrolidone)		*Sigma*	*P6755*	
Guanidine Thiocyanate	M : 118,2 g/mol	*Sigma*	*G6639*	
NaCl	M : 58.44g/mol	*Prolabo*	*28.244.295*	
N-Lauroyl Sarcosine	M : 293.38g/mol	*Sigma*	*L9150*	
Acétate de Potassium	M :98.14g/mol	*Sigma*	*P3542*	
Acétate de Sodium	M : 82.03g/mol	*Sigma*	*S7545*	
Acide acétique glacial	M : 60.05g/mol :	*Prolabo*	*20.104.243*	
Isopropanol = Propan-2-ol = 2-Propanol	M : 60.1 g/mol	*Merck*	*20842.298*	
Ethanol 100% (Analyse)	M : 46.07 g/mol	*Merck*	*1.00983.1000*	
Rnase (Ribonuclease)	10mg/mL	*Sigma* ou *Amersham*	*R6513* *E78020Y*	

MATERIALS

Filtres 0.22µm	Millipore	
Glass beads 0.1mm	Bioblock	B74471
Sterile screw-cap tubes 2mL, round bottom	ATGC	0214209510

Eppendorf Tubes 2 mL; 1,5 mL autoclaved

Tooth picks autoclaved

Bench-top Micro centrifuge

Refrigerated Centrifuge (4°C)

Heating blocks at 70°C & 37°C

Bead-Beater™ (Biospec Products,USA)

SOLUTIONS

Wear gloves at all times.

Use only MilliQ water (mq) sterile.

- Phosphate buffer pH 8, 0.1M

  - 9,32 mL Na2HPO4 : 14,2 g for 100mL H2O(dissolve using heating stirrer)
  - 0,68 mL NaH2PO4 (1M) : 12 g for 100mL H2O
  - 90 mL H2O pH ≈ 8
- EDTA, 2 H2O pH 8, 0.5M
  - 9,305g and H2O to 50 mL (dissolve using heating stirrer)
  - Adjust pH at 8 avec NaOH in pellets (~one)
- Tris-HCL (pH 7,5 at 1M or pH 8 at 1M)
  - 6,05 g Trizma base and H2O to 50mL.
  - Adjust pH at 7,5 or 8.0 with concentrated HCL
- TENP(50 mM Tris pH8, 20 mM EDTA pH8, 100 mM NaCl, 1% de PVPP)
  - 20 mL H2O (Falcon tube)
  - 1,5 mL de Tris-HCL pH 8, 1M
  - 1,2 mL EDTA pH 8, 0.5M
  - 0,6 mL NaCl, 5M
  - 0,3 g de PVPP (attention, unsoluble : it will remain in suspension)
- Guanidine Thiocyanate 4M
  - 12,37 g guanidine thiocyanate under hood in Falcon tube (attention toxic)
  - 13,5 mL H2O
  - 2,6 mL Tris-HCL 1M pH7.5
  - Let closed vessel protected from light, overnight in rotating device.
  - Adjust with H2O to 26.1 mL
  - Warm up at 60-70°C for 10min
  - Filter with 0,2 µm Millipore filter
  - Store at 4°C protected from light
- NaCl 5M
  - 14.6g and H2O to 50mL (in Falcon tube)
- N-Lauroyl Sarcosine 10%
  - 2g in 20mL H2O (inFalcon tube)
- N-Lauroyl Sarcosine 5%
  - 1g in 20mL phosphate tampon pH8 , 0,1M (in Falcon tube)
- TE pH 8 (10mM Tris.Cl pH8, 1mM EDTA pH8)
  - 200µl Tris HCL 1M pH8
  - 40µl EDTA 0,5M pH8
  - H2O to 20mL (in Falcon tube)
- Potassium Acetate (5M Acetate , 3M de Potassium)
  - 29,44g potassium acetate
  - + 11,5 mL glacial acetic acid
  - + 28,5 mL H2O
  - H2O to 100 mL
- Sodium Acetate 3M
  - 12,304g in 40mL H2O (in Falcon tube)
  - Adjust pH at 5,2 with glacial acetic acid.
  - H2O to 50 mL
- Isopropanol = Propan-2-ol = 2-Propanol
- Ethanol 100% for Analyses stored at –20°C
- Ethanol 70%
  - 70mL ethanol 100% + 30mL sterile H20, stored at –20°C
- Rnase (Ribonuclease) 10mg/mL
  - 20 mg Rnase,
  - + 20 µl Tris.Cl 1M pH 7,5
  - + 6 µl NACL 5M
  - H2O to 2 ml
  - If necessary, bring to 100°C, 15 min, and let cool at room temperature
  - Aliquot in 50µl, store at -20°C

ALIQUOTING BIOLOGICAL SAMPLES

Samples may be aliquotes in sterile Sarstedt screw cap tubes

- take frozen samples (-80°c) and work on a bed of dry-ice
- Use scalpel to cut pieces of sample on aluminium foil, to make 200mg
- Keep tubes on dry-ice or at -80°C

DNA EXTRACTION

- Turn on heating block set at 70°C
- in each tube containing 200mg stool aliquot, add 250µL Guanidine Thiocyanate
- Add 40µL N-Lauroyl sarcosine 10%
- Add 500µL N-Lauroyl sarcosine 5%
- Crush the stool aliquot with tooth pick (1 per sample)
- Vortex to homogeneity

It is possible to stop here and freeze samples overnight at –20°C

- Give a quick spin before incubation (centrifugation at 14000RPM for a few seconds)
- Incubate at 70°C in heating block for 1h (OK up to max 2h)

Note: the chemical treatment contributes to cell lysis and prevents degradation of nucleic acids. The Guanidine Thiocyanate inhibits nucleases and N-Lauroyl Sarcosine is a detergent lysing cells.

- During incubation, prepare Eppendorf tubes 2 ml containing ~750µL glass beads 0.1 mm. (1 per sample)
- weigh 15 mg PVPP on aluminium (1 per sample) + 300mg PVPP for the TENP solution
- at the end of incubation, add the beads to each sample.
- Shake in Bead-Beater™ for 5min (average speed)
- Let Bead-Beater™ still for 5min
- Shake again in Bead-Beater™ for 5min

Note: this ensures mechanical lysis of cells.

- Add 15mg PVPP per sample

Note: PVPP precipitates/adsorbs aromatic molecules

- Vortex : PVPP should not separate or float
- Centrifuge 3min at 14000RPM
- Collect supernatant in 2mL sterile tube with pipett

Note: possible to stop for up to 2 hours setting samples at 4°C

- Add 500µL TENP to the pellet (TENP homogenized before use)
- Vortex to fully resuspend the pellet
- Centrifuge 3min at 14000RPM
- Collect supernatant and pool with first one
- Repeat this TENP washing operation twice (3 washes total)
- Centrifuge the 2mL tube containing the pooled supernatants for 1min at 15000rpm

Note: this eliminates any particle in suspension.

- Dispense supernatant in two equal volumes in 2mL Eppendorf tubes (2 tubes per sample)
- Add 1mL isopropanol (propanol-2) in each tube

Note: isopropanol will precipitate nucleic acids

- mix gently by returning tubes a few times (DNA may be visible as a floculum)
- Leave for 10 min at room temperature (or overnight at 4°C)
- Centrifuge 5min at 14000RPM
- discard supernatant (using pipett)
- Dry tubes upside down on a kimwipe (making sure pellet is stable)
- Add 225µL phosphate buffer
- Add 25µL potassium acetate

Note : the latter precipitates proteins

- Vortex to resuspend the pellets
- Pool the 2 of the same sample in one of the 2 tubes
- Vortex and use a P200 pipett to dissolve
- leave 1h30 minimum on ice

Note: possible to stop for up to 2 hours at 4°C or overnight setting samples on ice

- Centrifuge 30min at 14000RPM at 4°C
- turn on heating blocks at 37°C
- collect supernatant containing DNA in 1.5mL Eppendorf tubes
- Add 2µL Rnase (10mg/mL)/vortex/spin
- Incubate 30 minutes at 37°C

Note : this ensures full digestion of RNA

- Add 50µL sodium acetate
- Add 1mL Ethanol 100% from –20°C
- mix gently by returning tubes a few times
- Leave 5 min at room temperature

Note: possible to stop overnight at 4°C or for longer periods at -20°C

Note: This step leads to DNA precipitation

- Prepare a 1.5mL tube with 500 µL cold ethanol 70% per sample
- If the DNA forms a floc, it is collected with the sterile tip of a Pasteur pipette and transferred to the tube containing 500µl ethanol 70%.
- Otherwise, collect the DN by centrifugation 3mns at 14000RPM remove ethanol and rinse with 500 µl Ethanol 70%, vortex
- Centrifuge 3 mns at 14000RPM and discard ethanol (wash 1)
- Do a second wash with 500 µl Ethanol 70%, vortex
- Centrifuge 3 mns at 14000RPM and discard ethanol (wash 2)
- Eliminate residual ethanol tapping tubes gently on a kimwipe
- Dry for 1h30 in laminar flow hood
- re suspend pellet in 200µL TE (more if necessary)
- leave at room temperature for 1 to 2h
- Vortex and re suspend completely using P200 pipette
- Store frozen at –20°C ; or at 4°C if to be used within a couple of days

IHMS DNA EXTRACTION PROTOCOL # 3

MoBio PowerSoil (HMP modification)

STEP 1:

- Measure out 2.0 grams stool sample in a 15ml Falcon tube
- Add 5 ml of Bead Solution
- Vortex until the stool is homogenized with the stool sample
- Centrifuge for 5 mins @ 1500 g
- Split the supernantant into 5 bead tubes (750µl/tube)
- Incubate sample 10 min at 65°C, then 10 min at 95°C.
- Place @ -80°C overnight

STEP 2:

If Solution C1 is precipitated, heat to 60°C until dissolved before use.

- Add 60µl of Solution C1 and vortex briefly.
- Secure PowerBead Tubes on a flat-bed vortex pad with tape. Vortex at maximum speed for 10 minutes.
- Spin tubes at 10,000 x g for 30 seconds at room temperature.
- Transfer the supernatant to a clean 2 ml Collection Tube. Expect ~400-500µl supernatant.
- Add 250µl of Solution C2 and vortex for 5 sec. Incubate at 4°C for 5 min.
- Spin the tubes at room temperature for 1 minute at 10,000 x g.
- Transfer no more than, 600µl of supernatant to a 2 ml Collection Tube.
- Add 200µl of Solution C3 and vortex briefly. Incubate at 4°C for 5 minutes.
- Spin tubes at 10,000 x g for 2 minutes at room temperature..
- Transfer no more than 750µl of supernatant to a 2 ml Collection Tube
- Add 1200µl of Solution C4 to the supernatant and vortex for 5 seconds.
- Load 675µl onto a Spin Filter and spin at 10,000 x g for 1 minute at room temperature. Discard the flow through. Repeat a total of three times to process all sample.
- Add 500µl of Solution C5 and spin for 30 seconds at 10,000 x g.
- Discard the flow through.
- Centrifuge again at room temperature for 1 minute at 10,000 x g.
- Place Spin Filter in a clean 1.5 ml tube avoiding splashing Solution C5 onto the Spin Filter.
- Add 100µl of sterile DNA-Free PCR Grade Water to the center of the white filter membrane.
- Spin for 30 seconds at 10,000 x g.
- Store frozen (-20° to -80°C).

IHMS DNA EXTRACTION PROTOCOL #4

- Turn on heating block to 75°C
- Label up screw-capped tubes, add 200 mg glass beads to each tube
- Take frozen fecal samples out of freezer to thaw
- Add 300 µL SLX buffer (from Omega Bio-Tek E.Z.N.A.® Stool DNA Kit) to each tube
- Add 10 µL proteinase K solution to each tube (20mg/mL proteinase K in 0.1mM $CaCl_2$)
- Invert tube 6 times to mix, add 200 mg of each fecal sample to prepared screw capped tubes
- Bead beat screw capped tubes 4 x 45 sec
- Incubate tubes at 70°C for 10 min (after 8 minutes has elapsed, turn heating block up to 95°C)
- Incubate for a further 5 min
- Incubate on ice for 2 min
- Add 100 µL Buffer P2 (from Omega Bio-Tek E.Z.N.A.® Stool DNA Kit), vortex for 30 sec
- Incubate on ice for 5 min
- Spin at 14500 x g for 5 min
- Remove supernatant to new 1.5 mL tube (discard pellet)
- Add 200 µL HTR reagent (from Omega Bio-Tek E.Z.N.A.® Stool DNA Kit) to each tube using a wide bore tip, vortex for 10 sec (mix HTR reagent bottle well before pipetting)
- Incubate at room temperature for 2 minutes (prepare Maxwell kit components (from Promega's Maxwell®16 DNA Purification Kit)while waiting)
- Spin at 14500 x g for 2 min
- Add supernatant to Maxwell cartridge (adding 300 µL Elution buffer to each blue collection tube), run through Maxwell 16 Instrument cycle.
- Transfer DNA from elution tube into labelled screw-cap 1.5 mL tube, store in -80°C freezer

IHMS DNA EXTRACTION PROTOCOL #5

Adapted from Zhongtang Yu and Mark Morrison, BioTechniques, 36:808-812.

This procedure is known as Repeated Bead-Beating (RBB) or the "double bead-beater procedure".

MATERIALS

- Gloves
- 1.5 ml eppendorf tubes (B74085-BIOplastics)
- 2.0 ml eppendorf tubes (623 201 Greiner)
- 2.0 ml screw cap tubes (B91211-BIOplastics)
- screw caps (B91303-BIOplastics)
- Glass beads 3mm
- Silicium / Zirkonium beads 0.5 mm (11079101 BioSpec)
- RNase-free Filtertips 10 (B95012-BIOplastics)
- RNase-free Filtertips 200 (4810-Corning)
- RNase-free Filtertips 1000 (B95210-BIOplastics)
- Nuclease free water (Promega-P1193)
- RNase H (Promega- M428A)
- Ethanol, >99% (Merck-)
- Ammonium acetate (Merck-)
- 2-Propanol (Merck 1.01040)
- Ethanol, pure (Merck 1.00983)
- QIAamp DNA stool Minikit (Qiagen 51504)

EQUIPMENT

- Thermoblock (<100oC)
- waterbath (<100oC)
- eppendorf centrifuge
- eppendorf centrifuge with cooling (5417R)
- Nanodrop-ND-1000
- Beat Beater (Precellys 24, Bertan Technologies)

SOLUTIONS

- Lysis buffer
- 500 mM NaCl,50 mM Tris-HCl (pH 8), 50 mM EDTA, 4 % SDS.
- 10 M ammonium acetate
- Measure 192.7gr $C_2H_7NO_2$ and fill to 250ml with water.
- 70% ethanol
- 35ml pure ethanol, add 15 ml water.

CELL LYSIS

1. Add 0,5g of 0,1mm zirconia beads and 4 glass beads (3 mm) to a 2,0ml screw-cap tube, then sterilize.
2. Weigh 0,25 g of faeces into the tube and add 1,0 ml of Lysis buffer.
    a. If buffer is precipitated heat at +70°C
3. Treat sample in FastPrep at room temperature (RT) at 5,5 ms for 3x 1min (cool samples on ice in between).

4. Heat at 95°C for 15 min mix samples shaking by hand every 5 min.
5. Centrifuge at +4°C for 5 min at full speed (to pellet stool particles).
    a. Transfer the supernatant into new 2ml eppendorf tube.
6. Add 300 ul of fresh lysis buffer to the lysis tube and repeat steps 3-4, then pool the supernatants.

PRECIPITATION OF NUCLEIC ACIDS:

7. Add 260 ul of 10 M ammonium acetate to each lysate tube, mix well, and incubate on ice for 5 min
8. Centrifuge at 4°C for 10 min at full speed in a cooled centrifuge. Discard pellet.
9. Transfer the supernatant to two 1,5 ml eppendorf tubes, add one volume of isopropanol and mix well, and incubate on ice for 30 min
10. Centrifuge at RT for 15 min at full speed, remove the supernatant by decanting. Wash nucleic acids pellet with 500 µl 70 % EtOH for 2 min. and dry the pellet to air with cups turned up-side down.
11. Dissolve the nucleic acid pellet in 200 ul (100 ul each) of TE buffer, leave at 4°C overnight and pool the two aliquots.

REMOVAL OF RNA, PROTEIN AND PURIFICATION (QIAMP DNA MINI KIT):

12. Add 2 ul of DNase-free RNase (10 mg/ml) and incubate at 37°C for 15 min.
13. Add 15 ul of proteinase K and 200 ul of Buffer AL mix well and incubate at 70°C for 10 min.
    a. Do not mix proteinase K and Buffer AL in advance!
14. Add 200 ul of EtOH and mix well. Transfer to a QIAmp column and centrifuge for 1 min at 13.000 rcf.
15. Discard the flow through, add 500 ul of Buffer AW1 and centrifuge for 1 min at RT at 13.000 rcf.
16. Discard the flow through, add 500 ul of Buffer AW2 and centrifuge for 1 min at RT at 13.000 rcf.
17. Dry the column by centrifugation at RT for 1 min, leaving the cup open.
18. Add 100 ul of Buffer AE and incubate at room temperature for 1 min. Then centrifuge at 13.000 rcf for 1 min.
19. Re-use the elute with the DNA, incubate for 1 min. Then centrifuge at 13.000 rcf for 1 min.

IHMS DNA EXTRACTION PROTOCOL #6

Protocol of the Repeated Bead Beating Plus Column (RBB+C) Method

(Improved extraction of PCR-quality community DNA from digesta and fecal samples. Zhongtang Yu and Mark Morrison. Short Technical Reports. BioTechniques 36:808-812 (May 2004)

CELL LYSIS:

1. Transfer 0.25 g of sample into a fresh 2-mL screw-cap tube. Add 1 mL of lysis buffer [500 mM NaCl, 50 mM Tris-HCl, pH 8.0, 50 mM EDTA, and 4% sodium dodecyl sulfate (SDS)] and 0.4 g of sterile zirconia beads (0.3 g of 0.1 mm and 0.1 g of 0.5 mm).
2. Homogenize for 3 min at maximum speed on a Mini-Beadbeater™ (BioSpec Products, Bartlesville, OK, USA).
3. Incubate at 70°C for 15 min, with gentle shaking by hand every 5 min.
4. Centrifuge at 4°C for 5 min at 16,000× g. Transfer the supernatant to a fresh 2-mL Eppendorf® tube.
5. Add 300 µL of fresh lysis buffer to the lysis tube and repeat steps 2–4, and then pool the supernatant.

PRECIPITATION OF NUCLEIC ACIDS:

6. Add 260 µL of 10 M ammonium acetate to each lysate tube, mix well, and incubate on ice for 5 min.
7. Centrifuge at 4°C for 10 min at 16,000× g.
8. Transfer the supernatant to two 1.5-mL Eppendorf tubes, add one volume of isopropanol and mix well, and incubate on ice for 30 min.
9. Centrifuge at 4°C for 15 min at 16,000× g, remove the supernatant using aspiration, wash the nucleic acids pellet with 70% ethanol, and dry the pellet under vacuum for 3 min.
10. Dissolve the nucleic acid pellet in 100 µL of TE (Tris-EDTA) buffer and pool the two aliquots.

REMOVAL OF RNA, PROTEIN, AND PURIFICATION:

11. Add 2 µL of DNase-free RNase (10 mg/mL) and incubate at 37°C for 15 min.
12. Add 15 µL of proteinase K and 200 µL of Buffer AL (from the QIAamp DNA Stool Mini Kit), mix well, and incubate at 70°C for 10 min.
13. Add 200 µL of ethanol and mix well. Transfer to a QIAamp column and centrifuge at 16,000× g for 1 min.
14. Discard the flow through, add 500 µL of Buffer AW1 (Qiagen), and centrifuge for 1 min at room temperature.
15. Discard the flow through, add 500 µL of Buffer AW2 (Qiagen), and centrifuge for 1 min at room temperature.
16. Dry the column by centrifugation at room temperature for 1 min.
17. Add 200 µL of Buffer AE (Qiagen) and incubate at room temperature for 2 min.
18. Centrifuge at room temperature for 1 min to elute the DNA.
19. Aliquot the DNA solution into four tubes. Run 2 µL on a 0.8% gel to check the DNA quality.
20. Store the DNA solutions at -20°C.

IHMS DNA EXTRACTION PROTOCOL # 7

Initial samples of 200mg stool => n=8, split into several aliquots of approximately 20mg each (see below).

DNA EXTRACTION PROCEDURE

- Take out the sample tube (feces pellet aliquot) from freezer.
- Add X µL Tris-SDS solution, homogeneize and dispatch in Y tubes (with cutted tips).
- Add 0.3 g glass beads, Z µl tris-SDS and 500 µl TE-saturated phenol

Sample name	Weight (mg)	X µl Tris-SDS	Y tubes	Z µL tris-SDS
A1-033	123	600 µl	6	200 µl
A1-083	185	900 µl	9	200 µl
A1-133	152	700 µl	7	200 µl
A1-183	183	900 µl	9	200 µl
B1-033	182	900 µl	9	200 µl
B1-083	203	1 ml	10	200 µl
B1-133	196	1 ml	10	200 µl
B1-183	175	800 µl	8	200 µl
C1-010	bacterial pool	300 µl	1	0
C1-030	from 0,9 ml	300 µl	1	0

- Shake the tube vigorously using FastPrep at 5.0 power level for 30 seconds to disrupt the bacterial cells.
- Centrifuge the tube (15,000 rpm × 5 min, 4˚C).
- Transfer 400 µl of the supernatant into a new 2ml screw cap micro tube. Add 400 µl phenol/chloroform/isoamyl alcohol (25:24:1).
- Shake the tube vigorously using FastPrep FP120 at 4.0 power level for 45 seconds.
- Centrifuge the tube (15,000 rpm × 5 min, 4˚C).
- Transfer 250 µl of the supernatant into a new 1.5 ml screw cap micro tube.
- Add 25 µl 3 M sodium acetate (pH5.2).
- Add 300 µl isopropanol, and mix the solution by inversion.
- Centrifuge the tube (15,000 rpm × 5 min, 4˚C).
- Discard the supernatant.
- Add 500 µl 70% ethanol.
- Centrifuge the tube (15,000 rpm × 5 min, 4˚C).
- Discard the supernatant.
- Dry the pellet by heating the tube on a heat block incubator at 60˚C.
- Add 100 µl TE. Vortex the solution by pulses to dissolve the pellet.
- Let the tubes one night at 4°C
- Add 1µl of RNase (20 mg/ml), heat 30 minutes at 37°C
- store at -80°C

IHMS DNA EXTRACTION PROTOCOL #8

NECESSARY MATERIAL

- Scale
- Water bath
- Heating blocks
- Homogenizer-Beater Retsch MM200
- Micro-Centrifuge
- Vortex
- Spatula
- Pipetts
- Sterile cones
- Sterile Microtubes 1,5 and 2 mL
- Lysozyme solution:
    - Lysozyme
    - Tris-HCl 1 M, pH 8
    - EDTA 0.5 M, pH 8
    - Triton X-100
    - H2O
- QIAamp DNA Stool Mini Kit (Qiagen)
- Zirconium beads 0,1 mm
- Absolute ethanol

METHOD

- Aliquot 300 mg zirconium beads per sample.
- Pre-heat heating block at 37°C take stool samples of the freezer.
- Prepare lysozyme :
- Weigh 100 mg lysozyme
- Add 100 µl Tris-HCL 1M pH8
    - 20 µl EDTA 0,5M pH8
    - 60 µl Triton X-100
    - 5 ml H_2O
- for 200 mg stool in a 2 mL screw-cap tube, give a short pulse of centrifuge to pellet stool material.
- Add 180 µl lysozyme solution to the 200 mg stool.
- Vortex to homogeneity of the suspension.
- Incubate 30 min at 37°C. ste the ASL buffer at 37°C at the same time.
- Add 1,220 ml of ASL buffer to each tube and vortex 15 seconds.
- Set the heating block at 95°C.
- Add 1 aliquot of zirconium beads (300 mg) to each tube and insert tubes in bead beater MM200 ; run during 3 min at 30 revolutions per second (maximum speed).
- Incubate 10 min at 95°C and vortex 15 sec.
- Centrifuge 1 min at 13000 rpm at room temperature.
- Transfer 1,2 ml of supernatant in an Eppendorf 2 mL tube.
- Add an InhibitEX pill to each tube, vortex untill complete dissolution (approximately 1 minute) and leave 1 min at room temperature.
- Set the heating block at 70°C.
- Centrifuge 6 min at 13200 rpm.
- Make sure that InhibitEX did pellet. If necessary, recentrifuge 6 min at 13200 rpm.
- Transfer supernatant to an Eppendorf 1,5 mL tube.
- Centrifuge 3 min at 13200 rpm.
- Dispense 15 µl of proteinase K in new Eppendorf 1,5 mL tubes.

- Transfer 200 µl of surnageant in tubes containing proteinase K; store remaining supernatant at -20°C.
- Add 200 µl of AL buffer and vortex to complete homogeneity.
- Incubate 10 min at 70°C and centrifuge a few seconds.
- Add 200 µL absolute ethanol and vortex. give a short pulse of centrifuge.

Tubes may be kept at 4°C for a maximum of 2 hours.

- Draw out mini-columns and label them.
- Set the content of the tubes on the columns (Volume is ~700µL).
- Centrifuge 2 min at 13200 rpm. Discard filtrate and reset in collection tube.
- Add 500 µL of AW1 buffer.
- Centrifuge 1 min at 13200 rpm. Discard filtrate and reset in collection tube.
- Add 500 µL of AW2 buffer.
- Centrifuge 1 min at 13200 rpm. Discard filtrate and reset in collection tube.
- Perform a second wash with 500µl of AW2 buffer.
- Centrifuge 3 min à 13200 rpm. Discard filtrate and transfer column to a 1,5 mL tube.
- Add 200 µL of AE buffer.
- Incubate 5 min at room temperature
- Centrifuge 1 min at 13200 rpm. Discard column.
- Store eluate at -20°C.

IHMS DNA EXTRACTION PROTOCOL #9

Fecal DNA extraction with Repated Beat Beating (RBB) method

CELL LYSIS

1. Add 0,25g of Ø 0,1mm zirconia/silica beads (BioSpec, Cat. No. 11079101z) and 3 glass beads (Ø 3 mm) into 2,0ml screw-cap tube (Sarstedt, 72.693).
2. Weigh 0,125g of faeces and add 0,5 ml of Lysis buffer (500 mM NaCl,50 mM Tris-HCl (pH 8), 50 mM EDTA, 4 % SDS) If buffer is precipitated heat at +70°C.
3. Treat sample in FastPrep®-24 Instrument (116004500) with CoolPrep Adapter (6002-528) (MP biomedicals) at 5,5 ms for 3x1 min (after every min wait for 5 min samples on ice to cool down the instrument). Use cool prep adapter with small amount of dry ice.
4. Incubate at 95°C for 15 min with gentle shaking by thermomixer.
5. Centrifuge at +4°C for 5 min at full speed (to pellet stool particles) Transfer the supernatant into new 2 ml eppendorf tube and store supernatant on ice.
6. Add 150 ul of fresh lysis buffer to the lysis tube and repeat steps 3-5, and then pool the supernatants.

