The interaction of genetics, inflammation and the microbiome in the human metaorganism

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Malte Christoph Rühlemann

Kiel, 2020

The interaction of genetics, inflammation and the microbiome in the human metaorganism

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Malte Christoph Rühlemann

Kiel, 2020

Erster Gutachter: Prof. Dr. Andre Franke

Zweiter Gutachter: Prof. Dr. Dr. h.c. Thomas C. G. Bosch

Tag der mündlichen Prüfung: 19.05.2020

Zum Druck genehmigt: 19.05.2020

Epirrhema

Müsset im Naturbetrachten
Immer eins wie alles achten.
Nichts ist drinnen, nichts ist draußen;
Denn was innen, das ist außen.
So ergreifet ohne Säumnis
Heilig öffentlich Geheimnis!

Freuet euch des wahren Scheins,
Euch des ernsten Spieles!
Kein Lebend'ges ist ein Eins,
Immer ist's ein Vieles.

- Johann Wolfgang Goethe

Zusammenfassung

Die evolutionäre Beziehung zwischen Tieren, von Schwämmen bis hin zu Säugetieren, und den mit ihnen assoziierten Mikroorganismen – dem Mikrobiom – hat biologische Einheiten hervorgebracht, die nicht ohne einander betrachtet werden sollten oder gar könnten: den Metaorganismus.

Insbesondere die Existenz und die Entwicklung des Menschen als Metaorganisms hat den Fokus wissenschaftlicher Arbeit auf sich gezogen, denn die human-assoziierten mikrobiellen Gemeinschaften haben mutmaßlich großen Einfluss auf das Wohlbefinden und die Gesundheit des Wirts . Speziell das intestinale Mikrobiom, welches das größte Reservoir an Mikroorganismen im menschlichen Körper bildet, hat das Interesse verschiedenster medizinischer Disziplinen auf sich gezogen. Interaktionen von Wirt und Mikroorganismen – dazu gehören Bakterien, Archaeen, Pilze, einzellige Eukaryoten und Viren – stehen im Zentrum vieler metabolischer Prozesse oder können diese modulieren.

Da viele der Faktoren, die diese human-assoziierten mikrobiellen Verbünde formen noch immer unbekannt sind, hatte diese Doktorarbeit das Ziel, zu einem tieferen Verständnis des menschlichen Metaorganismus beizutragen.

Das erste Kapitel stellt einen Überblick über die wichtigsten in dieser Arbeit diskutierten Konzepte, Technologien und Krankheitsbilder dar. Das zweite Kapitel betrachtet technologische Aspekte der 16S rRNA Gen-Amplikon-basierten Analyse von mikrobiologischen Verbünden und die Nutzung der 'Shotgun'-Metagenomsequenzierung. Außerdem wird diskutiert, wie diese Ansätze die Ergebnisse von Mikrobiom-Studien beeinflussen können. Durch die große Bandbreite an untersuchten Modell- und Nicht-Modell-Organismen im Sonderforschungsbereich 1182 'Ursprung und Funktionieren von Metaorganismen' konnte diese Studie die Erkenntnis gewinnen, dass Amplikon-basierte Strategien und Metagenomik sich gegenseitig komplementieren können. Die Technologie der Wahl sollte immer nur unter Betrachtung der jeweiligen mikrobiologischen Population identifiziert werden. Des weiteren konnte durch diese Studie gezeigt werden, dass der evolutionäre Übergang von aquatischen zu terrestrischen Lebensformen einen großen Einfluss auf die Wirts-assoziierte Mikrobiota hatte.

Im dritten Kapitel stehen die Betrachtung des Einflusses der genetischen Variabilität des Wirts auf die Zusammensetzung des Mikrobioms sowie die Präsenz und Abundanz einzelner Mitglieder der mikrobiellen Gemeinschaft im menschlichen Darm im Zentrum. In zwei Studien - die erste basierend auf 1.767 Individuen aus Norddeutschland, die zweite auf weiteren 7.275 Personen aus Nordost- und Süddeutschland – konnten vier bzw. 32 Abschnitte im menschlichen Genom mit Einfluss auf das Mikrobiom identifiziert werden. Zusätzlich konnten Erkenntnisse über die Bedeutung des ABo-Blutgruppensystems für die Interaktion mit Bakterien der *Bacteroides*

gewonnen und durch Anwendung Mendelscher Randomisierung kausative Zusammenhänge zwischen Darmbakterien und u.a. chronisch-entzündlichen Darmerkrankungen hergestellt werden.

Das letzte Ergebniskapitel, Kapitel 4, betrachtet krankheitsspezifische Veränderungen im menschlichen Mikrobiom. Diese zeigen sich bei primär sklerosierender Cholangitis (PSC) und *Colitis Ulcerosa* (UC) im Vergleich zu gesunden Kontrollpopulationen aus Norwegen und Deutschland. Insbesondere für die Gattungen *Veillonella* und *Coprococcus,* letztere steht oft im Zusammenhang mit einem gesunden Darm, konnte eine Erhöhung bzw. Depletion in der relativen Abundanz im Zusammenhang mit PSC-Erkrankung gefunden werden. Durch diese und weitere Veränderungen konnte eine diagnostische Signatur zur Differenzierung zwischen PSC-Patienten und gesunden Kontrollen ermittelt werden (AUC = 0.88). Auch das Darmmykobiom zeigte charakteristische Veränderungen im Zusammenhang mit PSC, ausgezeichnet durch eine erhöhte relative Abundanz der Gattungen *Candida* und *Trichocladium*. Zuletzt wurden krankheits-assoziierte Veränderungen im Hautmikrobiom von Patienten mit atopischem Exzem (AE) untersucht. In diesem Zusammenhang wurden Unterschiede in Hautregionen mit akuten und chronischen Läsionen sowie in nicht-läsionalen Hautbereichen von AE-Betroffenen im Vergleich zu gesunden Kontrollen gefunden. Läsionen gingen mit einer erhöhten relativen Abundanz von *Staphylococcus aureus* und anderen *Staphylococcus* spp. einher. Zusätzlich wurden epidermale Fettsäurezusammensetzungen und Mutationen im *Filaggrin (FLG)* Gen, einem zentralen Risikofaktor für AE, als Einflussgrößen auf das Hautmikrobiom erkannt.

Zusammenfassend legen diese Befunde einen starken Einfluss der Genetik des Wirts auf die Abundanz einzelner Mikroorganismen und die Zusammensetzung des Mikrobioms als Ganzem nahe. Die krankheitsassoziierten Veränderungen in verschiedenen Teilen der Mikrobiota implizieren metabolische Interaktionen von Wirt und Mikroorganismen. Ob es sich bei diesen um Ursache oder Folge der jeweiligen Krankheit handelt, kann nicht abschließend geklärt werden, jedoch deuten die Ergebnisse der Mendelschen Randomisierung auf einen aktiven Einfluss von Mikroorganismen auf komplexe Krankheitsbilder hin. Die vorgestellten Ergebnisse legen den Grundstein für eine tiefergehende Betrachtung von identifizierten Mikroorganismen (bspw. *Bacteroides* und *Trichocladium*) als Angriffspunkt für die Behandlung von chronischen Entzündungserkrankungen.

Summary

The evolutionary relationship between animals, from sponges to mammals, and their associated assemblages of microorganism – the microbiome – created biological entities that should not and cannot be considered without one another: the metaorganism.

Especially the existence and development of the human as a metaorganism has shifted into the focus of scientific research, as the human-associated microbial communities are thought to have great impact on the host's constitution and health. Here especially the intestinal microbiota, the largest reservoir of microorganisms associated with the human body, has gained large interest in a wide range of medical fields, from gastrointestinal to neurological disorders. Host-microbe interactions with these commensal microbes – bacteria, archaea, fungi, unicellular eukaryotes and viruses – are expected to integrate into and modulate a wide range of the host's metabolic processes.

As the forces shaping the human-associated microbial communities and the extent of these interactions are still largely unknown, this doctoral thesis aimed to contribute to a deeper understanding of the human metaorganism.

The first chapter introduces major concepts, technologies and diseases discussed in this thesis. Chapter 2 discusses how technological aspects of 16S rRNA gene amplicon-based surveys of microbial communities and shotgun metagenomic sequencing influence the outcome of microbiome studies. Using a wide range of microbial communities associated with model and non-model organisms investigated in the framework of the 'Collaborative Reseach Centre 1182 – Origin and Function of Metaorganisms', the study shows that amplicon-based and metagenomic approaches are capable of complementing each other, and that the choice of technology always depends on the community under investigation. In addition, the study demonstrates that the transition from aquatic to terrestrial lifestyle in the evolutionary history of animals had major impact on the community composition of the host-associated microbiome.

Chapter 3 investigates the effect of host-genetic variation on the community composition and the presence and abundance of individual members and taxonomic groups of the human intestinal microbiota. In two genome-wide association studies, one involving 1,767 individuals from Norther Germany and the other an additional 7,275 individuals from North-Eastern and Southern Germany, four and 32 human genomic loci with influence on microbial traits were identified, respectively. In addition, the second study reveals specific ABO histo-blood group related changes in members of *Bacteroides*, and causative effects of microbial features on inflammatory bowel disease as well as other inflammatory and non-inflammatory disorders using Mendelian randomization.

Chapter 4 summarizes findings on disease-associated changes in different parts of the human microbiome. Marked changes were found in the bacterial microbiota in connection with primary

sclerosing cholangitis (PSC) and ulcerative colitis (UC) in comparison with healthy control individuals from Germany and Norway. Especially an increase in the relative abundance of *Veillonella* and a decrease abundance of *Coprococcus,* a bacterium generally associated with intestinal health, were found to be robust in connection with PSC, together with a wide-ranging loss of diversity. In addition, bacterial signatures that are useful as diagnostic markers for the discrimination of PSC patients and healthy controls were obtained (AUC = 0.88). Also, the intestinal mycobiota exhibits clear differences between healthy individuals and PSC patients, displayed by disease-associated increases in relative abundances of the genera *Candida* and *Humicola/Trichocladium*. Lastly, disease-associated changes were found in the skin microbiota of individuals with atopic dermatitis (AD) at sites with acute and chronic lesions, as well as non-lesional skin across all sampled body sites. Lesions were associated with increased relative abundances of *Staphylococcus aureus* and changes of other *Staphylococcus* spp. In addition, epidermal lipid composition and mutations in the *filaggrin (FLG)* gene, a major known risk factor for AD, were found to be influencing skin microbiome composition.

In summary, the findings presented in this thesis suggest a strong impact of host-genetics on the abundance of individual members of the microbiome and the community composition as a whole. The disease associated changes in different parts of the host microbiota imply metabolic interactions of host and microbes. Whether these are causal for or consequence of disease remains uncertain, however, the results of Mendelian randomization suggest that members of the intestinal microbiota actively contribute to the modulation of complex disease phenotypes. These results may lead to an in-depth investigation of candidate microorganisms, such as *Bacteroides* and *Trichocladium*, for potential treatment targets of chronic inflammatory diseases.

Table of Contents

1 Introduction

1.1 Motivation and outline

The organisms of the human microbiota, especially in the gut, are assumed to be in close interaction with the host, having influence on physiologic processes of the host and the modulation of environmental stimuli. The developments in high-throughput sequencing technologies and the fulfillment of the computational demands connected to the processing of the generated data have opened vast opportunities to investigate these microbial communities at unprecedented scale. The investigation of host-microbe and host-microbiome interactions in health and disease are the basis to a new understanding of metazoan organisms not as one, but as many: a metaorganism. This novel view can create new options for intervention, especially in complex chronic diseases, in which to large extents disease etiology is still often unknown and potentially connected to an aberrant interaction of host and its associated microbes.

This thesis aims to contribute to a deeper understanding of the forces shaping the microbiome as part of the human metaorganism and to shed light on the interaction of host and microbes in health and disease.

Chapter 1 gives an overview of the fundamental concepts discussed in this thesis, introducing the central terminology surrounding microbiome research, the view of the host as metaorganisms in close interaction with its microorganisms, now and along its evolutionary history, and briefly describing the diseases that are subject of analysis in the upcoming chapters.

Chapter 2 focuses on technical influences when assessing microbial communities using next-generation sequencing techniques, investigating the influence of target-amplicon choice and the use of shotgun metagenomics on microbiome analysis across a broad range of hosts. The outcomes of this analysis were published as article in the journal *Microbiome* in September 2019 and were supported by efforts of the Collaborative Research Centre (CRC) 1182 - 'Origin and Function of Metaorganism'. This article is referred to as **Article A** in this thesis.

The focus in **Chapter 3** lies on the influence of variation in the host's genome on the composition of the human intestinal microbiota. Two articles form the backbone of this chapter: The first, referred to as **Article B**, was published in August 2017 in the journal *Gut Microbes* as addendum to an article by Wang *et al.* (2016) [1]. **Article B** introduces a permutation-free approximation of the distance-based F-test for the improved analysis of the influence of genetic variation on beta diversity in a cohort of almost 1,800 individuals from around Kiel, Schleswig-Holstein, Germany. This application reduced analysis time by orders of magnitude and identified new candidate loci for host-microbiome interactions. The second article in this chapter follows up along the line of associations of microbiome and host genetics, expanding the total number of participants included

in the analysis to 5-fold, investigating almost 9,000 individuals from Northern (Kiel), North-Eastern (Greifswald) and Southern (Augsburg) Germany in the to-date largest single-country microbiome GWAS. The results of this analysis were summarized in **Article C** and submitted to *Nature Genetics* in January 2020.

Chapter 4 changes focus towards an analysis of microbial communities in the context of chronic inflammatory disorders. **Article D** and **Article E** both focus on the disease-associated changes of the bacterial microbiota in the cholestatic liver disease primary sclerosing cholangitis (PSC) in comparison to healthy control individuals and individuals with ulcerative colitis (UC), whereas **Article F** introduces the fungal components of the microbiota as disease-associated factor in PSC. The **Articles D and F** were published in the journal *Gut* in May 2016 and October 2019, respectively. **Article E** was published in *Alimentary Pharmacology & Therapeutics* in June 2019. A fourth article, **Article G**, completes this chapter, analyzing the skin microbiota with regard to atopic inflammation and epidermal properties. This article was published in the *Journal of Allergy and Clinical Immunology* in January 2018.

The final **Chapter 5** attempts to put the **Articles A-G** into broader perspective, discussing the outcomes of the individual studies in the light of host-microbe and host-microbiome interactions in the human metaorganism, their implications on the host's evolutionary history and how all this impacts host health, also giving an outlook on ongoing and future perspectives in metaorganism-focused microbiome analysis.

1.2 The Human as 'Metaorganism'

1.2.1 From Microbiome to Metaorganism – Definitions

Biologist John L. Mohr is credited with coining the term 'microbiome' in a 1952 journal article discussing the extent to which protozoans can be used for monitoring pollution levels in fresh water [2]. Mohr's use of the term was short-hand to refer to the microbial organisms within a sampled environmental site. With the advent of '-omics' and high-throughput nucleotide sequencing technologies (see Chapter 1.4.1 for details), the meaning of the term 'microbiome' has broadened - it now explicitly encompasses the genomic material of the present microorganisms as well as their surrounding habitats - whereas the original, organism-focused meaning has been replaced by the term 'microbiota' [3].

The assemblages of microorganisms investigated in microbiome research include a broad range of taxonomic groups, from unicellular eukaryotes and fungi over prokaryotes (bacteria and archaea), to viruses. Many of these sub-entities are referred to by specific terms, *i.e.* the mycobiome (the collection of fungi and their genetic material in a specific environment), the bacteriome, the archaeome, and the virome (including also the phageome: the collection of bacteriophages).

These microorganisms don't simply co-exist, but rather they interact with each other in various ways (see Faust & Raes [4] for details). Such interactions between organisms can lead to neutral (0), positive (+) or negative (–) outcomes. An interaction with beneficial effects for both partners (+|+) is called 'mutualism'. If only one of the partners benefits from the interaction with no effect on the other partner, this is called 'commensalism' (+|0). If the positive effect for one of the actors leads to a negative effect for the other, it is called 'parasitism' or 'predation' (+|–). An interaction that is neutral for one partner and negative for the other (0|–) is called 'amensalism', while an interaction with negative effects on both partners is called 'competition' (–|–). Strictly neutral effects of the partners on each other are possible (0|0), however by definition this cannot be called an interaction. The interactions between all parties within an environment can be seen as complex networks or ecological communities.

The natural environments in which microbial communities are found are manifold. Large initiatives have been undertaken to survey and characterize these communities on a global scale. Notably, the Earth Microbiome Project (EMP) started in 2010 with the aim "to sequence microbes and microbial communities from every conceivable biome" [5], whereas the Human Microbiome Project (HMP), was specifically aimed at a large-scale survey of microbial communities on and in the human host [6].

The HMP addressed an additional layer of complexity on an unprecedented scale by considering the interaction of the host with its associated microbiota. All multicellular life on earth has developed from and in close contact with microorganisms, and thus the investigation of a system, for example a human, warrants the consideration of its associated microbial communities. The concept of interdependence between different species is a relatively old theory – it was introduced in 1877 by Karl Möbius with the term 'biocenosis'[7]. Yet, it took more than one hundred years until the term 'metaorganism' was coined by Graham Bell and subsequently established as a key concept describing the close relationship and interdependence of a host with its microbiota [8], [9]. This idea is also closely connected to the hologenome theory of evolution [10], which explicitly considers the host together with its symbiotic relationships and their influence on the host's fitness and *vice versa* as unit of selection.

Moreover, the development of the host together with its microbes over evolutionary timescales, a process called 'codiversification', can be investigated by comparing the phylogenies of hosts and their associated microbes, reflecting parallel trajectories. Codiversification can be driven by reciprocal adaptation of the two interacting partners to one another, this special case is termed 'coevolution' [11]. A well characterized example of symbiotic coevolution is the relationship between the pea aphid *Acyrthosiphon pisum* and the bacterial species *Buchnera aphidicola.* Over time, the aphid has evolved specialized organs for housing and vertical transfer of the bacteria to its

offspring, while the bacterium has evolved to have a reduced genome, specialized to provide essential amino acids to its host [12], indicating a deeply intertwined metabolic relationship.

Parallels in phylogenies can not only be reflected on the level of single microbial partners, but also on the level of the whole microbial community. This relationship is called 'phylosymbiosis' and was demonstrated to be a widespread phenomenon across the host evolutionary histories of corals [13], insects (Nasonia [14], Drosophila [15]), and primates [15], [16].

These interactions suggest a deep, functional relationship between microbes and their host and emphasize the need for a metaorganism-focused view in the analysis of microbial communities and host-health.

1.2.2 Human-associated microbial communities

The average male human body consists of an estimated 3×10^{13} human and 3.8×10^{13} bacterial cells, with bacteria outnumbering the host's own cells by about 1:1.3 (for females: 1:2.2)[17]. As mentioned earlier, large efforts were put into the survey of these human-associated microbial communities by the Human Microbiome Project [6]. A first landmark paper from the HMP reported the results gained from 4,788 specimens acquired from 242 adult individuals from the US, sampling up to 18 body habitats from five major body areas (urogenital, skin, nasal, gastrointestinal and oral)[18]. At the time of publishing, this resource represented the most comprehensive overview of host-associated microbial communities, highlighting the intra-personal variability between the sampled body sites, but also the clear habitat-specific community composition across individuals, with distinct bacterial taxa dominating any given habitat. A summary of bacterial abundances across different sampling sites inferred from metagenomic data of the Human Microbiome Project [19]can be found in Figure 1.1.

The gut is the largest reservoir of microorganisms in and on humans, with at least 95% of the total host-associated bacteria residing in the colon [17]. The most abundant phyla in the human intestinal microbiota are the mostly Gram-negative Bacteroidetes and the mostly Gram-positive Firmicutes, together usually comprising between 80-90% of the total relative abundance [18]. Genera in the phylum Bacteroidetes, frequently present in the intestinal community, include *Bacteroides* and *Prevotella*. Both bacteria have been found to be involved in the degradation of diet-derived complex carbohydrates and dietary fibers, such as cellulose, pectin and xylans [20], [21]. These degradation capacities were shown to be species-specific, for example *B. thetaiotaomicron* utilizes a vast range of plant and animal glucans [20], wheras *B. ovatus* is able to degrade hemicellulosic polysaccharides [22] and *P. copri* can ferment xylans [23]. The differing utilization of carbon sources has an effect on the output of metabolic compounds, for example the ratios in which the short-chain fatty acids (SCFAs) acetate, propionate and butyrate are produced [24].

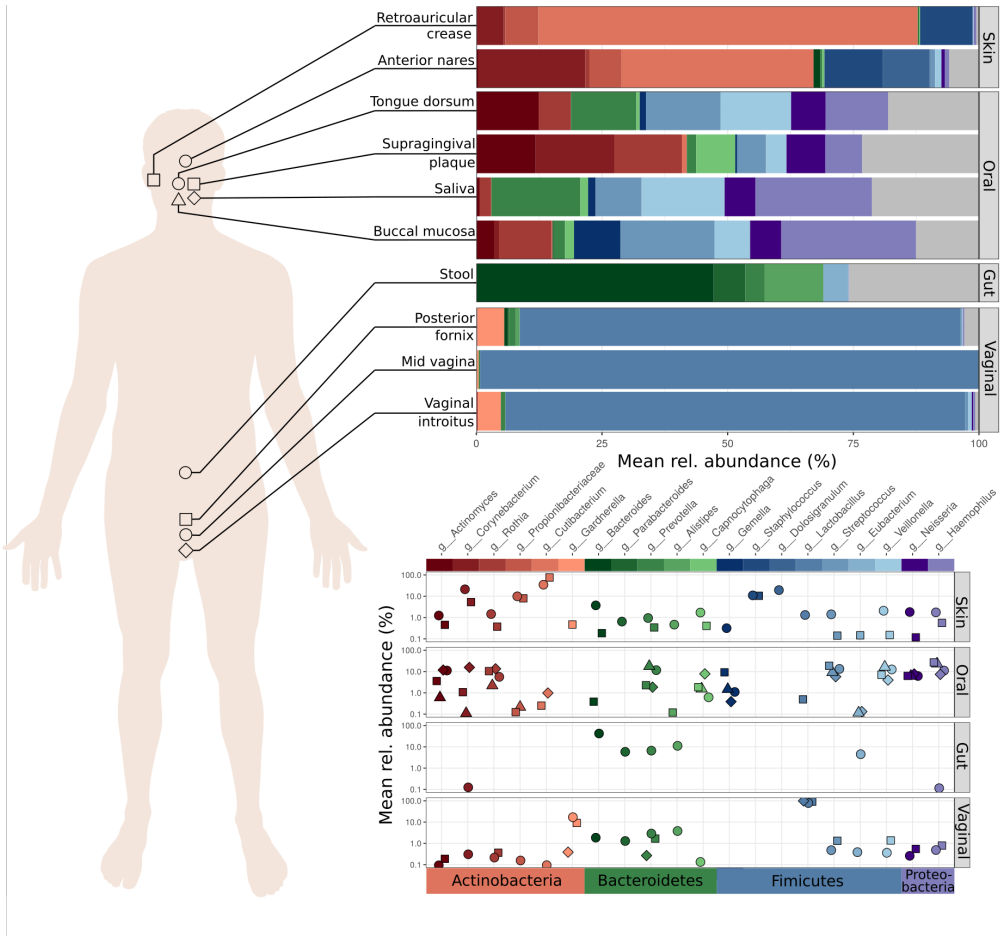


Figure 1.1: Bacterial microbiome composition across ten different human body sites in the Human Microbiome Project [19]. Genus-level abundances are shown for taxa present in at least 10% of samples and a surpassing a mean relative abundance of 5% at one sampling site. Primary colors reflect the phylum of the taxon: red: Actionobacteria; green: Bacteroidetes; blue: Firmicutes; purple: Proteobacteria. Shapes in the lower panel correspond to shapes in the upper panel of the respective sampling area (Skin, Oral, Gut, and Vaginal).

SCFAs play an important role in host-microbe signaling and host immune-cell regulation [25], and in addition, are crucial for gut barrier integrity and intestinal homeostasis [26], [27].

Firmicutes in the intestinal lumen and mucosa are more diverse than Bacteroidetes. Important members include the classes Clostridia (which is a very heterogenous, polyphyletic taxon in itself) and Lactobacillales (lactic acid bacteria; several of them used as probiotics, such as *Lactobacillus* spp. and *Enterococcus faecium*). Among Firmicutes, we also find potent producers of SCFAs (*Faecalibacterium, Roseburia*) [27] as well as bacteria converting primary to secondary bile acids (*Clostridium perfringens, Lactobacillus*) [28], an important step in the enterohepatic circulation [29]. Some members of Firmicutes were found to affect blood cholesterol levels (*Lactobacillus*) [30], others are generally regarded as associated with host-intestinal health (*Faecalibacterium*)[31].

Further prominent bacterial genera not belonging to either of the two major phyla are *Bifidobacterium* (phylum Actinobacteria), a potential probiotic promoting host health [32],

Akkermansia (Verrucomicrobia), a mucin-degrading bacterium with potential benefits for gut integrity [33], and a wide range of others belonging to the *Enterobacteriaceae* (Proteobacteria), which is a diverse family of commensal and transient gut bacteria with the ability for severe opportunistic pathogenicity [34].

The skin is the largest human organ and its bacteriota is dominated by three distinct genera, *Cutibacterium* (formerly *Propionibacterium [35]*), *Corynebacterium,* and *Staphylococcus* [18], [36]. These three bacteria make up 80-100% of the relative abundance in the community, with differing compositions depending on whether the sampled size is oily (*e.g.* the forehead), moist (the inside of the elbow), or dry (the volar forearm); other bacteria are present only in minor proportions [36]. The oral microbiome is very diverse in the relative abundances of the bacteria, depending on the sampling site, however members of the genera *Haemophilus, Prevotella, Veillonella,* and *Streptococcus* usually represent large parts of the microbial composition [18]. *Corynebacterium* and *Cutibacterium,* genera dominating the skin microbiome, are also found in comparably high abundances in dental plaque [18]. A healthy vaginal microbiota is characterized by dominance of usually one of three distinct *Lactobacillus* species, namely *L. crispatus, L. gasseri,* or *L. iners* [37].

While bacteria are the predominant organisms in all human-associated microbial communities, non-bacterial microorganisms have recently shifted into the focus of the assessment of human microbiomes. Targeted approaches surveying archaeal [38], fungal [39], and viral diversity [40], [41] suggest distinct and stable compositions across body sites.

1.2.3 Genetics, environment, evolution – forces shaping the human gut microbiome

The forces shaping the habitat specificity of host-associated microbiomes and especially the variability between individuals are still largely unknown and central of current scientific efforts. Studies investigating the fecal microbiome of monozygotic (MZ) and dizygotic (DZ) twins have shown that twins have a more similar microbiota compared to unrelated individuals [42]. These results were later confirmed in the TwinsUK cohort, which consisted of almost 1,000 twins [43]. Further, Goodrich *et al.* [43] demonstrated that MZ twins share even larger proportions of their microbiota than DZ twins, strongly suggesting an influence of the host's genetics on the microbiome. Additionally, a list of highly heritable taxa was compiled, identifying the genus *Christensenella* as the most heritable taxon within this sample collective [43]. These results sparked multiple independent, large-scale, and regional efforts to identify the genetic loci influencing the configuration of the human intestinal microbiota [1], [44]–[46]. Each of them found multiple associations of genomic loci and univariate microbial traits (see Table 1 for summary), however the overlap between the results was small, with only *SLIT3* replicating as influencing factor in two of the studies on a genome-wide significant level ($p<5\times10^{-8}$) [44], [46]. The little overlap between studies could possibly be explained by technical and methodological differences in the analysis

(Table 1). An additional study of 814 Israeli individuals did not yield any statistically robust genome-wide significant associations [47]. The participants in this study were genetically highly heterogeneous, belonging to five major and additional minor and admixed ethnic groups [47]. This genetic heterogeneity in combination with the (for a genome-wide association analysis) small sample size might explain the lacking results [47]. Nevertheless, the study by Rothschild *et al.* [47] emphasizes that additional factors, for example environmental or lifestyle elements, significantly contribute to host-associated community assembly.

A survey of 531 children and adults, including relatives and twins, from the US, rural Malawi and the Amazonas of Venezuela (Amerindians) revealed that a westernized lifestyle, *i.e.* consumption of processed food, high fat and sugar intake, and industrialized or office jobs, largely impacted the composition of the intestinal microbiome [49]. One of the main compositional differences accompanying westernization was demonstrated on one hand in the reduced diversity in the microbial composition of the US individuals, and on the other hand in the relative abundances of the genera *Bacteroides* and *Prevotella*. The genus *Prevotella* were found at mean relative abundances of > 20% in the non-westernized Malawians and Amerindians, while they were virtually absent in the individuals from the US surveyed in this study. Conversely the genus *Bacteroides* revealed a pattern with a mean of >10% relative abundance in the US study population and <1% mean in the Malawians and Amerindians [49]. This pattern was also found to be true in a study comparing the gut microbiome of the Hadza, an extant community of hunter-gatherers from Tanzania, and a study cohort from Italy as urban controls [50]. The shift from *Prevotella*- to *Bacteroides*-dominated intestinal communities in westernized populations is summarized in Figure 1.2.

Table 1: Overview of publications investigating human genetic associations with traits in a setting of a genome-wide association analysis based on 16S amplicon sequencing data (16S) or shotgun metagenomic sequencing (MGX).

Study	Population	Data	Model	Trait	Main findings
Blekhman *et al.* (2014) [51]	US Americans (N = 93)	16S	Linear regression	Beta diversity principal coordinates (PCos) as quantitative traits	No reporting of SNP vs. PCo results; No enriched pathways
				Taxon abundances as quantitative traits	8 coding variants associated with 4 taxa (Q<0.1)
Davenport *et al.* (2015) [52]	Hutterites (USA; N_{summer} = 91, N_{winter} = 93)	16S	Linear mixed model	Alpha diversity	No associations reported
				Taxon abundances as quantitative traits	95 independent associations with 32 taxa (Q<0.2)
Goodrich *et al.* (2016) *[43]*	United Kingdom ($N_{Twin\ pairs}$ = 1,126)	16S	Microbiome GWAS [53]	Beta diversity	Two independent associations ($P < 5 \times 10^{-8}$)
			Linear mixed model	Taxon abundances as quantitative traits	307 associations (Q<0.2) 28 associations ($P < 5 \times 10^{-8}$)
Bonder *et al.* (2016) [46]	The Netherlands (N = 1,514)	MGX	Rank-based Spearman correlation (zero truncated)	Taxon abundances as quantitative traits	9 associations ($P < 5 \times 10^{-8}$)
				Gene ontology (GO2000) annotation abundances as quantitative traits	12 associations ($P < 5 \times 10^{-8}$)
				MetaCyc pathways abundances as quantitative traits	21 associations ($P < 5 \times 10^{-8}$)
Turpin *et al.* (2016) [45]	Canada ($N_{discovery}$ = 1,098, $N_{replication}$ = 463)	16S	Generalized estimating equations (GEE)	Alpha diversity (log-normal) as quantitative trait	No associations ($P < 5 \times 10^{-8}$)
				Microbial dysbiosis index (log-normal) as quantitative trait	No associations ($P < 5 \times 10^{-8}$)
				Taxon abundances as quantitative traits (log-normal; two-part log normal)	58 associations with 33 taxa ($P < 5 \times 10^{-8}$); Of these, 4 replicating ($P < 0.05$)
Wang *et al.* (2016) [1]	Germany (N = 1,812)	16S	Non-parametric test of ordination fitting ('envfit')	Beta diversity	42 associations ($P < 5 \times 10^{-8}$)
			Generalized linear (hurdle) model (Negative binomial)	Taxon abundance as quantitative traits	40 associations with 24 taxa ($P < 5 \times 10^{-8}$)
Rothschild *et al.* (2018) [47]	Israel (N = 814)	MGX	Linear Mixed Model	Taxon abundance as quantitative (< 5% zeros) and qualitative traits (≥ 5% zeros)	43 associations with 35 taxa ($P < 5 \times 10^{-8}$); None significant after FDR correction.

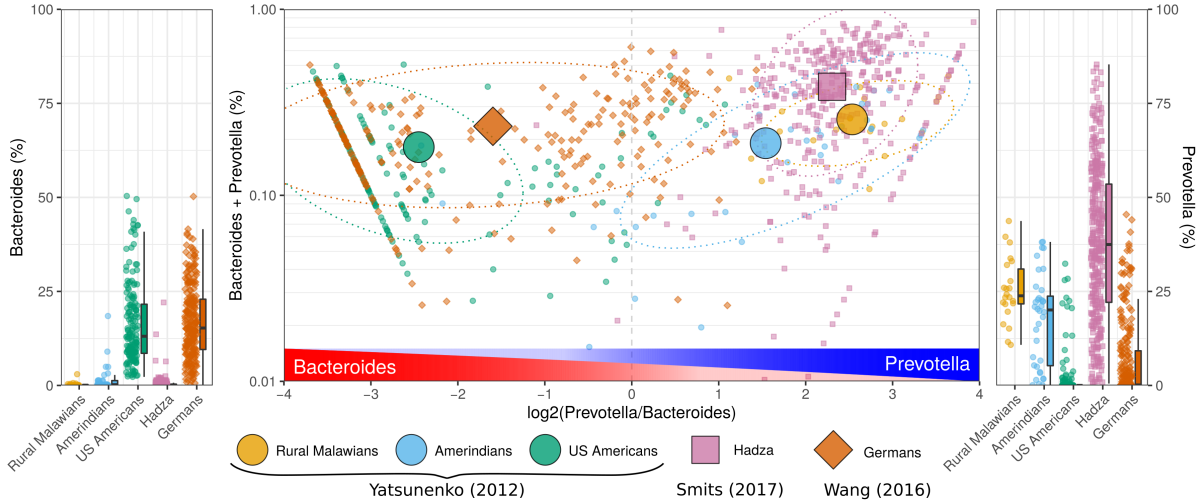


Figure 1.2: Relative abundance distribution and log-ratios of *Bacteroides* and *Prevotella* in five sample sets from three publications, covering non-westernized (Rural Malawians [49], Amerindians [49], Hadza [50]) and westernized (US Americans [49], Germans [1]) populations. The center panel shows the individual samples (small symbols) and group centroids (large symbols); ellipses show the 80% confidence intervals of the groups' multivariate t-distribution around the group centroids. The left and right panel display the individuals' relative abundances of *Bacteroides* and *Prevotella*, respectively. Group-wise distributions of the values are shown as box-and-whisker plots. Boxes display the range from the first to the third quartile, with group medians being marked by a black line within the respective boxes. Whiskers expand to the last individual datapoint within a range of 1.5 interquartile ranges above and below the groups' third and first quartiles, respectively.

Bacteroides and *Prevotella* are also central to the concept of the existence of 'enterotypes'. The enterotypes were first introduced as three distinct and stable clusters in the human intestinal microbiome, with each of them linked to high abundances of one respective bacterium, either the aforementioned *Bacteroides* or *Prevotella,* or the genus *Ruminococcus* [54]. Wu *et al. [55]* demonstrated the presence of the *Bacteroides-* and *Prevotella*-enterotypes in an independent cohort, linking the configuration to long-term dietary patterns either dominated by protein and animal fat intake or by carbohydrates, respectively. The *Ruminococcus*-enterotype could not be recovered in this study, likely due to a larger sample size, leading it to fuse with the *Bacteroides*-enterotype, no longer representing a separate cluster [55]. With the introduction of Dirichlet multinomial mixtures (DMM) for microbial community modeling, Holmes *et al.* [56] argue the existence of four enterotypes, with two of them being rather homogeneous and dominated by *Bacteroides,* and with the other two being more variable in their composition. Additionally, the results of the study of Holmes *et al.* suggested that ileal Crohn's disease, a subform of inflammatory bowel disease (IBD; details in chapter 1.3.1), is associated with the more variable community types and obesity is not showing strong connection to any of the enterotypes, but rather with the deviation therefrom [56]. The discussion about enterotypes remains controversial. Again others argue against any existence of enterotypes at all [57] or as them being analysis artifacts, introduced due to the compositional nature of microbiome sequencing data, which does not take quantitative assessments of abundance into account [58]. However, intriguingly similar patterns of enterotype clusters could also be demonstrated to be present in the gut microbiota of chimpanzees, suggesting an evolutionary conservation of stable community structures in primates

and possibly beyond [59]. Also, quantitative assessment in a cohort of infants from Denmark indicate an establishment of enterotypes as stable community states already very early in life [60]. Supporters are found for each of the presented models (see Costea *et al.* [61] for a summary), highlighting again how little is known about the processes shaping the intestinal microbiota.

Two large cohorts from Belgium (Flemish Gut Flora Project, FGFP; n = 1,106) and the Netherlands (LifeLines-DEEP, n = 1,135) were extensively phenotyped and surveyed about dietary habits and additional potentially influencing factors in an effort to clarify the extent to which diet and nutrition might affect the composition of the gut microbiome [62]. These analyses revealed effects associated with anthropometrics (age, gender, height, body-mass index; total combined effect size < 5%), diet (fruit intake, bread type preference, coffee, magnesium supplementation; total ~ 6%) and blood parameters (red blood cell count, hemoglobin, triglycerides; ~ 8%). However, the largest effects were found to be connected to medication intake, with the use of the antibiotic amoxicillin (~2%), osmotic laxatives, anti-allergic medication and steroidal contraceptives (all ~ 1%) having the strongest individual contributions. Faloney *et al.* [62] estimate the total combined effect sizes of all assessed factors to 16.4% of the total variation in the intestinal community. The effects of antibiotics use on the microbiome composition in the intestine are striking and long-lasting, with marked changes in some individuals still evident after six months compared to before the treatment [63]. Further systematic *in vitro* assessment of more than 1,000 marketed drugs on 40 representative bacterial strains commonly present in the intestinal microbiota found roughly one quarter of drugs targeting human metabolic pathways showed negative effects on the growth of the tested commensal bacteria [64]. These drugs did not include known 'antibacterials', such as antibiotics and antiseptics, for which 78% were shown to have anticommensal effects [64].

In addition to the presented more proximal or direct effects of genetics, nutrition and medication, the evolutionary history of humans and non-human primates has also raised interest in connection to the microbiome. Groussin and colleagues [65] could demonstrate that for a wide range of mammals changes in the host's genetics are paralleled by shifts in the microbiome. Looking at the fecal microbiota of humans and comparing it to our closest living relatives chimpanzees (*Pan troglodytes*) and bonobos (*P. paniscus*), as well as eastern and western gorillas (*Gorilla beringei* and *G. gorilla,* respectively), it was shown that the relation between the microbial communities follows that of the host's phylogeny, thus exhibiting a 'phylosymbiotic' pattern [66]. However, in contrast to diversification in African apes, in which the changes in the microbiota were following a steady trajectory, the evolutionary changes towards the human microbiota since the split from the *Pan* lineage at least 7-8 million years ago [67] exhibit signs of a strong acceleration accompanied by a loss of microbial diversity, possibly connected to changes in the diet [68]. Regardless, key bacteria of the human microbiota are shared with the microbiota of other

primates [59], and comparing extant species of the bacterial families *Bifidobacteriaceae* and *Bacteroidaceae* across hominids revealed patterns of inter-species transmission events and cospeciation of close relatives of *Bifidobacterium adolescentis* along the human-chimpanzee phylogeny [69]. However, adaptive evolution of host-associated bacteria does not only happen on timescales of millions of years, but also during the lifetime of healthy humans, as it could be shown for *Bacteroides fragilis* by Zhao *et al.* [70], indicating a fast adaptation of commensal intestinal bacteria to changes in the environment and promotion of long-term colonization abilities.

The rapid changes in the microbiota along the human lineage since the divergence from the common ancestor with the *Pan* genus is thought to impact population health today. The decreased diversity in comparison to other primates potentially leads to a community more sensitive and/or less resilient to external perturbations, as shown for other environments [71]. A low diversity of the intestinal microbiome has been associated with severity of irritable bowel syndrome [72] inflammatory bowel disease [73], and primary sclerosing cholangitis [74], and low diversities in other organs also showed connections to disease, for example of the lung microbiota in association with asthma [75]. Deviations in the intestinal microbiota from that of healthy individuals, often termed "dysbiosis", have been connected to a wide range of diseases with autoimmune/inflammatory components [76], [77] and cancer [78], as well as neurologic and psychiatric disorders [79]. These connections carry the opportunity and potential to target the microbiota as a whole or its individual members in the treatment of disease in a framework of personalized health and disease management. However, for this a deeper understanding of the interaction between host and microbiome and the assembly of the microbial communities is necessary.

1.2.4 The role of the gut microbiome in the human metaorganism

The role of the human gut microbiome is thought to be manifold. Members of the gut microbiota are involved in the breakdown of complex carbohydrates, releasing SCFAs, thus actively supporting and promoting intestinal health and also providing simpler carbohydrates as energy source available for enteric uptake by the host. By the conversion of primary to secondary bile acids, members of the microbiome play an essential role in the enterohepatic circulation, and thus in the hosts lipid metabolism. In addition, vitamins, vitamin precursors, and essential amino acids can be synthesized and provided by the microbiota [80], highlighting the mutualism between host and microbes and their metabolic interconnection, with the microbiota expanding the human host's genetic capacity by an estimated more than 100-fold [81].

Next to the metabolic interactions, the commensal microbiota actively contributes to host immunity. The integrity of a healthy microbiota and its metabolic processes, by secreting SCFAs, bacteriocins and bile acids, hinder pathogens from colonizing niches in the gastrointestinal tract

(GIT), a process called colonization resistance [82]. Work on germ-free and antibiotics-treated mice, *i.e.* laboratory mice depleted of microorganisms, showed the effects of the missing symbiotic interactions on the mammalian host and the maturation of its immune system. Kennedy *et al.* [83] reviewed and summarized effects on mouse physiology, encompassing widespread changes in a variety of immune cell populations including decreased migratory capacity of myeloid cells to peripheral tissues, general reduction of lymphoid cells and changes in cytokine levels. The effects on individual organs include a dramatically increased cecum size, a reduced spleen size, an increase in villus length, reduction in the production of antimicrobial peptides, diminished granules in Paneth cells and changed expression of Toll-like receptors, together with and likely connected to an impaired tolerance of commensal bacteria. These changes in physiology highlight the importance of host-microbe interaction for the maturation and homeostasis of the immune system and an adequate response to microbial encounters [84].

In humans, a disturbance in the microbiota has been associated to a 'leaky' gut barrier, leading to the uncontrolled influx of metabolites, toxic molecules and antigens into the bloodstream [85]. These compounds are hypothesized to trigger inflammation and autoimmunity in genetically susceptible individuals, paving the path to chronic inflammatory disorders [86].

1.2.5 The microbiome in personalized medicine

The term 'personalized' or 'precision medicine' describes a concept of individualized therapeutic approaches, tailored to the needs and symptoms of the patient to increase treatment effects and safety and decrease potential side-effects [87]. Especially for different types of cancer, the use of genotyping and sequencing technologies has opened individualized treatment options for cancers with specific frequent mutations [88], [89]. For complex inflammatory disorders, a targeted treatment and combination therapy with antibodies has shown to be effective, however positive treatment response remains in parts unpredictable [90]. Here, the microbiome has the potential as a potent modulator of pharmacological intervention and pharmacokinetics [91], [92].

As mentioned earlier, inflammatory and metabolic conditions are accompanied by a dysbiotic intestinal microbiota. Harnessing this dysbiosis and shifting the microbiota towards that characterized in non-diseased individuals, for example by supporting the expansion of SCFA-producing bacteria, could help in disease management and induction of remission [27], [93]. Recent advances also emphasize the importance of strain-level/genome-level characterization of bacteria, as the presence of specific genes can lead to changed pathogenicity, independent of phylogenetic relationship to non-pathogenic strains [94], thus, an in-depth and personalized characterization of the microbiota can open new strategies for disease diagnosis and treatment. Additionally, the non-invasive sampling of fecal material for microbiome analysis can be used for

disease diagnosis [95], stratification into sub-entities of disease [93] and prognostic disease monitoring [96].

Ultimately, the transfer of a microbiome itself from one person to another, *i.e.* in fecal microbiota transplantation (FMT), has been shown to be an effective treatment of recurrent *Clostridioides difficile* infection (rCDI) [97], although the exact molecular mechanisms of the treatment are still unknown [98]. The efficacy of FMT for other diseases is currently being evaluated. Pilot studies suggest therapeutic potential for ulcerative colitis [99], however flares of intestinal inflammation have also been reported as side-effect of FMT for rCDI [100].

A detailed review of the microbiome as target and source for personalized health and intervention can be found in Kashyap *et al.* [101].

1.3 Selected chronic inflammatory disorders discussed in this thesis

This thesis investigates connections between the microbiome and human chronic inflammatory disorders. For a better understanding, this chapter summarizes key aspects of the diseases discussed in the following chapters and how they differ from each other: inflammatory bowel disease (Crohn's disease, ulcerative colitis), autoimmune and cholestatic liver disease (primary sclerosing cholangitis, primary biliary cirrhosis, autoimmune hepatitis) and chronic inflammatory skin disease (atopic dermatitis, psoriasis).

1.3.1 Inflammatory bowel disease

Several complex, immune-mediated inflammatory disorders of the gastrointestinal tract are summarized under the umbrella of 'inflammatory bowel disease' (IBD), the main sub-entities being Crohn's disease (CD) and ulcerative colitis (UC), as well as indeterminate colitis (IC). A total of 6.8 million individuals world-wide are affected by IBD and the global age-standardized prevalence rate of IBD was 84.3 per 100,000 individuals in 2017 (increased from 79.5 per 100,000 in 1990)[102]. A clear connection can be drawn between disease prevalence and a high Sociodemographic Index (SDI), as age-standardized prevalence rate in 2017 was highest in the USA (464.5 per 100,000) and UK (449.6 per 100,000) and lowest in the Carribean (6.7 per 100,000) [102]. As for many complex diseases, the disease etiology for IBD is not fully understood. Multiple large genome-wide association studies (GWAS; see section 1.4.4 for details) in up to almost 68,000 individuals have been conducted to elucidate the genetic basis of IBD and its sub-entities, identifying more than 240 genetic susceptibility variants associated with IBD, which however collectively are estimated to only explain less than 30% of the disease risk [103]. Altogether, these findings strongly suggest an uncontrolled immune reaction to an environmental (or microbial) trigger in genetically susceptible individuals to be responsible for disease onset [104].

Phenotypically, the main difference between CD and UC is the location of the inflammation in the gastrointestinal tract (GIT). Individuals with UC are affected by inflammation restricted to different parts (proctitis, left-sided colitis) or the entire colon (pancolitis) [105], whereas CD is characterized by inflamed patches possible along the entire GIT, from mouth to the perianal area [106]. Both diseases are presented with multiple possible extra-intestinal comorbidities with inflammatory components, for example of the skin (psoriasis), the joints (rheumatoid arthritis) and the cardiovascular system, as well as psychological and psychiatric disorders [107].

As of today, IBD is not curable, can lead to development of gastrointestinal cancers and requires extensive disease management up to surgical removal of the affected parts of GIT [105], [106].

1.3.2 Autoimmune and cholestatic liver disease

Like IBD, autoimmune and cholestatic liver diseases are complex and their etiologies are unknown, likely connected to an environmental trigger for autoimmunity [108]. All autoimmune liver diseases share the presence of chronic inflammation in the liver, displayed by injury of plasma cells in autoimmune hepatitis (AIH), the involvement of small intrahepatic bile ducts in primary biliary cirrhosis (PBC) and inflammation and fibrosis of predominantly larger bile ducts in primary sclerosing cholangitis (PSC). Disease progression in all three ultimately leading to liver failure and the need for liver-transplantation [108]. Estimates of prevalence vary, ranging between 40-80 affected individuals in 100,000, with equal parts affected by PBC and AIH, and PSC being the rarest, taking 5-10% of the patients [109]. In line with other autoimmune disease, women are more frequently affected by AIH and PBC, however PSC is diagnosed in men twice as often as in women [110]. AIH was found to be present in 6-9% of PSC and PBC cases, a status termed 'overlap syndrome' [110]. Standard treatment for AIH includes the administration of corticosteroids and azathioprine to slow down disease progression. For PBC, treatment with the naturally occurring bile acid ursodeoxycholic acid (UDCA) has proven to prolong liver-transplant-free survival [111]. In PSC patients, UDCA is tested for standard treatment, however studies about the efficacy are still lacking and TNF-targeting antibodies also proved ineffective [112]. Among PSC patients, 70-80% are also diagnosed with pancolitis similar to ulcerative colitis, however classified as an additional sub-entity of IBD [113]. In addition, 7-13% of PSC patients develop cholangiocarcinoma (CCA), a diagnosis which is often delayed, as no sensitive diagnostic tests for screening are available [114]. The 5-year survival rate after CCA diagnosis ranges between 5-15% [114].

All three presented inflammatory liver diseases share disease associations with the presence of specific autoantibodies and *human leukocyte antigen (HLA)* alleles as risk factors [110].

1.3.3 Chronic inflammatory skin diseases

Two of the most prevalent inflammatory skin diseases in industrialized countries are atopic dermatitis (AD) and psoriasis. AD is a common disorder, affecting up to 20% of children before they enter school [115], but also later onset and persistence in adults is frequent, with prevalence up to 10% [116]. Clinical features of AD include a general dryness of the skin with redness and lesions [115]. In infants, lesions can be found on large parts of the body, including the face, shifting towards the flexural folds in children, and in adults, the shoulders and neck area, as well as the hands, wrists and ankles can be affected [115]. In twin studies, AD was found to be highly heritable [117], and although mutations in *filaggrin* (*FLG*) were identified as strong genetic risk factor [118], the etiology remains complex with contributions from genetics and environmental factors [115]. AD often occurs in flares and disease management is usually guided by supporting epidermal barrier repair with emollients and anti-inflammatory corticosteroid therapy [115], however complication can arise from infections by bacteria (*Staphylococcus aureus*) or fungi (*Malassezia sympodialis*) due to the disrupted barrier [119].

Psoriasis is less prevalent than AD, however with a global prevalence of 2-3% can still be considered common [120]. AD and psoriasis share the display of dry skin patches, however in psoriasis these are accompanied by scaling plaques [121]. Psoriasis can also involve inflammation of the joints (psoriatric arthritis) [122] and individuals with Psoriasis have a 4-fold relative risk of Crohn's disease, but not ulcerative colits [123]. Large-scale analysis of CD and psoriasis showed overlapping genetic risk loci, highlighting the shared (auto-)immunity mediated genetic susceptibility to both diseases [124].

1.4 Key methods and concepts

Current analysis of microbial communities heavily relies on high-throughput nucleotide sequencing, thus it is important to understand the technologies used for nucleotide sequencing data generation and processing. This chapter gives an overview of the developments in sequencing approaches and survey strategies for data generation as foundation for microbiome analysis, followed by an introduction of general approaches in (sequencing-)data processing and statistical methods for the bioinformatic analysis of microbiome data. The chapter concludes with an introduction to the concepts of genome-wide association analysis.

1.4.1 Nucleotide sequencing technologies

In 1953, Rosalind Franklin and Maurice Wilkins produced the crystallographic data of DNA from which James Watson and Francis Crick were able to derive its three-dimensional structure. After this, it took another 12 years until a first full nucleotide sequence, that of the alanine tRNA

from *Saccharomyces cerevisiae*, could be produced, using an approach relying on enzymatic degradation of RNA fragments [125]. A similar approach additionally using 2-D fractionation and radioactive labeling was developed by Fred Sanger, which in 1972 was used to derive the first complete protein-coding gene sequence [126]. The first success of a sequencing-by-synthesis (SBS) based nucleotide sequencing method using DNA polymerase to add sequentially administered radiolabeled nucleotides and measuring incorporation was achieved in 1974 [127]. The two-dimensional separation of DNA molecules was soon after replaced by polyacrylamide gel electrophoresis, separating DNA fragments by their length, providing higher resolution than previous fractionation. In 1975, Allan Maxam and Walter Gilbert developed a method to chemically cleave DNA fragments at specific nucleotides [128]. By using different chemicals on radiolabeled DNA, gel electrophoresis could be used to determine fragment lengths and thus infer the nucleotide sequence [128]. However, it was again Sanger, whom with first the 'plus and minus' system (together with Alan Coulson) and in 1977 with the dideoxy 'chain-termination' technique again revolutionized nucleotide sequencing and established the base to currently still in use sequencing technologies [129]. Sanger's approach relied on four parallel DNA polymerase reaction, each mixing the four deoxynucleotides (dNTPs) with one species of radiolabeled dideoxynucleotide (ddNTPs) in low concentration. ddNTPs are incorporated into the synthesized DNA molecule, however they lack the 3' hydroxyl group to which the next dNTP is normally attached, resulting in an incomplete DNA sequence. Through the low concentration of the respective ddNTP, each reaction produces all possible incomplete DNA sequences, all ending with the respective nucleotide. By using gel electrophoresis and readout of the radioactive labeling, the positions of each nucleotide in the DNA sequence can be determined. By replacing radioactive labels with ddNTP-specific fluorophores, the synthesis-reaction could be performed in a single vessel, paving the way for using single a capillary for the electrophoresis, and subsequently for automation of this now called "Sanger sequencing" for DNA fragments of up to 1kb on the first commercially available DNA sequencing machines. The Sanger sequencing remains the gold-standard for diagnostic applications up until today.