PRECIPITATION OF NUCLEIC ACIDS

7. Add 130 ul of 10 M ammonium acetate to each lysate tube, mix well, and incubate on ice for 5 min.
8. Centrifuge at +4°C for 10 min at full speed.
9. Transfer the supernatant to 2 ml eppendorf tube, add 750 ul of isopropanol and mix well, and incubate on ice for 30 min.
10. Centrifuge at +4°C for 15 min at full speed, remove the supernatant using aspiration, wash nucleic acids pellet with 70 % EtOH (0,5 ml) and dry the pellet under vacuum for 3 min.
11. Dissolve the nucleic acid pellet in 200 ul of TE buffer .

REMOVAL OF RNA, PROTEIN AND PURIFICATION: (WITH QIAAMP DNA MINIKIT (QIAGEN, 51306)

12. Add 2 ul of DNase-free RNase (10 mg/ml) (Roche, 10109142001) and incubate at +37°C for 15 min.
13. Add 15 ul of proteinase K and 200 ul of Buffer AL mix well and incubate at +70°C for 10 min.
14. Add 200 ul of EtOH and mix well. Transfer to a QIAmp column and centrifuge for 1 min at full speed.
15. Discard the flow through, add 500 ul of Buffer AW1 and centrifuge for 1 min at RT.
16. Discard the flow through, add 500 ul of Buffer AW2 and centrifuge for 1 min at RT.
17. Dry the column by centrifugation at RT for 1 min.
18. Add 100 ul + 100 ul of Buffer AE and incubate at room temperature for 1 min.
19. Centrifuge at RT for 1 min to elute the DNA.

IHMS DNA EXTRACTION PROTOCOL #10

LYSIS

- Resuspend fecal sample (100-200 mg) in 750 µl Lysis buffer and transfer to 2 ml screw- cap tube containing 300 mg of zirconium beads (0.1 mm, BioSpec Products).
- Incubate at 37˚C for 30 min.
- Add 85 µl 10 % SDS solution and 40 µl Proteinase K (15 mg/ml) and incubate 30 min at 60˚C.
- Add 500 µl of Phenol:Chloroform:Isoamyl alcohol (25:24:1).
- Disrupt cell that did not lyse with enzymatic lysis using a bead beater (BioSpec Products) set on high/homogenize for 2 min.
- Put on ice (2-5 min).
- Spin 13,000 rpm for 5 min, remove top layer and put into 1.5 ml eppendorf tube.
- Extract with Phenol:Chloroform:Isoamyl alcohol (25:24:1), vortex, spin at 13,000 rpm for 5 min and carefully remove upper phase.
- Extract once with Chloroform:Isoamylalcohol, vortex, spin, carefully remove top layer into new eppendorf tube.
- Precipitate DNA with 2.5 Vol ethanol and 1/10 vol 3 M sodium acetate (pH 5.2) and leave at -20 at least 1 hour (can be over night), or -80 for 30 min
- Spin at 13,000 rpm for 5 min to pellet DNA.
- Carefully discard Ethanol without disturbing DNA pallet (which may be brown) and air dry for 30 min.
- Resuspend DNA in 200 µl Tris Buffer (10 mM, pH 8) or H2O.
- *Comment: From here use the DNeasey Blood and Tissue kit and follow their instructions or instructions below.*
- Add 200 µl Buffer AL (without added ethanol). Mix thoroughly by vortexing.
- Ensure that ethanol has not been added to Buffer AL (see “Buffer AL”, page 18).
- Buffer AL can be purchased separately (see page 56 for ordering information).
- It is essential that the sample and Buffer AL are mixed immediately and thoroughly by vortexing or pipetting to yield a homogeneous solution.
- Add 200 µl ethanol (96–100%) to the sample, and mix thoroughly by vortexing. It is important that the sample and the ethanol are mixed thoroughly to yield a
- homogeneous solution.
- Pipet the mixture into the DNeasy Mini spin column placed in a 2 ml collection tube (provided). Centrifuge at _6000 x g (8000 rpm) for 1 min. Discard flow-through and collection tube.*
- Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl Buffer AW1, and centrifuge for 1 min at _6000 x g (8000 rpm). Discard flow-through and collection tube.*
- Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µl Buffer AW2, and centrifuge for 3 min at 20,000 x g (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube.
- It is important to dry the membrane of the DNeasy Mini spin column, since residual ethanol may interfere with subsequent reactions. This centrifugation step ensures that no residual ethanol will be carried over during the following elution. Following the centrifugation step, remove the DNeasy Mini spin column carefully so that the column does not come into contact with the flow-through, since this will result in carryover of ethanol. If carryover of ethanol occurs, empty the collection tube, then reuse it in another centrifugation for 1 min at 20,000 x g (14,000 rpm).
- Place the DNeasy Mini spin column in a clean 1.5 ml or 2 ml microcentrifuge tube (not provided), and pipet 200 µl Buffer AE directly onto the DNeasy membrane. Incubate at room temperature for 1 min, and then centrifuge for 1 min at _6000 x g (8000 rpm) to elute.

LYSIS BUFFER:

- 200 mM NaCl
- 100 mM Tris (pH 8.0)
- 20 mM EDTA
- Prepare and autoclave
- Add Lysozyme to 20 mg/ml before use
- PBS (per liter):
- 8 g NaCl
- 0.2 g KCl
- 1.44 g Na_2HPO_4
- 0.24 g KH_2PO_4

IHMS DNA EXTRACTION PROTOCOL # 11

OVERVIEW OF STEPS AND MATERIAL NEEDED

- lysis of stool samples in Buffer ASL
- adsorption of impurities to InhibitEX matrix
- purification of DNA on QIAamp Mini spin columns (QIAamp DNA Stool Mini Kit)

EQUIPMENT AND REAGENTS NEEDED – NOT IN KIT

- Ethanol (96-100%)
- BSA (for downstream PCR)
- 1.5 and 2 ml microcentrifuge tubes (locking caps/screw caps)
- Pipet tips with filter and barrier
- Microcentrifuge
- Water bath for incubation or heat block at 70˚ C
- Spatulas
- Vortex
- Ice

KIT CONTENTS

Number of preps 50

QIAamp Mini Spin Columns	50
Collection Tubes (2 ml)	200
InhibitEX® Tablets	50
Buffer ASL	140 ml
Buffer AL*	33 ml
Buffer AW1* (concentrate)	19 ml
Buffer AW2† (concentrate)	13 ml
Buffer AE	12 ml
Proteinase K	1.4 ml

- Buffer ASL (store at room temp)
    - mix buffer ASL by shaking
    - if precipitate has formed, incubate at 70˚ C
- Buffer AL (store at room temp)
    - mix buffer AL by shaking
    - if precipitate has formed, incubate at 70˚ C
    - DO NOT add proteinase K directly to Buffer AL
- Buffer AW2 (store at room temp)
    - Add 30 ml ethanol to Buffer AW2 concentrate (indicated on the bottle)
    - Mix thoroughly before use
- InhibitEX Tablets
    - Pop tablet directly out of packet into suitable 2 ml tube without touching

  - 2 ml tubes should be check to make sure mouth is wide enough to accomodate tablet
- QIAamp Mini Spin Columns - PRECAUTIONS
  - Carefully apply the sample or solution to the Mini spin column. Pipet into column without moistening the rim of the column.
  - Change pipet tips between all transfers (aerosol barrier pipet tips only)
  - Do not touch pipet tip to the Mini spin column membrane
  - After all vortexing steps, centrifuge tubes to remove drops from inside lid
  - Close Mini spin column before placing in microcentrifuge.
  - Remove both Mini spin column and collection tube from the centrifuge. Place Mini spin column into a new collection tube. Discard the filtrate and collection tube in Biohazard bin.
  - Open only one QIAamp Mini spin column at a time and avoid generating aerosols.
  - Set up a rack with multiple labeled collection tubes for transferring the Mini spin columns after centrifugation.
- Centrifugation
  - All centrifugation steps are carried out at room temperature.
  - Speed should be 20,000 x g (14,000 rpm). If centrifuge cannot reach 20,000 x g, increase time proportionately (e.g. as a proxy, at 10,000, centrifuge for double the time required at 20,000)

SET UP

- Check that 2 ml tubes used in step 5 are wide enough to accommodate InhibitEX tablet.
- Prepare all buffers according to label instructions.
- Deal with any precipitate in buffer solutions by placing in 70°C water bath until precipitate dissipates.
- Prepare a 70°C water bath for buffers and heat block for other steps (3 and 12).

PROCEDURE

LYSE CELLS IN BUFFER ASL

NOTE: Steps 1-8 involve material that is a biohazard. Use all precautions appropriate for handling biohazards.

1. Weigh 180–220 mg (best at 190-210 mg) stool in a 2 ml microcentrifuge tube, weigh the tube and place the tube on ice. If the sample is frozen, use a scalpel or spatula to scrape bits of stool into a 2 ml microcentrifuge tube on ice. Do not allow sample to thaw.
2. Add 1.4 ml Buffer ASL (700 ul and 700 ul) to each stool sample. Vortex continuously for 1 min or until the stool sample is thoroughly homogenized. Note: It is important to vortex the samples thoroughly and this step usually requires > 1 min. This helps ensure maximum DNA concentration in the final eluate.
3. Heat the suspension for 5 min at 70°C (up to 95°C for hard-to-lyse bacteria). Use heat block rather than water bath. This heating step increases total DNA yield 3- to 5-fold and helps to lyse bacteria and other parasites. The lysis temperature can be increased to 95°C for cells that are difficult to lyse (such as Gram-positive bacteria).
4. Vortex for 15 s and centrifuge sample at full speed for 1 min to pellet stool particles.
5. Pipet 1.2 ml (600 µl and 600 µl) of the supernatant into a new 2 ml microcentrifuge tube and discard the pellet. Note: The 2 ml tubes used should be wide enough to accommodate an InhibitEX Tablet. Transfer of small quantities of pelleted material will not affect the procedure.

ABSORB INHIBITORS WITH INHIBITEX

6. Add 1 InhibitEX Tablet to each sample and vortex immediately and continuously for 1 min or until the tablet is completely suspended. Incubate suspension for 1 min at room temperature to allow inhibitors to adsorb to the InhibitEX matrix.
7. Centrifuge sample at full speed for 3 min to pellet inhibitors bound to InhibitEX matrix.
8. Pipet all the supernatant into a new 1.5 ml microcentrifuge tube and discard the pellet. Centrifuge the sample at full speed for 3 min. Transfer of small quantities of pelleted material from step 7 will not affect the procedure.

DIGEST PROTEINS

9. Pipet 15 µl proteinase K into a new 1.5 ml microcentrifuge tube.
10. Pipet 200 µl supernatant from step 8 into the 1.5 ml microcentrifuge tube containing proteinase K.
11. Add 200 µl Buffer AL and vortex for 15 s. Note: Do not add proteinase K directly to Buffer AL. It is essential that the sample and Buffer AL are thoroughly mixed to form a homogeneous solution.
12. Incubate at 70°C for 10 min. Centrifuge briefly to remove drops from the inside of the tube lid (optional).
13. Add 200 µl of ethanol (96–100%) to the lysate, and mix by vortexing. Centrifuge briefly to remove drops from the inside of the tube lid (optional).

BIND DNA IN SPIN COLUMN

14. Label the lid of a new QIAamp spin column placed in a 2 ml collection tube. Carefully apply the complete lysate from step 13 to the QIAamp spin column without moistening the rim. Close the cap and centrifuge at full speed for 1 min.
15. Place the QIAamp spin column in a new 2 ml collection tube, and discard the tube containing the filtrate. Close each spin column in order to avoid aerosol formation during centrifugation. If the lysate has not completely passed through the column after centrifugation, centrifuge again until the QIAamp spin column is empty.

WASH SPIN COLUMN

16. Carefully open the QIAamp spin column and add 500 µl Buffer AW1. Close the cap and centrifuge at full speed for 1 min. Place the QIAamp spin column in a new 2 ml collection tube, and discard the collection tube containing the filtrate.
17. Carefully open the QIAamp spin column and add 500 µl Buffer AW2. Close the cap and centrifuge at full speed for 3 min. Discard the collection tube containing the filtrate. Note: Residual Buffer AW2 in the eluate may cause problems in downstream applications. Some centrifuge rotors may vibrate upon deceleration, resulting in the flow-through, which contains Buffer AW2, contacting the QIAamp spin column. Removing the QIAamp spin column and collection tube from the rotor may also cause flow-through to come into contact with the QIAamp spin column.
18. Recommended (definitely do): Place the QIAamp spin column in a new 2 ml collection tube and discard the old collection tube with the filtrate. Centrifuge at full speed for 1 min. This step helps to eliminate the chance of possible Buffer AW2 carryover.

ELUTE DNA

19. Transfer the QIAamp spin column into a new, labeled 1.5 ml microcentrifuge tube. Carefully open the QIAamp spin column and pipet 200 µl Buffer AE directly onto the QIAamp membrane. Close the cap and incubate for 1 min at room temperature, then centrifuge at full speed for 1 min to elute DNA.

IHMS DNA EXTRACTION PROTOCOL # 12

- Starting material: 200 mg solid faeces or 200 µL faeces-water suspension contained in a 2 mL Eppendorf tube.
- 400 µL NucleoSENS Lysis Buffer (Biomerieux) is added and mixed by vortexing.
- Next, shaking for 5 min at maximum speed on a TissueLyser (Qiagen).
- Centrifugation for 2 min at 13.000 rpm.
- 100 µL supernatant is mixed with 2 mL NucleoSENS Lysis Buffer and incubated at room temperature for 10 min.
- 2.1 mL sample is transferred to sample vessel on the Easymag (Biomerieux) apparatus.
- Easymag protocol: “Specific A” with 140 µL magnetic silica and 110 µL output volume.

IHMS DNA EXTRACTION PROTOCOL # 13

Protocol for Isolation of DNA from Stool Sample

(Qiagen stool DNA extraction kit)

Homogenize a fresh fecal sample (kept cold for a maximum of 12 hours) by kneading in a strong plastic bag (preferably at 4 °C, keep samples cold throughout). DNA extraction can be performed on previously frozen stool (avoid multiple freeze/thaw cycles).

1. Weigh 200-300 mg (half pea sized) solid stool in a 2 ml microcentrifuge tube and place the tube on ice. Use 300-500 µl of loose stool (more for watery stools)
2. Put 4-5 glass beads along with 1 ml 0.05 M phosphate buffer, vortex until the stool sample is thoroughly homogenized.
3. Spin down at full speed (table centrifuge >10.000 rpm) and save pellet (pellet needs to be clearly visible). Add 1 ml 0.05 M phosphate buffer to wash, vortex and spin down again, save the pellet.
4. Add 1.4 ml buffer ASL to each tube. Vortex continuously for 1 min or until the sample is completely suspended.
5. Heat the suspension for 5 min at 70° C.
6. Add 0.3 g zirconia beads and fill with ASL buffer, beat for 3 min on "homogenize".
7. Centrifuge the sample at full speed for 1 min to pellet the particles.
8. Pipet 1.2 ml of the supernatant into a new 2 ml microcentrifuge tube and discard the pellet.
9. Add 1 InhibitEX tablet to each sample and vortex immediately and continuously for 1 min or until the tablet is completely suspended. Incubate suspension for 1 min at room temperature to allow inhibitors to absorb to the inhibitEX matrix.
10. Centrifuge sample at full speed for 3 min to pellet inhibitors bound to InhibitEX.
11. Pipet the supernatant into a new 1.5 ml microcentrifuge tube and discard the pellet. Centrifuge the sample at full speed for 3 min.
12. Pipet 25 µl Proteinase K into a new 1.5 ml microcentrifuge tube.
13. Pipet 400 µl supernatant from step 8 into the 1.5 ml microcentrifuge tube containing Proteinase K.
14. Add 400 µl buffer AL and vortex for 15 s.
15. Incubate at 70° C for 10 min.
16. Add 400 µl ethanol of 200 proof to the lysate, and mix by vortexing.
17. Label the lid of TWO QIAamp spin columns placed in 2 ml collecting tubes. Carefully apply 610 ul of the complete lysate from step 16 to each of the two QIAamp spin column without moistening the rim. Close the cap and centrifuge at full speed for 1 min. Discard the filtrate. Place the QIAamp spin column in a new 2 ml collection tube, and discard the tube containing the filtrate.
18. Carefully open the QIAamp spin column and add 500 µl buffer AW1. Centrifuge at full speed for 1 min. Place the QIAamp spin column in a new 2 ml collection tube, and discard the collection tube containing the filtrate.
19. Carefully open the QIAamp spin columns and add 500 µl buffer AW2. Centrifuge at full speed for 3 min. Discard the collection tubes containing the filtrates.
20. Preheat the AE buffer at 65° C for 5-10 min.
21. Transfer the QIAamp spin columns into new, labeled 1.5 ml microcentrifuge tubes and pipet 100 µl preheated buffer AE directly onto the QIAamp membranes. Incubate for 5 min at room temperature, then centrifuge at full speed for 1 min to elute DNA.
22. Mix the two DNA samples together (200µl total) then split into on tube containing 175µl and the other containing 25µl. Ethanol precipitate the larger volume and dry the resulting pellet in a biosafety hood. Store the other sample (25µl) at -70°C.

IHMS DNA EXTRACTION PROTOCOL #14

gDNA extraction using FastDNA SPIN Kit for Soil (new instructions 2011)

- Remove samples for gDNA extraction from -80 °C freezer and place straight on ice, add 122 µl MT buffer and allow to thaw. Flick to get air out of beads, then add ~800 µl Sodium Phosphate Buffer into tube, do not add too much buffer or it will leak out of top of tube.
- Homogenize in the FastPrep® Instrument for 30 seconds at a speed setting of 6.5.
- Centrifuge at 14,000 x g for 5 minutes to pellet debris.
- Transfer supernatant to a clean 2.0 ml microcentrifuge tube. Add 250 µl PPS
- (Protein Precipitation Solution) and mix by shaking the tube by hand 10 times.
- Centrifuge at 14,000 x g for 5 minutes to pellet precipitate. Transfer supernatant to
- a clean 15 ml Corning tube.
- Resuspend Binding Matrix suspension and add 1.0 ml to supernatant in 15 ml tube.
- Place on rotator or invert by hand for 2 minutes to allow binding of DNA. Place tube
- in a rack for 3 minutes to allow settling of silica matrix.
- Remove and discard approx 1000 µl of supernatant being careful to avoid settled Binding
- Matrix.
- Resuspend Binding Matrix in the remaining amount of supernatant. Transfer approximately 600 µl of the mixture to a SPIN™ Filter and centrifuge at 14,000 x g for 1 minute. Empty the catch tube and add the remaining mixture to the SPIN™ Filter and centrifuge again at14,000 x g for 1 minute. Empty the catch tube again. Note – may take longer spins to get all liquid through.

- Add 500 µl prepared SEWS-M wash buffer using the force of the liquid from the pipette tip to gently resuspend the pellet.

NOTE: Ensure that ethanol has been added to the Concentrated SEWS-M.

- Centrifuge at 14,000 x g for 1 minute. Empty the catch tube and replace.
- Without any addition of liquid, centrifuge a second time at 14,000 x g for 2 minutes to “dry” the matrix of residual wash solution. Discard the catch tube and replace with a new, clean 1.5 ml eppendorf tube.
- Air dry the SPIN™ Filter for 5 minutes at room temperature.
- Gently resuspend Binding Matrix (above the SPIN filter) in 180 µl of DES (DNase/Pyrogen-Free Water). Incubate for 5 minutes at 55˚C in a heat block.
- Centrifuge at 14,000 x g for 1 minute to bring eluted DNA into the eppendorf tube. Discard the SPIN filter. DNA is now ready for PCR and other downstream applications.
- Store 50 µl aliquot at -20°C for backup and remainder at 4°C for use.

Do not store DNA in tubes supplied in the kit. Lids do not seal well and evaporation will occur.

IHMS DNA EXTRACTION PROTOCOL #15

Extraction of bacterial DNA from fecal samples using the QIAamp DNA stool kit (Qiagen, Hilden, Germany) with a modified protocol for cell lysis.

- Homogenization of 220 mg feces with 1.2 ml ASL lysis buffer of the kit by vortexing for 2 min in a 2 ml tube containing 0.75 g of sterile zirconia/silica beads (0.1 mm in diameter)
- Suspension is incubated for 15 min at 95°C with continuous shaking (1,400 min-1, Thermomixer 5436, Eppendorf)
- The sample is allowed to cool down on ice for 2 min
- Cells are mechanically lysed by running the Fastprep™ Instrument (Thermo Electron Corperation) for 8 min 15 sec
- After cooling down on ice for 2 min coarse particles cell debris and the zirconia/silica beads are spun down by centrifugation (20,000 x g, 4°C, for 1 min)
- Supernatant is transferred to a 2 ml tube
- The pellet is mixed with 350 µl ASL lysis buffer of the kit, vortexed for 1 min and incubated for 5 min at 95°C with continuous shaking as described above.
- After centrifugation at 20,000 x g and 4°C for 1 min the supernatants are combined
- InhibitEX tablet provided by the kit is added to the supernatant (DNA-damaging
- substances and PCR inhibitors adsorb to InhibitEX matrix, sample is vortexed immediately and continuously for 1 min, incubation of the suspension for 1 min at RT
    - The InhibitEX matrix is separated by centrifugation at 20,000 x g for 6 min at RT
    - The supernatant is collected and filled up to 1 ml with sterile phosphate-buffered
- saline (pH 7)
    - DNA was purified with the QIAcube machine (Qiagen) and eluted from the silica-based membrane with 200 µl ultra-pure water.
- QIAcube machine steps
- see Protocol QIAamp DNA Stool Handbook p. 16 from 9.-18.

IHMS DNA EXTRACTION PROTOCOL #16

DNA extraction with PSP® Spin Stool DNA Kit (Invitek)

1. SAMPLE HOMOGENIZATION AND PRELYSIS

- Weigh 50 mg of stool sample (frozen) into a 2.0 ml Safe-Lock-Tube and add 1.2 ml Lysis Buffer P to each stool sample .Vortex vigorously for 1 min.
- Incubate the homogenized sample for 10 min at 95°C in a thermomixer under continuous shaking at 900 rpm.
- Incubate the sample on ice for 3 minutes
- Add 5 Zirconia Beads II to the homogenate.
- Put the sample back to the 95°C thermo block, incubate for further 3 min at 95°C under continuous shaking at 900 rpm.
- Vortex the sample for 2 min.
- Centrifuge the sample at 13.400 x g (12.000 rpm) for 1 min to pellet solid stool particles.

2. REMOVAL OF PCR INHIBITIORS

- Transfer the supernatant into an InviAdsorb-Tube and vortex vigorously for 15 sec. Incubate

3. SECOND SAMPLE CLEANUP

- Transfer the supernatant completely into a new 1.5 ml Receiver Tube. Discard the pellet.
- Centrifuge the sample again at full speed for 3 min.

4. PROTEINASE K DIGESTION

- Transfer 25 µl Proteinase K into a new 1.5 ml Receiver Tube and pipet 400 µl of the supernatant from step 3 to the 1.5 ml Receiver Tube containing Proteinase K,
- mix shortly by vortexing and incubate the sample for 10 min at 70°C in a thermomixer under continuous shaking at 900 rpm.

5. BINDING OF THE DNA

- Add 200 µl of Binding Buffer P to the lysate and mix shortly by vortexing or by pipetting up and down several times.
- Transfer the mixture completely onto the membrane of the RTA Spin Filter. Incubate for 1 min at room temperature and centrifuge at 9.300 x g (10.000 rpm) for 2 min. Discard the filtrate and the RTA Receiver Tube.

6. WASHING STEPS

- Put the RTA Spin Filter in a new RTA Receiver Tube.
- Add 500 µl Wash Buffer I to the membrane of the RTA Spin Filter.
- Close the lid and centrifuge at 9.300 x g (10.000 rpm) for 1 min.
- Discard the filtrate and the RTA Receiver Tube.
- Put the RTA Spin Filter in a new RTA Receiver Tube.
- Add 700 µl Wash Buffer II to the membrane of the RTA Spin Filter.
- Close the lid and centrifuge at 9.300 x g (10.000 rpm) for 1 min.
- Discard the filtrate and put the RTA Spin Filter back to the same RTA Receiver Tube.

7. ETHANOL REMOVAL

- To eliminate any traces of ethanol, centrifuge again for 3 min at maximum speed, discard the RTA Receiver Tube

8. DNA ELUTION

- Place the RTA Spin Filter into a new 1.5 ml Receiver Tube
- Add 100 µl preheated (70°C) PCR-H2O.

- Incubate for 15 min. Centrifuge at 9.300 x g (10.000 rpm) for 1 min to elute the DNA.
- Finally discard the RTA Spin Filter.

IHMS DNA EXTRACTION PROTOCOL # 17

Fecal DNA extraction protocol with adapted G'NOME kit (BIO 101)

EQUIPMENT AND MATERIALS USED

- Centrifuge
- Speed Vacuum
- Screw-cap tubes & eppendorf tubes
- Glass beads 0.1 mm (Fisher Bioblock Scientific B74471)
- Water bath 55°C
- Bead-beater
- toothpicks

REAGENTS

- Cell Lysis/Denaturing solution (from kit)
- RNase Mix (from kit)
- Protease Mix (from kit)
- "Salt-Out" Mixture (from kit)
- PVPP (PolyVinylPolyPyrrolidone) (Sigma)
- TENP (50 mM Tris pH 8, 20 mM EDTA pH 8, 100 mM NaCl, 1% PVPP)
- Molecular water (Eurobio)
- Isopropanol
- Ethanol 100%
- Ethanol 70%
- TE buffer

SOLUTIONS PREPARATION

- EDTA pH 8, 0.5M
    - MM: 372.2 g.mol-1
    - 9.305 g for 50 mL H2O sterile
    - Adjust pH with NaOH
- Tris-HCl pH 8 1M
    - MM: 121.1 g.mol-1
    - 6.05 g for 50 mL H20 sterile
    - Adjust pH with HCl
- NaCl 5M
    - MM: 58,44 g.mol-1
    - 14.61 g for 50 mL H2O sterile
- TENP (50 mM Tris pH 8, 20 mM EDTA pH 8, 100 mM NaCl, 1% PVPP)
    - 20 mL H2O sterile
    - 1.5 mL Tris-HCl pH 8 1M
    - 1.2 mL EDTA pH 8 0.5M
    - 0.6 mL NaCl 5M
    - 0.3 g PVPP

EXTRACTION PROTOCOL

From 200 mg of feces sample (conserved at -80°C). :

- Add 550 µL of Cell suspension Solution (buffer). Use a toothpick to homogenize.
- Add 50 µL of RNase Mix (4°C). Vortex vigorously.
- Add 100 µL of Cell Lysis/Denaturing Solution. Vortex.
- Incubate at 55°C for 30 minutes.
- Add 25 µL of protease mix and vortex.
- Incubate 55°C for 120 minutes.
- Add 750 µL of glass beads 0.1 mm in each sample.
- Put samples for 10 minutes on Bead beater.
- Add 15 mg of PVPP and vortex vigorously.
- Centrifuge 3 minutes at 20 000g
- Retrieve the supernatant in new tubes (Lysate)
- Add 400 µL of TENP to the remaining pellet and vortex (to fully resuspended the pellet)
- Centrifuge 3 minutes at 20 000g.
- Pool the supernatant with the first one for each sample.
- Repeat the washing operation twice.
- Centrifuge 3 minutes at 20 000g all the retrieved supernatant.
- Transfer 750 µL of the supernatant in a new tub (almost half of the whole supernatant).
- Add 1mL of cold Isopropanol (-20 °C) and mix slowly. Let 10 minutes at -20°C. Centrifuge 5 minutes at 20 000g and discard the supernatant. Dry the pellet with the speed vaccum. Take back the DNA pellet with 400 µL of molecular H20.
- Add 100 µL of "Salt Out" mixture. Mix slowly. Let 10 minutes 4°C. Centrifuge 10 minutes at 20 000g. Put the supernatant in a new sterile tub.
- Add 1.5 mL of Ethanol 100% (-20°C) and mix slowly. Let 5 minutes at ambient temperature and centrifuge 5 minutes at 20 000g. Discard the supernatant.
- Take back the pellet with 1 mL of Ethanol 70% (-20°C). Centrifuge 5 minutes at 20 000g.
- Discard the supernatant. Residual ethanol should be removed using Speed Vaccum for 5
- minutes.
- Take back the pellet with 150 µL of TE buffer (10 mM pH 7.5, 1 mM EDTA)
- Transfer in a screw tub.