The second generation of DNA sequencing, also widely termed 'next-generation sequencing' (NGS), mostly stuck to the concept of sequencing-by-synthesis, however the major change compared to the Sanger sequencing and Maxam-Gilbert approach was that instead of electrophoresis, now sequencing was performed attached to a solid phase, either on paramagnetic beads or on the surface of solid-phase flow cells. Two major technologies became widely used, one was the pyrosequencing technology by 454 (later bought by Roche), the other the Solexa method (later bought by Illumina).

The Solexa/Illumina approach uses adapter oligonucleotides fused to the targeted DNA sequences with complementing oligos bound to a flow cell to capture and amplify DNA fragments

in a process called 'bridge amplification'. Bridge amplification is followed by the actual sequencing process, which utilizes fluorophore-labeled modified terminator nucleotides (similar to the ddNTPs used in Sanger sequencing) of which in each cycle, one is incorporated and by using a laser the fluorophores are excited and can be read. Using bridge amplification, clonal clusters of DNA sequences are created on the flowcell, amplifying the optical signal for readout and also enabling paired-end sequencing of DNA fragments starting from both sides of the clonal sequence. Millions of these clonal clusters can be bound to a single flowcell, enabling massively parallel sequencing of DNA. After reading the incorporated nucleotide, the label is cleaved off, reverting the terminator nucleotide back to normal, ready for the incorporation of the next labeled nucleotide in the cycle. This technology is the basis of all current Illumina sequencing machines.

The Roche/454 pyrosequencing used an entirely new approach for sequence readout. Similar to the Illumina technology, DNA fragments were captured by complementing oligos and clonally amplified, however not on flowcells, but on beads. These beads were washed over a reaction plate with wells, each well large enough for a single bead, thus for the readout of a single DNA fragment. The sequencing itself was performed by sequentially washing the four dNTPs over the reaction plate. If the "correct" nucleotide was available for the DNA polymerase, it was incorporated into the molecule, releasing pyrophosphate (PPi). The amount of PPi released is correlated to the number of successive occurrences of the nucleotide, *i.e.* a single "A" would produce less PPi than a sequence of "AA" or "AAA". The PPi is in a two-step process first converted to adenosine triphosphate (ATP), which is then by the enzyme luciferin translated to a light signal, which can be read for all wells in the reaction plate in parallel, resulting in the DNA sequence readout. While homooligomeric stretches correlated well with the emitted light signal, this can turn into a problem with longer stretches of homopolymers, which has become one of the major concerns with pyrosequencing through multiple applications [130], [131].

Both technologies advanced with improvements of high-resolution imaging devices and microfabrication, increasing the number of parallel reactions possible in a single sequencing run. Additional sequencing technologies were developed alongside 454 pyrosequencing and Solexa sequencing in the second generation of DNA sequencing, however none could establish itself in the competition, and since the discontinuation of pyrosequencing in 2016, the Illumina technology established a de-facto monopoly for DNA sequencing.

The third generation of DNA sequencing is widely defined as technologies capable of single molecule sequencing, without the need of prior amplification. The first device with wide acceptance was developed by Pacific Biosciences, introducing the single molecule real time (SMRT) technology [132]. Briefly, this technology relies on tiny holes in a metallic film, these nanostructures are called zero-mode waveguides (ZMWs), each of them having a single DNA polymerase enzyme attached to the bottom of it. Through the physical properties of the ZMWs and the use of

fluorescence labeled dNTPs, only the single dNTP that is incorporated by the polymerase is detected by a sensor for each ZMW, enabling signal detection in the moment of attachment to the produced DNA molecule and in addition, through the kinetics of this process, DNA modifications, *i.e.* methylation, on the template molecule can be assessed in addition to the nucleotide sequence [133].

The second technology of the third generation with widespread application is based on nanopores and was introduced by Oxford Nanopore Technologies (ONT). Nanopore DNA sequencing is not sequencing-by-synthesis based, but measures the nucleotide-species specific changes in currents when a DNA molecule is passing through a pore [134]. Both technologies are capable to sequence large DNA fragments up to hundreds of kilobases, giving the possibility of sequencing long stretches of DNA, for example for genome assembly, however, sequence data quality does not reach that of Illumina's short reads, giving both generations of DNA sequencing its niche in biological and biomedical research with complementary applications.

For a more detailed discourse on the development of DNA sequencing technologies through time, see Heather and Chain [135].

1.4.2 Molecular methods in surveying microbial communities

Early analyses of microbial communities were entirely dependent on microscopy and culturing of bacterial isolates, followed by extensive metabolic profiling on selective media and phenotype description, *e.g.* by Gram staining. However especially for complex communities from anaerobic environments, such as the intestinal microbiota, culturing is laborious and favors easily growing organism to be over-represented, rendering this approach biased [136]. Other members might not be culturable at all using standard media, resulting in vast differences in counts of total cells and the number of colonies from culturing, a phenomenon termed the 'great plate count anomaly' [137].

The possibilities of analyzing the composition and members of microbial and especially bacterial communities have since changed tremendously. Before DNA sequencing became widely available and, most importantly, affordable, different methods were used to determine community composition, two of which were 'terminal restriction fragment length polymorphism' (T-RFLP) [138] analysis and 'denaturing gradient gel electrophoresis' (DGGE) [139]. In T-RFLP, a selected marker gene, for example the 16S ribosomal RNA (rRNA) gene, was amplified by polymerase chain reaction (PCR) from the extracted DNA of a microbial community of interest using a primer pair in which one of the primers was labeled with a fluorescent dye. The amplified DNA was subsequently treated with one or multiple restriction enzymes, cutting the DNA at sites with a specific sequence, with the assumption that each bacterium in the community will result in a

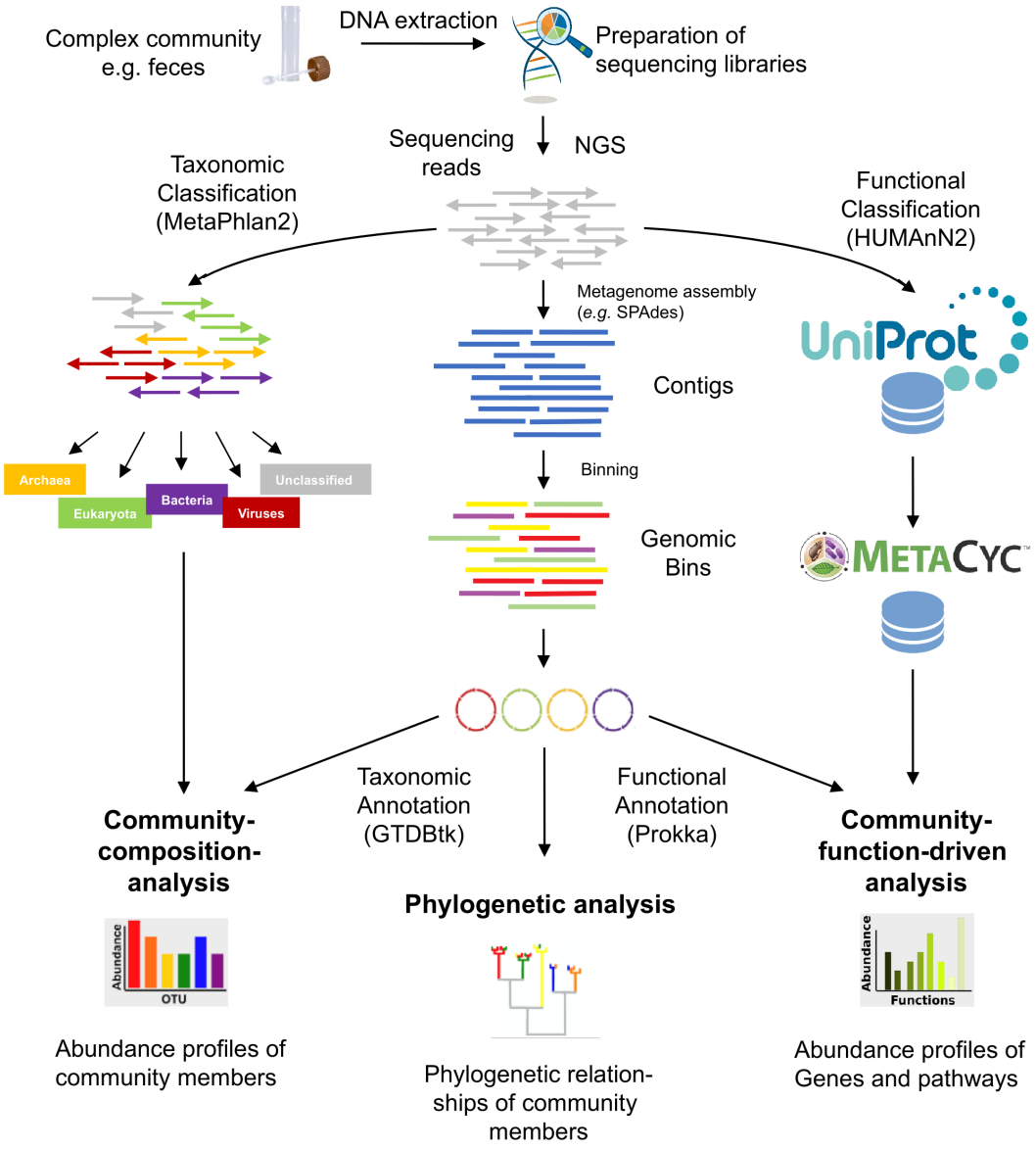


Figure 1.2: Overview of possible workflows for metagenome sequencing data. DNA extraction is followed by Illumina sequencing, resulting in short paired-end reads. After quality control of the data, these can be used for different processing branches. Left: The MetaPhlan2 pipeline [172] can used for relative abundance estimation of taxonomic groups in the community. Right: The HUMAnN pipeline [174] uses the UniProt database for read annotation and downstream mapping to the MetaCyc database for metabolic pathway abundance inference. Center: Sequencing data can be used for the assembly of larger fragments (contigs) of micorbiome community members (*e.g.* with SPAdes [175]). Using different properties of the contigs, bins can be created for the reconstruction of microbial genomes. These bins can be annotated for taxonomy (*e.g.* GTDBtk [167]) and functional capacity (*e.g.* Prokka [180]). Downstream analysis, depending on the chosen data processing steps, can include the analysis of taxonomic community composition, analysis of community function and the reconstruction of phylogenetic patterns of individual members of the community.

labeled fragment of a specific size. Using electrophoresis, the fragments were separated by size and fluorescence intensity could be used for quantitative estimations of community composition.

Just like T-RFLP, DGGE also relies on the generation of marker-gene amplicons from the community under investigation, however here, amplicon libraries were directly subjected to gel electrophoresis. The special properties of the used gel, using gradients of DNA denaturing agents, for example urea and formamide, resulted in differential properties of DNA migration before its denaturation, depending on the taxon-specific amplicon sequence. The resulting 'fingerprint' could then be used to determine the presence of a specific bacterial group in a community in comparison to other communities analyzed alongside. For both methods, a comparison of the resulting fingerprints with known restriction fragments or migration bands from reference bacteria could be used to infer taxonomic information of the community composition.

Advances in the taxonomic classification were crucial for the success of the use of DNA sequencing for community analysis. Especially the 16S rRNA gene was shown to be of high value. The 16S rRNA is an RNA molecule of ~1,500 bases and part of the prokaryotic ribosome. The 16S rRNA consists of conserved regions which assemble to a secondary structure with four domains, however in between conserved stretches, there are nine hypervariable regions (V1-V9) in which mutations can occur and can serve as molecular clock [140]. This mixture of conservation and variability can be used to infer phylogenetic relationships and taxonomic labels for microorganisms [141]. The Sanger sequencing technology soon led to databases of taxon-specific rRNA gene sequences, which could be used for an accurate estimation of community structure from DNA sequencing efforts of amplicon clone-libraries generated from environmental [142] and human associated microbial communities [143], highlighting the advantage over culture-based methods. The second generation of DNA sequencing technologies quickly increased the throughput and analysis depth of the analyzed communities and revolutionized the whole field of molecular community ecology, again, giving access to a so far underexplored "rare biosphere" [144]. Especially the pyrosequencing method (see previous section for details) quickly gained attention, as it combined high throughput (hundreds of thousands of sequences) with long reads (up to 1kb in later instruments, such as he GS FLX+), which made it valuable for also health related analysis of human associated microbes using 16S rRNA gene amplicons [145]. As interest in sequence-based analysis of communities increased, the technologic shortcomings of pyrosequencing regarding homopolymers became evident. Especially the Illumina technology proved to deliver the high quality data needed for the assessment of microbial communities [146], ultimately leading to a transition of the field to this technology, supported by well-established protocols developed for the earth microbiome project [147] and reduced costs by advanced multiplexing strategies [148]. The balance between high-quality data and low per-sample costs renders Illumina-based 16S rRNA amplicon sequencing the go-to technique for microbial community survey, with efforts aiming at

further decrease of sequencing cost still ongoing [149]. While amplicon sequencing is most popular for the assessment of bacterial communities using the 16S rRNA gene, complementary approaches are available targeting also archaea (16S rRNA gene, [38], [150]), fungi (18S rRNA gene or internally transcribed spacers (ITS), [151]) and other eukaryotes [152].

The generation and sequencing of "shotgun metagenomic libraries", meaning the untargeted sequencing of DNA fragments from a sample, has seen trials since almost two decades [153]. However, only recent advances in both sequencing technology and the development of novel algorithms to handle such data, combined with increased computational and storage capacities and decreasing costs opened this field to a broader range of scientist [76]. Shotgun metagenomics promise an unbiased representation of all members of microbial communities, not restricted to prokaryotes alone, but also potentially including eukaryotic microorganisms and viruses/phages, excluding RNA viruses which need additional enrichment and library preparation steps [154]. In addition to a taxonomic description of community composition, possible on a finer scale than 16S rRNA gene amplicon based approaches, shotgun metagenomics can be used for a description of the functional capacity of the community, capturing even strain-level genetic diversity of the community members [155], [156]. Recent studies of communities using shotgun metagenomics increasingly incorporate also data from high-throughput long-read technologies (see previous section for details) for increased quality of metagenome assembled genomes (MAGs) of the community members [157].

While these culture-independent approaches for the assessment of microbial communities were in the focus of microbial community ecology, also culture-based analyses have recently regained attention, as novel techniques for high-throughput selective culturing in combination with genomic sequencing of isolates have proven to be applicable and valuable for healthcare-based applications [158], paving the way for the emerging field of "culturomics" [159].

1.4.3 Molecular community ecology: microbiome analysis

Data processing steps before statistical analysis differ between amplicon-based sequencing data and shotgun metagenomic data. These steps depend on multiple factors, including also personal preferences of the data analyst.

For 16S rRNA gene amplicon data, multiple software collections have established themselves among researches, the most widely used are certainly MOTHUR (current version 1.44; [160]) and QIIME ("Quanitative Insights Into Microbial Ecology"; current version QIIME 2 v2019.10; [161], [162]), both aiming to provide out-of-the-box working data processing frameworks for the processing of amplicon data. While these frameworks are highly accessible for beginners in sequencing data processing, they might come at the cost of flexibility, a feature that is implemented in the USEARCH software (current version v11; [163]), which is also heavily focused on the

development of new algorithms for different steps in data processing. However, workflows usually follow the same succession of steps for state-of-the-art Illumina paired-end sequencing data: In a first step, the raw sequencing data is cleaned by removing low-quality parts of the sequence or discarding sequencing reads below a defined threshold altogether. Usually, the read length chosen for sequencing is adapted to the amplicon of choice, so that after initial quality control the read-pairs are overlapping and can be 'stitched' to form a complete amplicon sequence, with the same starting- end endpoint for all sequenced fragments. Widely used 16S rRNA gene amplicons are V1-V2 (primer pair: 27F-338R; amplicon size ~312bp; all amplicon sizes vary slightly between bacteria), V3-V4 (*e.g.* 341F-806R; ~450bp) and V4 (515-806R; ~292bp) (Johnson 2019).

Based on an assumption from 1994 [164], the clean amplicon sequences are binned into clusters of >97% sequence similarity, representing operational taxonomic units (OTUs) with the resolution of bacterial species, which could subsequently be subjected for taxonomic annotation. Several independent 16S rRNA sequence databases for annotation are available, the most widely used being GreenGenes [165], SILVA [166], the Ribosomal Database Project (RDP) [167] and the newly established Genome Taxonomy Database (GTDB) [168]. Recent estimates using up-to-date databases suggested the 97%-threshold is not representing species-level bins and should be corrected [169], which sparked the creation of new algorithm that by error correction models try to infer exact amplicon sequence variants (ASVs), such as DADA2 [170], UNOISE2 [171] and Deblur [172]. Independent of resolution, the processed sequencing data can be used to construct sample-by-feature count abundance tables from ASV/OTU-level and for bins of taxonomic groups on different taxonomic levels, which are the basis for the statistical analysis.

Just like for amplicon-based sequening data, metagenomic data can be processed in manifold ways. An exemplary workflow is depicted in Figure 1.2. Breifly, after DNA extraction from the sample, sequencing library preparation, short-read seqencing and initial quality control based on sequence quality data, the reads can be subjected to multiple independent, but partially orthogonal analysis workflows. Popular database-driven tools like MetaPhlan2 [173], Kraken [174] and HUMAnN2 [175] can be used to estimate taxonomic and functional composition of the community in the analysis directly from the annotation of the short sequencing reads. Alternatively, the reads can be used in *de novo* assembler software, *e.g.* SPAdes [176] or MEGAHIT [177], for the creation of contigs, meaning genomic stretches of the microorganisms in the community. Subsequently, genomic bins are created identifying and sorting contigs that belong to the same organism, *e.g.* with MaxBin2 [178] or MetaBAT2 [179]. If a genomic bin meets a set of defined criteria for completeness and contamination (CheckM) [180], it is called a metagenome assembled genome (MAG). Abundance of a MAG in a community can be estimated from its coverage and annotation of the MAG can be performed for taxonomy (*e.g.* GTDBtk) [168] and functional capacity (Prokka) [181]. Direct short read annotation is usually faster than assembly-

based approaches and working well for deeply studied communities. However, they might fall short for community members that are less well known or possibly novel. The information in MAGs can be valuable for the analysis of phylogenies, however assemblies will fail for low abundant microorganisms with too little coverage in the sequencing library.

Although technological advances massively increased sample processing throughput and data generation from microbial communities, the statistical methods used for the analysis are still largely based on methods developed for questions in classical ecology since the 1950s and even earlier [182], thus also terminology in modern numerical community ecology applied in microbiome analysis remains the same.

An analysis of a microbial community with regard to a trait, for example in a case/control setting or the influence of an environmental variable, is usually performed on three major levels: alpha diversity, beta diversity and the assessment of differential feature abundance, with 'features' being for example ASVs/OTUs, taxonomic groups or also microbial genes identified from metagenomic sequencing.

Alpha diversity (intra-individual or within-sample diversity), is a measure of community composition of a single community. Multiple measures for alpha diversity are available, each of them with different emphasis on specific properties of the community. One of the simplest measures for alpha diversity is the *species richness* (S), which is the number of distinct features present in a community. More complex estimates of alpha diversity include the *Shannon index* (H') [183], which is based in information theory and reaching maximum in communities with equal abundances of all features, the *Simpson index* (D) [184], emphasizing the influence of low-abundant features, or Faith's *phylogenetic diversity* (PD) [185], considering also the evolutionary relationships between features using a phylogenetic tree. Univariate tests can be used to assess differences in alpha diversity between discrete groups, depending on the properties of the chosen diversity's distribution, for example by non-parametric (*e.g.* Wilcoxon rank sum test) or parametric (Student's T-test) tests.

Beta diversity (inter-individual or between-sample diversity), is a measure for differences between two microbial communities, which can be calculated for all pairs of samples in the analysis. Like alpha diversity, beta diversity can be calculated based on different assumptions and it can take values between 0 (= both communities are equal with regard to the chosen diversity measure) and 1 (= both communities are maximally dissimilar). The *Jaccard similarity coefficient* (J) [186] is an unweighted measure, meaning it only accounts for presence or absence of features, but not their abundance, and is defined as the number of features shared by two communities, divided by the total number of features present in both samples combined. As diversities are usually measured as distances and not similarities, the *Jaccard distance* (d_J) is defined as $d_J = 1 - J$. One of the most popular weighted measure for beta diversity is the *Bray-Curtis dissimilarity* (BC) [187], which is

defined as the sum of relative feature abundances shared by two communities. Like in alpha diversity, also in beta diversity modern measures have been developed to incorporate phylogenetic information in the distance measure, such as the *UniFrac distance* [188], which is defined as the branches of a phylogenetic tree shared between two communities, and can be calculated based on presence and absence of branches (unweighted UniFrac) or considering the branch abundances (weighted UniFrac). The analysis of community differences using beta diversities can be performed in multiple ways. Popular choices are non-parametric permutational multivariate analysis of variance [189] or distance-based redundancy analysis (db-RDA) [190], a technique of constrained ordination, with multiple flavors and derivatives existing for both models.

The analysis of **differential feature abundance** can be used to identify microbial features with changes in prevalence or abundance between groups (*e.g.* healthy vs. disease) or connected to quantitative measurements (*e.g.* blood parameters or inflammation biomarkers). When analyzing microbiome data for differences between groups, one important thing to consider is the high sparsity seen in such datasets, meaning the high amounts of features with a value of zero. This property of zero-inflation exhibited by features investigated in microbiome research in combination with the usually non-normal distribution of values need to be accounted for when choosing the correct models for the analysis, as model assumptions of simple linear regression are violated in these data [191], [192].

While these analyses are central to many studies performed in the field of microbiome research, they can be extended by a multitude of additional approaches, for example the analysis of co-abundance and co-occurrence networks of microbial features [193] or the application of principles of compositional data analysis (CoDa) for microbiome data [194], giving the opportunity of exhaustive analysis, tailored to the question under investigation.

1.4.4 Genome-wide association analysis

With the publication of the first draft sequence of a human genome in 2001 [195], new opportunities opened for research of human genetics and it was through the efforts of the *International HapMap Consortium* [196] that detailed information about the block-like structure in the inheritance of genetic information. The general idea behind genome-wide association studies (GWAS) depends on the fact that recombination in cell division does not happen by pure chance, but a relatively specific position, creating stretches of DNA and the variation within to be inherited together in 'haplotypes'. These haplotypes are able to be identified by the allelic state of one or more tagging genetic variants that are in strong linkage disequilibrium (LD) with the surrounding, making it possible to use these tags to extract information about much larger proportions of the genome, which guided the development of comparably cheap and highly informative genotyping arrays for large scale genetic association studies.

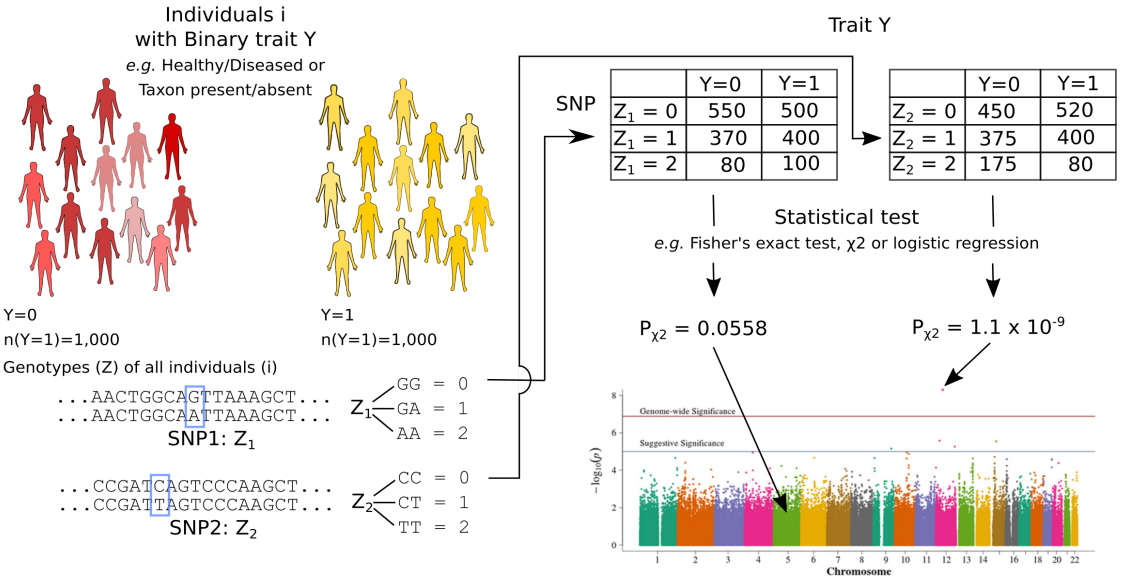


Figure 1.3 Schematic overview of a GWAS. Individuals (i) are assigned to groups or recruited based on a binary trait (Y). Genotyping information for genetic variants (Z) is acquired for example by genotyping. Genotype distributions for each SNP between traits are analyzed with statistical tests. Distribution of resulting P-values for all variants in the genome can be visualized using a Manhattan plot.

The first GWAS based on a genotyping array using chip technology to type 500,000 common single nucleotide polymorphisms (SNPs = genetic variants with a frequency of the less common (minor) allele in a population above a certain threshold, *e.g.* 1%) was published in 2007 [197], identifying 24 independent genetic loci associated with six diseases in 17,000 individuals. Figure 1.3 summarizes the steps of a GWAS. Briefly, all individuals included (i) are coded into binary phenotypes (Y) as being either affected by the trait under investigation (Y_i=1) or not (Y_i=0). Subsequently, for each genotyped SNP (Z_j) in an individual, the two possible alleles are coded as 0 (usually the more common (major) allele) and 1 (minor allele) and the occurrences of the minor allele per individual are counted, so that in a diploid genome, like the human genome, $Z_{j,i}$ can take the values $Z_{j,i}$=0 for individuals homozygous for the major allele, $Z_{j,i}$=2 for individuals homozygous for the minor allele and Z_i=1 for heterozygous individuals. Doing this for all individuals, this results in a 2x3-field contingency table per SNP which can be analyzed using Fisher's exact test or Chi-squared (χ^2) test for whether a change in allele frequency can be seen in association with the trait under investigation. More sophisticated and powerful statistical methods can be used to investigate associations, for example logistic regression, which allow the inclusion of potentially influencing covariates to be controlled for in the analysis, for example the individuals' age and sex or population stratification [198]. Further developments in genome-wide association analysis led to the possibilities to not only investigate binary traits, but to identify genetic loci influencing quantitative traits, like body height or blood pressure (quantitative trait loci, QTLs) [199], which can also be tissue specific genetic influences on gene expression (expression QTLs, eQTLs) [199].