IHMS DNA EXTRACTION PROTOCOL #18

PREPARATION OF FECAL SAMPLE SUSPENSION

Collected feces were immediately suspended in 20% glycerol (Wako) / phosphate buffer saline (PBS) solution (GIBCO), frozen in liquid nitrogen, and stored at -80 °C until use. In each experiment, 1.0 g of those of stool were used, respectively.

RECOVERY OF BACTERIA FROM FECAL SAMPLES

Frozen fecal sample (1.0 g of human feces) was thawed gradually on ice and suspended vigorously in a 50 mL tube (Falcon). The suspension of feces was filtered with a 100 μm-mesh nylon filter (Falcon) to separate bacterial cells from eukaryotic cells and other debris. The debris on the filter was washed using a glass or plastic bar with 10 ml of PBS buffer twice. The filtrate was centrifuged at 5,000 rpm for 10 min at 4 °C and the supernatant was then discarded. The bacterial pellet was rinsed with 35 ml of PBS buffer twice, and finally rinsed with 35 ml of TE10 (10mM Tris-HCl, 10mM EDTA, pH8.0) buffer. The bacterial pellet was used for microbial DNA isolation. Bacterial cells suspended in glycerol (20%)-PBS buffer were quickly frozen in liquid nitrogen and could be stored in freezer at -80 °C for at least half a year without degradation.

MICROBIAL CELL LYSIS AND DNA ISOLATION BY ENZYMATIC LYSIS METHOD

The bacterial pellet was suspended in 10 mL of TE10 buffer. The suspension was incubated with lysozyme (SIGMA, final conc. 15 mg/mL of cell suspension) at 37 °C for 1 h with gentle mixing. Purified achromopeptidase (Wako, final conc. 2,000 units/mL of cell suspension) was then added and the sample was incubated at 37 °C for 0.5 h. The sample was treated with 10 % (wt/vol) sodium dodecyl sulfate (SDS; final conc. 1 %) and proteinase K (Merck, final conc. 1 mg/mL of cell suspension) and incubated at 55 °C for 1 h. The solution was mixed with equal volume of phenol/chloroform/isoamyl alcohol (Invitrogen) and centrifuged at 5,000 rpm for 10 min. DNA was precipitated by adding 1/10 volume of 3 M sodium acetate (pH 4.5, Wako) and 2 volume of ethanol (Wako). DNA was pelleted by centrifugation at 5,000 rpm at 4 °C for 15 min. DNA pellet was rinsed with 75 % ethanol, dried in vacuum and dissolved in TE buffer.

IHMS DNA EXTRACTION PROTOCOL #19

Magna Pure LC DNA III Isolation Kit (Bacteria, Fungi), Cat. No. 03 264 785 001

All preparations are performed sterile under the laminar flow hood
- prepare Lysozyme (100mg/ml): 100mg ad 1ml 5% Glycerol/PBS

Stool: Take a peanut-size stool-sample and suspend it in 500µl PBS (2 ml tube)
Do NOT centrifuge!
Take 100µl of the stool-suspension into a MagnaLyser tube
Immediately (!) proceed with step 1) into Magnalyser Bead tube

PROCEDURE

- add 130µl Bacteria Lysis buffer (BLB) to 100µl sample into Magnalyser tubes, mix well
- homogenize at 6500rpm/20sec
- add 5,75µl lysozyme (100mg/ml) to 230µl BLB/sample mixture and mix well,
- Incubate at 37°C for 30min
- add 20µl Proteinase K (Roche ProteinaseK Magna Pure LC, dissolved in 1,2 ml Elution
- buffer)
- mix thoroughly, incubate for 10min at 65°C
- Incubate for a further 10min at 95°C, spin down, cool samples on ice (5min)
- centrifuge 2min at full speed (RT)
- transfer 100µl of the lysate supernatant into a MagnaPure sample tube

MAGNA PURE PREPARATIONS

- Use buffers free from precipitates, use buffers at room temperature.
- dissolve Prot.K (1,2ml Elution buffer); mix completely; RT!
- Magnetic Glass Particles: vortex immediately before use!, add the MGPs to the container just before starting the run
- mark 1,5ml-Eppis for Eluates, fill Magnalyser
- select the protocol: 'DNA Bacteria III' (sample volume: 100µl, elution volume: 100µl) and follow the instructions of the software
- transfer samples from Sample cartridge to marked 1.5mL tubes and store at -20°C

IHMS DNA EXTRACTION PROTOCOL # 20

Protocol modified from the QIAamp DNA stool handbook (Stool pathogen detection pp15-18)(Cat#51504)

Before starting: Make sure 70°C water bath is pre-heated.

1. Record the weight of the Dry Bead tubes (MoBio Cat# 12811-100-DBT).
2. Place a pea size sample (for human stool) into each bead tube.

NB. Protocol will work for smaller amounts of 50mg and up to 250mg.

3. Record the weight of the tube again to identify the amount of stool sample used.
4. Add 1.4 mL Buffer ASL to each stool sample. Mechanically lyse cells using Fastprep/homogenizer for 1 minute (repeated twice) at a frequency of 5.5.

NB. Make sure that ASL buffer has not precipitated. If it has, put in 70°C water bath to dissolve.

5. Heat the suspension for 5 minutes in a 70°C water bath.
6. Vortex for 15 sec and centrifuge sample at 15000 rpm for 1 min to pellet stool particles.
7. Add 1 InhibitEX Tablet into new 2.0 mL centrifuge tubes.
8. Pipet 1.2 mL of the supernatant from Step 6 into the 2.0 mL centrifuge tubes containing InhibitEX Tablet and vortex immediately until the tablet is completely suspended. Incubate suspension for at least 1 min at room temperature.
9. Centrifuge sample at 15000 rpm for 3 min to pellet inhibitors bound to InhibitEX matrix.
10. Pipet all the supernatant into a new 1.5 mL centrifuge tube and discard the pellet. Centrifuge the sample at 15000 rpm for 3 min.
11. Pipet 15 µL proteinase K into a new 1.5 mL centrifube tube.
12. Pipet 200 µL supernatant from Step 10 into the 1.5 mL centrifuge tube containing proteinase K.
13. Add 200 µL Buffer AL and vortex for 15 sec.
14. Heat the tubes for 10 minutes in a 70°C water bath.
15. Add 200 µL of 100% ethanol to the lysate and mix by vortexing.
16. Centrifuge sample for 30 sec to remove drops from inside of the tube lid.
17. Pipet the complete lysate from Step 13 to the QIAamp spin column without moistening the rim. Centrifuge at 15000 rpm for 1 min. Place the QIAamp spin column into a new 2 mL collection tube and discard the tube containing the filtrate.
18. Add 500 µL Buffer AW1 to the spin column. Centrifuge at 15000 rpm for 1 min. Place the QIAamp spin column into a new 2 mL collection tube and discard the tube containing the filtrate.
19. Add 500 µL Buffer AW2. Centrifuge at 15000 rpm for 3 min. Place the QIAamp spin column into a new 2 mL collection tube and discard the tube containing the filtrate. Centrifuge again at 15000 rpm for 1 min to eliminate possible Buffer AW2 carryover. Transfer the QIAamp spin column into new 1.5 mL centrifuge tube.
20. Add 30 µL Buffer AE. Incubate for at least 5 min at room temperature. Centrifuge at 15000 rpm for 1 min to elute DNA.

IHMS DNA EXTRACTION PROTOCOL #21

Standard Operating Procedure for isolation of genomic DNA from feces, in preparation for molecular analysis

PSP® Spin Stool DNA Kit (Invitek)

EQUIPMENT AND REAGENTS

- microcentrifuge
- thermomixer (for 70°C AND 95°C)
- bead beater (MagNA Lyser, Roche)
- measuring cylinder (250 ml)
- disposable gloves
- RNase-free Filtertips 10
- RNase-free Filtertips 200
- RNase-free Filtertips 1000
- reagents reservoirs for multichannel pipets
- Optional: RNase A (10 mg/ml)
- 96 - 100% ethanol
- Ultra pure water (Milli-Q)
- vortexer or other homogenizer
- Glass beads 3mm
- Silicium / Zirkonium beads 0.5 mm (BioSpec)
- 2.0 ml screw cap tubes
- screw caps
- PSP Spin Stool DNA Kit (Invitek)

PREPARING TUBES FOR BEAD BEATING

Add 0,5g of 0,1mm zirconia beads and 4 glass beads (3 mm) to a 2,0ml screw-cap tube, then sterilize.

PREPARING REAGENTS AND BUFFERS FOR THE PSP® SPIN STOOL DNA KIT

1. adjust the thermomixer to 70°C.
2. dissolve Proteinase K in ddH2O
3. warm up the needed amount of Elution Buffer D to 70°C, (200 µl Elution Buffer D are needed per sample).
4. heat heating blocks (e.g. thermomixer) to 70°C and 95 °C
5. label the needed amount of 2.0 ml RTA Spin Filter Sets
6. label the needed amount of 1.5 ml Receiver Tubes (per sample: 1 Receiver Tube), add the needed amount of ethanol to the Wash Buffer I and II

3 or 10 total DNA extractions:
add 250 µl ddH2O to Proteinase K, mix thoroughly and store the tube at -20°C Wash Buffer I and II are ready to use
50 total DNA-extractions:
add 1.5 ml ddH2O to Proteinase K, mix thoroughly and store the tube at -20°C add 30 ml 96-100% ethanol to the bottle Wash Buffer I add 42 ml 96-100% ethanol to each bottle Wash Buffer II mix thoroughly and always keep the bottle firmly closed
250 total DNA-extractions:

add 1.5 ml ddH2O to Proteinase K, mix thoroughly and store the tube at -20°C
add 80 ml 96-100% ethanol to each bottle Wash Buffer I
add 105 ml 96-100% ethanol to each bottle Wash Buffer II
mix thoroughly and always keep the bottle firmly closed

Important Notes:

The centrifugation steps were made with the Centrifuge 5415 D from Eppendorf.

The indicated settings refer to this centrifuge.

Preheat the Elution Buffer D to 70°C (e.g. transfer the needed volume into a tube and place it at the appropriate temperature into a thermomixer)

1. SAMPLE HOMOGENIZATION AND PRELYSIS

- Add 0,5g of 0,1mm zirconia beads and 4 glass beads (3 mm) to a 2,0ml screw-cap tube, then sterilize.
- Weigh 200 mg of stool sample (fresh or frozen) into the 2.0 ml screw-cap tube and add 1.2 ml Lysis Buffer P.

Important: If the sample is liquid, pipet 200 µl into the 2.0 ml screw-cap tube. Cut-off the end of the pipet tips to make pipetting easier (sterilize the tips after cutting!). If the sample is frozen, use a scalpel or spatula to scrape bits of stool into the provided 2.0 ml screw-cap tube on ice. Take care, that this samples do not thaw until Lysis Buffer P is added, otherwise the DNA in the sample may degrade. After addition of the buffer, the following steps can be performed at RT or like recommended.

- Treat sample in Magna Lyser at room temperature (RT) at 5,5 ms for 3x 1min (cool samples on ice in between).
- Incubate the sample for 10 min at 95°C in a thermomixer under continuously shaking at 900 rpm.
- Centrifuge the sample at 13.400 x g (12.000 rpm) for 1 min to pellet solid stool particles and beads.

2. REMOVAL OF PCR INHIBITORS

- Transfer the supernatant into an InviAdsorb-Tube and vortex vigorously for 15 sec.
- Incubate the suspension for 1 min at room temperature.
- Centrifuge the sample at full speed for 3 min.

3. SECOND SAMPLE CLEANUP

- Transfer the supernatant completely into a new 1.5 ml Receiver Tube.
- Discard the pellet.
- Centrifuge the sample again at full speed for 3 min.

4. PROTEINASE K DIGESTION

- Transfer 25 µl Proteinase K into a new 1.5 ml Receiver Tube and pipet 400 µl of the supernatant from step 3 to the 1.5 ml Receiver Tube containing Proteinase K.
- Mix shortly by vortexing and incubate the sample for 10 min at 70°C in a

thermomixer under continuous shaking at 900 rpm.

OPTIONAL: REMOVING TRACES OF RNA

Invisorb® RTA Spin filter can also purify low amounts of RNA besides DNA. For the elimination of RNA (if necessary) add 20 µl RNase A (10 mg/ml) before adding the Binding Buffer P. Vortex briefly and incubate the sample at room temperature for 5 minutes. Then go on as described in the protocol.

5. BINDING OF THE DNA

- Add 200 µl Binding Buffer P to the lysate and mix shortly by vortexing or by pipetting up and down several times.
- Transfer the mixture completely onto the membrane of the RTA Spin Filter. Incubate for 1 min at room temperature and centrifuge at 9.300 x g (10.000 rpm) for 2 min.
- Discard the filtrate and the RTA Receiver Tube.

6. WASHING STEPS

- Put the RTA Spin Filter in a new RTA Receiver Tube.
- Add 500 µl Wash Buffer I to the membrane of the RTA Spin Filter.
- Close the lid and centrifuge at 9.300 x g (10.000 rpm) for 1 min.
- Discard the filtrate and the RTA Receiver Tube.
- Put the RTA Spin Filter in a new RTA Receiver Tube.
- Add 700 µl Wash Buffer II to the membrane of the RTA Spin Filter.
- Close the lid and centrifuge at 9.300 x g (10.000 rpm) for 1 min.
- Discard the filtrate and put the RTA Spin Filter back to the same RTA Receiver Tube.

7. ETHANOL REMOVAL

- To eliminate any traces of ethanol, centrifuge again for 3 min at maximum speed, discard the RTA Receiver Tube.

8. DNA ELUTION

- Place the RTA Spin Filter into a new 1.5 ml Receiver Tube and add 100µl preheated (70°C) Elution Buffer D to the sample.
- Incubate for 3 min at RT.
- Centrifuge at 9.300 x g (10.000 rpm) for 1 min. to elute the DNA.
- Add another 100µl preheated (70°C) Elution Buffer D to the filter membrane
- Incubate for 3 min at RT.
- Centrifuge at 9.300 x g (10.000 rpm) for 1 min. to elute the DNA.
- Finally discard the RTA Spin Filter.

IHMS DNA EXTRACTION PROTOCOL Q

Fecal DNA extraction with the use of Qiagen QIAamp DNA stool kit

1. Homogenize the 150 to 200mg frozen feces with 1.0mL ASL lysis buffer of the kit by vortexing for 2min in a 2mL tube containing 0.3g of sterile zirconia beads Ø 0,1mm zirconia (BioSpec, Cat. No. 11079101z). [if buffer shows precipitate, heat at 70°C before use]
2. Incubate for 15min at 95°C.
3. Cells are mechanically lysed by running the Fastprep™ Instrument for 8min15sec (series of beating 1 min and resting 5 min are preferable).
4. Samples are allowed to cool down on ice for 2min.
5. Samples are centrifuged at 16000 x g, 4°C, for 5min.
6. Supernatant is transferred to a new 2mL tube.
7. The pellet is mixed with 300µL ASL lysis buffer of the kit, and steps 2-5 are repeated.
8. Supernatants are pooled in the new 2mL tube.

9. Add 260µl of 10M ammonium acetate to each lysate tube, mix well, and incubate on ice for 5 min.
10. Centrifuge at 16000 g, 4°C, for 10min.
11. Transfer the supernatant to two 1.5mL Eppendorf tubes, add one volume of isopropanol, mix well, and incubate on ice for 30 min.
12. Centrifuge at 16000 g, 4°C, 15min, remove the supernatant using aspiration, wash nucleic acids pellet with 70 % EtOH (0,5mL) and dry the pellet under vacuum for 3min.
13. Dissolve the nucleic acid pellet in 100µL of TE (Tris-EDTA) buffer and pool the two aliquots.
14. Add 2µL of DNase-free RNase (10mg/mL) and incubate at 37°C, 15 min.
15. Add 15µL proteinase K and 200µL AL buffer to the supernatant, vortex for 15sec and incubate at 70°C for 10 min.
16. Add 200µL of ethanol (96-100%) to the lysate, and mix by vortexing.
17. Transfer to a QIAamp spin column and centrifuge at 16000 g for 1min, at Room Temperature (RT).
18. Discard flow through, add 500µL buffer AW1 (Qiagen) and centrifuge at 16000 g for 1min, at RT.
19. Discard flow through, add 500µL buffer AW2 (Qiagen) and centrifuge at 16000 g for 1min, at RT
20. Dry the column by centrifugation at RT for 1min.
21. Add 200µL Buffer AE (Qiagen), incubate for 1min at RT
22. Centrifuge for 1min at 16000 g to elute DNA.

Quality control: use 1% agarose gel
Sample concentration: use Nanodrop or Qubit

PAPER 3

A fair comparison

To the Editor: Recently, Paulson *et al.*[1] introduced a normalization method, reporting that it improves clustering of meta-genomic abundance data, which is very important for many applications in the fast-growing area of microbiome research. However, in our view, the perceived improvement is due to a postprocessing procedure that is preferentially combined with some, but not all, normalizations included in their method comparison, rather than to the proposed normalization itself.

Paulson *et al.*[1] compared their normalization method to three existing ones using a data set from a study of microbial communities in the mouse gut and concluded that their method, called cumulative-sum scaling (CSS), "substantially improved" the separation between two known clusters present in the data[1]. As the authors kindly provided us with the source code, we were able to reproduce their first figure (**Supplementary Fig. 1**). However, this was possible only when we applied a logarithm transformation to the data normalized with their CSS method but not to the data normalized by the other methods. Combining the log transformation with each of the normalizations shows that differences in cluster separation are due mainly to this additional transformation and not to the normalization itself (**Fig. 1**). Thus, conceptually simpler methods, such as relative-abundance normalization (also called total-sum scaling (TSS)), should not be dismissed on these grounds.

To understand the large effect of the log transformation on this comparison, it is important to note that it is nonlinear, a feature that can fundamentally change the distribution of the data (skewing reduction, for example). Because the transformation is undefined for input values ≤0, one typically adds a small value (pseudocount) to non-negative input data to avoid log(0). However, owing to the nonlinearity of the log, this value also affects the transformation result (**Supplementary Fig. 2**). Paulson *et al.*[1] set the pseudocount to 1 as a way to preserve zero counts. However, as the four normalizations compared produce output values whose ranges differ by several orders of magnitude, the same pseudocount may not be optimal for all of them. It should instead be chosen to ensure a consistent treatment: for instance, by setting it to a value smaller than the minimum abundance value before transformation (**Supplementary Fig. 2** and **Supplementary Note**).

Methodological improvements are crucial in highly complex fields such as metagenomics. We feel, however, that in a comparison of different approaches, it is important to minimize the potential confounding sources by ensuring equal treatment of all methods under study.

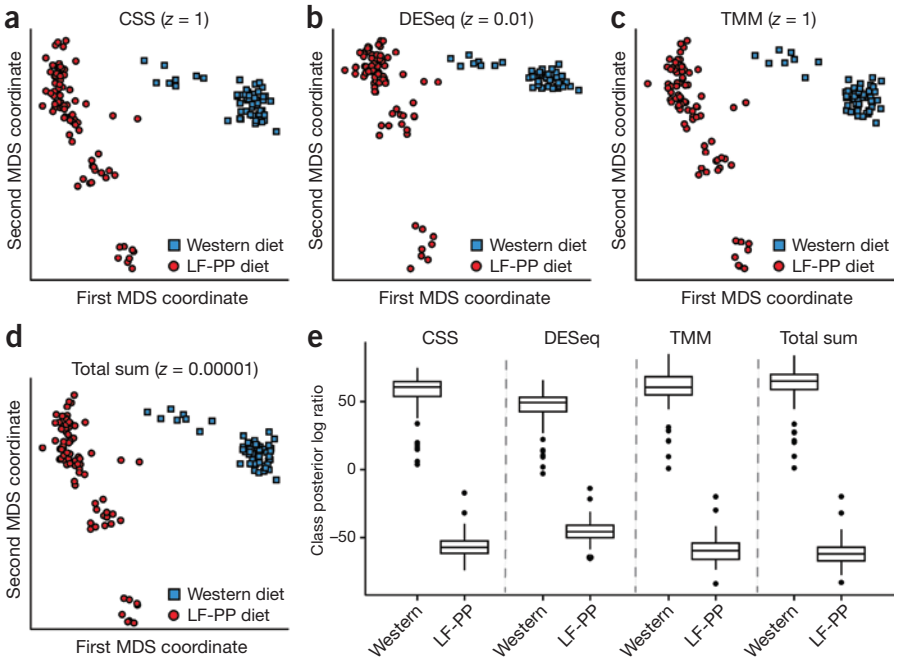


Figure 1 | Clustering analysis of different normalization methods. (**a**–**d**) First two principal coordinates of multidimensional-scaling (MDS) analysis of mouse stool data normalized by CSS (**a**), DESeq size factors (**b**), trimmed mean of *M*-values (TMM) (**c**) and total-sum scaling (**d**). The pseudocount (*z*) used with the log transformation is indicated in parentheses (**Supplementary Note**). Colors indicate clinical phenotype (diet). LF-PP, low-fat, plant polysaccharide–rich diet. All normalizations separate samples by diet. (**e**) Class posterior probability log ratio for Western diet obtained from linear discriminant analysis. Each box corresponds to the distribution of leave-one-out posterior probability of assignment to the 'Western' cluster across normalization methods. Samples were optimally distinguished by phenotypic similarity regardless of the method of normalization used. This figure corresponds to Figure 1 in Paulson *et al.*[1] (see also **Supplementary Fig. 1**).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2897).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Paul I Costea, Georg Zeller, Shinichi Sunagawa & Peer Bork

European Molecular Biology Laboratory, Heidelberg, Germany.
e-mail: bork@embl.de

1. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. *Nat. Methods* **10**, 1200–1202 (2013).

Paulson *et al.* reply: Costea *et al.*[1] challenge the fairness of the results presented in the first figure of our paper[2], which explored the effect of normalization and transformation procedures on clustering analysis of marker-gene survey

PAPER 4

ENTEROTYPES IN THE LANDSCAPE OF GUT MICROBIAL COMMUNITY COMPOSITION

Authors: Paul I. Costea[1,†], Falk Hildebrand[1,2,3,†], Manimozhiyan Arumugam[4], Fredrik Bäckhed[5,6], Martin J. Blaser[7], Frederic D. Bushman[8], Willem M. de Vos[9,10], S. Dusko Ehrlich[11,12], Claire M. Fraser[13], Masahira Hattori[14], Curtis Huttenhower[15], Ian B. Jeffery[16], Dan Knights[17,18], James D. Lewis[19], Ruth E. Ley[20], Howard Ochman[21], Paul W. O'Toole[16], Christopher Quince[22], David A. Relman[23,24,25], Fergus Shanahan[16], Shinichi Sunagawa[1], Jun Wang[5,26,27,28,29], George M. Weinstock[30], Gary D. Wu[31], Georg Zeller[1], Liping Zhao[32], Jeroen Raes[2,3,33,*], Rob Knight[34,35,36,37,*], Peer Bork[1,38,39,*]

† These authors contributed equally; * Corresponding authors

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
[2] VIB Center for the Biology of Disease, VIB, Belgium
[3] Laboratory of Microbiology, Vrije Universiteit Brussel (VUB), Brussels, Belgium
[4] The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.
[5] Wallenberg laboratory, Department of molecular and clinical medicine, Institute of medicine, University of Gothenburg, Sweden
[6] Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Receptology and Enteroendocrinology, Faculty of Health Sciences, University of Copenhagen, Copenhagen, DK-2200, Denmark
[7] New York University Langone Medical Center,550 First Avenue, Bellevue CD 689,New York, NY 10016
[8] Department of Microbiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104-6076, USA
[9] RPU Immunobiology, Department of Bacteriology & Immunology, University of Helsinki, Helsinki, Finland.
[10] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands.
[11] Metagenopolis, Institut National de la Recherche Agronomique, Jouy en Josas, France
[12] King's College London, Centre for Host-Microbiome Interactions, Dental Institute Central Office, Guy's Hospital, UK
[13] Institute for Genome Sciences at the University of Maryland School of Medicine, Baltimore, MD 21201, USA.
[14] Graduate School of Advanced Science and Engineering, Waseda University. 3-4-1 Okubo Shinjuku-ku, Tokyo 169-8555, Japan
[15] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA.
[16] Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland
[17] Biotechnology Institute, University of Minnesota, Saint Paul, MN 55108, USA
[18] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[19] Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
[20] Department of Microbiology, Cornell University, Ithaca, USA.
[21] Department of Integrative Biology, University of Texas, 2506 Speedway A5000, NMS 4.110, Austin TX 78712, USA
[22] Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK
[23] Department of Microbiology and Immunology, Stanford University, Stanford, California 94305, USA.
[24] Department of Medicine, Stanford University, Stanford, California 94305, USA.
[25] Veterans Affairs Palo Alto Health Care System 154T, 3801 Miranda Avenue, Palo Alto, California 94304, USA.
[26] Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark.
[27] Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia
[28] Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China
[29] Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong
[30] The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, USA
[31] Division of Gastroenterology, Perelman School of Medicine, University of Pennsylvania, 421 Curie Blvd Philadelphia, PA 19104, USA.
[32] Ministry of Education Key Laboratory for Systems Biomedicine, Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, PR China
[33] Department of Microbiology and Immunology, Rega Institute KU Leuven, Leuven, Belgium
[34] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA.
[35] Biofrontiers Institute, University of Colorado, Boulder, CO 80309, USA.
[36] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA.
[37] Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309, USA.
[38] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany
[39] Molecular Medicine Partnership Unit, 69120 Heidelberg, Germany

INTRODUCTION

The human body is colonized by trillions of microbes that contribute to human health and well-being. Different communities of microbes inhabit various anatomical regions (Figure 1). Inter-individual variation at each body site is considerable, but the separation among sites within individuals remains apparent [1]. The most densely populated habitat is the gut, with an estimated 1.5 kg of microbial biomass [2]. The gut harbors hundreds of bacterial and archaeal species, with Firmicutes and Bacteroidetes as dominant phyla [1,3–5]. Considerable variation in species composition has been described among individuals, for example in the US NIH Human Microbiome Project (HMP) [1] and the European Metagenomics of the Human Intestinal Tract project (MetaHIT) [3,6]. The gut microbial ecosystem shows a succession of different microbiota stages: community composition changes rapidly in early childhood, stabilizes in adults and subsequently degrades with age [7,8]. There is no satisfying description of this complex landscape across large populations and geographies, in part because some taxa seem to vary in abundance among individuals monotonously while others appear to have bimodal or more complex distributions [9]. Given the importance and complexity of the gut ecosystem, there is great interest in identifying compositional patterns and their underlying rules to understand human health and disease states. Such classification would potentiate microbiota-based diagnostics, therapies or prevention of disease, with the potential for personalized treatment through nutritional, microbial, and pharmaceutical interventions.