Since the introduction of GWAS, the method has quickly gained popularity for the assessment of genetic risk factors for a large variety of traits, especially for complex diseases, as the large sample number of samples included in GWAS can identify variants that have low individual effect size, however in sum can explain substantial amounts of heritable risk [200]. Through ongoing large-scale biobanking efforts, for example by the UK Biobank assessing deep phenotypic information and genotypes of 500,000 individuals [201], also rare diseases and rare genetic variants are accessible for analysis, with the largest association studies recently having superseded the mark of 1,000,000 individuals [202]. As mentioned earlier, individual effects identified in GWAS are usually small, needing additional post-GWAS analyses for the prioritization and interpretation of the results. These analyses can consist of different types of gene set enrichment analysis [203] to identify biological pathways or tissue specific effects or summary statistics can be used in Mendelian randomization (MR), to infer directional causal relationships between phenotypes [204], [205].

Briefly, MR mimics the design of a randomized controlled trial, using individual SNPs as instrumental variables. By observing the individuals' allelic status at a certain genetic locus, this assigns the individuals 'randomly' to control/placebo and treatment groups. Due to the nature of GWAS as population-based studies, this assumes equal, and thus negligible, exposure to potential confounders in all groups. By now looking at group dependent differences of one trait (outcome) depending on another trait (defined as exposure or risk factor), this can help to infer causal relationships[204]. Using 2-sample MR (2MR), effects of one population sample can be set into relation to a different independent populations sample [206], enabling to use summary statistics from a newly conduced GWAS together with summary statistics of previously published results to infer causal relationships of traits, using one as exposure and the other as outcome.

On the website of the regularly updated and curated *GWAS Catalog* (https://www.ebi.ac.uk/gwas/home) [207], results from GWAS are summarized and can be queried. As of February 2020, the summary statistics from 4,410 GWAS publications are available via the GWAS Catalog, totaling 172,351 SNP-versus-trait associations.

1.5 Publications and main findings

Table 2: Overview of publications forming the backbone of this thesis, summarizing the chapter (C) the respective article appears in, the article short reference (A-G), as well as aim, design, results and, conclusion drawn from each study.

C	Art.	Aim	Design	Results	Conclusion
2	A	Elucidating technical aspects in the analysis of metaorganisms.	Survey of 50 samples acquired from ten host-associated microbial communities using four 16S rRNA gene amplicon-based approaches and shotgun metagenomics.	Taxonomic similarities between amplicon-based approaches and shotgun sequencing. Functional profiles inferred from amplicons differ from shotgun results.	Choice of method depends on the community. Amplicons are useful tool next to shotgun sequencing.
3	B	Host-genetic effects on gut microbiome beta diversity.	Permutation-free distance-based F (DBF) test applied to microbiome data of 1,767 individuals from Northern Germany.	Four genomic loci replicate in two cohorts (P<0.05) and are genome-wider significant in meta analysis ($P<5 \times 10^{-8}$).	Genes involved in immune cell differentiation (*CPEB4*) and mucosal integrity (*NHE8*) influence gut microbiome.
	C	Host-genetic effects on gut microbiome beta diversity and individual microbial features.	DBF test and regression models applied to 8,956 individuals in five cohorts from Norther, North-Eastern and Southern Germany.	Variants in 32 loci show genome-wide significant association with microbiome in meta analysis ($P<5 \times 10^{-8}$).	ABO histo-blood group system is a potential modulator of human-associated microbial communities. Genes involved in TLR4-dependent immune regulation (*BLVRA*) have effect on commensals.
4	D	Replication of PSC-associated signals in gut microbiome.	Analysis of fecal microbiome of PSC (n=73), UC (n=88= and HC (n=98) individuals from Germany	Signatures found in Norwegian cohort partly replicate in affected Germans, especially *Veillonella* (PSC ↑) and *Coprococcus* (PSC ↓). Microbial profiles can be used to separate HC and PSC.	Signals found in Norwegian PSC cohort replicate in German cohort.
	E	Meta-analysis of PSC-associated microbiota of German and Norwegian individuals.	Analysis of fecal microbiome of PSC (n=137), UC (n= 118) and HC (n=133) individuals from Germany and Norway.	PSC patients differ in community composition and individual taxonomic groups from HC. Signals from previous studies replicate. New associations found with *Parabacteroides* (PSC ↑) and Protebacteria (PSC ↑). Widespread loss of taxa in association with PSC.	PSC-associated microbiota differs from HC, independent from country of origin of the sample. Associations are thus independent of environmental influence.
	F	Analysis of PSC-associated mycobiome.	Sequencing and analysis of ITS2 amplicon for fungal communities in PSC (n=65), UC (n=38) and HC (n=66) individuals from Germany.	Significant changes in the relative abundance of *Candida* and *Humicola/Trichocladium griseum* associated with PSC.	In additoin to bacteriome, also the mycobiome is altered in PSC.
	G	Identification of factors altering the skin microbiota.	Sampling of four nonlesional body sites in ten individuals with and without atopic dermatitis (AD), plus lesions in AD patients. Characterization of microbiome, lipid composition and fillagrin (*FLG*) mutations.	AD patients have higher relative abundances of Staphylococcus spp., also at non-affected body sites. Long-chain unsaturated fatty acid profiles differ. *FLG* influences microbiome composition.	Epidermal barrier integrity affects microbiome composition. Microbiome is altered in association with AD, independent of lesions.

C: Chapter; Art.: Article; HC: healthy controls; PSC: primary sclerosing cholangitis; UC: ulcerative colitis

2 Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms

Publication:

Philipp Rausch*, **Malte Rühlemann***, Britt M. Hermes, Shauni Doms, Tal Dagan, Katja Dierking, Hanna Domin, Sebastian Fraune, Jakob von Frieling, Ute Hentschel, Femke-Anouska Heinsen, Marc Höppner, Martin T. Jahn, Cornelia Jaspers, Kohar Annie B. Kissoyan, Daniela Langfeldt, Ateequr Rehman, Thorsten B. H. Reusch, Thomas Roeder, Ruth A. Schmitz, Hinrich Schulenburg, Ryszard Soluch, Felix Sommer, Eva Stukenbrock, Nancy Weiland-Bräuer, Philip Rosenstiel, Andre Franke, Thomas Bosch and John F. Baines: "Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms". *Microbiome*. 2019;7:133.

Article A: Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms

Rausch *et al. Microbiome* (2019) 7:133
https://doi.org/10.1186/s40168-019-0743-1

Microbiome

RESEARCH Open Access

Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms

Philipp Rausch[1,2,3*†], Malte Rühlemann[4†], Britt M. Hermes[1,2,5], Shauni Doms[1,2], Tal Dagan[6], Katja Dierking[7], Hanna Domin[8], Sebastian Fraune[8], Jakob von Frieling[9], Ute Hentschel[10,11], Femke-Anouska Heinsen[4], Marc Höppner[4], Martin T. Jahn[10], Cornelia Jaspers[11,12], Kohar Annie B. Kissoyan[7], Daniela Langfeldt[6], Ateequr Rehman[4], Thorsten B. H. Reusch[11,12], Thomas Roeder[9], Ruth A. Schmitz[6], Hinrich Schulenburg[7], Ryszard Soluch[6], Felix Sommer[4], Eva Stukenbrock[13,14], Nancy Weiland-Bräuer[6], Philip Rosenstiel[4], Andre Franke[4], Thomas Bosch[8] and John F. Baines[1,2*]

Abstract

Background: The interplay between hosts and their associated microbiome is now recognized as a fundamental basis of the ecology, evolution, and development of both players. These interdependencies inspired a new view of multicellular organisms as "metaorganisms." The goal of the Collaborative Research Center "Origin and Function of Metaorganisms" is to understand why and how microbial communities form long-term associations with hosts from diverse taxonomic groups, ranging from sponges to humans in addition to plants.

Methods: In order to optimize the choice of analysis procedures, which may differ according to the host organism and question at hand, we systematically compared the two main technical approaches for profiling microbial communities, 16S rRNA gene amplicon and metagenomic shotgun sequencing across our panel of ten host taxa. This includes two commonly used 16S rRNA gene regions and two amplification procedures, thus totaling five different microbial profiles per host sample.

Conclusion: While 16S rRNA gene-based analyses are subject to much skepticism, we demonstrate that many aspects of bacterial community characterization are consistent across methods. The resulting insight facilitates the selection of appropriate methods across a wide range of host taxa. Overall, we recommend single- over multi-step amplification procedures, and although exceptions and trade-offs exist, the V3 V4 over the V1 V2 region of the 16S rRNA gene. Finally, by contrasting taxonomic and functional profiles and performing phylogenetic analysis, we provide important and novel insight into broad evolutionary patterns among metaorganisms, whereby the transition of animals from an aquatic to a terrestrial habitat marks a major event in the evolution of host-associated microbial composition.

Keywords: Animal microbiome, Evolution, Phylosymbiosis, Holobiont, Metaorganism

* Correspondence: philipp.rausch@bio.ku.dk; baines@evolbio.mpg.de
†Philipp Rausch and Malte Rühlemann contributed equally to this work.
[1]Evolutionary Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany
Full list of author information is available at the end of the article

Background

Dynamic host-microbe interactions have shaped the evolution of life. Virtually all plants and animals are colonized by an interdependent complex of microorganisms, and there is growing recognition that the biological processes of hosts and their associated microbial communities function in tandem, often as biological partners comprising a collective entity known as the metaorganism [1]. For instance, symbiotic bacteria contribute to host health and development in critical ways, ranging from nutrient metabolism to regulating whole life cycles [2] and in turn benefit from habitats and resources the host provides. Moreover, it is well established that perturbations of the microbiome likely play an important role in many host disease states [3]. However, researchers have yet to elucidate the mechanisms driving these interactions, as the exact molecular and cellular processes are only poorly understood.

An integrated view on the metaorganism encompasses a cross-disciplinary approach that addresses how and why microbial communities form long-term associations with their hosts. Despite widespread agreement that the interdependencies of microbes and their hosts warrant study, there remains considerable incongruity between researchers regarding the best methodologies to study host-microbe interactions. The development of standardized protocols for characterizing and analyzing host-associated microbiomes across the tree of life is thus crucial to understand the evolution and function of metaorganisms without the issues of technical inconsistencies or data quality.

The rapidly growing interest in microbiome research has been bolstered by the ability to profile diverse microbial communities using next-generation sequencing (NGS). This culture-free, high-throughput technology enables identification and comparison of entire microbial communities, so-called metagenomics [4]. Metagenomics typically encompasses two particular sequencing strategies: amplicon sequencing, most often of the 16S rRNA gene as a phylogenetic marker; or shotgun sequencing, which captures the complete breadth of DNA within a sample [4].

The use of the 16S ribosomal RNA gene as a phylogenetic marker has proven to be an efficient and cost-effective strategy for microbiome analysis and even allows for the imputation of functional content based on taxon abundances [5]. However, PCR-based phylogenetic marker protocols are vulnerable to biases through sample preparation and sequencing errors. The choice of which hypervariable regions of the 16S rRNA gene are targeted for sequencing seems to be among the biggest factors underlying technical differences in microbiome composition [6–8]. Furthermore, 16S rRNA gene amplicon sequencing is typically limited to taxonomic classification at the genus level depending on the database and classifiers used [9], and provides only limited functional information [5]. These well-recognized limitations of amplicon-based microbial community analyses have raised concerns about the accuracy and reproducibility of 16S rRNA phylogenetic marker studies and have led to an increased interest in developing more reliable methods for amplicon library preparation and sequencing [8, 10].

Shotgun metagenomics, on the other hand, offers the advantage of species- and strain-level classification of bacteria. Additionally, it allows researchers to examine the functional relationships between hosts and bacteria by determining the functional content of samples directly [9, 11], and enables the exploration of yet unknown microbial life that would otherwise remain unclassifiable [12]. However, the relatively high costs of shotgun metagenomics and more demanding bioinformatic requirements have precluded its use for microbiome analysis on a wide scale [4, 9].

In this study, we set out to systematically compare experimental and analytical aspects of the two main technical approaches for microbial communities profiling, 16S rRNA gene amplicon and shotgun sequencing, across a diverse array of host species studied in the Collaborative Research Center 1182, "Origin and Function of Metaorganisms." The ten host species range from basal aquatic metazoans [*Aplysina aerophoba* (sponge) and *Mnemiopsis leidyi* (comb jelly)]; to marine and limnic cnidarians (*Aurelia aurita, Nematostella vectensis, Hydra vulgaris)*, standard vertebrate (*Mus musculus*), and invertebrate model organisms (*Drosophila melanogaster*, *Caenorhabditis elegans*); to *Homo sapiens*; and in addition to wheat (*Triticum aestivum*) and a standardized mock community. This setup provides a breadth of samples in terms of taxonomic composition and diversity. Conducting standardized data generation procedures on these diverse samples on the one hand provides a unique and powerful opportunity to systematically compare alternative methods, which display considerable heterogeneity in performance. On the other hand, this information enables researchers working on these or similar host species to choose the experimental (e.g., hypervariable region) or analytical pipelines that best suit their needs, which will be a valuable resource to the greater community of host-microbe researchers. Finally, we identified a number of interesting, broad-scale patterns contrasting the aquatic and terrestrial environment of metaorganisms, which also reflect their evolutionary trajectories.

Results

Our panel of hosts includes ten species, for which five biological replicates each were included (see

Additional file 1: Figure S1). The majority of hosts are metazoans, including the "golden sponge" (*Aplysina aerophoba)*, moon jellyfish (*Aurelia aurita*), comb jellyfish (*Mnemiopsis leidyi*), starlet sea anemone (*Nematostella vectensis)*, fresh-water polyp *Hydra vulgaris*, roundworm (*Ceanorhabditis elegans*), fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*), human (*Homo sapiens*), and the inclusion of wheat (*Triticum aestivum*), which can serve as an outgroup to the metazoan taxa. *Drosophila melanogaster* was additionally sampled using two different methods targeting feces and intestinal tissue. Nucleic acid extraction procedures were conducted according to the needs of the individual host species (see the "Methods" section and Additional file 1), after which all DNA templates were subjected to a standard panel of sequencing procedures. For 16S rRNA gene amplicon sequencing, we used primers flanking two commonly used variable regions, the V1 V2 and V3 V4 regions. Further, for each region, we compared a single-step fusion-primer PCR to a two-step procedure designed to improve the accuracy of amplicon-based studies [8]. Finally, all samples were also subjected to shotgun sequencing, such that five different sequence profiles were generated for each sample. While a single classification pipeline was employed for all four 16S rRNA gene amplicon sequence profiles, community composition based on shotgun data was evaluated using MEGAN [13], due to the advantage of simultaneously performing taxonomical and functional classification of shotgun reads and an overall good performance (for additional description, see Additional file 1).

Performance of data processing and quality control

All data generated from amplicons were subject to the same stringent quality control pipeline including read-trimming, merging of forward and reverse reads, quality filtering based on sequence quality and estimated errors, and chimera removal (see the "Methods" section). The one-step V1 V2 amplicon data showed the highest rate of read-survival (62.13 ± 23.90%; mean ± sd) followed by the corresponding two-step method (49.85 ± 23.90%; mean ± sd), in large part due to the greater coverage of this comparatively shorter amplicon (~ 312 bp). In contrast, 42.02 ± 16.41% and 36.88 ± 23.89% of the total reads were included in downstream analysis for the one-step and two-step V3 V4 data, respectively. The longer V3 V4 amplicon (~ 470 bp) was more affected by drops in quality at the end of the reads, which decreases the overlap of forward and reverse reads and thus increases the chances of sequencing errors (Additional file 1: Figure S2; for final sample sizes, see Additional file 2: Table S1). Overall, aside from chimera removal, each quality control step resulted in a comparatively greater loss of V3 V4 compared to V1 V2 data. On the other hand, the V3 V4 one-step method yields the lowest number of chimeras, suggesting a lower rate of chimera formation and/or detection in this approach (variable region $F_{1,214} = 3.8881$, $P = 0.0499$; PCR protocol $F_{1,214} = 8.1751$, $P = 0.0047$; variable region × PCR protocol $F_{1,214} = 6.4733$, $P = 0.0117$; linear mixed model with organism as random factor). Among all host taxa, we observe the highest proportion of retained reads in the V1 V2 one-step method and the lowest in the V3 V4 two-step method (Additional file 1: Figure S2B; variable region $F_{1,215} = 74.9989$, $P < 0.0001$; PCR protocol $F_{1,215} = 21.0743$, $P < 0.0001$; linear mixed model with organism as random factor). After quality filtering and the identification of bacterial reads, an average of 0.46 Gb of shotgun reads per sample was achieved (range 0.03–2.1 Gb) (Additional file 1: Figure S3A; for final sample sizes, see Additional file 2: Table S1). To provide an initial assessment and comparison between the amplicon and shotgun-based techniques, we plotted the discovered classifiable taxa and functions for the entire pooled dataset. Although the methods differ distinctly, each method shows a plateau in the number of discovered entities (see Additional file 1: Figures S3C, S3D).

Mock community

The analysis of standardized mock communities is an important measure to ensure general quality standards in microbial community analysis. In this study, we employed a commercially available mixture of eight bacterial and two yeast species. Comparison among the amplification procedures (one- and two-step PCR), 16S rRNA gene regions (V1 V2, V3 V4), and shotgun data reveals varying degrees of similarity to the expected microbial community composition (Fig. 1). One discrepancy is apparent due to the misclassification of *Escherichia*/*Shigella*, whose close relationship makes delineation at the genus level difficult based on the V1 V2 region and is subsequently classified to *Enterobacteriaceae* (Fig. 1a, Additional file 1: Figure S4). Classification of this bacterial group also differs based on the shotgun analysis employed, due to different naming and taxonomic standards of the respective databases (*Escherichia*, *Shigella*, and *Enterobacteriaceae* refer to the *Escherichia*/*Shigella* cluster) [14]. However, overall, the amplicon-based profiles show the closest matches to the expected community. The V3 V4 one-step method shows the lowest degree of deviation between observed and expected abundances of the focus taxa (Table 1; Additional file 1: Figure S4). In addition, the relative abundances of fungi in the mock community were relatively well predicted by MEGAN (see Fig. 1).

Next, we evaluated alpha and beta diversity across the different technical and analytical methods. Interestingly, most methods overestimate taxon richness but

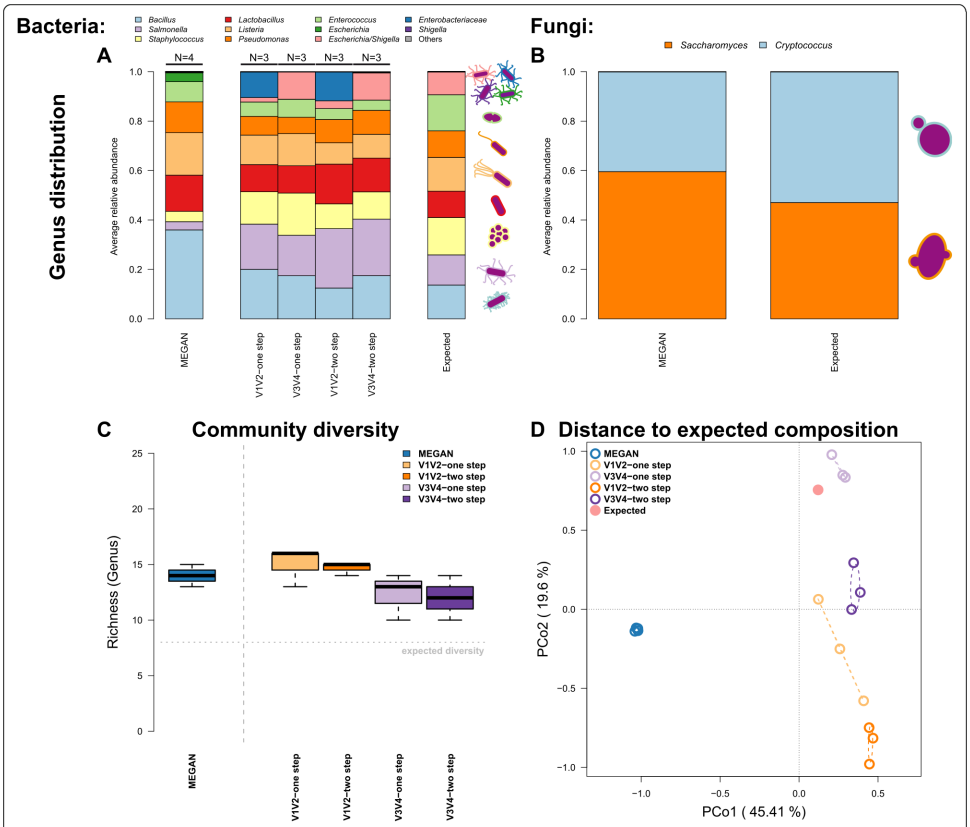


Fig. 1 Average community composition of bacteria (**a**) and fungi (**b**) in the mock community samples sequenced via metagenomic shotgun and 16S rRNA gene amplicon techniques (amplicon: V1 V2, V3 V4, one-step, two-step; shotgun: MEGAN based classification (short reads)). **c** Bacterial genus-level alpha diversity estimates in comparison to the expected community value. **d** Principle coordinate analysis of the Bray-Curtis distance between methods and the expected community. Ellipses represent standard deviations of points within the respective groups. Sample sizes for the different approaches are $N_{shotgun} = 4$, $N_{V1V2\text{-}one\text{-}step} = 3$, $N_{V1V2\text{-}two\text{-}step} = 3$, $N_{V3V4\text{-}one\text{-}step} = 3$, and $N_{V1V2\text{-}two\text{-}step} = 3$

underestimate complexity (as measured by the Shannon index) of the mock community, which could reflect biases arising from grouping taxon abundances based on slightly differing taxonomies (Fig. 1c, Additional file 1: Figures S4, S5A and Additional file 2: Table S2). Overall, the amplicon methods appear to more accurately reflect alpha diversity, although significant differences are present with regard to the amplified region (species richness: variable region $F_{1,10} = 6.3657$, $P = 0.0302$; Shannon H: method $F_{1,9} = 3.330$, $P = 0.1014$, variable region $F_{1,9} = 6.110$, $P = 0.0354$; ANOVA best model). With regard to beta diversity, the largest distance to the expected composition is observed for the shotgun-based data, while the amplicon-based techniques, in particular V3 V4, show the lowest distance (Fig. 1d, Additional file 1: Figure S5B). Pairwise tests show almost no differences between the amplicon-based techniques, while the shotgun-based data significantly differs from all amplicon profiles (Additional file 2: Table S3). Thus, in conclusion, shotgun-based analysis yields a higher degree of error compared to the amplicon-based approaches for the simple mock community used in our study.

Taxonomic diversity within and between hosts

To evaluate the performance of our panel of metagenomic methods over the range of complex host-associated communities in our consortium, we next employed a series of alpha and beta diversity analyses to

Table 1 Differences between expected and observed genus abundances in the mock communities ($N_{shotgun} = 4$, $N_{amplicon} = 3$) via a one-sample *t* test (two-sided) of relative abundances (*P* values are adjusted via Hommel procedure)

	Shotgun	Amplicon			
Members mock community	MEGAN	V1 V2 one-step	V3 V4 one-step	V1 V2 two-step	V3 V4 two-step
Staphylococcus	0.00002	0.52446	0.09200	0.03994	0.21564
Listeria	0.00395	0.34964	0.53267	0.03003	0.00545
Bacillus	0.00006	0.21420	0.02818	0.29671	0.30589
Pseudomonas	0.13668	0.36721	0.05776	0.38147	0.59037
Escherichia/Shigella[a]	NA	0.00462	0.45612	0.00237	0.59037
Shigella[a]	4.6372×10^{-10}	NA	NA	NA	NA
Escherichia[a]	0.00001	NA	NA	NA	NA
Enterobacteriaceae[a]	NA	0.87898	0.00004	0.19274	0.00055
Salmonella	3.8092×10^{-6}	0.34964	0.05838	0.09712	0.08851
Lactobacillus	0.00297	0.87898	0.53267	0.38147	0.59037
Enterococcus	0.00012	0.04816	0.03746	0.01159	0.00954

[a]*Escherichia/Shigella* relatives counted as equivalent

these samples, which also provides an opportunity to infer broad patterns across animal taxa based on a standardized methodology. Measures of alpha diversity display overall consistent values with respect to host species, although many significant differences between methods are present, which are mostly host-specific (Fig. 2a, b). However, several host taxa display high levels of consistency across methods including *A. aurita*, *C. elegans*, *D. melanogaster*, and *H. sapiens*, which show almost no significant differences between methods. Discrepancies and individual recommendations for each host species are discussed in Additional file 1: Figures S6–S16 and Additional file 2: Table S4. An intriguing observation is the tendency of aquatic hosts to display higher alpha diversity values than those of terrestrial hosts, which is supported by average differences between aquatic and terrestrial hosts and by relative consistent comparisons among single host species as well (Fig. 2c, d; Additional file 2: Table S5).

In order to investigate broad patterns of bacterial community similarity according to metagenomic procedure and host species, we performed beta diversity analyses including all host samples and each of their five different methodological profiles. This analysis reveals an overall strong signal of host species, irrespective of the method used to generate community profiles (Table 2; Fig. 3). Pairwise comparisons between hosts are significant in all cases except for samples derived from the V3 V4 two-step protocol, which did not consistently reach significance after correction for multiple testing (Additional file 2: Table S6). Further, complementary to the observations made for alpha diversity, we also find strong signals of community differentiation between the aquatic and terrestrial hosts (Table 2; Fig. 3b, d). The separation between these environments appears to be stronger based on amplicon data, whereas the separation between hosts is stronger based on shotgun-derived data (Table 2). To further evaluate the variability among biological replicates, we evaluated intra-group distances according to host species, which reveals organisms with generally higher community variability (i.e., *C. elegans*, *A. aurita*, *H. sapiens*, *H. vulgaris*, *T. aestivum*, and *M. leidyi*) than other host organisms in our study (*N. vectensis*, *M. musculus*, *D. melanogaster*, and *A. aerophoba*; Additional file 1: Figure S17A and C). Interestingly, intra-group distances also significantly differ between the aquatic and terrestrial environments, whereby aquatic organisms tend to display less variable communities than terrestrial ones (Additional file 1: Figure S17B and D). Thus, this suggests higher sample sizes may be necessary for experimental analysis of the higher variability/terrestrial taxa. The low performance of *T. aestivum* in subsequent analyses possibly originates from its commercial origin and low bacterial biomass relative to host material.

To identify individual drivers behind patterns of beta diversity, we performed indicator species analysis [15] at the genus level with respect to method, host species, and environment. Based on the amplicon data, we identified 56 of 313 indicators to display consistent associations across all four amplicon techniques, such as *Bacteroides*, *Barnesiella*, *Clostridium IV*, and *Faecalibacterium* in *H. sapiens* and *Helicobacter* and *Mucispirillum* in *M. musculus*, whereas other associations were limited to, e.g., only one variable region (Additional file 2: Tables S7 and S8). However, the overall pattern of host associations is largely consistent across methods (Additional file 1: Figure S18). We also identified numerous indicator genera for aquatic and terrestrial hosts (Additional file 2: Tables S9 and S10). Indicator analyses based on shotgun data reveals a smaller and less diverse set of host-specific indicators, which however show many congruencies with the amplicon-based data.

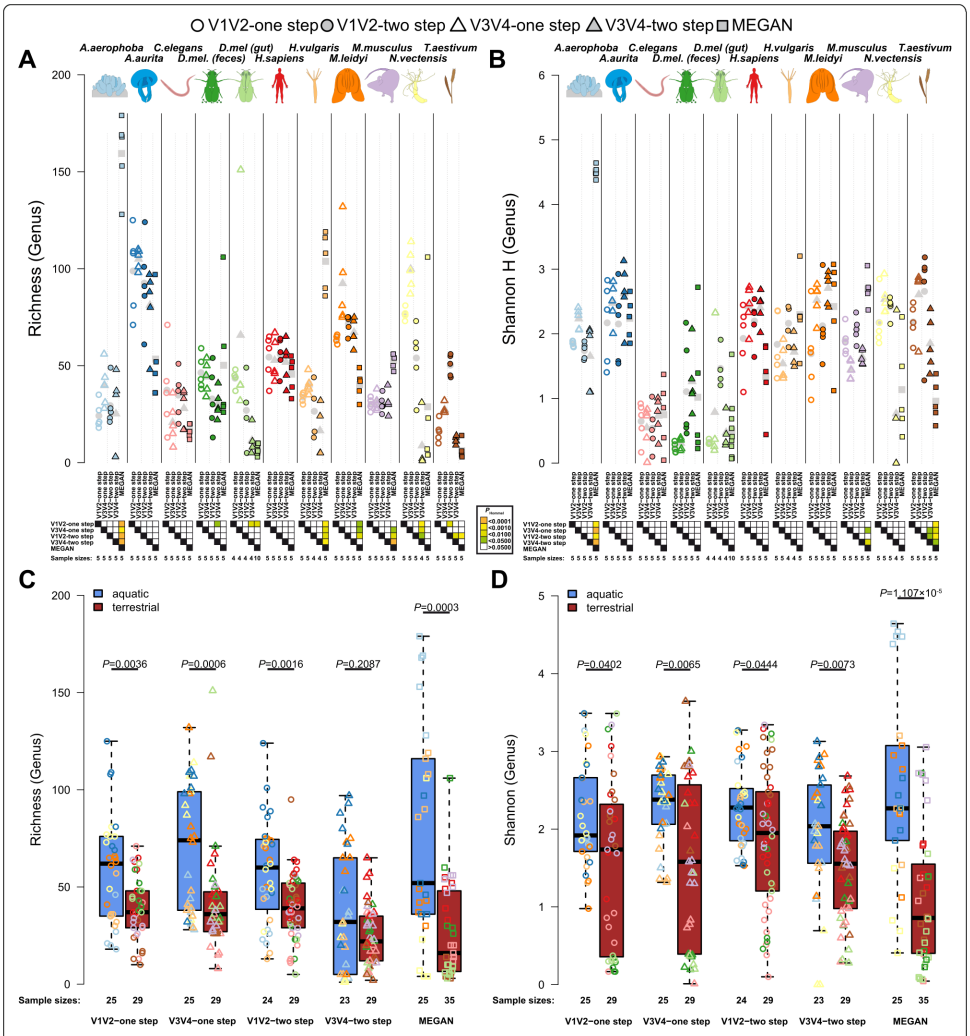


Fig. 2 Comparison of bacterial genus richness (**a**) and Shannon H (**b**) based on 16S rRNA gene amplicon and shotgun derived genus profiles based on MEGAN highlighting the differences between variable regions, amplification methods, and metagenomic classifier, as well as between the different host organisms. Colors show significance of amplification methods (**a**, **c**) or pairwise comparisons of methods (**b**, **d**) based on pairwise *t* tests with Hommel *P* value adjustment (**a**, **b**), and approximate Wilcoxon test for the comparison between environmental categories (**c**, **d**). Mean values are shown in gray symbols in plots **a** and **b**. Sample sizes are indicated below the samples

Functional diversity within and between hosts

To examine the diversity (gene richness) of metagenomic functions across host species, we evaluated EggNOG annotations (evolutionary genealogy of genes: Non-supervised Orthologous Groups [16]) to obtain a general functional spectrum (assembly-based and MEGAN), in addition to annotations derived from a database dedicated to functions interacting with carbohydrates (CAZY—Carbohydrate-Active enZYmes) [17]. Overall, the individual host communities differ

Table 2 Taxonomic distance-based PERMANOVA results for differences in community composition (genus level) between host species and host environments based on shared abundance (Bray-Curtis) and shared presence (Jaccard), based on whole genome shotgun and different amplicon strategies (*P* values are adjusted via Hommel's procedure)

Distance	Factor	Data	Classifier	*DF*	*F*	*P*	P_{Hommel}	R^2	adj. R^2
Bray-Curtis	Organism	Shotgun	MEGAN	10,49	6.3517	0.0001	0.0001	0.5645	0.4756
		Amplicon	V1 V2 one-step	10,43	7.1026	0.0001	0.0001	0.6229	0.5352
			V1 V2 two-step	10,42	4.2297	0.0001	0.0001	0.5018	0.3831
			V3 V4 one-step	10,43	7.8964	0.0001	0.0001	0.6474	0.5654
			V3 V4 two-step	10,41	3.7917	0.0001	0.0001	0.4805	0.3538
	Environment	Shotgun	MEGAN	1,58	5.8958	0.0001	0.0004	0.0923	0.0766
		Amplicon	V1 V2 one-step	1,52	6.1588	0.0001	0.0001	0.1059	0.0887
			V1 V2 two-step	1,51	4.6185	0.0001	0.0001	0.0830	0.0651
			V3 V4 one-step	1,52	5.4975	0.0001	0.0001	0.0956	0.0782
			V3 V4 two-step	1,50	3.3349	0.0001	0.0001	0.0625	0.0438
Jaccard	Organism	Shotgun	MEGAN	10,49	4.7458	0.0001	0.0001	0.4920	0.3883
		Amplicon	V1 V2 one-step	10,43	3.6867	0.0001	0.0001	0.4616	0.3364
			V1 V2 two-step	10,42	2.9760	0.0001	0.0001	0.4147	0.2754
			V3 V4 one-step	10,43	4.0248	0.0001	0.0001	0.4835	0.3633
			V3 V4 two-step	10,41	2.9343	0.0001	0.0001	0.4171	0.2750
	Environment	Shotgun	MEGAN	1,58	4.3872	0.0001	0.0004	0.0703	0.0543
		Amplicon	V1 V2 one-step	1,52	3.8714	0.0001	0.0001	0.0693	0.0514
			V1 V2 two-step	1,51	3.6541	0.0001	0.0001	0.0669	0.0486
			V3 V4 one-step	1,52	4.3213	0.0001	0.0001	0.0767	0.0590
			V3 V4 two-step	1,50	3.6646	0.0001	0.0001	0.0683	0.0497

drastically in gene richness (EggNOG genes (MEGAN) $\chi^2 = 52.202$, $P < 2.10 \times 10^{-16}$; EggNOG genes (assembly) $\chi^2 = 49.986$, $P < 2.10 \times 10^{-16}$; CAZY $\chi^2 = 48.815$, $P < 2.10 \times 10^{-16}$; approximate Kruskal-Wallis test). Although the values also differ considerably between methods, overall, the functional repertoires are most diverse in the vertebrate hosts, while only *H. vulgaris* and *A. aerophoba* as aquatic hosts carry comparably diverse functional repertoires (Fig. 4a, : Figure S19). Interestingly, in contrast to taxonomic diversity, we observe no difference in functional diversity between aquatic and terrestrial hosts.

Next we examined community differences (beta diversity) at the functional level, which are overall more pronounced (average adj. $R^2 = 0.5084$; Fig. 4) than those based on taxonomic (genus level) classification (shotgun adj. $R^2 = 0.4756$, amplicon average adj. $R^2 = 0.4594$; see Tables 2 and 3; Figs. 3 and 4, Additional file 1: Figure S20). On the functional level, aquatic and terrestrial hosts are considerably less distinct than observed at the taxonomic level (taxonomic shotgun data $R^2 = 0.0766$, taxonomic amplicon average adj. $R^2 = 0.0690$, functional shotgun data $R^2 = 0.0441$; see Tables 2 and 3; Fig. 4, Additional file 1: Figure S20). Variability of the functional repertoires was lowest in *A. aerophoba*, *D. melanogaster* feces, and *M. musculus* gut contents, while *H. vulgaris*, *C. elegans*, and *D. melanogaster* gut samples displayed the highest intra-group distances, which translates to a higher amount of functional heterogeneity between replicates (Additional file 1: Figure S21). This reflects in large part the patterns we observed in taxonomic variability of those host-associated communities (Additional file 1: Figure S17).

Indicator functions

To identify specific functions that are characteristic of individual hosts, we applied indicator analysis to genomic functions. General functions in EggNOG reveal several interesting patterns, including CRISPR-related genes in *A. aerophoba*, *H. sapiens*, and *H. vulgaris*, suggesting a particular importance of viruses in these communities. Further, most species show characteristic genes mainly involved in energy production and conversion, amino acid transport and metabolism, replication, recombination, and repair, as well as cell wall/membrane/envelope biogenesis (Additional file 2: Tables S11–S13).

Analysis of carbohydrate-metabolizing functions based on CAZY [17] (Carbohydrate-Active enZYmes) reveals the highest number of characteristic glycoside hydrolases (GH) in *H. sapiens* and *M. musculus*, whereas polysaccharide lyases (PLs) for non-hydrolytic cleavage of

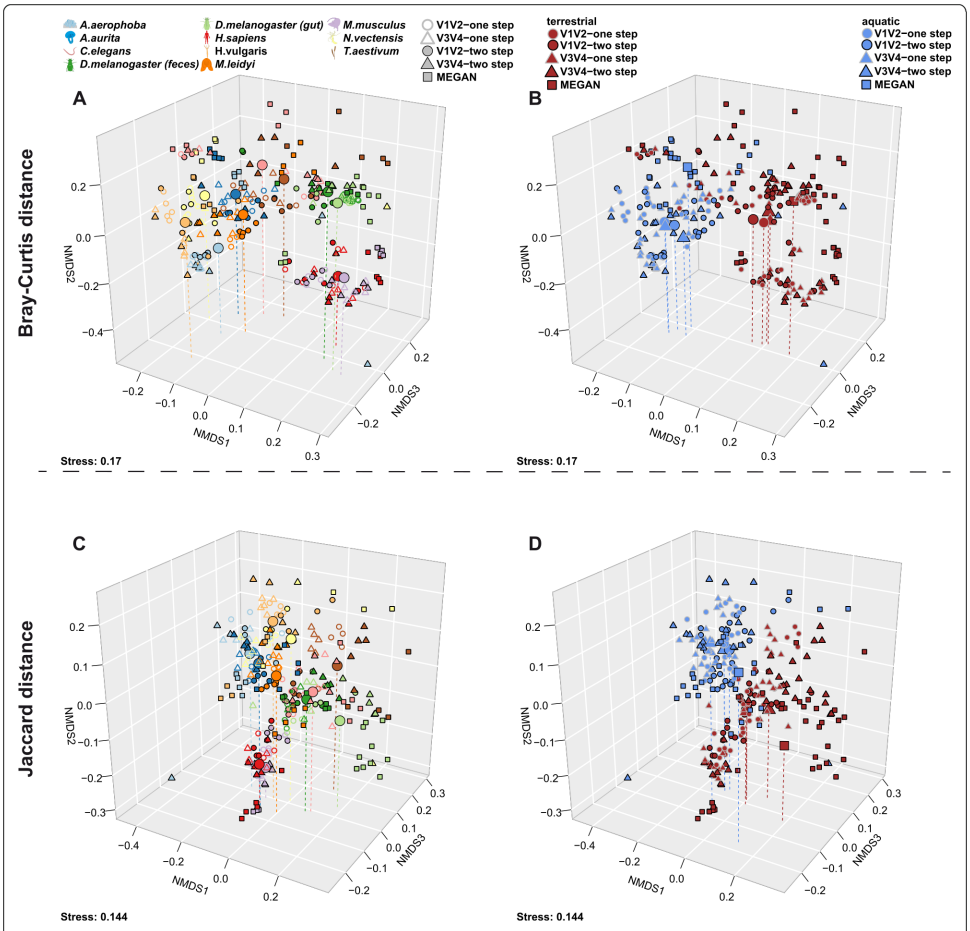


Fig. 3 Non-metric multidimensional scaling of Bray-Curtis distances based on genus abundance profiles derived from the different 16S rRNA gene amplicon methods (V1 V2/V3 V4, one-step/two-step) and shotgun-derived genus profiles highlighting **a** host differences and **b** differences between host environments (terrestrial/aquatic; see Table 2). Non-metric multidimensional scaling of Jaccard distances based on genus presence/absence profiles derived from the different 16S rRNA gene amplicon methods and shotgun-derived genus profiles highlighting **c** host taxon differences and **d** differences between host environments (terrestrial/aquatic; see Table 2). Both panels show a separation based on host organisms and environments and not by method. Large symbols indicate the centroid of the respective host groups and vertical lines help to determine their position in space. Sample sizes are equal to Fig. 2 (see also Additional file 2: Table S1)

glycosidic bonds are present in *A. aerophoba* and *H. sapiens* (Additional file 2: Table S14). Interestingly, parts of the cellulosome are only associated to *A. aerophoba*, while the freshwater polyp *H. vulgaris* carries characteristic auxiliary CAZYs involved specifically in lignin and chitin digestion, which may reflect adaptations of the host microbial communities to their diets (e.g., *Artemia nauplii*).

Performance of metagenome imputation from 16S rRNA gene amplicon data using PICRUSt across metaorganisms

Researchers often desire to obtain the insight gained from functional metagenomic information despite being limited to 16S rRNA gene data, for which imputation methods such as PICRUSt can be employed [5]. However, due to their dependence on variable region and database coverage [5], these imputations should be

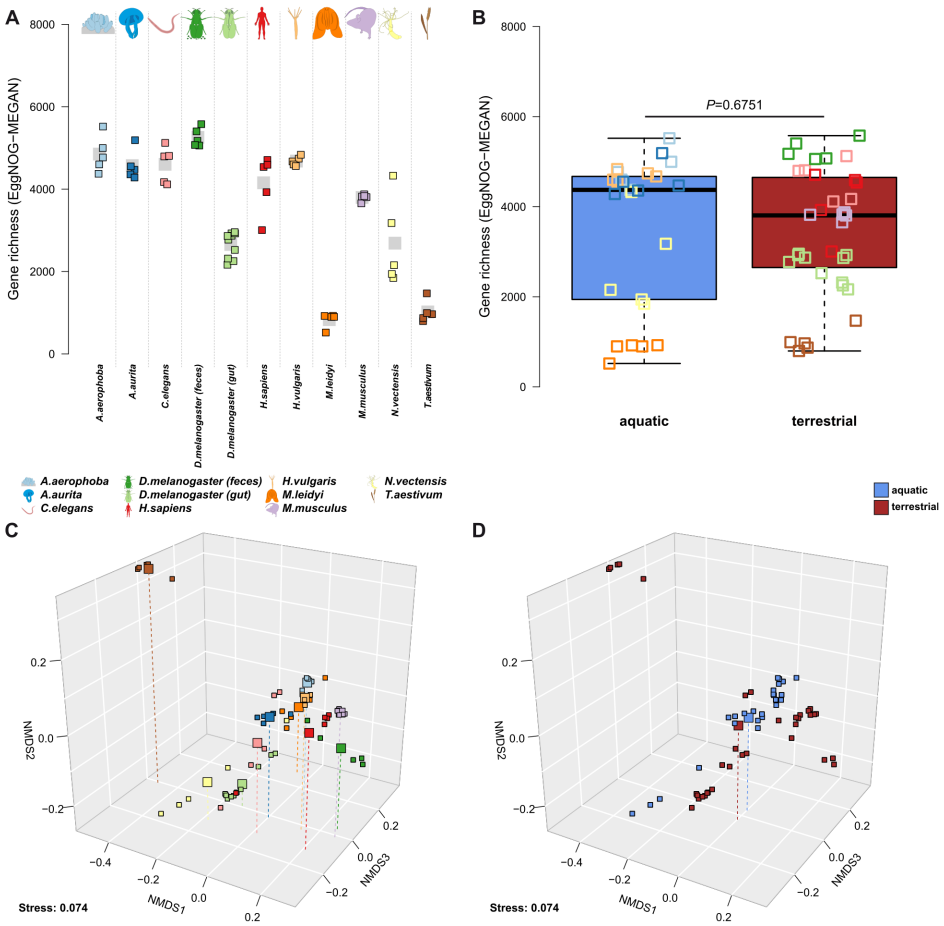


Fig. 4 Functional diversities were derived from the number and abundances of MEGAN-based EggNOG annotations. Functional richness between **a** host organisms and **b** host environmental groups based is displayed, as well as functional differences between hosts (**c**) and environmental groups (**d**). Non-metric multidimensional scaling is based on Bray-Curtis distances on the differences in functional composition between the host organisms displayed (**c**, **d**; see Table 3). Large symbols indicate the centroid of the respective groups. Functional variation of communities based on pairwise Bray-Curtis distances within host organism groups and environmental groups. Sample sizes for the host taxa are $N = 5$, except for *D. melanogaster* gut tissue ($N = 10$; see Additional file 2: Table S1)

viewed with caution. Given our dataset of both 16S amplicon and shotgun metagenomic sequences, we systematically evaluated the performance of PICRUSt predictions across hosts and amplicon data type (V1 V2/V3 V4, one-step/two-step protocol). Beginning with the mock community, the V1 V2 region displays lower performance for imputing functions compared to V3 V4, as indicated by a higher weighted Nearest Sequenced Taxon Index (NSTI) ($t = 17.812$, $P = 1.119 \times 10^{-7}$; Additional file 1: Figure S22A). High NSTI values imply low availability of genome representatives for the respective sample, due to either large phylogenetic distance for each OTU to its closest sequenced reference genome or a high frequency of poorly represented OTUs [5]. Comparing the distribution of functional categories based on Clusters of Orthologous Groups (COG) [18] between

Table 3 Functional distance-based PERMANOVA results for differences in general functional community composition (EggNOG) and carbohydrate-active enzymes (CAZY) between host species and host environments based on shared abundance (Bray-Curtis) and shared presence (Jaccard) of functions (*P* values are adjusted via Hommel procedure)

Distance	Factor	Data	*DF*	*F*	*P*	P_{Hommel}	R^2	adj. R^2
Bray-Curtis	Organism	CAZY	10,47	7.3323	0.0001	0.0001	0.6094	0.5263
		EggNOG categories	10,49	5.6088	0.0001	0.0001	0.5337	0.4386
		EggNOG gene + description	10,49	4.4454	0.0001	0.0001	0.4757	0.3687
		EggNOG (MEGAN categories)	10,49	12.2594	0.0001	0.0001	0.7144	0.6562
		EggNOG (MEGAN gene)	10,49	8.2788	0.0001	0.0001	0.6282	0.5523
	Environment	CAZY	1,56	5.4257	0.0001	0.0007	0.0883	0.0721
		EggNOG categories	1,58	2.5429	0.0195	0.0195	0.0420	0.0255
		EggNOG gene + description	1,58	3.0662	0.0001	0.0007	0.0502	0.0338
		EggNOG (MEGAN categories)	1,58	3.7703	0.0015	0.0030	0.0610	0.0448
		EggNOG (MEGAN gene)	1,58	3.7271	0.0002	0.0012	0.0604	0.0442
Jaccard	Organism	CAZY	10,47	3.9098	0.0001	0.0001	0.4541	0.3380
		EggNOG categories	10,49	3.7179	0.0001	0.0001	0.4314	0.3154
		EggNOG gene + description	10,49	2.5275	0.0001	0.0001	0.3403	0.2057
		EggNOG (MEGAN categories)	10,49	7.7781	0.0001	0.0001	0.6135	0.5346
		EggNOG (MEGAN gene)	10,49	5.4989	0.0001	0.0001	0.5288	0.4326
	Environment	CAZY	1,56	2.5866	0.0003	0.0021	0.0442	0.0271
		EggNOG categories	1,58	1.4180	0.1442	0.1442	0.0239	0.0070
		EggNOG gene + description	1,58	1.9535	0.0004	0.0024	0.0326	0.0159
		EggNOG (MEGAN categories)	1,58	3.0425	0.0460	0.0920	0.0498	0.0335
		EggNOG (MEGAN gene)	1,58	3.1222	0.0001	0.0009	0.0511	0.0347

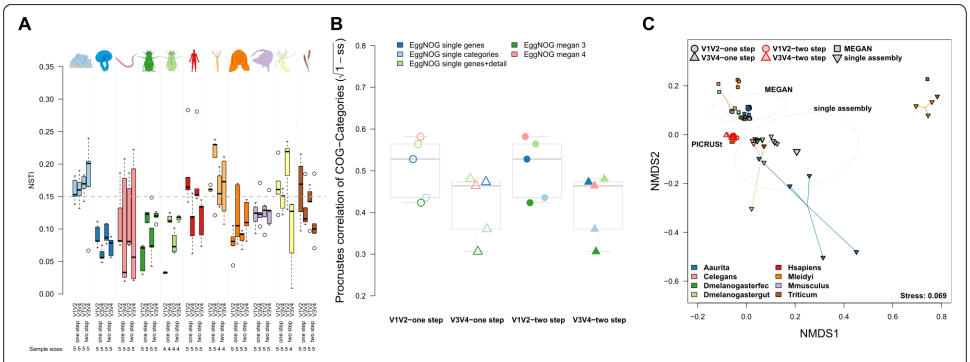


Fig. 5 a Differences in the Nearest Sequenced Taxon Index (imputation success) between variable regions (average: $Z = 0.3869$, $P = 0.7017$, approximate Wilcoxon test; probability: odds ratio = 1.5941, $P = 0.1402$, Fisher's test) and amplification method ($Z = 0.0667$, $P = 0.9472$, approximate Wilcoxon test; probability: odds ratio = 1.5511, $P = 0.1436$, Fisher's test). **b** Procrustes correlation of imputed- and shotgun-based COG categories among different techniques, with significantly higher correspondence between imputed and measured functional profiles in the V1 V2 compared to the V3 V4 region ($F_{1,18} = 7.8537$, $P = 0.0118$, ANOVA). **c** Non-metric multidimensional scaling displays Bray-Curtis distances based on functional category abundances (COG categories) derived from PICRUSt (V1 V2/V3 V4, one-step/two-step) and shotgun-based approaches (MEGAN, single assembly). Ellipses represent standard deviations of points within the respective groups

the different imputations (no cutoff applied) and the actual shotgun-based repertoires reveals considerable overlap except categories R (general function prediction only) and S (function unknown) (Additional file 1: Figure S22B).

Next we evaluated functional imputations for the different host species and amplification methods. We found no significant difference in average NSTI values or prediction success (NSTI < 0.15) between amplification protocols or variable region. However, approximately a third (31.8%) of the samples are lost due to incomplete imputation (NSTI > 0.15; Fig. 5a). Notable problematic host taxa are *A. aerophoba* and *H. vulgaris*, for which no sample remained below the NSTI cutoff value. Other host taxa displayed clear differential performance with regard to the variable region used, whereby *H. sapiens*, *N. vectensis*, and *T. aestivum* were successfully predicted based on V3 V4, but not V1 V2. However, when we employ Procrustes tests to compare community functional profiles based on shotgun sequencing (single assembly, MEGAN) and functional imputations at the COG-category level, we find a lower correspondence of the V3 V4-based imputations compared to those based on V1 V2 (Fig. 5b), while the amplification methods displayed no significant difference. A similar pattern is observed when we correlate community differences based on shotgun results and lower level (single functions) COG annotations based on PICRUSt, although the difference is not significant ($F_{1,18} = 0.6172$, $P = 0.4423$; ANOVA).

To investigate the similarities among methods in more detail, we merged shotgun and PICRUSt based annotations at the level of COG categories. Principle coordinate analysis reveals only small differences between imputations with regard to amplification method or variable region (Fig. 5c). However, large differences exist between the PICRUSt and shotgun-based functional repertoires, as well as between the shotgun techniques (MEGAN, single assembly). Differences between the shotgun techniques were significant but smaller than their distance to the imputed functional spectra (Fig. 5c; Additional file 2: Table S15), a pattern also found in the relative abundances of functional categories (Additional file 1: Figure S23).

In summary, the PICRUSt-imputed functional repertoires significantly differ from actual shotgun profiles. While variation in imputation success is largely dependent on the composition of the particular host community, V3 V4 appears to more often yield successful imputations. However, when successful, V1 V2-derived imputations display closer similarity to actual functional profiles. Finally, the amplification method (one-step, two-step) appears to have no significant effect on the quality of functional imputations. These data therefore support the notion that metagenome imputations should be evaluated with care, as they depend on the underlying variable region and sample source.