Reproducible patterns of variation in the microbiota have been observed in the adult gut. Separated into community clusters, they have been termed “enterotypes” [10]. Ecosystem types have also been described in the vagina [11] and other body sites [12–14]. In the gut, multiple studies have built on the original approach [10], while others have proposed different methodology or questioned the stratification methods. Interpretations of the results include variation in cluster numbers and proposals for gradients. Alternative interpretations are often discussed in the literature, indicating the difficulty of finding a clear-cut solution (e.g. [7,8,12,15,16]).

Here, we review gut microbial community compositions in light of vastly increased amounts of data. We use refined methodologies and perform meta-analyses to propose an adapted concept of enterotypes, with the goal of reconciling divergent viewpoints, and illustrate the practical value of reproducible stratification. This approach does not diminish the need to pursue other analyses and avenues for interpretation, as stratification only captures some of the dimensions of microbiota complexity.

ENTEROTYPES: EVOLUTION OF A CONCEPT FOR STRATIFYING GUT COMMUNITY COMPOSITION

In 2011, clustering human fecal metagenomic samples based on their taxonomic composition resulted in the proposal of three “enterotypes” [10]. These “densely populated areas in a multidimensional space of community composition” showed similar taxonomic properties in datasets from 6 nationalities (US-American, Danish, French, Italian, Japanese and Spanish) using three sequencing technologies (Illumina, 454, Sanger), as well as 16S rRNA profiling data, and were independent of age, gender, cultural background, and geography.

Each enterotype was characterized by an ecological network centered around one indicator (driver) taxon: enterotype 1, here denoted ET B for clarity, is a *Bacteroides*-dominated cluster; ET P (enterotype 2) is dominated by *Prevotella*, a genus whose abundance is inversely correlated with *Bacteroides*; and ET F (enterotype 3), which is distinguished by an overrepresentation of Firmicutes, most prominently *Ruminococcus* (Arumugam et al. 2011 and supplement therein). Analyses were performed at genus level, where microbial ecological niches are hypothesized to be reflected [17]. One caveat of this approach is that species- and strain-level variations are neglected, although they contribute to functional differences between individuals and can be important in a clinical context [18,19].

In Japanese individuals, a study focused on drivers identified via principal component analysis, reported the same genera as those representing the enterotypes [20]. A large scale, diet-focused study in a U.S. cohort reported a link between long term-diet and *Bacteroides*, and proposed that a gradient of *Bacteroides* and *Prevotella* may also explain the observed pattern [15]. One of the clusters shared similar dominant taxa with ET P, while the other is similar to a merge of ETs F and B. Following this, a study of individuals from Venezuelan and Malawian rural areas, and US metropolitan areas also emphasized the importance of *Prevotella* and *Bacteroides* as driving taxa, as well as a strikingly different composition in infants, with communities dominated by Bifidobacteria and Proteobacteria [8]. This highlights the need for caution when extrapolating overall community patterns from a potentially biased sampling of the world population at different ages. With similar methods, subsequent studies on the HMP 16S rRNA data [14], and a meta-analysis of four large studies [21], also recovered three enterotypes.

Using Dirichlet multinomial mixture models (DMMs), four optimal clusters were identified in one of the datasets analyzed in the original enterotype paper [16]. Two of the clusters resembled ET B and P, while a third showed an increased prevalence of *Ruminococcus* and other *Firmicutes* genera. The last cluster was driven by a high fraction of unidentified taxa, but was close to the *Firmicutes* cluster. DMMs have also been used to identify three optimal clusters in a healthy Swedish cohort, again showing compositions similar to ET B and ET P, with one additional cluster dominated by unknown taxa [22]. A further study that applied the same method to the HMP 16S rRNA data, found that the gut microbiome could be separated into four optimal clusters [13], similar to those previously identified [22]. When applying DMMs to the MetaHIT metagenomics dataset, we identify four clusters, consistent in their relationships to the three previously proposed enterotypes (Figure 2, Suppl. Fig. 1).

An ensemble-based network approach confirmed that the three dominant gut taxa that contribute to enterotype clustering (*Prevotella*, *Bacteroides* and *Ruminococcaceae*) in the HMP 16S rRNA data are hubs of three co-occurrence network clusters and showed that their abundances are mutually negatively correlated (Ref. 1 and Suppl. Figure 4 therein). This negative correlation could also be shown with qPCR data of 35 signature taxa [23]. Three distinct networks were found in adult Amish individuals [24]; the dominant genera contributing to these networks overlap partly with the driver taxa of the original enterotypes. Similarly, six species co-abundance groups (CAGs) were reported in a dataset consisting of Irish adults and elderly individuals, with the healthy ones mostly associated with three networks that contained similar taxa to those separating the enterotypes [7,25]. Thus, independent of clustering approaches, bacterial co-abundance networks provide an underlying species network that may explain the existence of enterotypes.

Enterotype-like structures have also been reported with varying levels of confidence in some studies on mammalian gut microbiota, such as non-human primates [26,27], mice [28,29] and pigs [30]. Although they harbor compositionally distinct gut microbial communities, some of the same taxonomic groups of bacteria appear to drive clustering in both human and non-human mammals. Should compositionally similar enterotypes occur across a large range of different host species, it would suggest that these community patterns coevolved with mammalian diversification. Enterotypes might thus be a widespread attribute of mammalian gut communities present in different forms in a multitude of datasets (Figure 2 and Suppl. Table 1).

CHALLENGES IN DEFINING MICROBIAL COMMUNITY TYPES

Many recent studies have discussed whether the compositional landscape is best represented by 2, 3, 4 or even 6 clusters or ecological "states" (Figure 2, Suppl. Table 1). Assessing clustering in fecal microbiota profiles is non-trivial, given that demonstration of alternate states is debated in disciplines from ecology to philosophy [12,31]. Changing taxonomic levels, distance metrics, clustering algorithms or cluster optimality scores can yield different numbers of clusters (Suppl. Fig. 2 and Supplementary Material), even on the same dataset (e.g. [12]). However, it remains possible that confounding factors could obfuscate the underlying structure (Suppl. Fig. 3). Although numerous studies have reported on clustered enterotypes (Figure 2), some have questioned even the validity of enterotypes as states and proposed a gradient-based interpretation of microbial community compositions instead, arguing that clustering with weak separation among clusters may be misleading [8,12,15,24,32]. Proponents of "enterogradients" [33] hold that a gradient between *Bacteroides*, *Prevotella,* and *Firmicutes* abundance sufficiently explains the data.

Regardless of clustering outcome or modelling assumptions, an analysis of the MetaHIT data [6], backed by reports from the literature (Suppl. Table 1), reveals that the local substructure is always similar (Figure 2C), i.e. a three cluster model finds *Bacteroides*, *Prevotella* and *Firmicutes*-dominated clusters, and a two cluster model identifies *Prevotella*- and *Bacteroides*-dominated subsets of samples. Partitioning of the gut microbiota is thus considered stable in the sense that related cluster compositions are recovered if the number of clusters is pre-specified, reconciling many studies in which different numbers were reported. However, a similar outcome would also be observed if the underlying data represented a gradient, bisected by an artificial division.

There is agreement that there are distinct areas within the microbial composition landscape in which the respective communities show biological differences. The concept of enterotypes can help capture such differences, although defining meaningful and robust boundaries is challenging. This is analogous to clustering of macro-biomes, which faces similar problems. One example includes the differences between Treeless, Savannah, and Forest ecosystems in sub-Saharan Africa; these states could equally be represented as a gradient in response to mean precipitation [34] or as contrasting stable states [35]. Are data from a single time point sufficient (e.g. examining the distribution of these three zones across Africa), or should explicit community modeling and transitions between states guide the approach? Although this question has not yet been conclusively resolved, it seems intuitive to consider the ecosystem clustering (Treeless, Savannah, Forest) when describing the local fauna, even though the boundaries between states may not be sharp.

Many of the conclusions from previous studies, especially those relating to the strength of stratification, are based on the original clustering approach [10]. We assessed the power of this

method by clustering 16S rRNA data from the HMP project [1], trying to separate samples according to body-site. Although the originally proposed Jensen-Shannon distance (JSD) at genus level and the weighted UniFrac on the OTU level find the expected four clusters corresponding to skin, stool, vaginal, and oral microbiomes, they do so with little "statistical support" (Suppl. Fig. 2). With simplifying assumptions, a modeling approach, such as the DMMs, may prove better at resolving such separation (Suppl. Fig. 1). Regardless of methodological and clustering considerations most studies of the human gut report *Prevotella*-enriched samples as forming a separate cluster (Figure 2B, Suppl. Table 1). This is in line with recent reports [9] of bimodal *Prevotella* distributions among individuals, also visible in the datasets analyzed here (see Figure 3A, Suppl. Fig. 4, Suppl. Table 1 and references therein). While this implies at least two community types, the discussion about gradients or the number of types remains relevant and centers around the difficulty of reliably separating ET B and ET F, based on the current, mostly cross-sectional datasets (Suppl. Table 1). If ET P samples are discarded, the abundance of *Bacteroides* is distributed in a gradient across typical gut samples. However, bimodal distributions of *Parabacteroides* and less abundant genera such as *Methanobrevibacter* are also present (Figure 3A, Suppl. Fig. 4); although the latter is not detected by most bacterial 16S rRNA primers and can thus escapes analysis in many datasets. ET B and ET F may represent two extremes of a *Bacteroides*/*Firmicutes* gradient within a merged ET B+F cluster. However, given that some distance-based clustering approaches do not capture an intuitively strong clustering structure exemplified by different body sites (Suppl. Fig. 2). Clustering based on explicit underlying data models such as co-occurrence networks and DMMs may be able to identify a more refined substructure of 3-6 groups [13,16,25] that subdivides ET F + B (Figure 2, Suppl. Fig. 1).

Given the practical challenges in accurately determining the gut community structure, such as overcoming batch effects, taking into account confounders (Suppl. Fig. 3) and accounting for temporal variation, an objective number of stable states are difficult to determine. Still, in the (mostly Western) subjects studied cross-sectionally, *Bacteroides* and *Prevotella* act as driving taxa that explain inter-individual differences, and delineate the main sources of variation regardless of the technique used. The extremes of the gradient within the ET B+F cluster are substantially different in composition and diversity. These are discussed in the following sections in terms of function, ecology, disease and diet. While the three enterotypes may not always be the best explanation of the data, it is the model that has been most used and provides a framework that we use below.

FUNCTIONAL PROPERTIES OF ENTEROTYPES

The functional potential of enterotypes appears to differ among the extremes of the enterotype landscape (Figure 3B). In the following, we describe patterns of functional composition and their potential implications.

The gut microbiota can harvest energy for the host by catabolizing and digesting polysaccharides using carbohydrate-active enzymes (CAZymes). When analyzing driver taxa of enterotypes Purushe et al. [36] found that the number of genes encoding CAZymes was higher in genomes of taxa within the phylum Bacteroidetes (which includes both *Bacteroides* and *Prevotella*) compared to those of Firmicutes [37]; this suggests that Bacteroidetes may have the potential to metabolize more diverse polysaccharides, facilitating host access to complex glycans (Suppl. Table 2). Communities enriched in Firmicutes have differences in energy harvesting behaviours with implications for the host.

For the MetaHIT samples [6], most KEGG Orthologs (KOs) and Clusters of Orthologous Groups (COGs) showed significantly different abundances (FDR<0.1) between enterotypes in any of the two-, three- or four-enterotype stratifications. On average 14,291 (SD=2,988) of 22,029 tested COGs and 4,316 (SD=694) of 5,623 analyzed KOs were significantly different (FDR<0.1, methods as ref. [38], Suppl. Table 3). The two enterotype clustering gave the least number of significant COG/KO differences (10,848/2,493), while the three enterotype clustering had the highest number of significantly different KOs (3,474) and the four enterotype model the highest number of COGs (16,211), suggesting that finer-grained divisions of the microbiome reveal more functionally distinct communities (Suppl. Fig. 5). Functionally, ET F is most rich, perhaps contributing to community stability through functional redundancy. COG categories representing broad functional categories were significantly different between three enterotypes (23 out of 25, Suppl. Table 3), although the effect size was small compared to taxonomic levels (Suppl. Fig. 6). Consistent with the CAZymes analysis, Carbohydrate metabolism is overrepresented in ET B. Taken together these data suggest there are considerable functional differences between enterotypes.

ECOLOGICAL PROPERTIES OF ENTEROTYPES

Differences in taxonomic and functional composition suggest that enterotypes also differ in ecological properties; here, we discuss properties of enterotypes that are relevant to microbial community ecology. High richness can contribute to community resilience by, for example, compensating for perturbational challenges through functional redundancy. This may make ecosystems less susceptible to invasion because niches are already filled by resident species [39]. Conversely, reduced richness can have a negative impact by making communities more susceptible to dysbiosis upon perturbation [40]. The latter is likely to apply to gut communities, as several diseases have been associated with reduced bacterial diversity and richness, including inflammatory bowel disease (IBD) [41], irritable bowel syndrome (IBS) [42], *C. difficile* infection [43], and obesity [4,6,44].

Taxonomic richness was found to differ among three community clusters identified in an Amish population [24], with ET F having the highest and ET B the lowest values. This also was the case for enterotypes derived from HMP 16S rRNA gene data and shotgun metagenomes [1] as well as a Chinese dataset [45]. A diversity gradient exists between the extremes of these two enterotypes in all datasets (Figure 3C). ET F, which is mostly dominated by the phylum *Firmicutes*, represents the community state with the highest species richness and also the densest co-occurrence network [46]. Simulations of community compositions have yielded enterotype-like clusters over a range of species interaction strengths, consistent with the notion that enterotypes can be an emergent feature of a community [47]. Differences in genetic, functional and metabolic richness also exist among different enterotypes (Suppl. Fig. 5, 7).

Gut community composition in healthy adults is relatively stable over long time periods [48]. Studies investigating enterotypes within a time-series generally reported high enterotype stability over several months/years in healthy human [24,49] or chimpanzee hosts [26]. Two subjects showed stationary dynamics in 75-88% of the OTU's recorded daily for more than a year [50]. In an 8 to 10 years longitudinal study, enterotypes were stable, with only a few switches possibly occurring due to life style changes [51]. These observations are in line with two possible interpretations of gut community dynamics: a) the existence of globally applicable attractors (i.e., enterotypes, Suppl. Fig. 8A) or b) individual-specific attractors that exist mostly due to temporal autocorrelation of that

individual's gut community (Suppl. Fig. 8B). The response of the microbial community to different perturbations needs to be better studied because at present only limited data on gut community perturbations are available—e. g. antibiotics, fecal microbiota transplantation (FMT) and diet.

Only a few studies have addressed the impact of antibiotics on the gut microbiota. Antibiotic (ciprofloxacin) treatment of three human subjects resulted in incomplete recovery of the microbiota at the end of a 10 month period that encompassed repeated treatment, and the responses varied between individuals [52]. Long-term exposure to sub-therapeutic antibiotic doses profoundly changed the mouse gut microbiota composition [53], and short-term therapeutic doses induced substantial, partially recoverable shifts in humans [52,54]. Antibiotics usage in humans can result in prolonged, recurring infections with *C. difficile*, which is treatable by FMT (e.g. [43]). We have only begun to gain insight into the potential mechanisms underlying such interventions [55]. Although not yet studied, enterotype assignment may help guide donor choice in FMT or aid in promoting microbial community recovery following antibiotics exposure.

Diet is also known to influence microbiome composition, and has been investigated both within and outside the realm of enterotypes. For example, *Prevotella* enrichment has been linked to non-western and/or fiber-rich diets [8,56,57], also reported as overrepresentation of ET P [15,24,58,59]. This may be due to the ability of *Prevotella* hydrolases to metabolize plant fibers [36], in contrast with ET B associations with a diet rich in animal proteins and saturated fats [15,57]. In murine models, high-fat dietary intervention can decrease *Bacteroidetes* and favor *Firmicutes* [60,61]. Subjects on controlled diets for a 10-day period retained their original enterotype [15], as was also observed in travelling subjects [57]. However, dietary interventions over a year had a strong impact on the *Bacteroidetes*/*Firmicutes* ratio [15,62], which could lead to enterotype switches. Using the ratio from qPCR of *Prevotella* to *Bacteroides* as a proxy, enterotypes remained stable within the 6 months of a food intervention study, although cholesterol levels increased after intervention in individuals with ET P [23], suggesting broad implications of enterotypes in human health.

Whether extrinsic or intrinsic factors can lead to shifts in enterotypes is not presently clear, as the strength and direction of a response to a given perturbation is still controversial. Nonetheless, the composition of the microbiota can be modulated and may be exploited therapeutically.

ENTEROTYPES AND DISEASE

Enterotypes and their driving taxa have been repeatedly associated with diseases (Figure 3D). For example, a pilot study showed an increased prevalence of ET P in healthy individuals who had the heterozygous form of a Crohn's Disease (CD) risk allele [22]. Similarly, *Prevotella* was associated in mouse knockout models with colitis [63], highlighting that host genotypes can influence gut composition. However, in a study of more than 1,000 genotyped twin pairs, no heritability of enterotypes was apparent [64].

Overall, ET F appears linked to high microbiota diversity and decreased host inflammatory status (Figure 3C and D); this follows from multiple studies that associated a diseased host state to either *Prevotella*- or *Bacteroides*-enriched gut communities; CD patients harbor more *Prevotella* [65], while both *Prevotella* and *Bacteroides* were increased in the mucosa of UC patients [66]. Increased *Bacteroides* and *Prevotella* relative abundance has been linked to patients with colorectal cancer, coeliac disease and acute intestinal graft-versus-host disease in mice [67–69]. Furthermore, long term

antibiotic usage [70], rheumatoid arthritis [71], HIV infection [72] as well as Type 2 Diabetes[73] have also been linked to increased *Prevotella* abundance.

Increase in *Bacteroides* or ET B itself has been linked to NASH [44], colorectal cancer [58,69,74], celiac disease [68], immune-senescence and constant low-grade inflammation. The Bacteroides dominated CAG was correlated with increased C-reactive protein (CRP) levels [25]. *Bacteroides* abundance has also been linked to inflammatory parameters (insulin resistance, lymphocyte counts, CRP), fatty liver and insulin resistance [6]. In reanalysis of the MetaHIT dataset we find lymphocyte counts and CRP to be significantly increased in ET B compared to ET F (FDR<0.1), with ET F samples on average lower in insulin levels and Insulin resistance index (HOMA IR), (FDR=0.107 for both) (Suppl. Table 4). Mice with an ET B-like enterotype had increased local inflammation in the gut, as measured by calprotectin levels [28]. Thus far, ET F has only been associated with atherosclerosis [75], though an increase in the Firmicutes-Bacteroidetes ratio has been observed in IBD patients [76]. The causality of these observations is still in question because it is not clear whether the disease state selects for an altered microbiota or vice versa. Furthermore, some of these associations may manifest at higher resolution taxonomical units, with specific species or even strains being the affected unit.

CLINICAL RELEVANCE OF STRATIFICATION

The many factors that can influence the gut microbiota suggest that an enterotype classification by itself will not be sufficiently specific as a stand-alone diagnostic marker of a relevant state [77]. Thus, enterotypes should rather be viewed as a potential indicator of increased risk and complications and used to complement other clinical markers. Stratification could improve the efficacy of drugs, nutrients, or diets in a more personalized setting. Such approaches can be implemented largely independently of a gradient- or cluster-centric view, as any clinical stratification can potentially be useful if accurate. For example, defined cutoffs for body mass index (BMI) are important guides to patient disease risk, although such cutoffs are imposed on a continuous gradient [78].

In clinical research, one role of enterotyping should be to define the boundaries of 'normal' gut communities and identify individuals outside of them, serving as a health indicator. A model of six-species communities showed striking distribution patterns among patients; the networks resembling the three enterotypes most strongly were overrepresented in healthy patients, while two of the new states were overrepresented in frail, elderly patients [7].

An intriguing case is the newly reported enterotype H, enriched in *Enterobacteriaceae*, but compositionally located "between" ET B and ET P, that was linked to obesity, NASH, high blood ethanol levels and high reactive oxygen species (ROS) [44]. ET H could be an early example of a dysbiotic enterotype. A single pathogenic strain may be changing the ecosystem by inducing a strong immune reaction, as in the case of *C. difficile* [79] or *Salmonella* infections [50]. Similarly, since ET B has been associated with host inflammation (section "Enterotypes and disease"), that knowledge could be exploited in evaluation of health status

Further applications lie in supporting biomarker discovery and intervention studies. Possibly, some diseases will have different etiologies, depending on enterotype. Enterotype stratification could allow discovery of underlying, weaker signals that are dominated by the large variation in microbial communities between individuals. In one mouse study, for example, only an enterotype stratification allowed discovery of genotype-microbiome and cage-microbiome associations [28].

Similarly, stratifying patients into 8 ecotypes helped identify medical parameters that correlated with microbial composition [25], and clustering before classifying *C. difficile* samples significantly improved accuracy [80]. Because enterotypes can be associated with disease, immune function and diet, they might provide a proxy for the risks linked to patient lifestyle. In clinical trials, determining enterotype at entry might reduce microbiome-induced response variation. Therefore, defined microbiota profile clusters can be a useful clinical tool.

TOWARDS STANDARDS FOR ENTEROTYPING

Standardization of enterotyping is essential to enable comparable research and clinical efforts. In addition to the technical challenges mentioned above, an inherent property of clustering is that assignments of single samples depend on which other samples are analyzed at the same time. An enterotype defined this way makes comparisons across studies difficult. For example, if by chance the majority of samples in a single study are ET B or ET F and only a few are ET P, the optimal cluster score might indicate two or one cluster(s). Nevertheless, one might identify these few ET P samples, based on the knowledge that such a state exists. Combining data from multiple studies is often challenging, because differences in DNA extraction methods, sample handling, sequencing technology, primer choice (for 16S rRNA gene amplification) and data processing (e.g. 16S rRNA clustering, copy number correction and chimera reduction) influence the proportions of bacteria detected, which may lead to a per-study clustering and a bias in detecting enterotype clusters [81]. Extreme rigor is needed in standardizing these steps, perhaps in conjunction with artificial "mock" communities that span a large proportion of the phylogenetic spectrum of microbes found in the gut and enable comparability between standard and clinical samples. Furthermore, there is a need for more longitudinal studies involving larger population cohorts across multiple continents to identify additional confounding factors. Indeed, several consortia such as IHMS[40], MBQC[41] and GSC[42] are already trying to set standards for metagenomics and identify sources of variation.

We propose two synergistic approaches geared towards de-novo enterotype assignment and recovery of structure defined in a reference dataset (Figure 4). For the latter, we propose a classification standard that both circumvents many of the problems outlined above and provides more comparable results as well. While we do not want to limit other explorations of the data or novel analysis options, any different scheme should at least be compared with the results from the procedure we describe here. Based on the MetaHIT data [6] set, we have trained a classifier at genus level on taxonomic and functional features, that recovers intrinsic clustering observed in the Chinese type 2 diabetes study [45] and in the HMP [1] dataset (Suppl. Fig. 10). The classifier is available at [http://enterotypes.org] and can easily be obtained following the instructions therein. If the results of a de-novo clustering differ from the classifier results, we recommend caution in naming the stratification outcome as enterotypes.

CONCLUSIONS

[40] http://www.microbiome-standards.org/
[41] http://www.mbqc.org/
[42] http://gensc.org/

Identification and characterization of the major patterns related to human gut microbiota configurations remain challenging. Given an array of available approaches, each with their advantages and caveats, the number of recovered enterotype states and their statistical support will vary. With more standardization and control of sample processing and data analysis, increased concordance among different studies can be achieved. Enterotype attribution must be refined with a wider range of samples and contextual information, extending beyond the industrialized world to better represent the global human population.

Despite the difficulties outlined above, multiple studies have found enterotypes with similarities in compositional properties (Figure 2). While not clearly discrete and confounded by various factors, they differ in taxonomic, functional and ecological properties and can be recovered across large datasets (Suppl. Fig. 10). To improve comparability among future studies, we propose here a standard enterotyping procedure that can serve as a reference to guide the interpretation of results.

Enterotypes may be relevant in various clinical settings, ranging from direct disease associations to guidance about compatible donor-recipient pairs in fecal microbiota transplantation. They may also become important for more personalized dietary interventions or other gut modulation treatments. We believe, despite our still limited knowledge, that enterotypes are a useful concept to describe the human microbial community landscape.

REFERENCES

1. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
2. Luckey, T. D. Introduction to intestinal microecology. *Am. J. Clin. Nutr.* **25,** 1292–1294 (1972).
3. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
4. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–4 (2009).
5. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).
6. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).
7. Jeffery, I. B., Claesson, M. J., O'Toole, P. W. & Shanahan, F. Categorization of the gut microbiota: enterotypes or gradients? *Nat. Rev. Microbiol.* **10,** 591–592 (2012).
8. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–7 (2012).
9. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M. & de Vos, W. M. Tipping elements in the human intestinal ecosystem. *Nat. Commun.* **5,** 4344 (2014).
10. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–80 (2011).
11. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl ,** 4680–4687 (2011).
12. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9,** e1002863 (2013).
13. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509,** 357–60 (2014).
14. Zhou, Y. *et al.* Exploration of bacterial community classes in major human habitats. *Genome Biol.* **15,** R66 (2014).
15. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–8 (2011).
16. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One* **7,** e30126 (2012).
17. Dethlefsen, L., Eckburg, P. B., Bik, E. M. & Relman, D. a. Assembly of the human intestinal microbiota. *Trends Ecol. Evol.* **21,** 517–23 (2006).
18. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).
19. Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* **16,** 82 (2015).
20. Morotomi, N. *et al.* Evaluation of Intestinal Microbiotas of Healthy Japanese Adults and Effect of Antibiotics Using the 16S Ribosomal RNA Gene Based Clone Library Method. *Biol. Pharm. Bull.* **34,** 1011–20 (2011).