Phylogenetic patterns in microbial community composition

The term "phylosymbiosis" refers to the phenomenon where the pattern of similarity among host-associated microbial communities parallels the phylogeny of their hosts [19]. Highly divergent hosts with drastic differences in physiology and life history might be expected to overwhelm the likelihood of observing phylosymbiosis, which can typically be observed within a given host clade [19]. However, the factors driving differences in composition among our panel of hosts may also be expected to vary in terms of the bacterial phylogenetic scale at which they are most readily observed [20]. Thus, we evaluated the degree to which bacterial community relationships (beta diversity) reflect the underlying phylogeny of our hosts at a range of bacterial taxonomic ranks, spanning from the genus to the phylum level.

In order to assess the general overlap between beta diversity and phylogenetic distance of the host species, we performed Procrustes analysis [21]. These analyses reveal that the strongest phylogenetic signal is observed when bacterial taxa are grouped at the order and/or family level, whereby the one-step protocols and the V3 V4 region display greater correlations to phylogenetic distance (Fig. 6). A similar pattern is observed for shotgun-based community profiles (i.e., MEGAN), although its fit increases again at the genus level. Measuring beta diversity based on co-occurrence of bacterial taxa between hosts (Jaccard; Fig. 6a) displays a weaker correspondence to host phylogeny than the abundance-based measure (Bray-Curtis; Fig. 6b).

To assess the fit of individual host taxa, we examined the residuals of the correlation between community composition and phylogenetic distance. This reveals a large variation in correspondence among host taxa, with *M. musculus*, *M. leidyi*, *H. sapiens*, and *D. melanogaster* (feces) displaying the highest, while *H. vulgaris*, *C. elegans*, and *A. aerophoba* display the lowest correspondence between their microbiome composition and phylogenetic position (largest residuals; Additional file 1: Figure S24), pointing towards increased environmental influences on these microbial communities. Furthermore, terrestrial hosts display an overall better correspondence between co-occurrences of bacterial genera and host relatedness (V1 V2 one-step: $Z = 2.9578$, $P = 0.0025$), as do measurements based on V3 V4 (one-step: $Z = 2.7496$, $P = 0.0054$; two-step: $Z = 2.8097$, $P = 0.0046$; approximate Wilcoxon test).

Next, given the peak of correspondence between bacterial community composition and host phylogeny

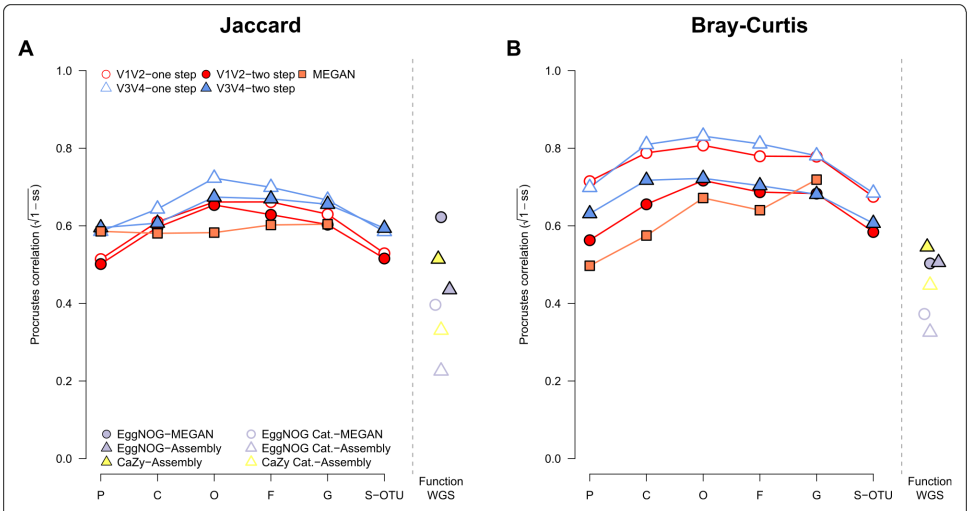


Fig. 6 Multivariate correlation (Procrustes analyses) of phylogenetic distance among host organisms and community distances based on (**a**) shared presence among samples or (**b**) differences in abundance in 16S rRNA gene amplicon or shotgun-derived community profiles at different taxonomic cutoffs (Phylum to Genus, additional species level OTUs for amplicon-based profiles). Similar results are shown for the correspondence between phylogenetic distances among samples and their distances based on (**a**) shared presence or (**b**) abundance differences in their functional composition. The functional composition was derived from COGs and COG categories imputed from PICRUSt, EggNOG-derived COG categories and genes, and CAZY functions. All correlations are significant at $P \leq 0.05$ (10,000 permutations)

observed at the order and/or family level, we set out to identify individual community members whose abundances best correlate to host phylogenetic distance using Moran's *I* eigenvector method [22]. This reveals 41 bacterial families and 36 orders with significant phylogenetic signal based on one or more amplicon data set, whereby 16 families and 18 orders display repeated associations across methods (e.g., *Clostridia*, *Bacteroidales*, *Desulfovibrionales*; Additional file 2: Table S16; Additional file 1: Figures S25 and S26). Analyzing communities based on shotgun data on the other hand identifies 75 bacterial families and 19 orders associated with phylogenetic distances, whereby 17 and 20 display repeated associations, respectively (Additional file 2: Table S16; Additional file 1: Figure S27). The combined results of these analyses identify several families and orders with strong and consistent phylogenetic associations, in particular for the vertebrate hosts (e.g., *Bacteroidaceae*/*Bacteroidales*, *Bifidobacteriaceae*/*Bifidobacteriales*, *Desulfovibrionaceae*/*Desulfovibrionales*, *Ruminococcaceae*/*Clostridiales*; see Additional file 2: Table S16). Other individual examples include bacteria related to *Helicobacteraceae*/*Campylobacterales* in *A. aurita*, which are observed in other marine cnidarians and may be involved in sulfur oxidation [23]. *Alcanivoracaceae*, an alkane-degrading bacterial group, is strongly associated to the coastal cnidarian *N. vectensis*. This association might originate from adaptation to a polluted coastal environment [24]. *Acidobacteria Gp6* and *Gp9* specifically occur in *A. aerophoba* and are commonly associated to the core microbial community of sponges [25].

Phylogenetic patterns in functional community composition

In order to contrast the patterns observed at the taxonomic level to those based on function, we used Procrustes correlation to measure the overlap between phylogenetic distance and community distance based on the panel of functional categories in our analyses. Interestingly, the two functional categories displaying the greatest correspondence to host phylogeny are the CAZY and single EggNOG-based functions (Fig. 6). The remainder of patterns between phylogeny and bacterial functional spectra differed among the host species and functional categories (Additional file 1: Figure S28), and *T. aestivum* and *D. melanogaster* (feces) display the lowest correspondence, while *C. elegans*, *M. musculus*, *and H. sapiens* display the best correspondence (smallest residuals; Additional file 1: Figure S24) between their functional repertoire and phylogenetic position. As observed for the taxonomic analyses, terrestrial hosts again display a slightly better correlation than aquatic hosts (smaller residuals), in particular for the co-abundance of

EggNOG categories ($Z = 2.2116$, $P = 0.0267$), CAZY ($Z = 2.0393$, $P = 0.0414$), and the co-occurrence of EggNOG categories ($Z = 2.7377$, $P = 0.0061$) and genes ($Z = 3.3062$, $P = 0.0007$; approximate Wilcoxon test) among hosts.

Finally, to reveal individual functions correlating to host phylogeny, we used the aforementioned Moran's *I* eigenvector analyses with additional indicator analyses to narrow the potential clade associations. Interestingly, most functions that correlate to a specific host taxon/clade (1–3 host taxa) are mainly restricted to vertebrate hosts or in combination with a vertebrate host (Additional file 2: Tables S17–S20). This pattern is repeated across all functional annotations used in this study. Examples include fucosyltransferases, fucosidases, and polysaccharide-binding proteins, as well as different lyases for hyaluronate, xanthan, and chondroitin that stem from CAZY (see Additional file 1: Figure S28; Additional file 2: Table S17). These functions are related to glycan and mucin degradation and interaction, which mediate many intimate host-bacterial interactions and are also observed in subsequent analyses based on general functional databases (EggNOG; Additional file 2: Tables S18–S20). Many other phylogenetically correlated functions appear to be driven by the vertebrate hosts as well, which likely reflects the high functional diversity within this group (Fig. 4 and Additional file 1: Figure S21). Only *LPXC* and *LPXK* (EggNOG), genes involved in the biosynthesis of the outer membrane, are exclusively associated to the non-vertebrate hosts (LPXC, UDP-3-O-acyl-*N*-acetylglucosamine deacetylase; LPXK, Tetraacyldisaccharide 4′-kinase), as is an oxidative damage repair function (MSRA reductase) associated to *H. vulgaris* (Additional file 2: Table S19; Additional file 1: Figure S28). Finally, antibiotic resistance genes and virulence factors also show frequent phylogenetic and host-specific signals (Additional file 2: Tables S18 and S19; Additional file 1: Figure S28).

Discussion

Despite the great number of metagenomic studies published to date, which range in their focus on technical, analytical, or biological aspects, our study represents a unique contribution given its breadth of different host samples analyzed with a panel of standardized methods. In particular, the trade-offs between 16S rRNA gene amplicon and shotgun sequencing concerning amplification bias, functional information, and both monetary and computational costs warrant careful consideration when designing research projects. While 16S rRNA gene amplicon-based analyses are subject to considerable skepticism and criticism, we demonstrate that in many aspects similar, if not superior characterization of bacterial communities is achieved by these methods. We also show, however, that important insight can be gained through the combination of taxonomic and functional profiling, and that imputation-based functional profiles significantly differ from actual profiles. Our findings thus provide a guide for selecting an appropriate methodology for metagenomic analyses across a variety of metaorganisms. Finally, these data provide novel insight into the broad-scale evolution of host-associated bacterial communities, which can be viewed as particularly reliable given the repeatability of observations (e.g., differences between aquatic and terrestrial hosts, indicator taxa) across methods.

Given the concerns regarding the accuracy of 16S rRNA gene amplicon sequencing, other studies such as that of Gohl et al. [8] performed systematic comparisons of different library preparation methods and found superior results for a two-step amplification procedure. This method offers the additional advantage that one panel of adapter/barcode sequences can be combined with any number of different primers. Our first analyses were based on a standard mock community including Gram-positive and Gram-negative bacteria from the Bacilli and Gamma-Proteobacteria (eight species), as well as two fungi, which did not support an improvement of performance based on the two-step protocol. However, a number of changes were made to the Gohl et al. [8] protocol to adapt it to our lab procedures (e.g., larger reaction volumes, polymerase, variable region, heterogeneity spacers) that may contribute to these discrepancies, in addition to our different and diverse set of samples and other factors with potential influence on the performance of amplicon sequencing [6–8, 26–28]. The complexity of the mock community, i.e., the number of taxa, distribution, and phylogenetic breadth, may also have an influence on the discovery of clear trends in amplification biases or detection limits for certain taxonomic groups [29]. Thus, the even and phylogenetically shallow mock community in our study may be less suited than the staggered and diverse mixtures used in other studies [8] but still provides valuable information on repeatability, primer biases, and accuracy [29]. Nonetheless, when applied to our range of complex host-associated communities, we also found that significant differences in most parameters were due to the variable region rather than amplification method, and in many cases, biological signals were either improved or limited to the one-step protocol. Thus, in combination with the less complex laboratory procedures associated with the one-step protocol, we would generally recommend this procedure over two-step protocols.

Additional sources of variation influencing the outcome of our 16S rRNA gene amplicon-based community profiling are nucleic acid extraction procedures and the bioinformatic pipelines we employed. For the former, extraction procedures differed between host species due to

specific optimizations required for individual host species. Thus, certain differences in taxonomic and functional composition may be influenced by the specific protocols employed, as observed elsewhere [30]. Differences in the latter range from trimming and merging to clustering and classification, which are stringent and incorporate more reliable de novo clustering algorithms [31] as well as different classification databases [32]. Heterogeneity among the different amplicon approaches is however smaller than the differences between the amplicon and shotgun methods, as observed in other benchmarking studies [27]. Differences between shotgun approaches have been investigated in detail and also yield varying performances among classifiers, but in general, find a comparatively high performance of MEGAN-based approaches [9, 33, 34], which we also confirm in our study.

Given the limited number of studies that have compared imputed- and shotgun-derived functional repertoires [5, 35], our study also provides important additional insights. As imputation by definition is data-dependent, the differential performance and prediction among hosts in our study may in large part be explained by the amount of bacteria isolated, sequenced, and deposited (16S rRNA or genome) from these hosts or their respective environments. This seems to be most critical for the aquatic hosts. Furthermore, we observe a clear effect of variable region on the prediction performance, which is most obvious based on the mock community. The PICRUSt algorithm was developed and tested using primers targeting V3 V4 16S rRNA, and thus optimization of the imputation algorithm might be biased towards this target over the V1 V2 variable region. Although these performance differences, in particular the bias towards model organisms compared to less characterized communities (e.g., hypersaline microbial mats), were previously shown [5], our study provides additional, experimentally validated guidelines for a number of novel host taxa.

Interestingly, the strongest correspondence between bacterial community similarity and host genetic distance was detected at the bacterial order level for most of the employed methods. This may on the one hand reflect the deep phylogenetic relationships between our host taxa, such that turnover of bacterial taxa erodes phylosymbiosis over time [19, 20]. On the other hand, some of the more striking observations made among our host taxa are the differences between aquatic and terrestrial hosts, both at the level of alpha and beta diversity. Based on a molecular clock for the 16S rRNA gene of roughly 1% divergence per 50 million years [36], bacterial order level divergence corresponds well with the timing of animal terrestrialization (425–500 MYA) [37, 38]. Although evolutionary rates can widely vary among bacteria species [39], other studies of individual gut microbial lineages such as the *Enteroccoci* indicate that animal terrestrialization was indeed a likely driver of diversification [40]. Specifically, the changing availability of carbohydrates in the host gut can be seen as a main driver of this diversification, which is consistent with the association of CAZY-based functional repertoires correlating to phylogenetic distance in our data set [19, 41].

In contrast to the patterns observed based on 16S rRNA gene amplicon-based profiles, the differentiation of bacterial communities according to host habitat was less pronounced based on functional genomic repertoires. This raises the possibility that the colonization of land by ancient animals required the acquisition of new, land-adapted bacterial lineages to perform some of the same ancestral functions. The overall observation of increased beta diversity among terrestrial compared to aquatic hosts (Additional file 1: Figure S19) could in part reflect differential acquisition among host lineages after colonizing land, although dispersal in the aquatic environment may on the other hand act as a greater homogenizing factor among aquatic hosts. The stronger correspondence between bacterial community and host phylogenetic distance among terrestrial hosts is also generally consistent with this hypothesis. However, the higher alpha diversity and the slightly lower correspondence with the phylogenetic patterns in aquatic hosts may also indicate a higher influence of environmental bacteria or a lack of physiological control over bacterial communities.

Bacterial taxa and functions involved in carbohydrate utilization were among the most notable associations to individual hosts, groups of hosts, and/or host phylogenetic relationships. Taxa such as *Bacteroidales*, *Ruminococcaceae*/*Ruminococcales*, and *Clostridia* associated to humans and/or mice include members known for a mucosal lifestyle, and these hosts also display the most diverse and abundant repertoire of carbohydrate-active enzymes (particularly glycosylhydrolases) in their microbiome. Other examples include sialidases, esterases, and fucosyltransferases, as well as different extracellular structures that appear to be specific to aquatic hosts, indicating differences in mucus and glycan composition according to this host environment. Glycan structures provide a direct link between the microbial community and the host via attachment, nutrition, and communication [42, 43], and the composition of mucin and glycan structures themselves show strong evolutionary patterns and are distinct among taxonomic groups [41]. Thus, a high diversity of glycan structures within and between hosts may determine the specific sets carbohydrate-facilitating enzymes of the respective microbial communities.

In addition to the bacterial carbohydrate hydrolases that digest surrounding host and dietary carbohydrates,

we also identified a number of glycosyltransferases associated with capsular polysaccharide synthesis (Additional file 2: Tables S19 and S20). This type of glycosylation is an important facilitator for host association and survival [44] and plays a crucial role in infections [45] in mutualists and pathogens alike [44, 46]. Thus, capsular and excreted glycan structures are important for the successful colonization and persistence in different environments [47, 48] and host organisms [44, 48].

Conclusions

In summary, the systematic comparison of five different metagenomic sequencing methods applied to ten different holobiont yielded a number of novel technical and biological insights. Although important exceptions will exist, we demonstrate that broad-scale biological patterns are largely consistent across these varying methods. As many aspects of differential performance in our study are host-specific (more detailed description of individual hosts can be found in Additional file 1), future development and benchmarking analyses would also benefit from including a range of different host/environmental samples.

Methods

DNA extraction and 16S rRNA gene amplicon sequencing

Protocols for each host type are described in Additional file 1: Figures S18–S28. Each library (16S rRNA gene amplicon, shotgun) included at least one mock community sample based on the ZymoBIOMICS™ Microbial Community DNA Standard (Lot.: ZRC187324, ZRC187325) consisting of eight bacterial species (*Pseudomonas aeruginosa* (10.4%), *Escherichia coli* (9.0%), *Salmonella enterica* (11.8%), *Lactobacillus fermentum* (10.3%), *Enterococcus faecalis* (14.1%), *Staphylococcus aureus* (14.6%), *Listeria monocytogenes* (13.2%), *Bacillus subtilis* (13.2%)) and two fungi (*Saccharomyces cerevisiae* (1.6%), *Cryptococcus neoformans* (1.8%)).

The 16S rRNA gene was amplified using uniquely barcoded primers flanking the V1 and V2 hypervariable regions (27F–338R) and V3 V4 hypervariable regions (515F–806R) with fused MiSeq adapters and heterogeneity spacers in a 25-µl PCR [28]. For the traditional one-step PCR protocol, we used 4 µl of each forward and reverse primer (0.28 µM), 0.5 µl dNTPs (200 µM each), 0.25 µl Phusion Hot Start II High-Fidelity DNA Polymerase (0.5 Us), 5 µl of HF buffer (Thermo Fisher Scientific, Inc., Waltham, MA, USA), and 1 µl of undiluted DNA. PCRs were conducted with the following cycling conditions (98 °C, 30 s; 30 × [98 °C, 9 s; 55 °C, 60 s; 72 °C, 90 s]; 72 °C, 10 min; 10 °C, infinity) and checked on a 1.5% agarose gel. Using a modified version of the recently published two-step PCR protocol by Gohl et al. 2016, we employed for the first round of amplification fusion primers consisting of the 16S rRNA gene primers (V1 V2, V3 V4) and a part of the Illumina Nextera adapter with the following cycling conditions in a 25- µl PCR reaction (98 °C, 30 s; 25 × [98 °C, 10 s; 55 °C, 30 s; 72 °C, 60 s]; 72 °C, 10 min; 10 °C, infinity) [8]. Following the PCR, the product was diluted 1:10 and 5 µl were used in an additional reaction of 10 µl (98 °C, 30 s; 10 × [98 °C, 9 s; 55 °C, 30 s; 72 °C, 60 s]; 72 °C, 10 min; 10 °C, infinity) utilizing the Nextera adapter overhangs to ligate the Illumina adapter sequence and individual MIDs to the amplicons, following the manufacturer's instructions. The PCR protocol we used was 1 µl of each forward and reverse primer (5 µM), 0.3 µl dNTPs (10 µM), 0.2 µl Phusion Hot Start II High-Fidelity DNA Polymerase (2 U/µl), 2 µl of 5 × HF buffer (Thermo Fisher Scientific, Inc., Waltham, MA, USA), and 5 µl of the diluted PCR product. The concentration of the amplicons was estimated using a Gel Doc™ XR+ System coupled with Image Lab™ Software (BioRad, Hercules, CA USA) with 3 µl of O'GeneRulerTM 100 bp Plus DNA Ladder (Thermo Fisher Scientific, Inc., Waltham, MA, USA) as the internal standard for band intensity measurement. The samples of individual gels were pooled into approximately equimolar sub-pools as indicated by band intensity and measured with the Qubit dsDNA br Assay Kit (Life Technologies GmbH, Darmstadt, Germany). Sub-pools were mixed in an equimolar fashion and stored at − 20 °C until sequencing.

Library preparation for shotgun sequencing was performed using the NexteraXT kit (Illumina) for fragmentation and multiplexing of input DNA following the manufacturer's instructions. Amplicon sequencing was performed on the Illumina MiSeq platform with v3 chemistry (2 × 300 cycle kit), while shotgun sequencing was performed on an Illumina NextSeq 500 platform via 2 × 150 bp Mid Output Kit at the IKMB Sequencing Center (CAU Kiel, Germany).

Amplicon analysis

The respective V1 V2 and V3 V4 PCR primer sequences were removed from the sequencing data using cutadapt (v.1.8.3) [49]. Sequence data in FastQ format was quality trimmed using sickle (v.1.33) in paired-end mode with default settings and removing sequences dropping below 100 bp after trimming [50]. Forward and reverse read were merged into a single amplicon read using VSEARCH allowing fragments with a length of 280–350 bp for V1 V2 and 350–500 bp for V3 V4 amplicons [51]. Sequence data was quality controlled using fastq_quality_filter (FastX Toolkit) retaining sequences with no more than 5% of per-base quality values below 30 and subsequently with VSEARCH discarding sequences with more than one expected error [51, 52]. Reference-guided chimera removal was performed using the gold.fa

reference in VSEARCH (v2.4.3). The UTAX algorithm was used for a fast classification of the sequence data in order to remove sequences not assigned to the domains Bacteria or Archaea and exclude amplicon fragments from Chloroplasts [53]. Notably, only a total of 15 sequences were assigned to the domain Archaea, all found in two samples of human feces, accounting for less than 0.1% of the clean reads in these samples. The entire cleaned sequence data was concatenated into a single file and dereplicated and processed with VSEARCH for OTU picking using the UCLUST algorithm [54] using a 97% similarity threshold. OTUs were again checked for chimeric sequences, now using the de novo implementation of the UCHIME algorithm in VSEARCH [51, 54, 55]. All clean sequence data of the samples were mapped back to the cleaned OTU sequences using VSEARCH. OTU sequences and clean sequences mapping to the OTUs were taxonomically annotated using the RDP classifier algorithm with the RDP training set 14 [56, 57]. Sequence data were normalized by selecting 10,000 random sequences per sample. Taxon-by-sample abundance tables were created for all taxonomic levels from Phylum to Genus, as well as for OTUs.

PICRUSt functional imputations

Species-level OTUs (97% similarity threshold) were further classified using the GreenGenes (August 2013) database [58] via RDP classifier as implemented in mothur (v1.39.5) and merged with the abundances into a biome file which was uploaded to the Galaxy PICRUSt v1.1.1 pipeline (http://galaxy.morganlangille.com/) to derive functional imputations (COG predictions) [5]. To achieve accurate functional predictions, samples with NSTI ≤ 0.15 (weighted Nearest Sequenced Taxon Index) were pruned from the data set, as recommended by the developers.

Shotgun sequencing

Raw demultiplexed sequences were trimmed via Trimmomatic (v0.36) for low-quality regions with a minimum length of 50 bp as well as for adaptor and remaining MID sequences [59]. After trimming reads were mapped to host-specific genome databases and *ΦX* with additional retention databases containing all fully sequenced bacterial and metagenomic genomes (5 September 2015) via DeconSeq (v0.4.3) [60]. Single and paired sequences were repaired using the BBTools (v37.28) repair function [61]. Combined sequences were searched against the non-redundant NCBI database (28 July 2017) via DIAMOND [62] with (*E* value cutoff 0.001, v0.8.28) and MEGAN [13] classifying hits by functions (EggNOG—October 2016) and taxa (May 2017) (v6.6.1). For assemblies of single samples, we used metaSPADES [63] (v3.9.1) using paired reads in addition to unpaired reads left from the previous steps. PROKKA (v1.12) was used for gene calling and initial genome annotation [64] using the metagenome option with additional identifying rRNA and snRNA via barnap, ARAGORN [65], and Infernal [66]. ORFs were further annotated via EggNOG annotation via HMMER models implemented in the EggNOG-mapper (v0.12.7) [16, 67], CAZY database via dbCAN (v5, July 24, 2016), and HMMER3 [17, 68]. Gene abundances were derived from mapping the all reads back to the predicted ORF via bowtie2 (v2.2.6) [69] and calculated TPM (transcripts per kilobase million) via SamTools (v1.5) [70].

18S rRNA genes were obtained from NCBI GeneBank and aligned via ClustalW (v1.4) [71] for host tree construction, which includes *A. aerophoba* (gi:51095211, AY5917991), *M. leidyi* (gi:14517703, AF2937001), *H. vulgaris* (gi:761889987, JN5940542), *A. aurita* (gi:14700050, AY0392081), *N. vectensis* (gi:13897746, AF2543821), *T. aestivum* (gi:15982656, AY0490401), *M. musculus* (gi:374088232, NR_0032783), *H. sapiens* (gi:36162, X032051), *D. melanogaster* (gi:939630477, NR_1335591), and *C. elegans* (gi:30525807, AY2681171). Phylogenetic distance was calculated via DNADIST (v3.5c) [72] and a maximum likelihood tree was constructed via FastTree v2.1 CAT+Γ model [73]. Accuracy was improved via increased minimum evolution rounds for initial tree search [−spr 4], more exhaustive tree search [−mlacc 2], and a slow initial tree search [−slownni].

Statistical analysis

Statistical analyses were carried out via R (v3.4.3) [74]. Alpha diversity indices (richness, Shannon-Weaver index) and beta diversity metrics based on the shared presence (Jaccard distance) or abundance (Bray-Curtis distance) of taxa were calculated in the *vegan* package [75] and ordinated via Principal Coordinate Analysis (PCoA, avoiding negative eigenvalues), or via non-metric multidimensional scaling (NMDS) using a maximum of 10,000 random starts to obtain a minimally stressed configuration in three dimensions. Clusters were fit via an iterative process (10,000 permutations) and tested for separation by direct gradient analysis via distance-based redundancy analyses and permutative ANOVA (10,000 permutations) [76, 77]. Univariate analyses were carried out with approximate Wilcoxon/Kruskal tests as implemented in *coin* [78] (10,000 permutations). Procrustes tests were used to relate pairwise community distances based on either different data sources such as functional repertoires or taxonomic composition, as well as phylogenetic distances [21, 79]. Moran's *I* eigenvector technique was employed to correlate bacterial community members and their functions to phylogenetic divergence, as implemented in *ape* (10,000 permutations) [22, 80].

Indicator species analysis, employing the generalized indicator value (*IndVal.g*), was used to assess the predictive value of a taxon for each respective host phenotype/category as implemented in *indicspecies* [15]. Linear mixed models, as implemented in *nlme* were used to compare the influence of amplification method or variable region without the influence of the organism of origin [81]. We employed the Hommel and Benjamini-Yekutieli adjustment of *P* values when advised [82, 83].

Additional files

Additional file 1: Supplementary Materials. (PDF 6900 kb)
Additional file 2: Supplementary Tables. (ZIP 1765 kb)

Acknowledgements
We thank Katja Cloppenborg-Schmidt, Melanie Vollstedt, and Dr. Sven Künzel for the excellent assistance and help during the development of the project and their constant drive to improve its quality.

Authors' contributions
PRa, PRo, AF, TB, and JFB conceived and designed research. PRa and MR performed data analyses. PRa, MR, BH, SD, and JFB interpreted the results and wrote the manuscript. PRa, MR, TD, KD, HD, SD, SF, JF, UHH, FAH, BH, MH, MJ, CJ, KABK, DL, AR, TBHR, TR, RAS, HS, RS, FS, ES, NWB, PRo, AF, TB, and JFB generated and interpreted host-specific data and gave intellectual input. All authors read and approved the final manuscript.

Funding
This work was funded by the DFG Collaborative Research Centre (CRC) 1182 "Origin and Function of Metaorganisms" subproject Z3 and the Max-Planck-Society.

Availability of data and materials
Sequence and meta-data are accessible under the study identifier PRJEB30924 ("https://www.ebi.ac.uk/ena"). Remaining DNA from non-human samples can be made available upon request. All human samples and information on their corresponding phenotypes have to be obtained from the PopGen Biobank Kiel (Schleswig-Holstein, Germany) through a Material Data Access Form. Information about the Material Data Access Form and how to apply can be found at "https://www.uksh.de/p2n/Information+for+Researchers.html".

Ethics approval and consent to participate
Human samples
Study participants were randomly recruited from inhabitants of Schleswig-Holstein (Germany) who were recruited for the PopGen cohort. Five individuals from the PopGen biobank (Schleswig-Holstein, Germany) were randomly selected among the healthy and unmedicated individuals and included in the study without corresponding meta-information. Study participants collected fecal samples at home without conservation buffers in standard fecal tubes (sterile feces container 76 × 20 mm, Sarstedt) and shipped them immediately at room temperature or brought them to the collection center (within 24 h). Samples were stored at − 80 °C until processing. Human feces ($N = 4$) were sampled and extracted following the procedures as described in Wang et al. 2016 [84]. A biopsy sample of the sigmoid colon was taken from a healthy control individual without macro- or microscopical inflammation ($N = 1$) and DNA was extracted as described in Rausch et al. [85]. Investigators were blinded to sample identities and written informed consent was obtained from all study participants before the study. All protocols were approved by the Ethics Committee of the Medical Faculty of Kiel and by the data protection officer of the University Hospital Schleswig-Holstein in adherence with the Declaration of Helsinki Principles.
Animal and plant samples
Wild-derived, hybrid mice were sacrificed according to the German animal welfare law and Federation of European Laboratory Animal Science Associations guidelines. Hybrid breeding stocks of wild-derived *M. m. musculus* × *M. m. domesticus* hybrids captured in 2008 are kept at the Max Planck Institute Plön (11th lab generation). The approval for mouse husbandry and experiment was obtained from the local veterinary office "Veterinäramt Kreis Plön" (Permit: 1401-144/PLÖ-004697). All samplings, including invertebrate and plant samples, were performed in concordance with the German animal welfare law and Federation of European Laboratory Animal Science Associations guidelines. Further details for each host type are provided in Additional file 1.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Evolutionary Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany. [2]Institute for Experimental Medicine, Kiel University, Kiel, Germany. [3]Department of Biology, Laboratory of Genomics and Molecular Biomedicine, University of Copenhagen, Copenhagen Ø, Denmark. [4]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany. [5]Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany. [6]Institute of General Microbiology, Kiel University, Kiel, Germany. [7]Department of Evolutionary Ecology and Genetics, Zoological Institute, Kiel University, Kiel, Germany. [8]Zoological Institute, Kiel University, Kiel, Germany. [9]Molecular Physiology, Zoological Institute, Kiel University, Kiel, Germany. [10]Marine Ecology, Research Unit Marine Symbioses, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany. [11]Kiel University, Kiel, Germany. [12]Marine Ecology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany. [13]Environmental Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany. [14]Environmental Genomics, Botanical Institute, Kiel University, Kiel, Germany.

Received: 8 February 2019 Accepted: 23 August 2019
Published online: 14 September 2019

References
1. Bosch TCG, McFall-Ngai MJ. Metaorganisms as the new frontier. Zoology. 2011;114(4):185–90.
2. McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci. 2013;110(9):3229–36.
3. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis. 2015;26(1):26191.
4. Morgan XC, Huttenhower C. Chapter 12: human microbiome analysis. PLoS Comput Biol. 2012;8(12):e1002808.
5. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814–21.
6. Hiergeist A, Glasner J, Reischl U, Gessner A. Analyses of intestinal microbiota: culture versus sequencing. ILAR J. 2015;56(2):228–40.
7. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol. 2015;6:771.
8. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016;34(9):942–9.
9. Walsh AM, Crispie F, O'Sullivan O, Finnegan L, Claesson MJ, Cotter PD. Species classifier choice is a key consideration when analysing low-complexity food microbiome data. Microbiome. 2018;6(1):50.
10. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, et al. The long-term stability of the human gut microbiota. Science. 2013;341(6141):1237439.
11. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol. 2016;7(459):459.

12. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431–7.
13. Huson D, Auch A, Qi J, Schuster S. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
14. Hong Nhung P, Ohkusu K, Mishima N, Noda M, Monir Shah M, Sun X, Hayashi M, Ezaki T. Phylogeny and species identification of the family Enterobacteriaceae based on dnaJ sequences. Diagn Microbiol Infect Dis. 2007;58(2):153–61.
15. De Cáceres M, Legendre P, Moretti M. Improving indicator species analysis by combining groups of sites. Oikos. 2010;119(10):1674–84.
16. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016; 44(1):286–93.
17. Cantarel BL. The carbohydrate-active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 2009;37(Database issue):233–8.
18. Fink C, von Frieling J, Knop M, Roeder T. Drosophila Fecal Sampling. Bio-protocol 2017;7:e2547.
19. Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. PLoS Biol. 2016;14(11): e2000225.
20. Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun. 2017;8:14319.
21. Peres-Neto P, Jackson D. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia. 2001;129(2):169–78.
22. Gittleman JL, Kot M. Adaptation: statistics and a null model for estimating phylogenetic effects. Syst Zool. 1990;39(3):227–41.
23. Murray AE, Rack FR, Zook R, Williams MJM, Higham ML, Broe M, Kaufmann RS, Daly M. Microbiome composition and diversity of the ice-dwelling sea anemone, Edwardsiella andrillae. Integr Comp Biol. 2016;56(4):542–55.
24. Schneiker S, dos Santos VAPM, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, et al. Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis. Nat Biotechnol. 2006;24(8):997.
25. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol. 2012;10(9):641–54.
26. Wu JY, Jiang XT, Jiang YX, Lu SY, Zou F, Zhou HW. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. BMC Microbiol. 2010;10:255.
27. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics. 2016;17(1):55.
28. Fadrosh D, Ma B, Gajer P, Sengamalay N, Ott S, Brotman R, Ravel J. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome. 2014;2(1):6.
29. Highlander S. Mock Community Analysis. In: Nelson EK, editor. Encyclopedia of Metagenomics. New York: Springer New York; 2013. p. 1–7.
30. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35(11):1069–76.
31. Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ. 2015;3:e1487.
32. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J. 2012;6(1):94–103.
33. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nat Methods. 2017;14(11):1063–71.
34. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:19233.
35. Xu Z, Malmer D, Langille MGI, Way SF, Knight R. Which is more important for classifying microbial communities: who's there or what they can do? ISME J. 2014;8(12):2357–9.
36. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. Proc Natl Acad Sci U S A. 1999;96(22):12638–43.
37. Benton MJ. The origins of modern biodiversity on land. Philos Trans R Soc B. 2010;365(1558):3667–79.
38. Rota-Stabelli O, Daley Allison C, Pisani D. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr Biol. 2013;23(5):392–8.
39. Kuo CH, Ochman H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol Direct. 2009;4:35.
40. Lebreton F, Manson AL, Saavedra JT, Straub TJ, Earl AM, Gilmore MS. Tracing the Enterococci from Paleozoic origins to the hospital. Cell. 2017;169(5):849–861.e813.
41. Bishop JR, Gagneux P. Evolution of carbohydrate antigens—microbial forces shaping host glycomes? Glycobiology. 2007;17(5):23R–34R.
42. Pickard JM, Maurice CF, Kinnebrew MA, Abt MC, Schenten D, Golovkina TV, Bogatyrev SR, Ismagilov RF, Pamer EG, Turnbaugh PJ, et al. Rapid fucosylation of intestinal epithelium sustains host-commensal symbiosis in sickness. Nature. 2014;514(7524):638–41.
43. Schwartzman JA, Koch E, Heath-Heckman EAC, Zhou L, Kremer N, McFall-Ngai MJ, Ruby EG. The chemistry of negotiation: rhythmic, glycan-driven acidification in a symbiotic conversation. Proc Natl Acad Sci. 2015;112(2): 566–71.
44. Martens EC, Chiang HC, Gordon JI. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. Cell Host Microbe. 2008;4(5):447–57.
45. Boulnois GJ, Roberts IS. Genetics of capsular polysaccharide production in bacteria. Curr Top Microbiol Immunol. 1990;150:1–18.
46. Mahdavi J, Pirinccioglu N, Oldfield NJ, Carlsohn E, Stoof J, Aslam A, Self T, Cawthraw SA, Petrovska L, Colborne N, et al. A novel O-linked glycan modulates Campylobacter jejuni major outer membrane protein-mediated adhesion to human histo-blood group antigens and chicken colonization. Open Biol. 2014;4(1):130202.
47. Tounkang S, Premkumar D, Gustavo S, Nathalie B, Yann B, Patricia C, Florence L, Olivier N, Brigitte G, Anne L, et al. Capsular glucan and intracellular glycogen of Mycobacterium tuberculosis: biosynthesis and impact on the persistence in mice. Mol Microbiol. 2008;70(3):762–74.
48. Roberts IS. The biochemistry and genetics of capsular polysaccharide production in bacteria. Annu Rev Microbiol. 1996;50(1):285–315.
49. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10.
50. Joshi N, Fass J: Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files 1.33 edn. 2011. https://github.com/najoshi/sickle.
51. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
52. Hannon G: FASTX-Toolkit. In. http://hannonlab.cshl.edu/fastx_toolkit; 2010.
53. Edgar RC. UTAX algorithm; 2015.
54. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.
55. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16): 2194–200.
56. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
57. Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, et al. The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res. 2003; 31(1):442–3.
58. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610–8.
59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
60. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011; 6(3):e17288.
61. Bushnell B, Rood J: BBTools bioinformatics tools, including BBMap. In., 37.28 edn. http://sourceforge.net/projects/bbmap; 2017.

62. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
63. Nurk S, Meleshko D, Korobeynikov A, Pevzner P: metaSPAdes: a new versatile de novo metagenomics assembler. arXiv preprint arXiv: 160403071 2016.
64. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
65. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–6.
66. Kolbe DL, Eddy SR. Fast filtering for RNA homology search. Bioinformatics. 2011;27(22):3102–9.
67. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 2017;34(8):2115–22.
68. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012; 40(1):445–51.
69. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The sequence alignment/ map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
71. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–80.
72. Felsenstein J. DNADIST -- Program to compute distance matrix from nucleotide sequences. 3.5c edn; 1993.
73. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
74. Team RC. R: A language and environment for statistical computing. In: R Foundation for Statistical Computing. 3.3.2 edn; 2016.
75. Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H: vegan: Community Ecology Package 1. 17-6 edn: 2011 http://CRAN.R-project.org.
76. Legendre P, Anderson MJ. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol Monogr. 1999;69(1):1–24.
77. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26(1):32–46.
78. Hothorn T, Hornik K, Van de Wiel MA, Zeileis A. A Lego system for conditional inference. Am Stat. 2006;60(3):257–63.
79. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26(11):1463–4.
80. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.
81. Pinheiro J, Bates D, DebRoy S, Sarkar D, Team RDC: nlme: Linear and Nonlinear Mixed Effects Models. 2011 http://CRAN.R-project.org.
82. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;75(2):383–6.
83. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–88.
84. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen F-A, Ruhlemann MC, Szymczak S, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet. 2016; 48(11):1396–406.
85. Rausch P, Rehman A, Künzel S, Häsler R, Ott SJ, Schreiber S, Rosenstiel P, Franke A, Baines JF. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. Proc Natl Acad Sci. 2011;108(47):19030–5.

Publisher's Note

3 Host-genetic influence on the human intestinal microbiome

Publications:

Malte C. Rühlemann, Frauke Degenhardt, Louise B. Thingholm, Jun Wang, Jurgita Skieceviciene, Philipp Rausch, Johannes R. Hov, Wolfgang Lieb, Tom H. Karlsen, Matthias Laudes, John F. Baines, Femke-Anouska Heinsen, and Andre Franke*: "*Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in *SLC9A8* (NHE8) and 3 other loci". *Gut Microbes.* 2018;9(1):68-75.

Malte Christoph Rühlemann, Britt Marie Hermes, Corinna Bang, Shauni Doms, Lucas Moitinho-Silva, Louise Bruun Thingholm, Fabian Frost, Frauke Degenhardt, Michael Wittig, Jan Kässens, Frank Ulrich Weiss, Annette Peters, Klaus Neuhaus, Uwe Völker, Henry Völzke, Georg Homuth, Matthias Laudes, Wolfgang Lieb, Dirk Haller, Markus Maximilian Lerch, John Baines, Andre Franke: "ABO histo-blood groups influence gut microbiome, with causal relationship between Bacteroides and inflammatory bowel disease". *Manuscript under review in Nature Genetics.*

Article B: Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in *SLC9A8* (NHE8) and 3 other loci

GUT MICROBES
2018, VOL. 9, NO. 1, 68–75
https://doi.org/10.1080/19490976.2017.1356979

Taylor & Francis
Taylor & Francis Group

ADDENDUM

OPEN ACCESS Check for updates

Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in *SLC9A8* (NHE8) and 3 other loci

Malte C. Rühlemann[a], Frauke Degenhardt[a], Louise B. Thingholm[a], Jun Wang[b,c,†], Jurgita Skiecevičienė[a], Philipp Rausch[b,c], Johannes R. Hov[d,e,f,g], Wolfgang Lieb[h], Tom H. Karlsen[d,e,f,g,i], Matthias Laudes[j], John F. Baines[b,c], Femke-Anouska Heinsen[a], and Andre Franke[a]

[a]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; [b]Evolutionary Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany; [c]Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Kiel, Germany; [d]Norwegian PSC Research Center, Division of Surgery, Inflammatory Medicine and Transplantation, University Hospital Rikshospitalet, Oslo, Norway; [e]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [f]Research Institute of Internal Medicine, Oslo University Hospital Rikshospitalet, Oslo, Norway; [g]Section of Gastroenterology, Department of Transplantation Medicine, Oslo University Hospital Rikshospitalet, Oslo, Norway; [h]Institute of Epidemiology, Christian-Albrechts-University of Kiel, Kiel, Germany; [i]Department of Clinical Medicine, University of Bergen, Bergen, Norway; [j]Department of Internal Medicine I, University Hospital S.-H. (UKSH, Campus Kiel), Kiel, Germany

ABSTRACT

Factors shaping the human intestinal microbiota range from environmental influences, like smoking and exercise, over dietary patterns and disease to the host's genetic variation. Recently, we could show in a microbiome genome-wide association study (mGWAS) targeting genetic variation influencing the β diversity of gut microbial communities, that approximately 10% of the overall gut microbiome variation can be explained by host genetics. Here, we report on the application of a new method for genotype-β-diversity association testing, the distance-based F (DBF) test. With this we identified 4 loci with genome-wide significant associations, harboring the genes *CBEP4, SLC9A8, TNFSF4*, and *SP140*, respectively. Our findings highlight the utility of the high-performance DBF test in β diversity GWAS and emphasize the important role of host genetics and immunity in shaping the human intestinal microbiota.

ARTICLE HISTORY
Received 31 March 2017
Revised 11 July 2017
Accepted 11 July 2017

KEYWORDS
β diversity; GWAS; human gut microbiota; immunity; IBD

Introduction

The human gut microbiota as an important focus of medical research within the past few years, has been investigated in the context of numerous inflammatory and non-inflammatory disorders of the intestine, but also in other systemic diseases, rendering gut health and the underlying host-microbiota interactions as a key component of well-being. While changes in α- and β diversity, as well as changes in the presence or absence and the abundance of specific microbial taxa have been shown to be associated with numerous diseases, the processes and factors shaping a 'healthy' gut microbiota are still largely understudied. First studies could show connections between host genotypes and changes in the abundance of specific taxa. These studies were either rather underpowered, investigating only roughly one hundred individuals,[1,2] or based on candidate genes to reduce multiple testing burden.[3,4]

An analysis approach, focusing on host-genetic influences on β diversity using the microbiomeGWAS framework,[5] which uses linear models to correlate genotype distance data with pairwise β diversity data, correcting for skewness and kurtosis of the results, identified 2 loci on chromosome 9 and chromosome 4 to be associated with variation in weighted UniFrac distance and Bray-Curtis dissimilarity, respectively.[4]

Recently, we estimated in a host-microbiome genome-wide association study (mGWAS), linking β diversity to host genetic variation, that roughly 10% of the variation in the gut microbiota is explained by the host's genetic architecture (model with 42 loci) in a Northern German study population.[6] This proportion

CONTACT Prof. Dr. *rer. nat.* Andre Franke a.franke@mucosa.de Institute of Clinical Molecular Biology, Rosalind-Franklin-Str. 12, 24105 Kiel, Germany.
[†]Current address: Department of Microbiology and Immunology, KU Leuven & Center for the Biology of Disease, VIB, Leuven, Belgium

Addendum to: Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen F-A, Rühlemann MC, Szymczak S, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nature genetics. 2016; 48(11):1396-1406.

of explained variation has about the same order of magnitude as the proportion explained by non-genetic factors (such as dietary and lifestyle factors) described elsewhere.[7,8] Additionally, we could show correlations of serum bile- and fatty acids with the abundance of microbial traits. Especially variants in the gene encoding for the transcription factor Vitamin D Receptor (*VDR*), among whose ligands are also bile acids, were found to play an essential role in shaping of intestinal communities.[6]

Here, we present the application of an alternative analytical approach for the investigation of β diversity host-genomic associations with shaping the gut microbiota, which does not rely on extensive permutations, thus massively reducing the computational burden, while exhibiting high concordance with comparable permutation-based approaches.

Our findings highlight the role of the host's immune functions and signaling in the assembly and homeostasis of gut-associated microbial communities in humans. In addition, our identified loci are located near known inflammatory bowel disease (IBD) genetic susceptibility loci, previously identified through case-control GWAS, implicating the host-microbiome interplay in IBD disease etiology.

Approximate inference of null distribution as an alternative to extensive permutative tests in β diversity GWAS

Permutative distance-based analysis of variance,[9] as implemented in the *adonis* function of the *vegan* package[10] for R,[11] is an widely used approach to investigate differences in β diversity based on categorical variables. However, approaches relying on permutation are slow regarding computation time, and thus, not applicable to large data sets comprising several hundreds of samples and millions of genetic variants. The method of moment matching tries to overcome these problems by approximating an unknown null distribution based on known distributions. In this case a Pearson Type III distribution, and parameters estimated from the data itself,[12] provide the opportunity to analyze large data sets in a GWAS setting comparably fast using this distance-based F test (DBF test). The Pearson Type III distribution was chosen as its properties as a 3-parameter Gamma distribution makes modeling of a multitude of other distributions possible, using its first 3 moments calculated from the data: mean, variance and skewness.

While the DFB test has been shown to be applicable to different types of data sets and distance measures,[12,13] it has not been used in large-scale studies investigating factors shaping microbial communities. We applied this method on β diversity data represented as Bray-Curtis dissimilarity on genus level abundance data, in analogy to the input data used in our previous publication.[6] The genotype information used was the same as described in the previously published article.[6] The data set consisted of 2 independent cohorts, PopGen and FoCus, from Northern Germany, comprising 830 and 937 individuals, respectively, and 1767 individuals in total. To account for influences of nutrition and anthropometrics, the Bray-Curtis dissimilarity was corrected for the covariates total energy intake, alcohol consumption, and water intake, as well as age, gender, and body mass index, respectively. Furthermore, β diversity data was corrected for variation in the first 3 genetic principal components. This was done fitting a distance-based Redundancy Analysis[9] (*capscale* function of the *vegan* package[10] for R[11]) using the aforementioned covariates as constraints. The residual variation of this model was subsequently used as distance matrix in the DBF-test. The DBF-test was performed in R[11] using the *snpStats* package[14] to import genotype data in *plink* format[15] and applying the *DBF.test* function imported from the R source code file accompanying the original article describing the DBF test (https://wwwf.imperial.ac.uk/~gmontana/software/dbf/dbf_test.R).[12] To ensure the detection of robust signals and to account for the different sample sizes, a meta-analysis was performed only using genotype-information overlapping in both cohorts and using a weighted Z-score based test.[16] Association results were classified as "significant," if the meta-analysis P-value passed the genome-wide significance threshold of $P < 5 \times 10^{-8}$ in the meta-analysis, and both cohorts displayed a significant P-value ($P < 0.05$).

Genes involved in host-immunity are associated with shifts in β diversity

Using the afore-mentioned significance criteria, 4 loci were found as significantly associated with variation in β diversity in the meta-analysis. The locus with the strongest signal is located on chromosome 5 (rs67909753; chr5:173306058; $P_{meta} = 3.61 \times 10^{-9}$; Fig. 1A in strong LD with the *CPEB4* gene (Cytoplasmic Polyadenylation Element Binding Protein 4).

CPEB4 is an effector by which RORγt, a key determinant in the cell differentiation of Th17 cells, inhibits proliferation of thymocytes.[17] One variant at this locus (rs7705502; $R^2_{LeadSNP} = 0.928$) has previously been reported to be associated with Crohn's disease[18,19] and obesity-related traits.[20] The second signal is located on chromosome 20 (rs113738363; chr20:48449631; $P_{meta} = 1.54 \times 10^{-8}$; Fig. 1B). A variant at this locus in strong linkage disequilibrium with the lead SNP (rs4809760; $R^2 = 0.765$) has been identified in our previous mGWAS[6] and is located in an intronic area of the *SLC9A8* gene, encoding for NHE8 (cation proton antiporter 8). This protein is expressed in goblet cells in the intestine[21] and is known to be essential for mucosal integrity, with loss of expression leading to increased bacterial adhesion and inflammation in mice following dextran sodium sulfate (DSS) treatment.[22] Additionally, this locus was previously found to be associated with psoriasis,[23-25] a chronic disorder of the skin with proposed links to the intestinal microbiota.[26]

Our third hit is located on chromosome 2 (rs11678791; chr2:231223975; $P_{meta} = 1.19 \times 10^{-8}$; Fig. 1C) harboring the SP140 Nuclear Body Protein and the SP140L genes. This locus was previously associated with Crohn's disease[19] and SP140L is a key regulator of the macrophage transcriptional program, whose depletion leads to a severely impaired microbe-induced activation.[27] The fourth and last association finding is located on chromosome 1 (rs11811788; chr1:173150727; $P_{meta} = 2.1 \times 10^{-8}$; Fig. 1D). This locus harbors the *TNFSF4* (*OX40L; CD252*) gene that is located 2.1 kbp downstream of rs11811788. The OX40-OX40L signaling pathway has been shown to regulate cytokines in T-cells, antigen-presenting cells (APCs), NK cells and NKT cells, thus plays a central role in inflammation.[28]

Permutation-based analysis

To confirm the validity of the signals, permutation based testing was performed for the 4 variants

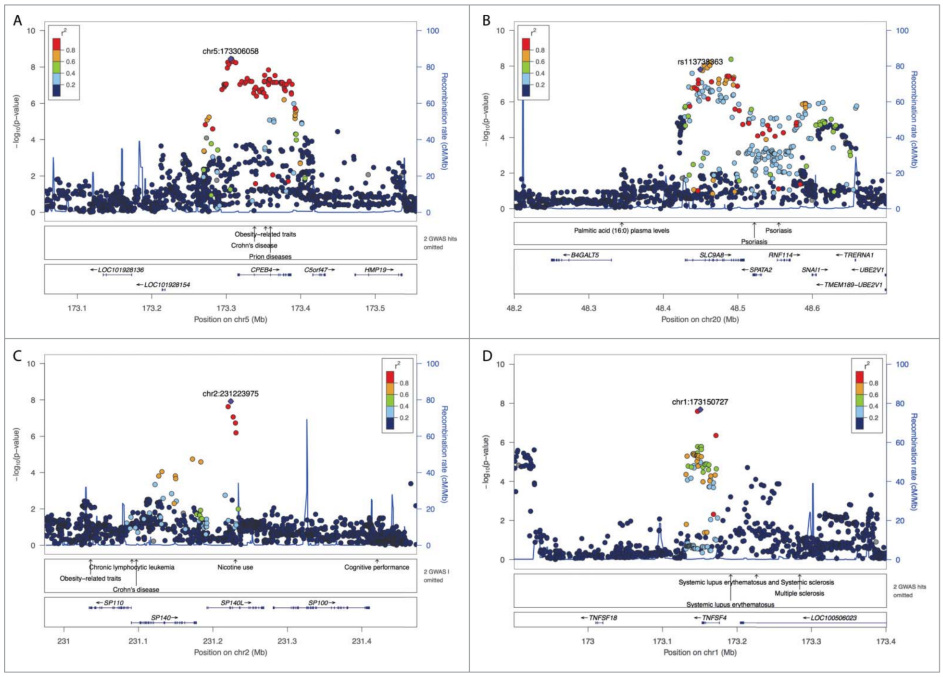


Figure 1. Regional association plots of the β diversity meta-analysis. (A) *TNFSF4/OX40L*, Chromosome 1: 173Mb-173.4Mb, $P_{meta} = 2.1 \times 10^{-8}$; (B) *SP140* and *SP140L*, Chromosome 2: 231Mb-231.4Mb, $P_{meta} = 1.19 \times 10^{-8}$; (C) *CBEP4*, Chromosome 5: 173.1Mb-173.5Mb, $P_{meta} = 3.61 \times 10^{-9}$; (D) *SLC9A8/NHE8*, Chromosome 20: 48.2Mb-48.7Mb, $P_{meta} = 1.54 \times 10^{-8}$.

identified as genome-wide significant in the analysis based on approximate inference. Using the *adonis* function from the *vegan*[10] package for R^{11} and 10^6 random permutations of the genotypes, the ΔF distribution was determined empirically. Comparing P-Values from DBF test and permutation based test, we see a large congruency of the results (Table 1). We could not find any systematic deviations exhibited by the permutation-free method, as all P-values are in the same order of magnitude as those obtained from a classical and widely used permutational approach (Table 1). This is also made evident by the good concordance of the empirical distribution with the approximated probability density function obtained from the DBF test for each of the respective variants under investigation (Fig. 2). While 10^6 permutations only allow to calculate P-values larger than 10^{-6}, all variants with P-values below this threshold in the DBF test showed no permutations with stronger signals than the actual genotype.