21. Karlsson, F. H., Nookaew, I. & Nielsen, J. Metagenomic Data Utilization and Analysis (MEDUSA) and Construction of a Global Gut Microbial Gene Catalogue. *PLoS Comput. Biol.* **10,** e1003706 (2014).

22. Quince, C. *et al.* The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study. *Gut* **0,** 2012–2014 (2013).

23. Roager, H. M., Licht, T. R., Poulsen, S. K., Larsen, T. M. & Bahl, M. I. Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet. *Appl. Environ. Microbiol.* **80,** 1142–9 (2014).

24. Zupancic, M. L. *et al.* Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. *PLoS One* **7,** e43052 (2012).

25. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–84 (2012).

26. Moeller, A. H. *et al.* Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat. Commun.* **3,** 1179 (2012).

27. Moeller, A. H. *et al.* Stability of the gorilla microbiome despite simian immunodeficiency virus infection. *Mol. Ecol.* **24,** 690–7 (2015).

28. Hildebrand, F. *et al.* Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* **14,** R4 (2013).

29. Wang, J. *et al.* Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. *Pnas* **111,** E2703–10 (2014).

30. Mach, N. *et al.* Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environ. Microbiol. Rep.* (2015). doi:10.1111/1758-2229.12285

31. Scheffer, M. & Carpenter, S. R. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends Ecol. Evol.* **18,** 648–656 (2003).

32. Huse, S. M., Ye, Y., Zhou, Y. & Fodor, A. A. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS One* **7,** e34242 (2012).

33. Bäckhed, F. *et al.* Defining a Healthy Human Gut Microbiome: Current Concepts, Future Directions, and Clinical Applications. *Cell Host Microbe* **12,** 611–622 (2012).

34. Sankaran, M. *et al.* Determinants of woody cover in African savannas. *Nature* **438,** 846–9 (2005).

35. Staver, a C., Archibald, S. & Levin, S. Tree cover in sub-Saharan Africa: rainfall and fire constrain forest and savanna as alternative stable states. *Ecology* **92,** 1063–72 (2011).

36. Purushe, J. *et al.* Comparative genome analysis of Prevotella ruminicola and Prevotella bryantii: insights into their environmental niche. *Microb. Ecol.* **60,** 721–9 (2010).

37. Kaoutari, A. El, Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.* **11,** 497–504 (2013).

38. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).

39. Elmqvist, T. *et al.* Response diversity, ecosystem change, and resilience. *Front. Ecol. Environ.* **1,**

488–494 (2003).

40. Virgin, H. W. & Todd, J. A. Metagenomics and Personalized Medicine. *Cell* **147,** 44–56 (2011).
41. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).
42. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G799–807 (2011).
43. van Nood, E. *et al.* Duodenal infusion of donor feces for recurrent Clostridium difficile. *N. Engl. J. Med.* **368,** 407–15 (2013).
44. Zhu, L. *et al.* Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology* **57,** 601–9 (2013).
45. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
46. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8,** e1002606 (2012).
47. Gibson, T. E., Bashan, A., Cao, H.-T., Weiss, S. T. & Liu, Y.-Y. On the Origins and Control of Community Types in the Human Microbiome. *arXiv* (2015).
48. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).
49. Mardanov, A. V *et al.* Metagenomic Analysis of the Dynamic Changes in the Gut Microbiome of the Participants of the MARS-500 Experiment, Simulating Long Term Space Flight. *Acta Naturae* **5,** 116–125 (2013).
50. David, L. a *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).
51. Rajilić-Stojanović, M., Heilig, H. G. H. J., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15,** 1146–1159 (2012).
52. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4554–61 (2011).
53. Cho, I. *et al.* Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* (2012). doi:10.1038/nature11400
54. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16,** 73 (2015).
55. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517,** 205–8 (2015).
56. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14691–6 (2010).
57. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).
58. Ou, J. *et al.* Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. *Am. J. Clin. Nutr.* **98,** 111–20 (2013).

59. Nakayama, J. *et al.* Diversity in gut bacterial community of school-age children in Asia. *Sci. Rep.* **5,** 8397 (2015).

60. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1,** 6ra14 (2009).

61. Hildebrandt, M. A. *et al.* High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* **137,** 1716–24.e1–2 (2009).

62. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Human gut microbes associated with obesity. *Nature* **444,** 1022–3 (2006).

63. Elinav, E. *et al.* NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145,** 745–57 (2011).

64. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159,** 789–799 (2014).

65. Kleessen, B., Kroesen, A. J., Buhr, H. J. & Blaut, M. Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. *Scand. J. Gastroenterol.* **37,** 1034–41 (2002).

66. Lucke, K., Miehlke, S., Jacobs, E. & Schuppler, M. Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. *J. Med. Microbiol.* **55,** 617–24 (2006).

67. Heimesaat, M. M. *et al.* MyD88/TLR9 mediated immunopathology and gut microbiota dynamics in a novel murine model of intestinal graft-versus-host disease. *Gut* **59,** 1079–87 (2010).

68. De Palma, G. *et al.* Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. *BMC Microbiol.* **10,** 63 (2010).

69. Sobhani, I. *et al.* Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One* **6,** e16393 (2011).

70. Jernberg, C., Löfmark, S., Edlund, C. & Jansson, J. K. Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* **156,** 3216–23 (2010).

71. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2,** e01202–e01202 (2013).

72. Noguera-Julian, M. *et al.* Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* 1–12 (2016). doi:10.1016/j.ebiom.2016.01.032

73. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5,** e9085 (2010).

74. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).

75. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* **3,** 1245 (2012).

76. Rajilić-Stojanović, M. *et al.* Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. *Gastroenterology* **141,** 1792–801 (2011).

77. Knights, D. *et al.* Rethinking 'Enterotypes'. *Cell Host Microbe* **16,** 433–437 (2014).

78. Flegal, K. M., Kit, B. K., Orpana, H. & Graubard, B. I. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *JAMA* **309,** 71–82 (2013).

79. Fuentes, S. *et al.* Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. *ISME J.* 1–13 (2014). doi:10.1038/ismej.2014.13

80. Schubert, A. M. *et al.* Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *MBio* **5,** e01021–14 (2014).

81. Wu, G. D. *et al.* Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.* **10,** 206 (2010).

ACKNOWLEDGEMENTS

The authors are grateful to the members of the Bork group at EMBL for discussions and assistance. The research leading to these results has received funding from EMBL, the European Research Council via the CancerBiome project (project reference 268985), MicrobesInside (250172) and the European Community's Seventh Framework Programme via the MetaHIT (HEALTH-F4-2007-201052), the METACARDIS project (FP7-HEALTH-2012-INNOVATION-I-305312), the European Union's Horizon 2020 research and innovation programme (Marie Sklodowska-Curie grant 600375), Metagenopolis grant ANR-11-DPBS-0001 and the IHMS project (FP7-HEALTH-2010-single-stage-261376).

COMPETING INTEREST STATEMENT

The authors declare no competing financial interests.

FIGURES

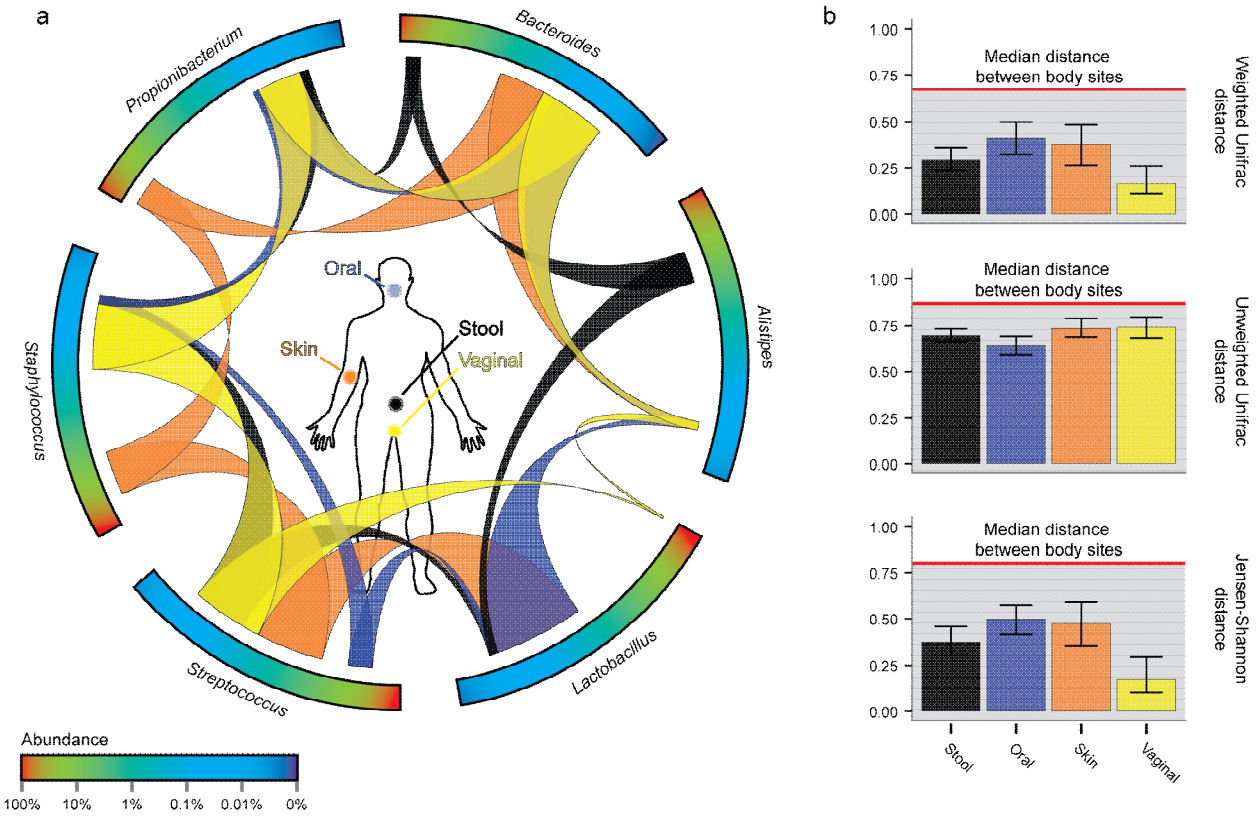


Figure 1: The microbiota of distinct body locations within the healthy human is separable at the genus level.

(A) From 2,910 HMP samples, the six most discriminating genera between body sites are represented. Their interquartile range (indicated by the width of the ribbon) within each body site, show clear differences in site abundance for each. (B) Median inter-sample distances (error bars ranging from the 25th to the 75th quantile) compared to the median between all body-sites (red line) illustrate that the most widely used distance measures are able to capture similarities and differences between these biomes, albeit with different effectiveness.

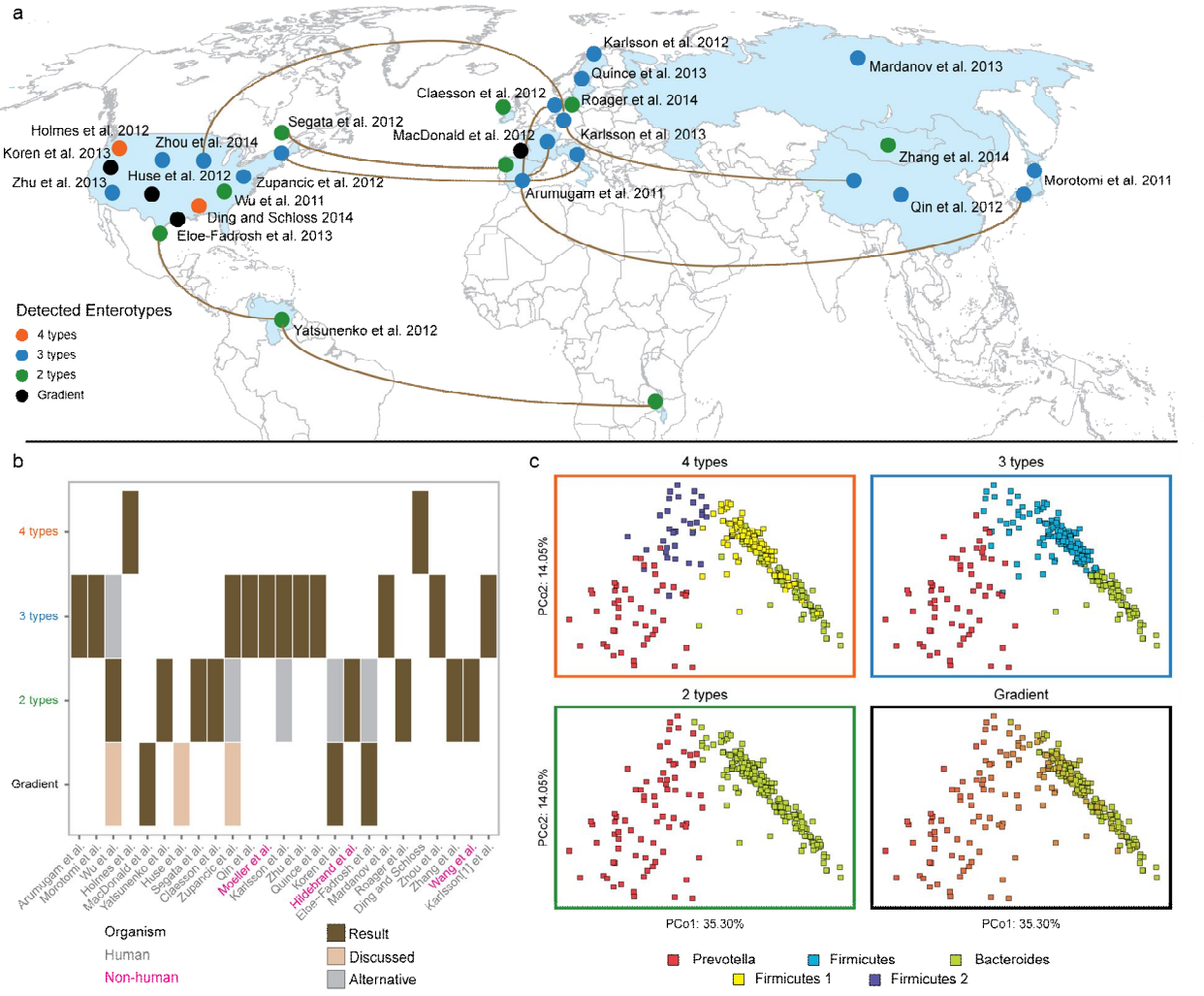


Figure 2: Overview of publications that specifically address enterotypes, published before January 2015.

(A) The geographical distribution of samples from which enterotypes have been derived (Suppl. Table 1), are colored according to the number of microbial clusters they report. The location on the map indicates the country that the samples were collected from. The links between locations represent samples belonging to one single study. The overrepresentation of "western" countries is a well-known bias and probably misses a portion of global variation in other human societies. (B) Number of enterotypes derived from a study as "Result", and less likely "Alternative" enterotype model that is often proposed within the same study or cannot be ruled out ("Discussed"). (C) Projection onto a set of 278 Danish samples[6] of the three most frequent enterotype classification schemes, based on different methods, and including the Prevotella/Bacteroides gradient. This shows a split into a gradient and two, three (distance based clustering) or four enterotypes (Dirichlet multinomial mixture models). The local structure is preserved regardless of the method applied and Prevotella (ET P) remains separated, suggesting the methods mostly differ in dividing the area between ET B and ET F.

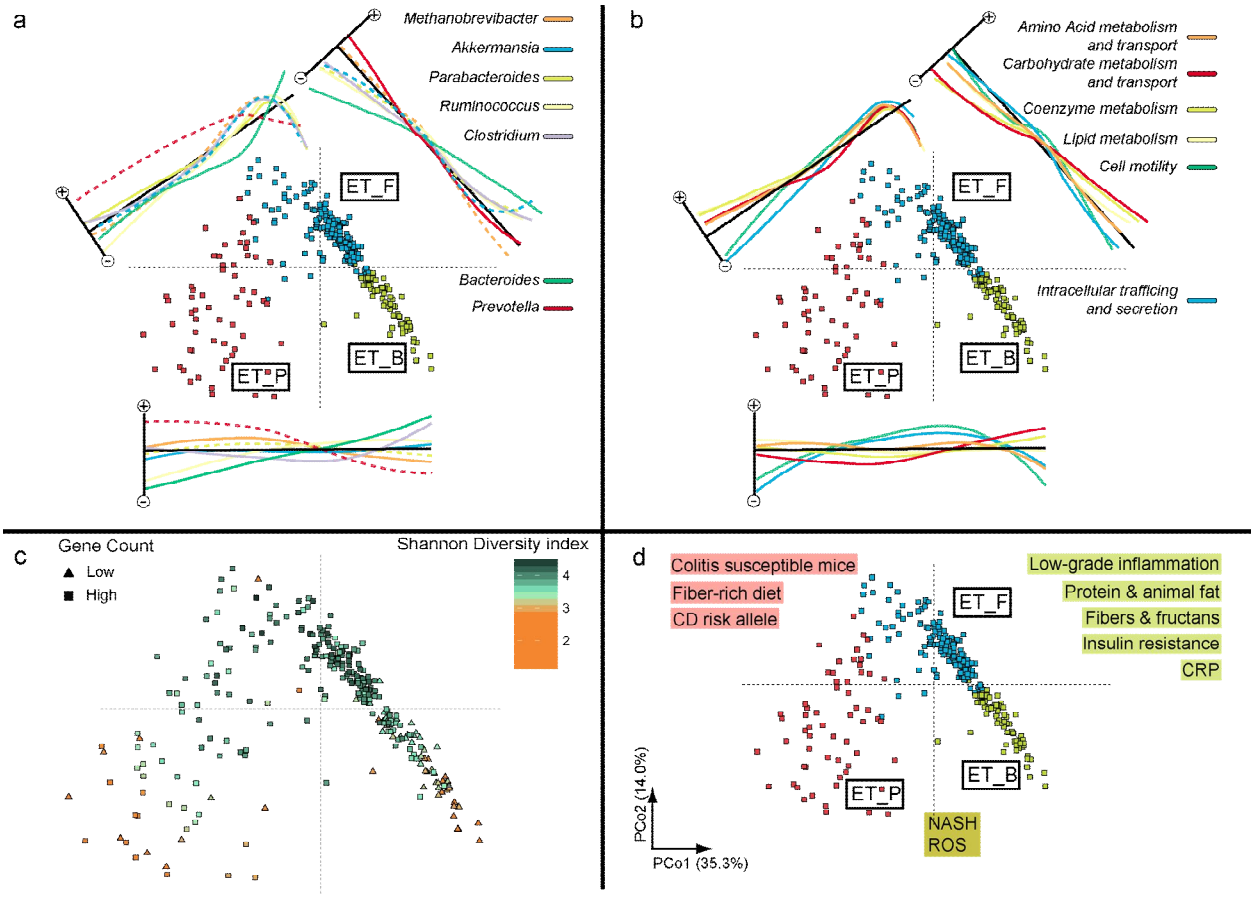


Figure 3: Microbiota of human fecal samples has local substructure.

Ordination of 278 Danish[6] samples on Jensen-Shannon distance transformed space. For orientation, a three-enterotype model is illustrated by color in A, B and D. (A) The log-transformed relative abundance of the most significantly differing genera. On the adjacent axis, the projected abundance changes between the respective community types are shown. Bimodal abundance profiles (dotted lines, dip test p-value < 0.05) as well as gradual abundance changes (solid lines) can be identified, supporting a gradient or cluster model, respectively. (B) Abundance changes of selected COG categories were projected onto the ordination, illustrating that functional composition differs between enterotypes. (c) mOTU level Shannon diversity index and gene richness (low gene count (LGC) is considered for subjects with less than 480k genes according to Le Chatelier et al., 2013; all other subjects have high gene count (HGC)) are significantly different between enterotypes (Suppl. Fig. 7), mostly following gradual changes over the whole enterotype space. (D) Summary of the diseases and dietary constituents that have been associated with *Prevotella*, Firmicutes or *Bacteroides* -dominated gut communities (Suppl. Table 5).

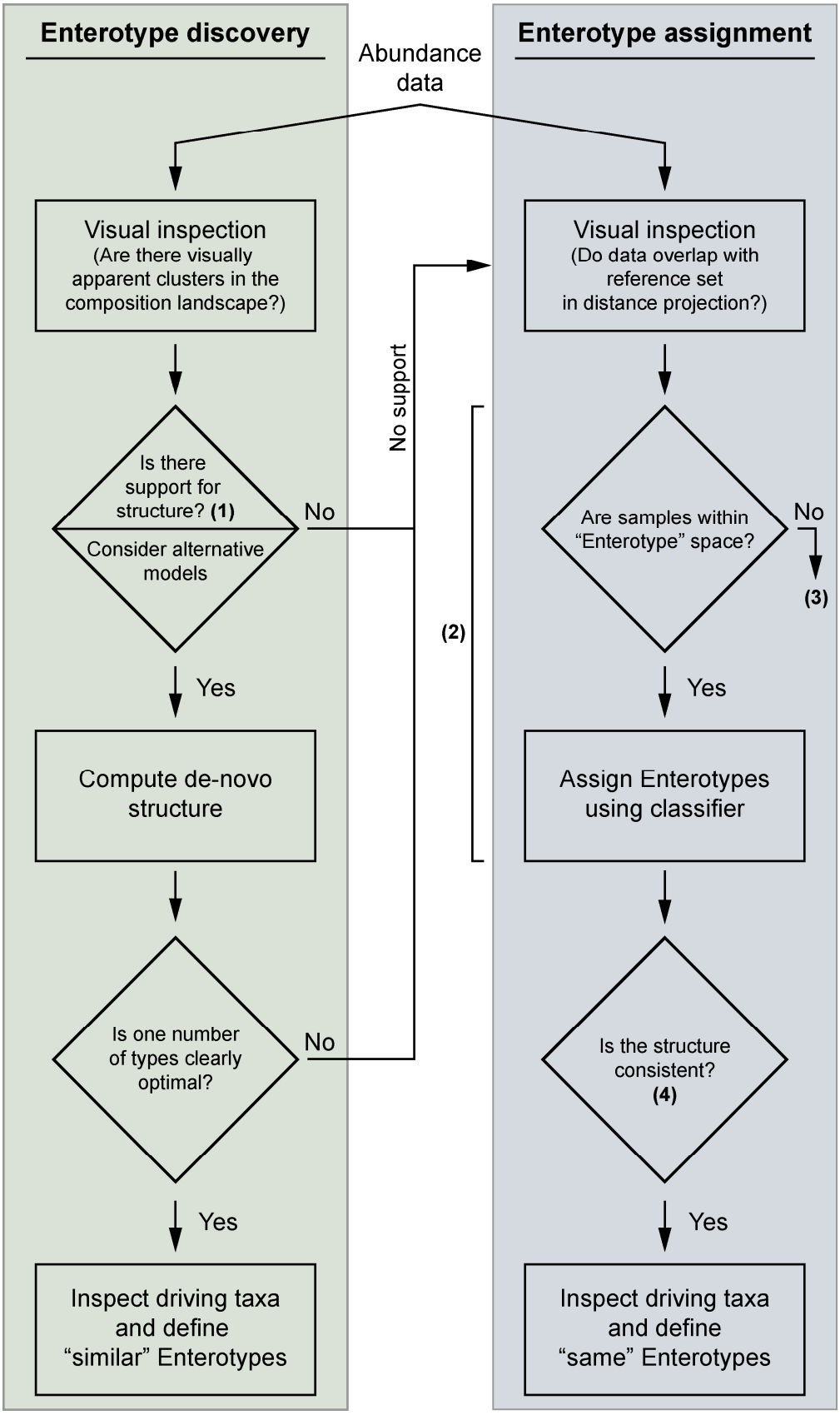

Enterotype discovery
Abundance data
Enterotype assignment
Visual inspection (Are there visually apparent clusters in the composition landscape?)
Visual inspection (Do data overlap with reference set in distance projection?)
No support
Is there support for structure? (1)
Consider alternative models
No
Are samples within "Enterotype" space?
No
(3)
(2)
Yes
Yes
Compute de-novo structure
Assign Enterotypes using classifier
Is one number of types clearly optimal?
No
Is the structure consistent? (4)
Yes
Yes
Inspect driving taxa and define "similar" Enterotypes
Inspect driving taxa and define "same" Enterotypes

Figure 4: Determination of Enterotype structure.

Flow diagram including recommended steps for determining enterotype assignment based on microbial abundance data. Two main routes to obtain enterotype assignments are depicted: de-novo identification of enterotypes (discovery) and assignment based on a reference dataset. The suitability of existing models imposed on the data to describe the composition landscape (1) can be assessed by either determining the existence of cluster structure, using one of the proposed clustering strength measures (Suppl. Fig. 2) or by using a DMM modeling framework [16]. Other models might also be useful in capturing the structure in the data, although an exact implementation is not yet available. Determining whether samples are within the enterotype space (2) is based on similarity in composition to adult human stool samples from the HMP [1] and MetaHIT [6] studies. This suitability check and a respective classifier are available at [http://enterotypes.org]. There are many explanations for the different compositional structure (3); for example, they may come from non-western individuals, or from infants. Technical issues such as DNA extraction, PCR primers, and/or bioinformatics preprocessing, may skew the analysis. The consistency of the separation (4) obtained from the classifier may be determined using a Silhouette index.

SUPPLEMENTARY MATERIAL

DATA SOURCE

All analyses were performed on previously published data. 16S rRNA gene abundance data from 2910 samples, comprising 1266 oral, 357 airways, 187 stool, 836 skin and 264 vaginal, originating from the Human Microbiome Project (HMP)[1] were downloaded from their web resource (http://www.hmpdacc.org/). An additional 139 stool metagenomic shotgun sequenced samples were downloaded from the same location. Additional metagenomic shotgun sequencing data originate from samples (368 Chinese samples and 278 samples from the MetaHIT project) described in Qin et al. (2012) and Le Chatelier et al. (2014), respectively. Three additional samples from the US were used, which are described in Schloissnig et al. (2013).

Data, together with code for generating the main figures can be found at: https://hub.docker.com/r/costeapaul/enterotype_figures/. Instructions for pulling and running the docker can also be found there.

TAXONOMIC AND FUNCTIONAL ANALYSIS

For 16S rRNA gene-based taxonomic composition analysis, we used operational taxonomic unit (OTU) or genus level relative abundances. Genus level abundance matrices were calculated by adding relative abundances of all taxonomically annotated OTUs. OTUs not annotated at genus level are considered "unclassified" and their relative abundances were agglomerated into that category.

For cross-study analyses, shotgun sequencing reads were mapped to a database of selected single copy phylogenetic marker genes (mOTU.v1.padded)[82] and summarized into species-level (mOTU) and genus-level relative abundances. Functional profiles of clusters of orthologous groups (COGs) and KEGG orthology groups (KO), including both those of eukaryotic and bacterial origin, for MetaHIT, Chinese, and HMP samples were computed using MOCAT[83] by mapping shotgun sequencing reads to an annotated reference gene catalogue as described in Voigt et al.[54]. COG category abundances were calculated by summing the abundance of the respective COGs belonging to each category per sample, excluding NOGs.

Weighted and unweighted UniFrac distances were downloaded from the HMP web resource (*http://www.hmpdacc.org/*). Jensen-Shannon distances were computed on genus level relative abundance matrices, as described in the enterotyping tutorial (*http://enterotype.embl.de/enterotypes.html*).

For determining the optimal number of clusters on all data matrices, three different measures were used from the R fpc package[84] (version 2.1.9). The Calinski-Harabasz index[85] and the silhouette index[86] of a given distance matrix and a set number of clusters were computed using the function *pamk* with default parameters. The prediction strength[87] was calculated with a modified version of the *prediction.strength* function, allowing a distance matrix as a parameter, with the dataset being randomized 50 times. For all the different measures, we varied the number of clusters between two and ten and considered the cluster number with the highest value for each measure to be the optimal one.

ORDINATION

Visualization of distance matrices was performed using unsupervised ordination methods. Principal coordinate analysis was performed with the R ade4 package (version 1.6.2), using the *dudi.pco* function.

Parallel to each pair of enterotypes, we plotted the abundances of selected genera and functional categories (Figure 3). For each combination, we performed principal component analysis on the 2-dimensional PCoA coordinates, to identify the axis that explains the greatest variation between enterotypes (i.e. the first eigenvector). This component forms the x-axis for the distributional plots. Subsequently, we log transformed abundance of genera/ functions, then scaled and centered them. The plotted line is a smoothed spline fitted to these transformed abundances, using the R base function *smooth.spline.* To test for significant bimodal distribution of feature abundance, Hartigan's dip test statistic from R package *diptest* was used.

FEATURE SIGNIFICANCE TESTING

Univariate testing for differential abundances of taxonomic and functional features between two or more groups was tested using a Wilcox-Rank Sum test or Kruskal-Wallis test (p-value), respectively, corrected for multiple testing using the Benjamini-Hochberg false discovery rate (q-value). COGs and KOs occurring in less than four samples or with an average normalized count abundance < 30 across were excluded from the univariate analysis.

DIVERSITY ANALYSIS

Richness and Shannon diversity index for taxonomic and functional features, mOTU and OTU were calculated after rarefaction of matrices to 3 000 and5 000, units per sample, respectively, using the *vegan* R package. Rarefactions of COG, KO and gene matrices were done using a C++ program developed for the Tara project [88] with a per sample rarefaction depth of 6 900 000. In total we performed 30 repetitions, in each of which we measured the richness and Shannon diversity metrics within a rarefaction. The median value of these was taken as the respective richness/ diversity measurement for each sample. These thresholds were chosen to include most samples.

PARAMETER-DEPENDENCE OF CLUSTERING

Clustering algorithms have been developed and employed for nearly 100 years (e.g. Driver and Kroeber 1932) and have more recently been applied to analyze microbial compositions, especially those of the human gut. However, it is challenging to determine whether there are actual clusters present, and if so, how many? A number of clustering optimality measures, as well as distance measures were employed for determining the number of microbial clusters that may be present in the human gut microbiota. To describe inter-sample differences, most studies use a combination of weighted and unweighted UniFrac distances or Jensen-Shannon distance (JSD) (Suppl. Table 1). To determine the optimal number of clusters in the space described by the distance measure, the CH-index[85], silhouette index[86] and prediction strength[87] are commonly used.

Different distance metrics will give a different weight to community features. For example, the UniFrac distance[90], be it weighted (taking abundance into account) or unweighted (presence-

absence only), is based on the importance of phylogenetic distance between the components of the community. In contrast, the JSD distance does not take phylogenetic information into account, and measures the mutual information shared between two samples. For both weighted UniFrac and Jensen-Shannon, the underlying hypothesis is that variation in highly abundant members is the most relevant feature for describing similarities. The hypothesis of unweighted analysis is that community membership is the most important feature. While these distances are conceptually very distinct, they may result in the same outcome, though they exhibit quantitatively and qualitatively different properties (Suppl. Fig. 9). Another property worth considering is the absolute numbers of microbes in any given sample; it may be that observed fluctuations in composition poorly reflect the actual cell counts of the members, as the total amount varies considerably. This may constitute a further confounder when trying to disentangle compositional properties.

Within the Human Microbiome Project (HMP) dataset, 2910 samples are available from a range of human body sites[1]. We used this dataset to benchmark the aforementioned distances and optimality criteria, based on the assumption that human body-sites are inhabited by different microbial populations and that their separation should be clear (Figure 1). The PCoA projections of the distance space into two dimensions shows that the largest part of the variation does not separate the body sites properly, except in the case of the JSD distance on genus level (Suppl. Fig. 2a). This metric and weighted UniFrac both recovered the four expected clusters in conjunction with PS or Silhouette index (Suppl. Fig. 2b). However, even when recovered, the separation appears not to be very strong, with silhouette values being low (0.4 at best). Since often three or less clusters are chosen to be the optimal cluster number, we conclude that the clustering approach is underpowered.

CROSS-STUDY ENTEROTYPE COMPARISON

Comparability of the structure across multiple datasets is a necessary characteristic of the enterotype concept. However, although similar genera were reported as being most abundant in gut stratifications (Suppl. Table 1), this does not automatically imply similar communities or structure. To test the assumption of comparability, we used three unrelated large datasets, with a different sampling procedures (US HMP[1], Chinese diabetes type 2 study[45], European MetaHIT consortium[6]) and clustered these with the PAM clustering algorithm on a JSD distance at genus level[10]. The obtained clusters had an overrepresentation of *Prevotella*, *Bacteroides* or Firmicutes (the latter represented by *Ruminococcus*, *Eubacterium* and *Subdilogranulum* respectively), as expected.

Although we do not exclude alternative scenarios (see Figure 2), we first trained a LASSO logistic regression classifier [91] to recover the three enterotypes within the MetaHIT samples. This was then used to classify samples from the other two studies. The respective ROC-AUC was high (Figure 4), meaning that the classifier and unsupervised clustering mostly assigned the same cluster memberships to samples. One difference is that the classifier can be used on any arbitrarily small dataset. This approach could also be expanded to classify single samples based on other machine learning techniques, e.g. a trained DMM model.

Furthermore, if enterotypes reflect community compositions and not just differences in the driver species, i.e. are reflecting different ecological networks, we expect the classification to remain pertinent after removing *Bacteroides* and *Prevotella* from the data. Indeed, although with lower

accuracy than when including these two taxa, the classification still captures the initial enterotypes in all datasets.

Further, using the abundance of gene families within each respective enterotype, the prediction of enterotype state is even stronger in cross validation than when using taxa abundances (Figure 4). For this analysis, we used only commonly represented functional categories (i.e., COG's that have representative genes in at least five of the 50 most abundant genera), ensuring that the classifier does not exploit functional categories which are restricted to taxonomic subgroups.

DETERMINING IF SAMPLES ARE WITHIN ENTEROTYPE SPACE

Using the HMP dataset[1], we compute the distance between all stool samples using a genus summarized OTU table. This allows us to define the expected distance distribution of stool samples. For any novel sample, we compute the distance to all stool samples in the HMP data and consider it to be in the enterotyping space if its average distance is within one standard deviation of the stool distance distribution. Using this approach, we correctly identify western-like stool sample and reject all other body-site samples as not being in the enterotyping space. Furthermore, we also correctly classify infant samples as being outside the enterotyping space (data not shown).

SUPPLEMENATRY INFORMATION

Supplementary Table 1: Microbial community studies researching the presence of enterotypes (ET). Abbreviations: F=ET F (Firmicutes enriched), B=ET B (Bacteroides enriched), P=ET P (Prevotella enriched), CH= Calinski–Harabasz pseudo F–statistic, SIL= Silhouette internal cluster optimality criterion.

Study	Year	Technology	ET reported	Optimal Cluster number	Notes
[10]	2011	454 rRNA, illumina WGS, Sanger WGS	B, F, P	CH (3)	First study to show ET's
[20]	2011	Sanger rRNA	B, F, P	visual	clone library
[15]	2011	454 rRNA	(F+B), P	CH (3), SIL (2)	Diet relation to ET's
[24]	2012	454 rRNA	B, F, P	[92] (3), SIL (2)	Species network based ET identification
[8]	2012	454 rRNA	P, (F+B), *Bifidobacteria*	SIL (2)	Includes children that form a separate cluster
[32]	2012	454 rRNA	gradient	visual	Analysis not based on clustering, HMP
[12]	2013	454 rRNA, illumina WGS	various	SIL(2), CH(3) (rDNA); 2 (WGS)	Extensive testing of methodology, HMP
[7,25]	2012		P, B, *Ruminococcus* , *Oscillibacter, Alistipes, Odoribacter*	CAGs (6), SIL(2), CH(2)	co-abundance groups
[45]	2012	illumina WGS	B, F, P	SIL (3)	
[75]	2012	illumina sg	B, F, P	CH (3), SIL (2)	Atherosclerosis associated to ET F
[16]	2012	Sanger rRNA	Similar to F, P, B, F2	Dirichlet Multinomial Mixtures (4)	
[26]	2012	illumina rRNA	B, F, P	CH (3), SIL(2/3)	Chimpanzee
[22]	2013	454 rRNA	B, F, P	[16] (3)	Association of ET P to CD risk allele
[44]	2013	454 rRNA	B, F, P	Based on composition	obesity and NASH in adolescents
[28]	2013	Illumina	F, B	CH (2), SIL (2)	Mouse; ET B shows

		rRNA			links to inflammation
[49]	2013	454 rRNA	B, F, P	Based on composition	Time series
[13]	2014	454 rRNA	Similar to B, F, P, F2	Dirichlet Multinomial Mixtures (4)	HMP reanalysis
[14]	2014	454 rRNA	B, F, P	Complete linkage, Bray-Curtis clustering, SIL(3)	HMP reanalysis
[23]	2014	qPCR	B, P	Prevotella to Bacteroides ratio	Time series on Food trials
[93]	2014	454 rRNA	B, P	Weighted Unifrac SIL(2)	
[29]	2014	454 rRNA	B, Robinsella (Firmicutes dominated)	CH (2)	Wild mice; predictable ET switch after capture
[21]	2014	Illumina WGS	B, F, P	CH(3), SIL(3)	4 datasets combined [1,3,45,94]
[27]	2015	Illumina rRNA	Similar to F, P	CH(2)	Gorilla, no association to SIV infection
[30]	2015	454 rRNA	F, P	CH(2), SIL(2)	Swine, juvenile development into adult enterotypes
[95]	2016	Illumina rRNA	F,B,P	JSD clustering (2,3), DMM(4)	3984 Samples from US and Europe

Supplementary Table 2: Percentage of CAZY enzymes annotated within 8 substrate categories on a selected subset of gut specific bacterial genomes as published in [37]. Bacteroides contains 15 genomes of genus Bacteroides, Firmicutes are 104 genomes of phylum Firmicutes and Prevotella contains 3 genomes of genus Prevotella. Note that due to multiple substrate specificities, percentage do not add up to 100%.

CAZY category	Bacteroidetes	Firmicutes	Prevotella
Plant.Cell.Wall.Carbohydrates	50%	35%	42%
Chitin	0%	0%	0%
Alpha.glucans	5%	20%	10%
Animal.Carbohydrates	50%	28%	35%
Bacterial.Cell.Wall.Carbohydrates	4%	23%	12%
Fructans	1%	4%	1%
Fungal.Carbohydrates	11%	7%	9%
Dextran	0%	0%	0%

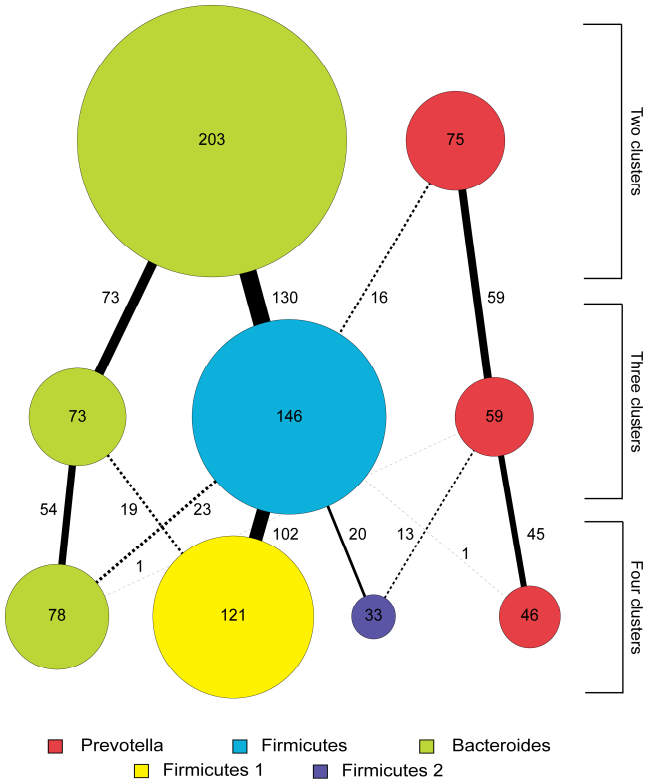
Supplementary Table 3: Functional differences between 3 different enterotype models: 2 Types represents a model comparing ET P against a combined ET F+ ET B, 3 Types compares the three first reported enterotypes (ET B, ET F, ET P) and 4 Types are enterotypes as determined by DMM modelling. Used gene families are derived from COG [96] and KEGG [97] annotations.

Supplementary Table 4: Associations between obesity related parameters reported in [6] and ET state, split for the 2, 3 and 4 clusters.

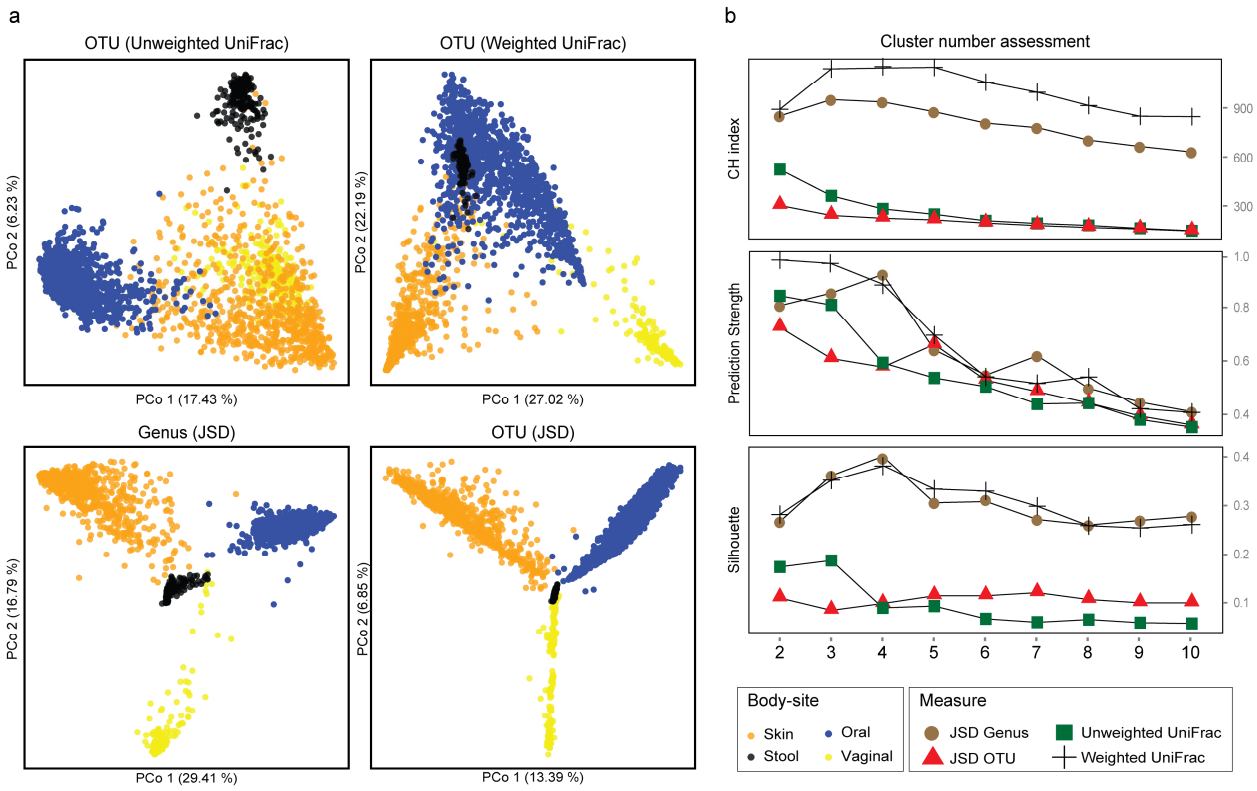
Supplementary Table 5: Studies reporting associations between enterotype drivers and host states

Study	Enriched driver	Phenotype
75	Bacteroides	Atherosclerosis
4,61	Bacteroides	High-fat diet
22	Prevotella	CD risk allele
63,98	Prevotella	Colitis susceptible mice
8,15,56–58	Prevotella	Fiber-rich diet
44	Bacteroides	NASH and ROS
15,57	Bacteroides	Protein & animal fat
99	Bacteroides	Fibers & fructans
6,25,28,100,101	Bacteroides	Low-grade inflammation, CRP and insulin resistance

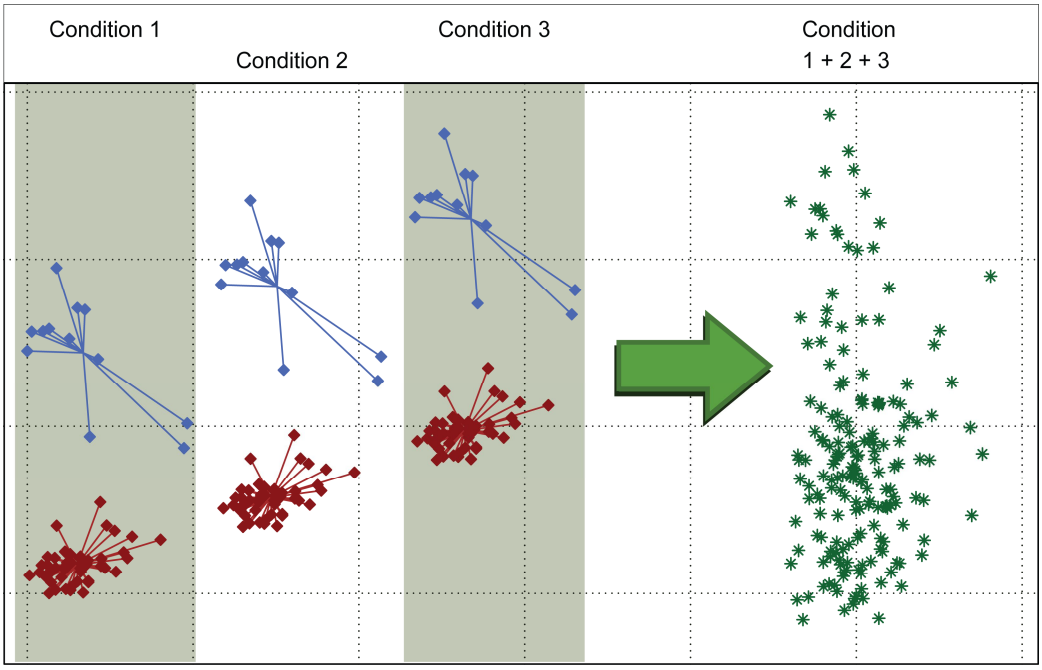
SUPPLEMENTARY FIGURES



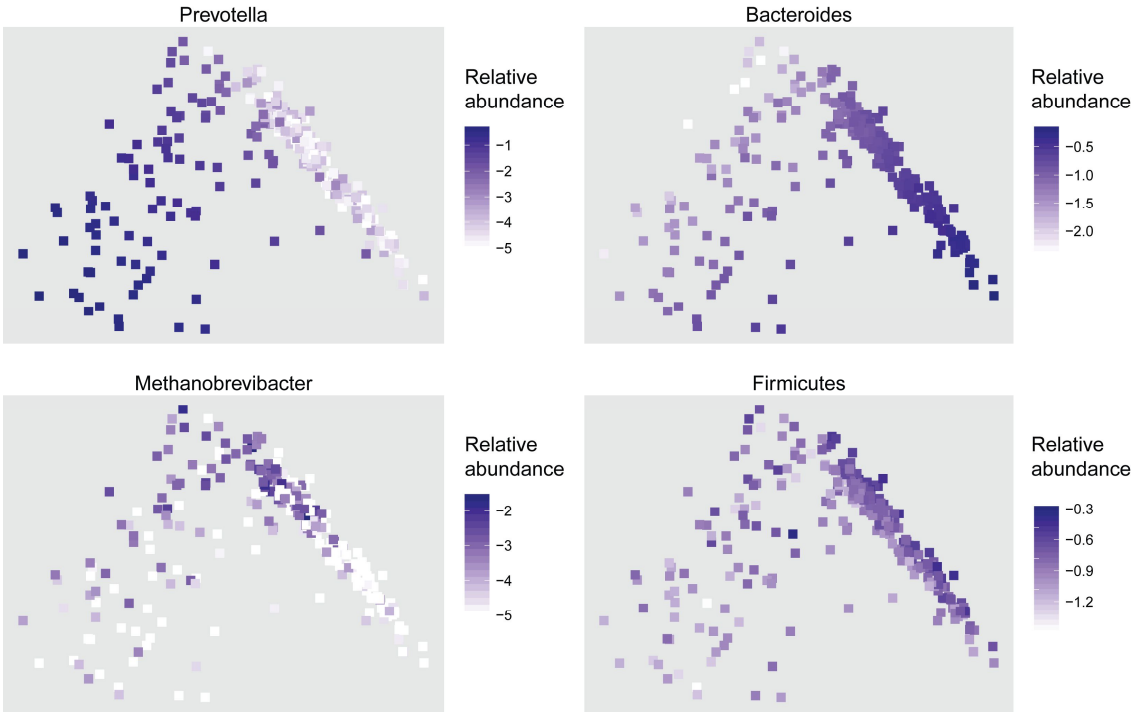
Supplementary Figure 1: Hierarchical structure of different clusterings using MetaHIT[6] samples. Each circle represents a cluster, as obtained by PAMk for 2 and 3 clusters and DMM for 4 clusters. The connecting lines show the number of samples that overlap between the cluster definitions. Overall, the different clusterings are highly associated, forming a hierarchical structure.



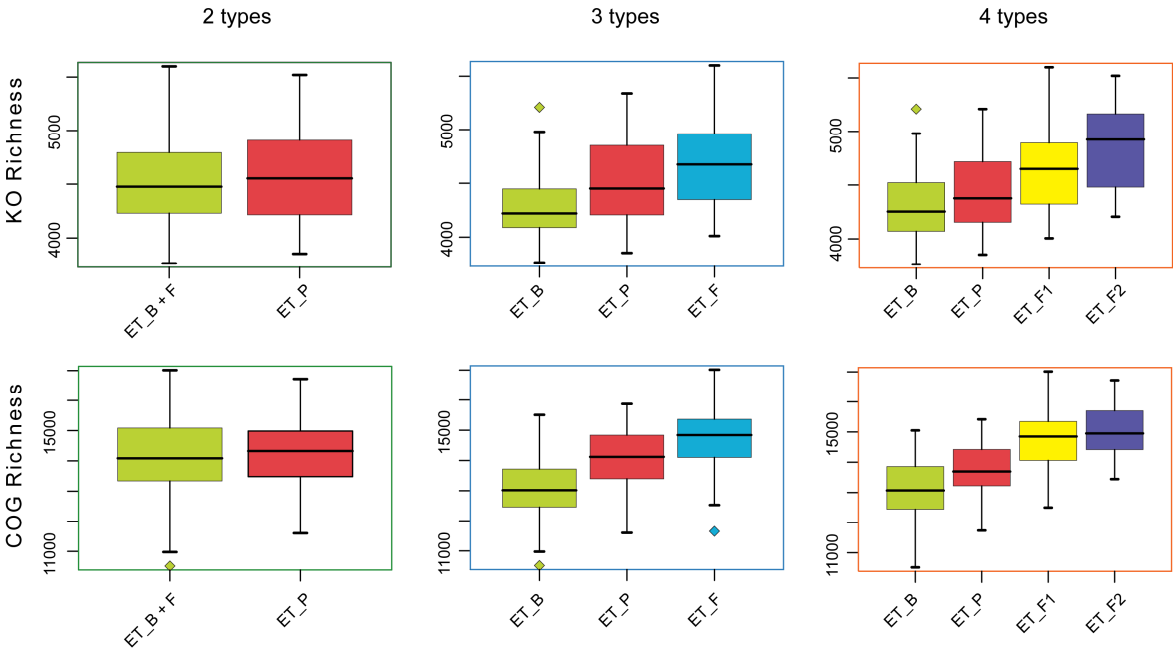
Supplementary Figure 2: Clustering of human body sites, based on their genus level abundance composition. Body-site separation as in Figure 1, using frequently used enterotype clustering methods. (a) Ordination of the HMP 16S rRNA (v35) dataset using four common inter-sample distance measures. (b) The optimal cluster number calculated within each distance measure using common clustering optimality measures. Body site separation was recovered by Jensen Shannon divergence (JSD) distance and weighted UniFrac.



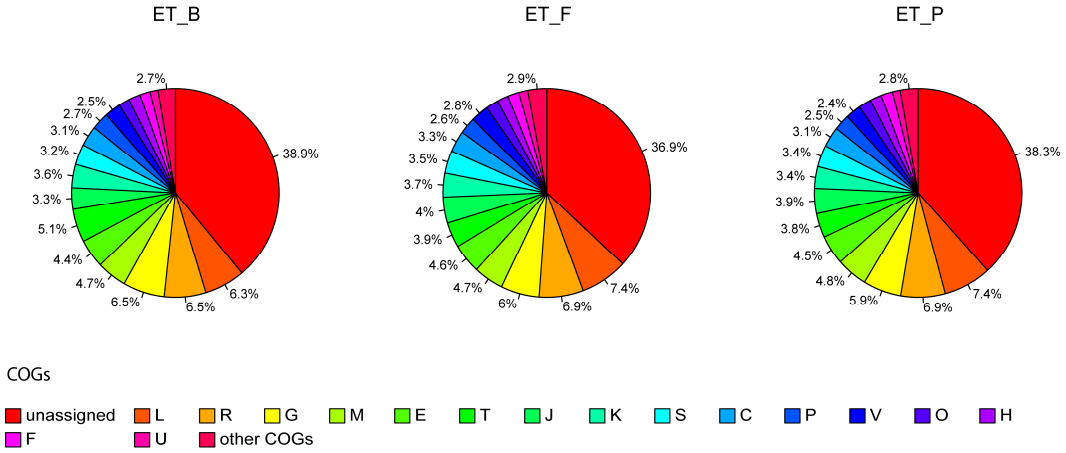
Supplementary Figure 3: Systematic shifts in the microbiota composition under different conditions can obscure clusters in the data. These confounders could be immune system or food related, but also technical biases, such as different DNA extraction methods. For the mouse microbiota [28], a discrete separation between ET F and ET B is possible. Due to the design of this study, more factors were controlled for than possible in human studies, like diet and environment. The figure illustrates how confounders may impact the observed landscape (not based on real data).



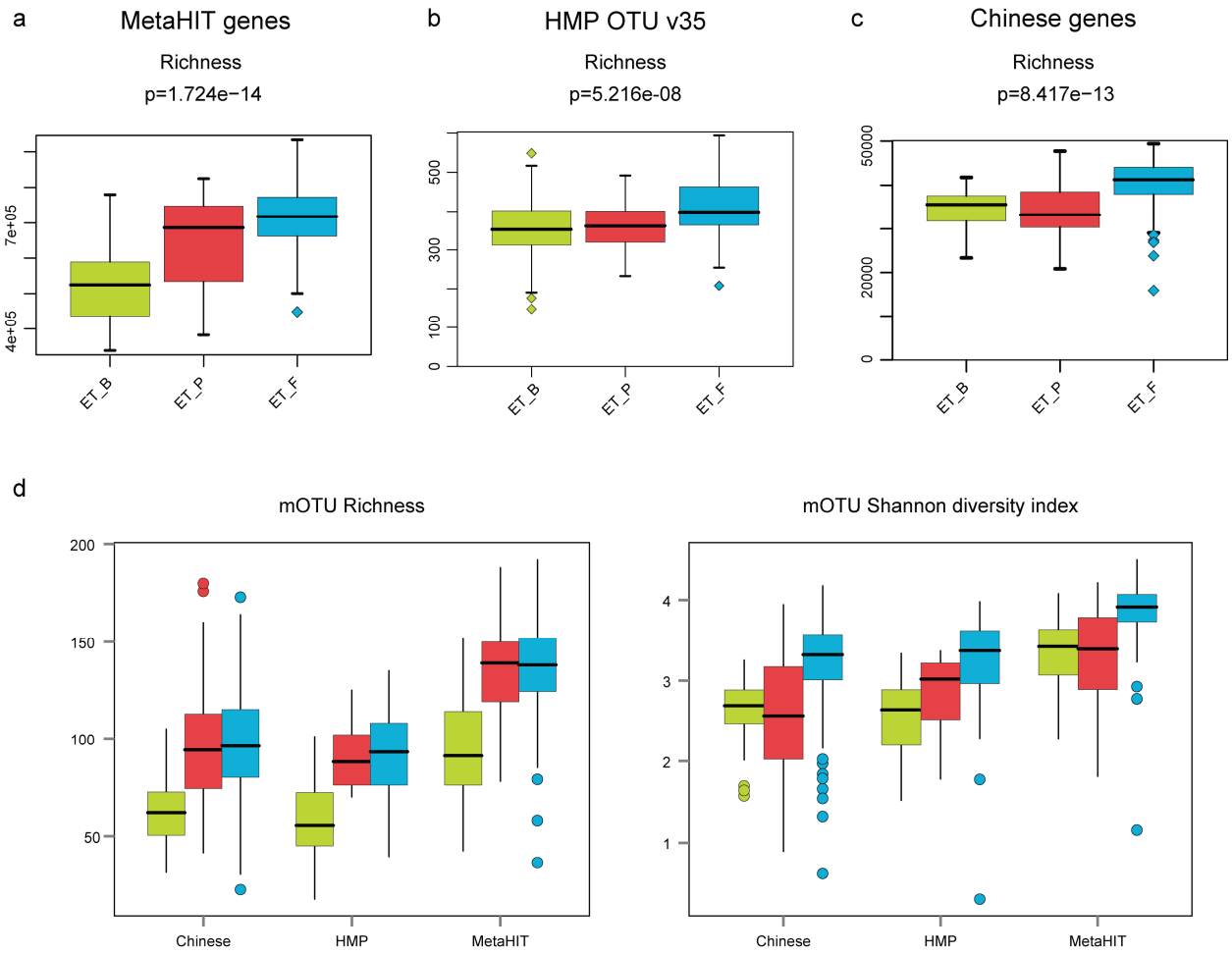
Supplementary Figure 4: Log10 transformed relative abundance of relevant genera, superimposed onto the PCoA ordination of the MetaHIT dataset[6], showing the bimodal distributions of *Prevotella* and *Methanobrevibacter* and the unimodal distribution of *Bacteroides* and *Firmicutes*, similar to Figure 3a.



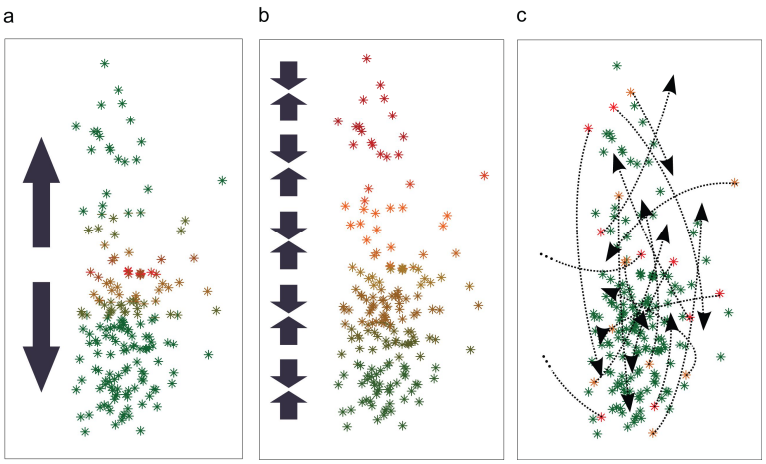
Supplementary Figure 1: Functional richness differs substantially between enterotypes defined on the MetaHIT dataset, as measured on COG [96] and KO [97] level data.



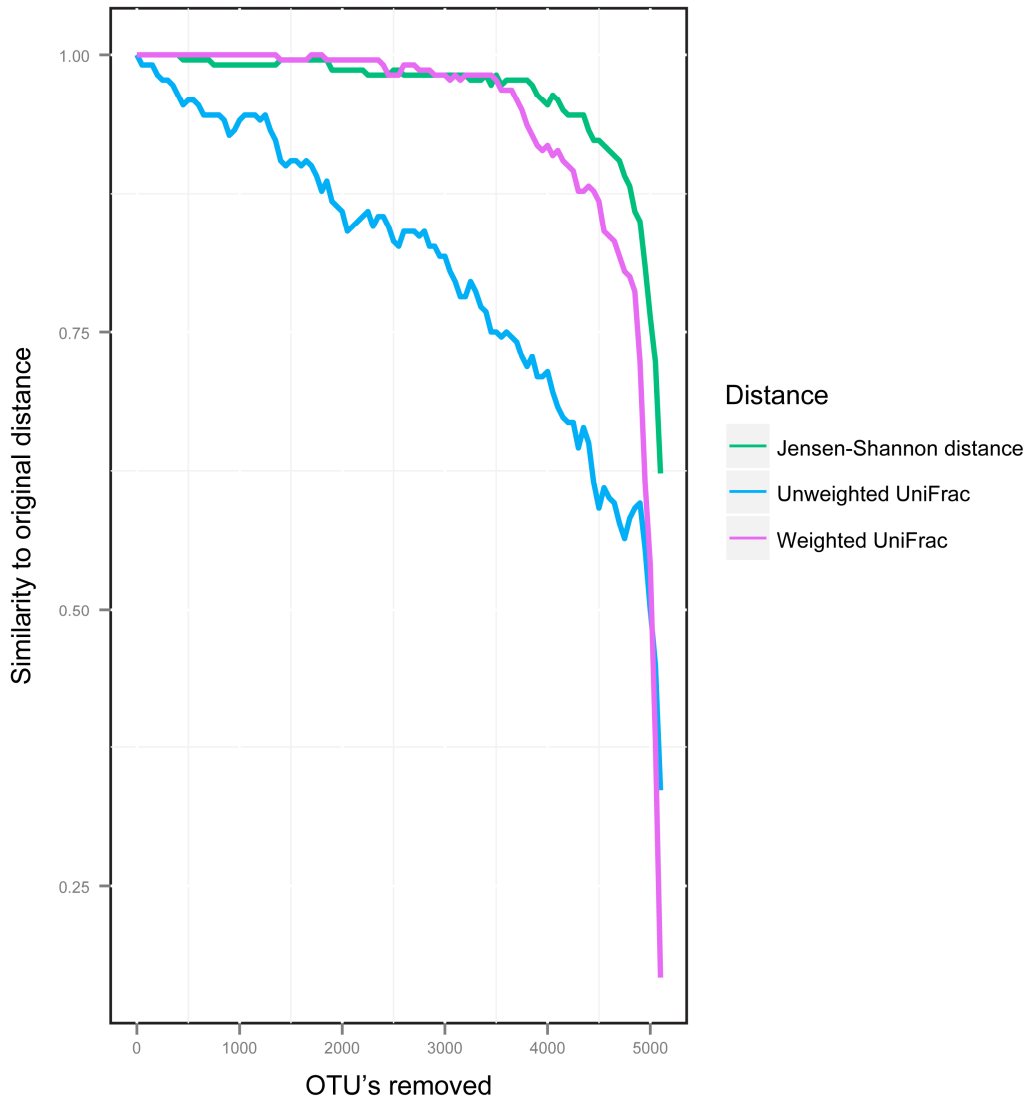
Supplementary Figure 6: Although 23/25 COG categories are significantly different between three enterotype (Suppl. Table 3), the overall composition remains relatively stable between enterotypes. This is probably due to consistently small effect size differences between enterotypes in their functional composition that is more stable than the taxonomic composition.



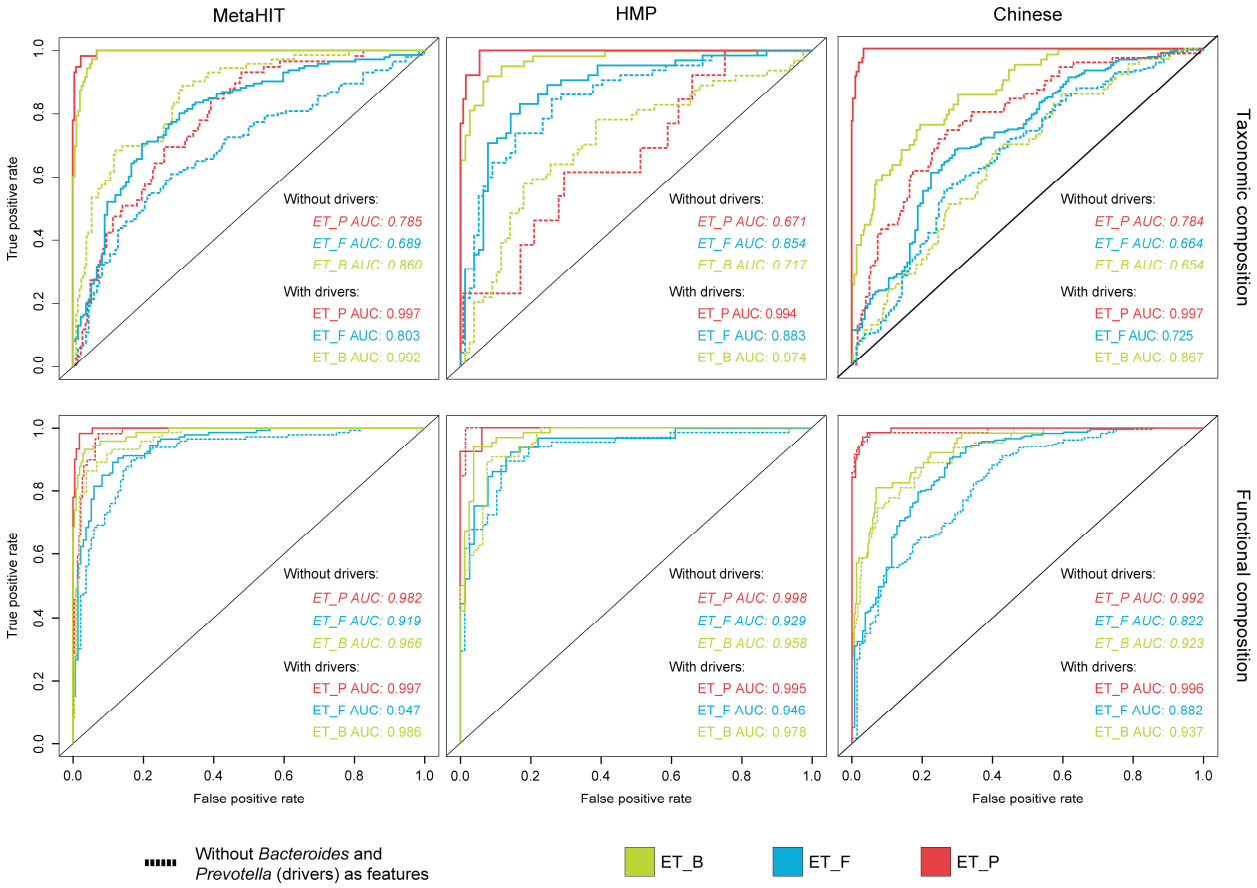
Supplementary Figure 7: ET richness differences calculated on (a) MetaHIT gene catalog [6] (b) HMP v35 OTUs [1] (c) Chinese samples [45]. Note that MetaHIT genes do encompass functional as well as taxonomic diversity. Ecological properties of the enterotypes, such as richness and diversity, derived using marker genes (97) recapitulate the same trends within each of the datasets (d).



Supplementary Figure 8: Three hypotheses that could account for the observed enterotype gradient in temporal data: (a) the Bacteroides/Firmicutes gradient could be driven by 2 global optimal states or (b) temporal samples are auto correlating to an optimum specified by each microbiota individually, but not driven by global attractors. Last, (c) the Bacteroides/Firmicutes gradient does not reflect an ecological pattern and is subject to strong temporal changes, but this is unlikely to apply to the majority of samples.



Supplementary Figure 9: Sensitivity of distance measures to low abundant features. OTU's are sorted from low to high abundant and removed one by one. Each resulting neighbor constellation is compared to the original one (computed using all the OTU's) showing the weight features get in the final nearest neighbor determination. JSD and weighted Unifrac are not influenced by low abundant features, while the unweighted Unifrac approach is.



Supplementary Figure 10: Robust classification of enterotypes across studies. A three enterotype model classifier was trained on genus level abundances of the MetaHIT[6] samples. This model also recovers enterotypes in the HMP and Chinese metagenomic datasets. The receiver operating characteristic area under the curve (ROC-AUC) for classifier performance on the MetaHIT (internal cross-validation), Chinese and HMP datasets are shown, with the clustering ground truth being estimated using unsupervised clustering of samples in the respective dataset. Although there are known batch effects between these datasets [94], the properties of the enterotypes are comparable and recoverable. Furthermore, the classification is possible even when removing the genera *Bacteroides* and *Prevotella* from the feature set ("Without drivers"). The classification of enterotypes on functional (COG) abundances in almost all cases outperforms the taxonomic classification across all three datasets. In the functional context, "Without drivers" represents a dataset where COGs that contain a gene from either the *Bacteroides* or *Prevotella* genus were removed prior to training and subsequent classification.

SUPPLEMENTARY REFERENCES

1. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).

2. Luckey, T. D. Introduction to intestinal microecology. *Am. J. Clin. Nutr.* **25,** 1292–1294 (1972).

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

4. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–4 (2009).

5. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).

6. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–6 (2013).

7. Jeffery, I. B., Claesson, M. J., O'Toole, P. W. & Shanahan, F. Categorization of the gut microbiota: enterotypes or gradients? *Nat. Rev. Microbiol.* **10,** 591–592 (2012).

8. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–7 (2012).

9. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M. & de Vos, W. M. Tipping elements in the human intestinal ecosystem. *Nat. Commun.* **5,** 4344 (2014).

10. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–80 (2011).

11. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl ,** 4680–4687 (2011).

12. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9,** e1002863 (2013).

13. Ding, T. & Schloss, P. D. Dynamics and associations of microbial community types across the human body. *Nature* **509,** 357–60 (2014).

14. Zhou, Y. *et al.* Exploration of bacterial community classes in major human habitats. *Genome Biol.* **15,** R66 (2014).

15. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–8 (2011).

16. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS One* **7,** e30126 (2012).

17. Dethlefsen, L., Eckburg, P. B., Bik, E. M. & Relman, D. a. Assembly of the human intestinal microbiota. *Trends Ecol. Evol.* **21,** 517–23 (2006).

18. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).

19. Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* **16,** 82 (2015).

20. Morotomi, N. *et al.* Evaluation of Intestinal Microbiotas of Healthy Japanese Adults and Effect of Antibiotics Using the 16S Ribosomal RNA Gene Based Clone Library Method. *Biol. Pharm. Bull.* **34,** 1011–20 (2011).

21. Karlsson, F. H., Nookaew, I. & Nielsen, J. Metagenomic Data Utilization and Analysis (MEDUSA) and Construction of a Global Gut Microbial Gene Catalogue. *PLoS Comput. Biol.* **10,** e1003706 (2014).

22. Quince, C. *et al.* The impact of Crohn's disease genes on healthy human gut microbiota: a pilot study. *Gut* **0,** 2012–2014 (2013).

23. Roager, H. M., Licht, T. R., Poulsen, S. K., Larsen, T. M. & Bahl, M. I. Microbial enterotypes, inferred by the prevotella-to-bacteroides ratio, remained stable during a 6-month randomized controlled diet intervention with the new nordic diet. *Appl. Environ. Microbiol.* **80,** 1142–9 (2014).

24. Zupancic, M. L. *et al.* Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. *PLoS One* **7,** e43052 (2012).

25. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–84 (2012).

26. Moeller, A. H. *et al.* Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat. Commun.* **3,** 1179 (2012).

27. Moeller, A. H. *et al.* Stability of the gorilla microbiome despite simian immunodeficiency virus infection. *Mol. Ecol.* **24,** 690–7 (2015).

28. Hildebrand, F. *et al.* Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* **14,** R4 (2013).

29. Wang, J. *et al.* Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. *Pnas* **111,** E2703–10 (2014).

30. Mach, N. *et al.* Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environ. Microbiol. Rep.* (2015). doi:10.1111/1758-2229.12285

31. Scheffer, M. & Carpenter, S. R. Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends Ecol. Evol.* **18,** 648–656 (2003).

32. Huse, S. M., Ye, Y., Zhou, Y. & Fodor, A. A. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS One* **7,** e34242 (2012).

33. Bäckhed, F. *et al.* Defining a Healthy Human Gut Microbiome: Current Concepts, Future Directions, and Clinical Applications. *Cell Host Microbe* **12,** 611–622 (2012).

34. Sankaran, M. *et al.* Determinants of woody cover in African savannas. *Nature* **438,** 846–9 (2005).

35. Staver, a C., Archibald, S. & Levin, S. Tree cover in sub-Saharan Africa: rainfall and fire constrain forest and savanna as alternative stable states. *Ecology* **92,** 1063–72 (2011).

36. Purushe, J. *et al.* Comparative genome analysis of Prevotella ruminicola and Prevotella bryantii: insights into their environmental niche. *Microb. Ecol.* **60,** 721–9 (2010).

37. Kaoutari, A. El, Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat. Rev. Microbiol.* **11,** 497–504 (2013).

38. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).

39. Elmqvist, T. *et al.* Response diversity, ecosystem change, and resilience. *Front. Ecol. Environ.* **1,** 488–494 (2003).

40. Virgin, H. W. & Todd, J. A. Metagenomics and Personalized Medicine. *Cell* **147,** 44–56 (2011).

41. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55,** 205–11 (2006).

42. Carroll, I. M. *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am. J. Physiol. Gastrointest. Liver Physiol.* **301,** G799–807 (2011).

43. van Nood, E. *et al.* Duodenal infusion of donor feces for recurrent Clostridium difficile. *N. Engl. J. Med.* **368,** 407–15 (2013).

44. Zhu, L. *et al.* Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology* **57,** 601–9 (2013).

45. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).

46. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8,** e1002606 (2012).

47. Gibson, T. E., Bashan, A., Cao, H.-T., Weiss, S. T. & Liu, Y.-Y. On the Origins and Control of Community Types in the Human Microbiome. *arXiv* (2015).

48. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).

49. Mardanov, A. V *et al.* Metagenomic Analysis of the Dynamic Changes in the Gut Microbiome of the Participants of the MARS-500 Experiment, Simulating Long Term Space Flight. *Acta Naturae* **5,** 116–125 (2013).

50. David, L. a *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).

51. Rajilić-Stojanović, M., Heilig, H. G. H. J., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15,** 1146–1159 (2012).

52. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4554–61 (2011).

53. Cho, I. *et al.* Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* (2012). doi:10.1038/nature11400

54. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16,** 73 (2015).

55. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517,** 205–8 (2015).

56. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 14691–6 (2010).

57. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).

58. Ou, J. *et al.* Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. *Am. J. Clin. Nutr.* **98,** 111–20 (2013).

59. Nakayama, J. *et al.* Diversity in gut bacterial community of school-age children in Asia. *Sci. Rep.* **5,** 8397 (2015).

60. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1,** 6ra14 (2009).

61. Hildebrandt, M. A. *et al.* High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology* **137,** 1716–24.e1–2 (2009).

62. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Human gut microbes associated with obesity. *Nature* **444,** 1022–3 (2006).

63. Elinav, E. *et al.* NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145,** 745–57 (2011).

64. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159,** 789–799 (2014).

65. Kleessen, B., Kroesen, A. J., Buhr, H. J. & Blaut, M. Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. *Scand. J. Gastroenterol.* **37,** 1034–41 (2002).

66. Lucke, K., Miehlke, S., Jacobs, E. & Schuppler, M. Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. *J. Med. Microbiol.* **55,** 617–24 (2006).

67. Heimesaat, M. M. *et al.* MyD88/TLR9 mediated immunopathology and gut microbiota dynamics in a novel murine model of intestinal graft-versus-host disease. *Gut* **59,** 1079–87 (2010).

68. De Palma, G. *et al.* Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. *BMC Microbiol.* **10,** 63 (2010).

69. Sobhani, I. *et al.* Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One* **6,** e16393 (2011).

70. Jernberg, C., Löfmark, S., Edlund, C. & Jansson, J. K. Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology* **156,** 3216–23 (2010).

71. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2,** e01202–e01202 (2013).

72. Noguera-Julian, M. *et al.* Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* 1–12 (2016). doi:10.1016/j.ebiom.2016.01.032

73. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5,** e9085 (2010).

74. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).

75. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* **3,** 1245 (2012).

76. Rajilić-Stojanović, M. *et al.* Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. *Gastroenterology* **141,** 1792–801 (2011).

77. Knights, D. *et al.* Rethinking 'Enterotypes'. *Cell Host Microbe* **16,** 433–437 (2014).

78. Flegal, K. M., Kit, B. K., Orpana, H. & Graubard, B. I. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *JAMA* **309,** 71–82 (2013).

79. Fuentes, S. *et al.* Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection. *ISME J.* 1–13 (2014). doi:10.1038/ismej.2014.13

80. Schubert, A. M. *et al.* Microbiome data distinguish patients with Clostridium difficile infection and non-C. difficile-associated diarrhea from healthy controls. *MBio* **5,** e01021–14 (2014).

81. Wu, G. D. *et al.* Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. *BMC Microbiol.* **10,** 206 (2010).

82. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–9 (2013).

83. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One* **7,** e47656 (2012).

84. Hennig, C. fpc: Flexible procedures for clustering. (2014).

85. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* **3,** 1–27 (1974).

86. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).

87. Tibshirani, R. & Walther, G. Cluster Validation by Prediction Strength. *J. Comput. Graph. Stat.* **14,** 511–528 (2005).

88. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).

89. Driver, H. E. & Kroeber, A. L. Quantitative expression of cultural relationships. *Univ. Caltfornia Publ. Am. Archeol. Ethnol.* **31,** 211–256 (1932).

90. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–35 (2005).

91. Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B* **58,** 267–288 (1994).

92. Zhou, J. *et al.* Functional molecular ecological networks. *MBio* **1,** (2010).

93. Zhang, J. *et al.* Mongolians core gut microbiota and its correlation with seasonal dietary changes. *Sci. Rep.* **4,** 5001 (2014).

94. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99–103 (2013).

95. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science (80-.).* **352,** 560–564 (2016).

96. Tatusov, R. L., Koonin, E. V & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–7 (1997).

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

98. Brinkman, B. M. *et al.* Caspase deficiency alters the murine gut microbiome. *Cell Death Dis.* **2,** e220 (2011).

99. Sonnenburg, E. D. *et al.* Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141,** 1241–52 (2010).

100. Rath, H. & Herfarth, H. Normal luminal bacteria, especially Bacteroides species, mediate chronic colitis, gastritis, and arthritis in HLA-B27/human beta2 microglobulin transgenic rats. *J. Clin. …* 945–953 (1996).

101. Bloom, S. M. *et al.* Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell Host Microbe* **9,** 390–403 (2011).

PAPER 5

DELINEATION AND CHARACTERIZATION OF SUBSPECIES IN THE GLOBAL HUMAN GUT MICROBIOME

Authors: Paul I. Costea[1], Georg Zeller[1], Luis Pedro Coelho[1], Shinichi Sunagawa[1], Falk Hildebrand[1], Jamie Huerta Cepas[1], Kristoffer Forslund[1], Robin Muench[1], Almagul Kushugulova[2], Peer Bork[1,*]

* - Corresponding author

ABSTRACT

The concept of prokaryotic species has been a topic of discussion in microbiology for over a century. However, with the advance of sequencing technologies, operational definitions of species have become more widely used and have proven crucial to our continued understanding of microbial ecology. Here, for the first time, we tested for substructure within species in large, natural habitats, without isolation and cultivation biases. We surveyed the variation landscape of 73 prevalent microbial species in 2144 human fecal metagenomes, and show that the majority, accounting for 70% of the known abundance, can be further stratified into subspecies. Individuals are usually dominated by only one such subspecies (per species), as expected from ecological theory, although we also find co-occurrence in rare cases. The geographical distribution of these subspecies reveals phylogenetic differences in dispersal patterns, ranging from globally distributed Bacteroidetes subspecies to several geographically restricted Firmicutes subspecies. To show their functional significance, we perform reference independent pan-genomic analysis, determining and comparing the core and accessory genomes based on co-variation in metagenomes. With this approach we find, for example, that some *Eubacterium rectale* subspecies specifically harbor a flagellum operon and associate with lower community diversity, higher host BMI and higher blood insulin levels.

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

[2] Center for Life Sciences, Nazarbayev University, Astana, Kazakhstan

INTRODUCTION

While there is still long-standing debate on whether there is a coherent species concept in the prokaryotic world[1–5], modern molecular technologies offer multiple operational definitions of species that are being successfully applied to understanding microbial ecology. They are based on molecular considerations, some genome-wide, such as DNA-DNA hybridization[6] (DDH) or pair-wise average nucleotide identity[4] (ANI) while others are restricted to a specific marker gene (i.e. 16S rRNA gene, or a variable region therein) or multiple marker genes[7] (i.e. multilocus sequence typing). Ideally, these operational definitions would identify a measurable unit, within which there is little to no phenotypic difference. More generally though, they define a space of low genotypic variation, which is sometimes only poorly predictive of phenotypic outcomes[8]. As phenotypic differences can be observed within operationally defined species, the concept of subspecies, a taxonomic rank subordinate to species, has been discussed in the literature as early as the 1950s[9], in the case of *Bacillus cereus*, for which there exist virulent and non-virulent subspecies. Sometimes, differently adapted members of the same bacterial species were called "ecotypes". For example in the case of the ocean-dwelling Procholorococus[10], multiple subspecies exist that are specialized for different temperatures and light conditions, but the ecotype concept has been advocated in a range of habitats, indicating this to be a global property of prokaryotes. Pan-genomic analysis of a small number of species has sometimes confirmed the existence of population structure within species, but such datasets can be biased by differences in the ability to isolate and culture individual subspecies[11].

In 1997, Palys proposed a framework for determining and classifying ecological diversity based on clusters of DNA sequence data similarity[12], which has not been applied as a large enough sample is lacking. In order to quantify the occurrence of subspecies in this framework in a natural setting, we here use gut metagenomes from 2144 fecal samples of a wide geographic range, encompassing 9 countries from three continents. We demonstrate that substructure within species is the rule rather than the exception, analyze global subspecies dispersal patterns, show that subspecies are mutually exclusive in most individuals and illustrate the utility of the concept by associating particular ones to specific genes and host phenotypes, with implications for disease. Moreover, these finding support the ecotype view of bacterial evolution, suggesting the existence of cohesive clusters of strains with the same fundamental dynamic properties, bound together in their evolutionary fate[13].

SUBSPECIES DELINEATION

Building on recently developed methodology to study metagenomic samples at strain level[14], we here explore genomic variation to identify population structure within species of the human gut microbiome, the ecosystem that is most extensively sampled by metagenomics. To this end, we combined DNA shotgun data from 2144 deeply sequences human stool metagenomes (X Gb +- Y) from 9 countries, spanning 3 continents by including published[15–19] as well as newly generated data (see Methods). We profiled and assessed the variation of single nucleotide polymorphisms (SNP) of 73 microbial species with sufficient read coverage and prevalence in our set of gut metagenomes (Methods and Supplementary Table 1). These accounted for an average of 95.5% (SD=6%) of the sequencing reads per sample that can be mapped to reference genomes, covering 50% of the total reads (Figure 1c, see Methods).

The existence of subspecies was assessed by computing a per-species pairwise distance matrix of SNP profiles between all samples with enough coverage over the respective representative

genome. For the latter we used a Manhattan distance between non-reference allele frequencies (see Methods). This distance matrix is then the basis for assessing the existence of subspecies clusters, by using a prediction strength measure[20]. We imposed a very stringent cutoff and operationally defined the number of subspecies as being the highest number of clusters with prediction strength above 0.8 (see Methods). This way we identified subspecies in 39 of the 73 species, accounting for 70% (SD=3%) of the reads mapping to the representative genomes. Thus, our approach detects subspecies in the majority of highly abundant members of the human gut microbiome.

Once subspecies had been identified, we determined a set of unique alleles which unambiguously identify each one; the number of such "genotyping" positions ranges from 100 to 10000. Using only these genotyping positions, we could assign subspecies in a substantially expanded set of samples. This greatly expands our power to investigate the relation between subspecies and host properties and allows subspecies level abundances estimations in novel samples.

To assess which of the subspecies have a sequenced reference genome, we mapped all NCBI genomic sequences into our variation landscape and assigned them to the corresponding subspecies genotypes. Out of the 39 bacterial species with clear substructure, only 12 are completely covered, while for the remaining 27 at least one subspecies is without a representative reference genome (Figure 1). In some cases, a reference sequence is lacking for the subspecies most commonly found in the sampled population thus limiting the applicability of classical reference-based pan-genomic analysis to major gut microbial species.

In order to identify gene content differences between subspecies, we developed a new method based on the co-abundance concept, previously used to pool genes into metagenomic species independent of reference genomes[21]. Using the allele frequencies at the genotyping positions of each subspecies, we could identify genes that consistently correlated with these (i.e. subspecies accessory genes), while also being able to determine the set of genes that highly correlated with the species abundance, which we term core genes (see Methods). For these pan-genome reconstructions we used a human gut reference gene catalog[22], consisting of nearly 10 million genes to which the vast majority of gut metagenomics reads could be mapped. To evaluate the accuracy of this approach we used *Bacteroides vulgatus/dorei,* for which multiple reference genomes are available. For a global assessment of the accuracy of our pan-genomic reconstruction we considered the genus level annotation of the gene catalog (the highest taxonomic resolution for which confident assignments have been made) and found that the median accuracy for the core is 0.99 and 0.96 for the accessory components. Comparing the number of genes that are unique to a subspecies with the number of genotyping SNPs we find both measures to correlate, suggesting proportionality between acquired mutations and phenotypic distance (Figure 1b).

To quantify these functional differences and their potential consequences, we tested for enrichment of functional annotations between the accessory gene content of conspecific subspecies. Using homology-based KEGG annotation and, we found at least one significant ($P < 0.05$, corrected for multiple testing within each species) pathway enrichment between subspecies in 11 of the 39 species (see Methods). We note that on average 70% of the subspecies specific genes did not have an annotation at the KEGG pathway level, highlighting the need for better functional annotations to understand microbial diversity and explaining why more significant differences were not found. Generally, significant differences were found in nitrogen and carbon metabolism, and in chemotaxis-related pathways. Of note, we find significant differences in invasion potential in two of the four *E. coli* subspecies and show, using strain pathogenicity information from PATRIC[23], a highly significant enrichment in disease annotations. Metagenomic data from a recent outbreak[24] indicate that the offending strain

comes from one of the two subspecies enriched in invasion associated potential (Supplementary Figure). This highlights the utility of the subspecies concept as a way of stratifying bacterial populations.

SUB-SPECIES DISPERSAL

As we analyzed metagenomics samples from nine countries and three continents (Austria, China, Denmark, France, Germany, Kazakhstan, Spain, Sweden, and USA), we could assess the global geographic range of each subspecies. While for many there was no clear geographical signal, some do show striking regional enrichments (Figure 1). For example, for *Eubacterium rectale*, one subspecies (*E. rectalis* spp. MG3) was found exclusively in Chinese samples, while the other two were found in all other countries, including Kazakhstan. To validate our finding, we profiled an additional set of 300 samples from a Chinese cohort, not included in our initial dataset and confirmed that all individuals exclusively harbored *E. rectale* spp. MG3 (and none of the other subspecies of *E. rectale*). In contrast, when investigating 13 individuals from US American studies that are reported to be of Asian descent we found none of them to harbor MG3. This, together with the fact that other samples Asian samples (from Kazakhstan) do not harbor the Chinese subspecies, suggests very low dispersal This may also explain why associations between *E. rectale* and host physiology have often proven to be unstable when testing in different cohorts, as these subspecies could not be profiled previously.

For most other species, geo-stratification appeared less extreme, with certain subspecies observed predominantly, but not exclusively in certain countries. This might reflect an adaptation to specific environmental factors, many of which are more prevalent in some geographic regions, but not exclusively found there. Conversely, the observed structure could be the result of drift and differential dispersal potential among gut microbial sub-species (Figure 1). Overall, the strongest geographical restrictions across subspecies were observed in the Chinese samples, followed by the ones from Kazakhstan (Figure 1), while European and American samples appeared more similar in their gut subspecies compositions. When comparing geographic ranges across bacterial taxa, members of the Firmicutes phylum show significantly more geographic restrictions compared to other phyla (one-sided Wilcoxon test p-value = 0.0004). *Escherichia coli*, for example, shows an almost uniform geographic distribution, indicating pervasive dispersal, in line with previous observations[26].

SUB-SPECIES DOMINANCE AND PERSISTENCE IN INDIVIDUALS

Having considered global trends in subspecies distribution and their potential dispersal limitation, we further interrogated their properties when confined to the gut of a single individual. Here, overall strain-level stability has been shown[14], allowing for an individual to be confidently identified at a later time point based on a SNP profile over all species. However, strain variation and complex co-occurrence patterns have been reported[27] when an individual undergoes fecal microbial transplant, indicating that the strain level stability can be disturbed.

In order to study the population structure of conspecific subspecies and their temporal stability, we recorded the allele frequency of each subspecies in each sample, for 124 individuals for which we have longitudinal data (Fig.2). For all 39 stratifiable species, we saw a clear dominance of one conspecific subspecies per individual, in line with ecological theory that predicts that among closely related taxa (species) the most adaptable one outcompetes the others in the same ecological niche. Only, in very few cases and only for some of the studied species were more than one conspecific subspecies observed in the same individual. In these rare cases we noted considerable fluctuation in the relative abundance of these sub-species over time, suggesting

cohabitation/co-occurrence to be an unstable constellation (Figure 2). However, in general we observed sub-species composition to be stable exchanges of one subspecies for another to be very uncommon when considering multiple time-points for hundreds of individuals up to three years apart. A striking exception to this was seen in an individual who changed microbial composition dramatically and persistently after antibiotics treatment[28] with a corresponding switch of subspecies (Figure2). Such rare exceptions aside, the combinatorics of subspecies and their apparent persistence over time provides another layer contributing to gut microbial individuality.

SUBSPECIES ASSOCIATIONS WITH HOST PHENOTYPES

Given persistence and dominance of subspecies within individuals over long periods of time, we asked whether their presence was associated with community and host properties. Accounting for differences between studies and non-random distribution of samples across subspecies, our statistical analysis identified eleven associations with microbial community diversity, two with host BMI, three with gender and one with type II diabetes. The latter appears to be a strongly protective subspecies of *B. coprocola*, which is highly significant in a Chinese type II diabetes cohort[17] but does not show the same effect in the Swedish[16] or MetaHIT[29] diabetic cohorts.

The strongest association with host BMI is presented by *E. rectale,* which is split into 3 subspecies, one of which (*spp.* MG3) is almost completely restricted to samples from China (Figure 3). The remaining human populations are colonized by one of the other two subspecies, with some samples containing a combination of the two. We note that the only reference genome available is representative of *E. rectale spp.* MG2, meaning no reference-dependent method can be applied to this case. Using our reference-independent reconstruction of each subspecies' gene content, as well as analysis of the coverage over the reference genome (Figure 3), we can relate functional differences to BMI. Specifically, *E. rectale spp.* MG1 is missing at least 18 genes related to bacterial chemotaxis and flagellar assembly. These genes are necessary for bacterial motility and also represent an important signal for host immune activation[30]. The TLR5 receptor in human epithelial cells (in this case colonocytes) recognizes flagellin and induces a downstream cascade which results in initiation of pro-inflammatory pathways and secretion of IL8. Flagellin from *E. rectale* has been directly shown to induce inflammation in human colonocytes[30]. Such low-grade inflammation has been repeatedly linked to obesity, increased insulin resistance and diabetes[31]. Using the MetaHIT[29] and Swedish[16] type 2 diabetes cohorts, we showed that BMI and insulin resistance are significantly higher in individuals who predominantly harbour the flagellum-carrying sub-species. Furthermore, their microbiomes appeared to have lower community diversity (Shannon species index), in line with previous observations of a negative correlation between community diversity and host BMI. Similar observations can be made for *Eubacterium eligens,* which splits into four subspecies. At least 8 genes related to bacterial chemotaxis and flagellar assembly are specific only to *E. eligens spp.* MG3. Only few samples from the MetaHIT and Swedish T2D studies contain this subspecies, which explains why none of the observed phenotypic differences are significant (Figure 3.b). However, all of the group median differences are in the same direction as for *E. rectale*.

We note that there is no association between flagellum positive *E. eligens* and *E. rectale*, indicating that both flagella gene containing subspecies from different species have independent effects. Many genes have been shown to be transferred across lineages and mechanisms like conjugation also exist for transfer within a species. Yet, we clearly observe restrictions as judged from this diverse cohort of 1800 different individuals implying clear transmission limitations of genes that are functionally relevant for the host. The existence of such clear-cut functional difference between subspecies has implication for disease treatment as the respective

"detrimental" subspecies could be replaced or outcompeted by subspecies without the respective gene cassettes.

DISCUSSION

We have developed a classification of strain populations into prevalent subspecies of the human gut and investigated their geographic, ecological, and functional properties illustrating that this level of resolution can be relevant to unravelling specific interactions between microbiota and host physiology. As most of the abundant species in the human gut can be split into higher-resolution clusters of genomic variation, this might also provide a further handle towards more personalised diagnosis and treatment strategies. Here we have restricted our analysis to the most common and clearly identifiable subspecies (using stringent cut-offs), likely underestimating their true number. We further hypothesize the differing functional potential of conspecific subspecies may explain some of the unclear associations between taxonomic composition and host phenotypes reported at species, genus and higher taxonomic levels.

Our approach identified subspecies commonly found in human gut microbiomes that are typically distinguished from conspecific subspecies. While this approach identifies population structure below the species level, considerable allelic variation can still be observed within a sub-species. However, we consider this lower resolution to be a strength, as it enables powerful comparisons across samples. Ultimate resolution, entailing a distinction between two genomes when just one SNP is present, would make comparison impossible, as no two strains are likely to be the same. Any other arbitrary definition of how many positions should be distinct for the determination of a different taxonomic unit would most likely not hold across species (as is the case for most 16S based identity cutoff definitions). Furthermore, we believe this level to be a natural and informative taxonomic unit to consider, because it arises from external evolutionary pressures which have shape separate bacterial populations. This implies a qualitative difference between sub-species and makes then the likely unit of selection.

REFERENCES

1. Rosselló-Mora, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25,** 39–67 (2001).

2. Doolittle, W. F. & Zhaxybayeva, O. On the origin of prokaryotic species. *Genome Res.* **19,** 744–56 (2009).

3. Gevers, D. *et al.* Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3,** 733–739 (2005).

4. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2567–72 (2005).

5. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6,** 431 (2008).

6. Stackebrandt, E. *et al.* Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* **52,** 1043–7 (2002).

7. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10,** 881–4 (2013).

8. Koeppel, A. F. & Wu, M. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.* **41,** 5175–88 (2013).

9. DAVENPORT, R. & SMITH, C. Panophthalmitis due to an organism of the Bacillus subtilis group. *Br. J. Ophthalmol.* **36,** 389–92 (1952).

10. Johnson, Z. I. *et al.* Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311,** 1737–40 (2006).

11. Gordienko, E. N., Kazanov, M. D. & Gelfand, M. S. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J. Bacteriol.* **195,** 2786–92 (2013).

12. Palys, T., Nakamura, L. K. & Cohan, F. M. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int. J. Syst. Bacteriol.* **47,** 1145–56 (1997).

13. Cohan, F. M. Bacterial Species and Speciation. *Syst. Biol.* **50,** 513–524 (2001).

14. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).

15. Chatelier, E. Le *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).

16. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498,** 99–103 (2013).

17. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).

18. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).

19. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–14 (2012).

20. Tibshirani, R. & Walther, G. Cluster Validation by Prediction Strength. *J. Comput. Graph. Stat.* **14,** 511–528 (2005).

21. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).

22. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32,** 834–841 (2014).

23. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42,** D581–91 (2014).

24. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA* **309,** 1502–10 (2013).

25. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).

26. Petty, N. K. *et al.* Global dissemination of a multidrug resistant Escherichia coli clone. *Proc. Natl. Acad. Sci.* **111,** 5694–5699 (2014).

27. Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352,** 586–9 (2016).

28. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16,** 73 (2015).

29. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528,** 262–266 (2015).

30. Neville, B. A. *et al.* Pro-inflammatory flagellin proteins of prevalent motile commensal bacteria are variably abundant in the intestinal microbiome of elderly humans. *PLoS One* **8,** e68919 (2013).

31. Gregor, M. F. & Hotamisligil, G. S. Inflammatory Mechanisms in Obesity. *http://dx.doi.org/10.1146/annurev-immunol-031210-101322* (2011).

FIGURES

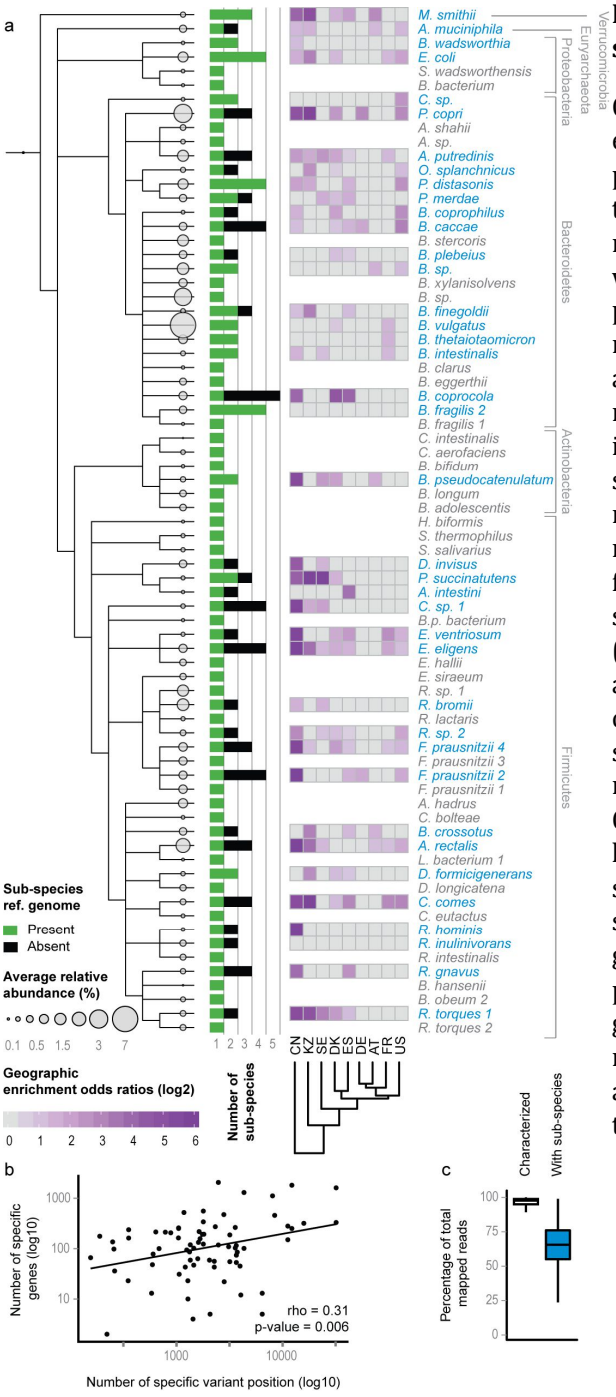


Figure 1. Overview of gut microbial subspecies.

Gut microbial species explored for the existence of sub-species show wide phylogenetic spread according to NCBI taxonomy (a) and include the main archaeal member of the human gut microbiome, as well as representatives of all abundant phyla. The circles on the cladogram represent each species' average abundance across 2144 gut metagenomes. Bars represent the number of sub-species identified in each, with '1' indicating no subspecies. The black portion of the bar represents sub-species for which no representative genome sequence is available from NCBI. Geographic preferences of subspecies are displayed as heatmap (showing significant enrichment per country as maximum log-odds ratio across conspecific subspecies), highlighting that subspecies with a restricted geographic range are predominantly found in the Chinese and Kazakh population sampled here. The number of variant positions specific to a sub-species (defining genotypes, see main text) and the number of accessory genes unique to this subspecies show a positive correlation (b). Our assessment of gut microbial subspecies included the vast majority of abundant community members and identified subspecies in the majority of these (c).

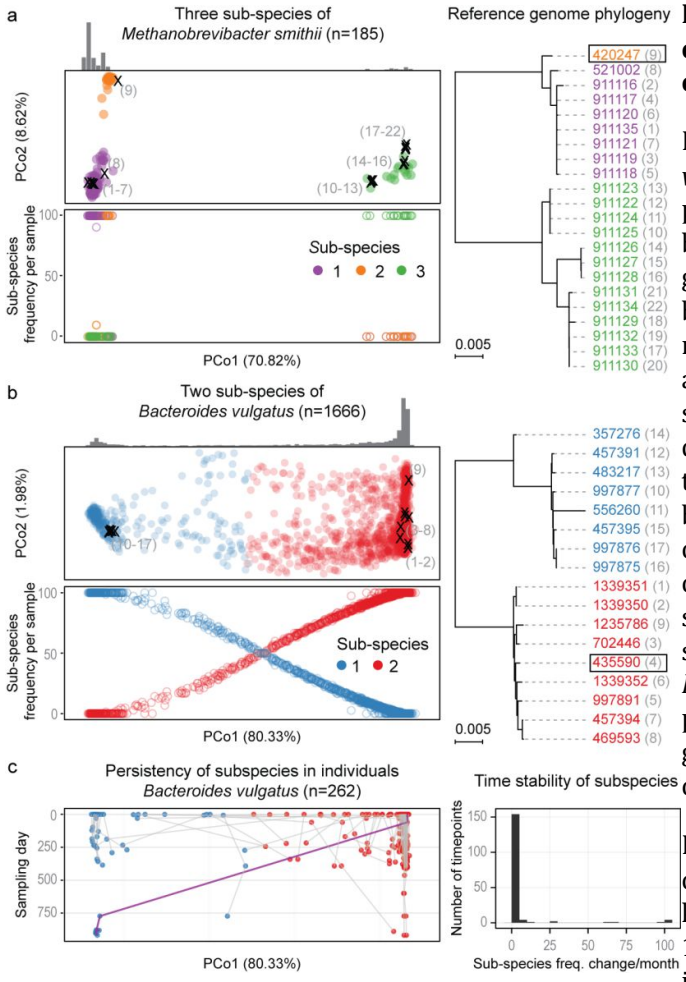


Figure 2. Sub-species co-occurrence and phylogenetic consistency.

For two species (*M. smithii* and *B. vulgatus*) we represent a PCoA projection of the between-sample and between samples and reference genomes distances (top plots in a and b). We chose these two examples, as many representative genomes are available for them, covering all of the subspecies. For each sample, we quantify the frequency of each one of the subspecies (bottom bplot in a and b) and show that for *M. smithii* only one sample has two sub-species co-occurring while all the other have a single dominating one. More samples show this co-occurrence pattern for *Bacteroides vulgatus*. We further present a phylogeny of the reference genomes, fully consistent with the observed clustering.

Importantly, sub-species are stable over time, with most individuals (x%) keeping their sub-species for up to 1000 days (c). The highlighted individual (purple line), who switches from one sub-species to the other, underwent at least one antibiotic intervention before the switch.

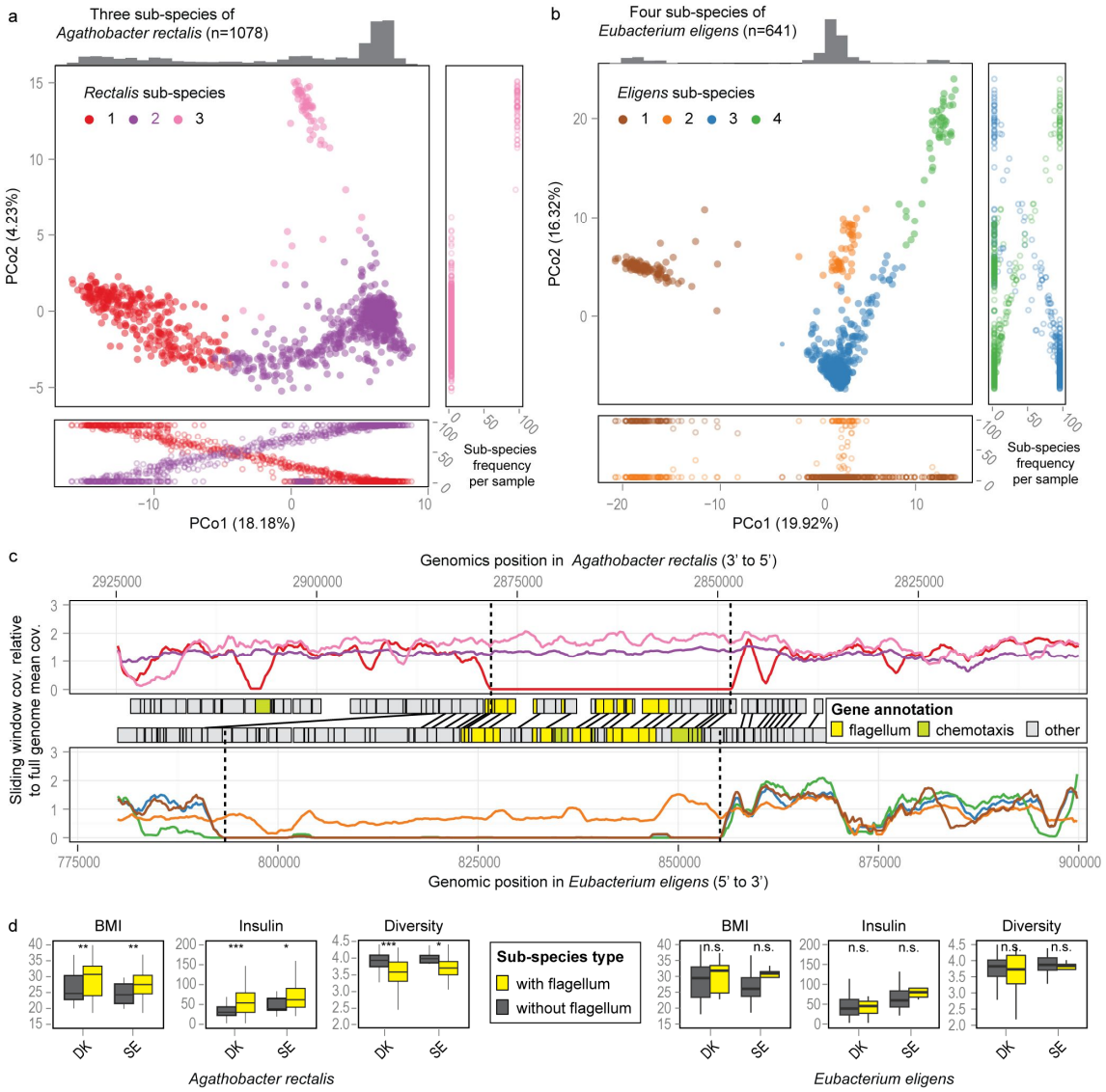


Figure 3. Gene complement differences between sub-species and their implication for the host.

PCoA projections of two Eubacterium species, *E. rectale* (a) and *E. eligens* (b) show the existence of three and four sub-species, respectively. Very few individuals have more than one of these sub-species present at one time as seen in plots at bottom showing sub-species frequencies. Functionally, the main distinction between some of the sub-species is the deletion of multiple flagellum and chemotaxis related genes, relative to the representative genome (c). Grouping *Eubacterium rectale* samples based on this deletion, shows a significant increase in BMI and blood insulin levels as well as a decrease in overall diversity in samples dominated by the flagellum carrying sub-species. The direction is recovered when considering *Eubacterium elignes*, though the difference is not significant (d).