Replication of 42 loci identified in mGWAS

The boundaries of the loci provided in Table 1 in Wang *et al.*[6] were evaluated for their replicability using the DBF test. The major difference between both approaches is that the DBF test is based directly on the β diversity matrix, while the previously published approach is based on the ordination of this distance matrix. For 41 of the 42 loci we obtained a nominally significant P-value ($P < 0.05$) at the exact respective position of the lead SNPs. As mentioned earlier, the *SLC9A8* locus on chromosome 20 shows a genome-wide significant association in both analysis strategies (see Table 2). Three more of the lead SNPs showing significant associations in the original article have P-values $<10^{-5}$, and another 5 loci reached this threshold when considering SNPs in the neighborhood – using physical boundaries obtained from the DEPICT analysis – of the lead SNP of the original analysis (see Table 2). Among these loci is one that spans the *BANK1* (B-Cell Scaffold Protein With Ankyrin Repeats 1; chr4:102901822) gene, which was previously reported to be associated with IBD[19] and which is in line with the reported loci reaching genome-wide significance. One locus on chromosome 8 (rs138022915; chr8:19885934) covers the *LPL* (Lipoprotein Lipase) gene. Gene expression of *LPL* was shown to be influenced by the microbiota through altered expression of fasting-induced adipose factor (*Fiaf*) in mice. The only lead SNP not exhibiting a significant P-value < 0.05 is the variant rs225153 (chr11:8853177), however, within the only 0.94 kb spanning locus another variant reaches at least nominal statistical significance (chr11:8852400; $P_{meta} = 2.38 \times 10^{-2}$).

Discussion

The effect of host-genetic variation on the complex phenotype of β diversity of the intestinal microbiota is still largely unknown. We could show, that our adapted method is applicable to microbiome data and yields results in line with classical permutation approaches, without the need of doing millions of permutations per variant, as at least 2×10^7 permutations would be needed to approach the threshold of genome-wide significance. For a typical data set of several millions of imputed genetic variants, this number would easily exceed 10^{14} necessary permutations.

By applying this new method, the DBF test, to β diversity data of 2 independent Northern German cohorts, consisting of a total of almost 1,800 individuals, we could show that variants in genes primarily involved in immune related functions and inflammatory processes showed an association with changes in the gut microbial community. While all for loci are sensible targets with respect to the interactions

Table 1. Comparison of DBF test based [P(DBF)] and permutation based analysis [P(Perm)] of the 4 variants showing significant associations to changes in β diversity in 2 independent Northern-German cohorts. In the case that none of the permutations resulted in a larger ΔF than the actual genotype, P(Perm) is set to $<10^{-6}$. Positions are given as chromosome and position (chr:pos) and are based on the hg19 version of the human genome annotation.

		Focus			Popgen			
rsID	chr:pos	ΔF	P(DBF)	P(Perm)	ΔF	P(DBF)	P(Perm)	Meta P(meta)
rs11811788	chr1:173150727	0.0071569	1.08×10^{-8}	$< 10^{-6}$	0.0034576	0.035664	0.033779	2.10×10^{-8}
rs11678791	chr2:231223975	0.0052987	1.50×10^{-5}	2.5×10^{-5}	0.0050288	0.00019994	0.000234	1.19×10^{-8}
rs67909753	chr5:173306058	0.0073541	4.10×10^{-9}	$< 10^{-6}$	0.0036817	0.01813936	0.017608	1.45×10^{-8}
rs113738363	chr20:48449631	0.0073984	5.82×10^{-9}	$< 10^{-6}$	0.0035011	0.03922766	0.037279	1.54×10^{-8}

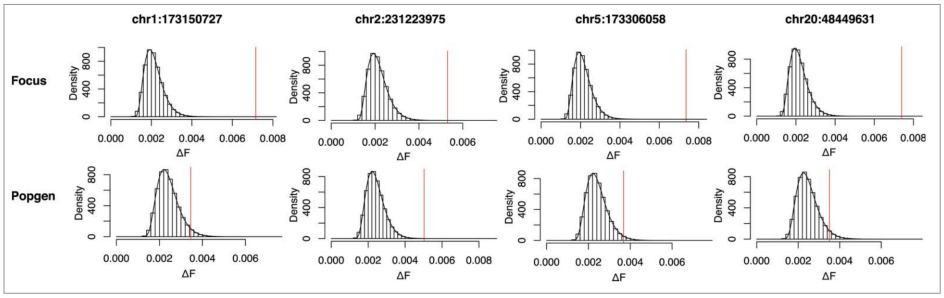


Figure 2. Comparison of the empirical distribution of ΔF from 10^6 permutations of each of the 4 variants in both cohorts with probability density function approximated by using moment matching to Pearson Type III distribution. Red lines indicate the ΔF of the actual genotype distribution in the cohorts.

between host and associated microbes, especially the *SLC9A8/NHE8* gene locus is an intriguing candidate for future studies. This is due to its high expression in goblet cells,[17] its crucial role for mucosal integrity[22] and its potential role in selective bacterial adherence.[29]

The association signal in the *TNFSF4* locus and its role in regulation of cytokines is in line with recent findings underlining the links of the gut microbiota to cytokine production.[30]

Furthermore, 3 of the 4 loci found in our re-analysis are also known to be overlapping with loci associated to different kinds of chronic inflammatory disorders, namely Crohn's disease and psoriasis. Especially for Crohn's disease it was proposed, that host-microbe interactions were, and probably are, a driving factor in the manifestation of the disorder.[18] Moreover, it was shown, that loci associated with Crohn's disease and psoriasis are overlapping to a certain extent[31] and comorbidities of the 2 diseases are widely reported.[32]

Our findings emphasize the role of gut microbes as potential triggers of these diseases, and possibly additional chronic disorders.

The observed differences in significance of the results highlight the difficulties and challenges accompanying mbQTL (microbiome quantitative trait) association analyses of, for example, microbial diversity in connection to host-genetics. The ordination-based analysis described in Wang *et al.*[6] reduces the dimensions of the high-dimensional data to principal coordinates, which has the benefit of removing stochastic noises and pathways with relatively smaller contributions, and reveals the most important pathways affecting the major variable patterns of microbial β diversity, in this case, vitamin-related pathways and bile-acid related genes centered by VDR. However, variation not necessarily displayed by the 2 major axes of the ordination might not be detected by this method. Thus, the DBF test serves as an addition to the previously published results on the connection between β diversity and host-genetics, strengthening especially the importance of those loci exhibiting strong to intermediate results in both analyses.

However, while these results are intriguing, they should mainly serve as a starting point and perspective for subsequent analyses in larger and hence better powered cohorts, investigating the genetic effects of host-microbiota interactions, leading to additional and potentially more robust signals for the complex trait of β diversity, overcoming the challenges of small effect sizes, sensitivity to technical differences and confounding environmental factors. In a recent review, Zhernakova and colleagues further discuss the phenomenon that there is little overlap in the findings between all the mbQTL studies with more than 1000 samples analyzed published so far, likely because there were many significant differences between the data sets and methods that were used.[33] In summary, classical GWAS methodology cannot be used for mbQTL studies, given the complexity of the trait under study, and the development of best-practice workflows and stringent thresholds are in its infancy. As shown in this study, the DBF test deserves a careful consideration for future studies.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Table 2. Replication of the 42 genome-wide significant loci previously found to be associated with β diversity. We modified Table 1 from Wang *et al.*[6] as follows: *Lead SNP* corresponds to position and P-value from the Meta-Analysis of the DBF-test applied to the Popgen and FoCus cohorts. *Best in Locus*: Position and lowest P-value of the DBF test meta-analysis in the locus defined by the columns 'Locus Start' and 'Locus End'. *Positions* are based on the hg19 version of the human genome annotation. P-values in bold font indicate a value below of $P < 0.05$. Additional italic fonts indicate P-values $< 10^{-5}$.

SNP_ID	Chr	A1	A2	Locus Start	Locus End	Nearest Gene	Genes in Locus	Effect size Wang *et al.*	Lead SNP Position	P(Meta DBF)	Best in Locus Position	P(Meta DBF)
rs804427	1	A	C	33538964	33623510	*AK2*	*ADC; TRIM62; AK2*	0.79%	chr1:33538964	**4.96×10^{-3}**	chr1:33595212	**2.09×10^{-3}**
rs1288616	1	G	A	53885577	53965248	*DMRTB1*	*DMRTB1*	0.76%	chr1:53952777	**1.58×10^{-4}**	chr1:53946485	**9.29×10^{-5}**
rs1102737	1	G	A	172700868	172779833	*FASLG*		0.66%	chr1:172777616	**2.24×10^{-3}**	chr1:172747021	**1.97×10^{-3}**
rs72853661	2	T	C	25323083	25453968	*POMC*	*POMC; EFR3B*	0.79%	chr2:25439262	**6.57×10^{-5}**	chr2:25439758	***1.42×10^{-6}***
rs7567349	2	A	G	61384324	61853037	*XPO1*	*AHSA2; USP34;XPO1; KIAA1841*	0.76%	chr2:61839853	***1.49×10^{-6}***	chr2:61486628	***2.22×10^{-7}***
rs2010917	2	T	C	135172338	135197891	*MGAT5*	*MGAT5*	0.74%	chr2:135194856	**5.04×10^{-5}**	chr2:135183686	***6.13×10^{-6}***
rs71415332	2	G	A	102309520	102616128	—	*IL1R2; MAP4K4*	0.68%	chr2:102499952	**2.56×10^{-5}**	chr2:102529630	***2.85×10^{-6}***
rs4670302	2	T	G	33808725	34068392	*FAM98A*	*FAM98A*	0.92%	chr2:34068392	**6.53×10^{-3}**	chr2:34033733	**7.43×10^{-4}**
rs6711771	2	C	G	34339420	34491584	—	—	0.71%	chr2:34339420	**1.77×10^{-2}**	chr2:34421584	**8.92×10^{-4}**
rs13099587	3	G	A	146250561	146275555	*PLSCR1*	*PLSCR1*	0.70%	chr3:146268616	**3.60×10^{-3}**	chr3:146275555	**1.09×10^{-3}**
rs9647379	3	G	C	171759410	171833266	*FNDC3B*	*FNDC3B*	0.75%	chr3:171785168	**8.98×10^{-5}**	chr3:171785168	**8.98×10^{-5}**
rs143050036	3	C	T	49898318	50208819	*SEMA3F*	*RBM5; MST1R; CAMKV; MON1A; RBM6; SEMA3F*	0.75%	chr3:50071965	**1.15×10^{-2}**	chr3:49987475	**1.33×10^{-5}**
rs60500975	4	A	T	102769693	102929034	—	*BANK1*	0.82%	chr4:102901822	***2.03×10^{-6}***	chr4:102885147	***1.67×10^{-6}***
rs62367773	5	A	G	74171398	74220999	*FAM169A*		0.67%	chr5:74179975	**1.55×10^{-4}**	chr5:74193565	**6.08×10^{-5}**
rs1292672	6	C	T	87217958	87509434	*HTR1E*		0.70%	chr6:87432577	**9.91×10^{-5}**	chr6:87242812	**4.85×10^{-5}**
rs35148810	7	C	T	151515842	151530983	—	*PRKAG2*	0.83%	chr7:151520485	**8.69×10^{-4}**	chr7:151520550	**3.77×10^{-4}**
rs12705241	7	A	C	104219681	104381102	—	*LHFPL3*	0.76%	chr7:104258313	**2.01×10^{-3}**	chr7:104258313	**2.01×10^{-3}**
rs13260600	8	C	T	3705807	3713004	*CSMD1*	*CSMD1*	0.77%	chr8:3705807	**8.45×10^{-4}**	chr8:3705807	**8.45×10^{-4}**
rs138022915	8	T	C	19815256	19939049	*LPL*	*LPL*	0.73%	chr8:19885934	**2.19×10^{-4}**	chr8:19876234	***4.45×10^{-6}***
rs11986935	8	T	A	10576753	10732050	*SOX7*	*SOX7; PINX1*	0.97%	chr8:10691549	**1.05×10^{-5}**	chr8:10695125	***6.63×10^{-6}***
rs7818750	8	G	A	135273640	135299611	*ZFAT*		0.74%	chr8:135274269	**1.83×10^{-3}**	chr8:135273640	**4.42×10^{-4}**
rs1325919	9	C	T	37626956	37650386	*FRMPD1*		0.67%	chr9:37642802	**4.34×10^{-3}**	chr9:37638047	**1.93×10^{-3}**
rs7082134	10	A	G	87865009	87884110	*GRID1*	*GRID1*	0.84%	chr10:87865009	**4.05×10^{-4}**	chr10:87884110	**3.71×10^{-4}**
rs2251536	11	G	C	8852239	8853177	—	*ST5*	0.76%	chr11:8853177	1.57×10^{-1}	chr11:8852400	**2.38×10^{-2}**
rs4472950	11	C	T	120798714	120853675	—	*GRIK4*	0.69%	chr11:120807892	**4.56×10^{-4}**	chr11:120798714	**3.16×10^{-4}**
rs7974353	12	T	C	48256280	48270596	—	*VDR*	0.75%	chr12:48269798	**4.69×10^{-3}**	chr12:48263162	**1.22×10^{-3}**
rs4760399	12	T	C	93011759	93081307	*C12orf74*		0.67%	chr12:93047282	**1.80×10^{-2}**	chr12:93021626	**2.30×10^{-3}**
rs6573564	14	T	A	65119676	65157187	*PLEKHG3*		0.73%	chr14:65142395	**1.72×10^{-5}**	chr14:65141759	**1.72×10^{-5}**
rs12910631	15	G	T	26603288	26622999	—		0.79%	chr15:26606605	**1.42×10^{-4}**	chr15:26606605	**1.42×10^{-4}**
rs8040493	15	T	G	101414167	101418682	—		0.65%	chr15:101418659	**6.27×10^{-4}**	chr15:101418335	**5.03×10^{-5}**
rs293377	15	G	C	89623490	89635268	*ABHD2*	*ABHD2*	0.70%	chr15:89634414	**3.83×10^{-3}**	chr15:89623490	**1.89×10^{-3}**
rs8055365	16	T	C	84566729	84581275	*KIAA1609*	*KIAA1609*	0.70%	chr16:84580531	**8.98×10^{-5}**	chr16:84580531	**8.98×10^{-5}**
rs59986499	16	G	A	3065924	3097940	*CLDN6*	*MMP25; TNFRSF12A; CLDN6; CCDC64B; HCFC1R1; THOC6*	0.68%	chr16:3069752	**8.73×10^{-3}**	chr16:3082157	**6.93×10^{-3}**
rs12931878	16	A	G	11031741	11207817	*CLEC16A*	*DEXI; CLEC16A*	0.65%	chr16:11042194	**2.02×10^{-3}**	chr16:11082874	**9.66×10^{-5}**
rs62085746	17	T	C	66166300	66213540	*AMZ2*		0.69%	chr17:66196145	**2.04×10^{-3}**	chr17:66196145	**2.04×10^{-3}**
rs16969051	17	C	T	32248813	32258877	*ACCN1*	*ACCN1*	0.65%	chr17:32258877	**5.12×10^{-4}**	chr17:32258877	**5.12×10^{-4}**
rs12601692	17	A	G	782416	794333	—	*NXN*	0.68%	chr17:782416	**1.57×10^{-2}**	chr17:782416	**1.57×10^{-2}**
rs2267922	19	C	G	18217350	18289634	*IFI30*	*MAST3; IFI30;PIK3R2*	0.77%	chr19:18278766	***3.32×10^{-7}***	chr19:18278766	***3.32×10^{-7}***
rs273647	19	C	G	51739767	51766748	*C19orf75*	*CD33; C19orf75*	0.84%	chr19:51751858	**1.38×10^{-3}**	chr19:51766748	**1.97×10^{-5}**
rs4809760	20	A	G	48428863	48591125	*SLC9A8*	*RNF114; SLC9A8; SPATA2*	0.85%	chr20:48454671	***9.28×10^{-9}***	chr20:48490801	***4.15×10^{-9}***
rs2835692	21	A	G	38657572	38704886	*DSCR3*		0.68%	chr21:38670335	**2.11×10^{-4}**	chr21:38657572	**1.13×10^{-4}**
rs9917541	22	C	A	31520338	31531133	*PLA2G3*	*PLA2G3; INPP5J*	0.71%	chr22:31529043	**1.30×10^{-2}**	chr22:31529043	**1.30×10^{-2}**

Funding

This work was supported by the German Research Foundation (DFG) Collaborative Research Center (CRC) 1182, "Origin and Function of Metaorganisms" and the DFG Excellence Cluster 306, "Inflammation at Interfaces," and the German Federal Ministry of Education and Research (BMBF) project CP3 in "SysINFLAME."

ORCID

Femke-Anouska Heinsen http://orcid.org/0000-0003-3652-6402

References

[1] Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biology. 2015; 16(1):191. PMID:26374288. doi:10.1186/s13059-015-0759-1

[2] Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. Genome-wide association studies of the human gut microbiota. Plos One. 2015; 10(11):e0140301. PMID:26528553. doi:10.1371/journal.pone.0140301

[3] Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, et al. Human genetics shape the gut microbiome. Cell. 2014; 159(4):789-99. PMID:25417156. doi:10.1016/j.cell.2014.09.053

[4] Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. Genetic determinants of the gut microbiome in UK twins. Cell Host & Microbe. 2016; 19(5):731-43. doi:10.1016/j.chom.2016.04.017

[5] Hua X, Song L, Yu G, Goedert JJ, Abnet CC, Landi MT, Shi J. MicorbiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition. BioRxiv 2015; doi:10.1101/031187

[6] Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen F-A, Rühlemann MC, Szymczak S, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genetics. 2016; 48(11):1396-406. PMID:27723756. doi:10.1038/ng.3695

[7] Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, et al. Population-level analysis of gut microbiome variation. Science. 2016; 352(6285):560-4. PMID:27126039. doi:10.1126/science.aad3503

[8] Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila A V, Falony G, Vieira-Silva S, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 2016; 352(6285):565-9. PMID:27126040. doi:10.1126/science.aad3369

[9] McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001; 82(1):290-7. doi:10.1890/0012-9658(2001)082%5b0290:FMMTCD%5d2.0.CO;2

[10] Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens M, Wagner H. vegan: community ecology package. R package version 2.0-10. R package version 2013; 1

[11] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna Austria. 2016; 0:{ISBN} 3-900051-07-0

[12] Minas C, Montana G. Distance-based analysis of variance: Approximate inference. Statistical Analysis and Data Mining: The ASA Data Sci J. 2014; 7(6):450-70. doi:10.1002/sam.11227

[13] Winkler AM, Ridgway GR, Douaud G, Nichols TE, Smith SM. Faster permutation inference in brain imaging. NeuroImage. 2016; 141:502-16. PMID:27288322. doi:10.1016/j.neuroimage.2016.05.068

[14] Clayton D. snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.26.0. 2015

[15] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly MJ, et al. PLINK: a tool set for whole-genome and population-based linkage analyses. Am J Hum Genetics. 2007; 81:559-75. doi:10.1086/519795

[16] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26(17):2190-1. PMID:20616382. doi:10.1093/bioinformatics/btq340

[17] Xi H, Schwartz R, Engel I, Murre C, Kersh GJ. Interplay between RORgammat, Egr3, and E proteins controls proliferation in response to pre-TCR signals. Immunity. 2006; 24(6):813-26. PMID:16782036. doi:10.1016/j.immuni.2006.03.023

[18] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491(7422):119-24. PMID:23128233. doi:10.1038/nature11582

[19] Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature Genetics. 2015; 47(9):979-86. PMID:26192919. doi:10.1038/ng.3359

[20] Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, Strawbridge RJ, Pers TH, Fischer K, Justice AE, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015; 518(7538):187-96. PMID:25673412. doi:10.1038/nature14132

[21] Xu H, Li Q, Zhao Y, Li J, Ghishan FK. Intestinal NHE8 is highly expressed in goblet cells and its expression is subject to TNF-α regulation. Am J Physiol Gastrointestinal Liver Physiol. 2016; 310(2):G64-9. doi:10.1152/ajpgi.00367.2015

[22] Wang A, Li J, Zhao Y, Johansson MEV, Xu H, Ghishan FK. Loss of NHE8 expression impairs intestinal mucosal integrity. Am J Physiol Gastrointestinal Liver Physiol. 2015; 309(11):G855-64, ajpgi.00278.2015

[23] Capon F, Bijlmakers M-J, Wolf N, Quaranta M, Huffmeier U, Allen M, Timms K, Abkevich V, Gutin A, Smith R, et al. Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene. Hum Mol Genetics. 2008; 17 (13):1938-45. doi:10.1093/hmg/ddn091

[24] Stuart PE, Nair RP, Ellinghaus E, Ding J, Tejasvi T, Gudjonsson JE, Li Y, Weidinger S, Eberlein B, Gieger C, et al. Genome-wide association analysis identifies three psoriasis susceptibility loci. Nat Genetics. 2010; 42(11):1000-4. PMID:20953189. doi:10.1038/ng.693

[25] Baurecht H, Hotze M, Brand S, Büning C, Cormican P, Corvin A, Ellinghaus D, Ellinghaus E, Esparza-Gordillo J, Fölster-Holst R, et al. Genome-wide comparative analysis of atopic dermatitis and Psoriasis gives insight into opposing genetic mechanisms. Am J Hum Genetics. 2015; 96(1):104-20. doi:10.1016/j.ajhg.2014.12.004

[26] Scher JU, Ubeda C, Artacho A, Attur M, Isaac S, Reddy SM, Marmon S, Neimann A, Brusca S, Patel T, et al. Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. Arthritis Rheumatol. 2015; 67(1):128-39. doi:10.1002/art.38892

[27] Mehta S, Cronkite DA, Basavappa M, Saunders TL, Adiliaghdam F, Amatullah H, Morrison SA, Pagan JD, Anthony RM, Tonnerre P, et al. Maintenance of macrophage transcriptional programs and intestinal homeostasis by epigenetic reader SP140 Sci Immunol. 2017; 2(9). doi:10.1126/sciimmunol.aag3160

[28] Croft M, So T, Duan W, Soroosh P. The significance of OX40 and OX40L to T-cell biology and immune disease. Immunol Rev. 2009; 229(1):173-91. PMID:19426222. doi:10.1111/j.1600-065X.2009.00766.x

[29] Liu C, Xu H, Zhang B, Johansson ME V, Li J, Hansson GC, Ghishan FK. NHE8 plays an important role in mucosal protection via its effect on bacterial adhesion. AJP: Cell Physiol. 2013; 305(1):C121-8

[30] Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, Horst R ter, Jansen T, Jacobs L, Bonder MJ, et al. Linking the human gut microbiome to inflammatory cytokine production capacity. Cell. 2016; 167(7):1897. PMID:27984736. doi:10.1016/j.cell.2016.11.046

[31] Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, Park YR, Raychaudhuri S, Pouget JG, Hübenthal M, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nature Genetics. 2016; 48 (5):510-8. PMID:26974007. doi:10.1038/ng.3528

[32] Lee FI, Bellary S V, Francis C. Increased occurrence of psoriasis in patients with Crohn's disease and their relatives. Am J Gastroenterol. 1990; 85(8):962-3. PMID:2375323

[33] Kurilshikov A, Wijmenga C, Fu J, Zhernakova A. Host genetics and gut microbiome: challenges and perspectives. Trends Immunol. 2017. [Epub ahead of print] PMID:28669638. doi:10.1016/j.it.2017.06.003

Article C: ABO histo-blood groups influence gut microbiome, with causal relationship between *Bacteroides* and inflammatory bowel disease

ABO histo-blood groups influence gut microbiome, with causal relationship between *Bacteroides* and inflammatory bowel disease

Authors and affiliations:
Malte Christoph Rühlemann[1], Britt Marie Hermes[2,3,4], Corinna Bang[1], Shauni Doms[2,3], Lucas Moitinho-Silva[1,5], Louise Bruun Thingholm[1], Fabian Frost[6], Frauke Degenhardt[1], Michael Wittig[1], Jan Kässens[1], Frank Ulrich Weiss[6], Annette Peters[7,8], Klaus Neuhaus[9], Uwe Völker[10], Henry Völzke[10], Georg Homuth[10], Matthias Laudes[11], Wolfgang Lieb[12], Dirk Haller[9,13], Markus Maximilian Lerch[6], John Baines[2,3], Andre Franke[1]

[1] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany
[2] Evolutionary Genomics, Max-Planck-Institute for Evolutionary Biology, Plön, Germany
[3] Institute of Experimental Medicine, Kiel University, Kiel, Germany
[4] Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany
[5] Department of Dermatology, Kiel University, Kiel, Germany
[6] Department of Medicine A, University Medicine Greifswald, Greifswald, Germany
[7] Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany
[8] German Center for Diabetes Research (DZD), Neuherberg, Germany
[9] ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany
[10] Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
[11] Department of Internal Medicine 1, Kiel University, Kiel, Germany
[12] Institute of Epidemiology, Kiel University, Kiel, Germany
[13] Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany

Correspondence should be addressed to: Prof. Dr. Andre Franke, Institute of Clinical Molecular Biology (IKMB), Kiel University, Rosalind-Franklin-Str. 12, 24106 Kiel, Germany. Email: a.franke@mucosa.de

Abstract

The intestinal microbiome is implicated as an important modulating factor in multiple inflammatory,[1,2] neurologic,[3] and neoplastic diseases.[4] Recent genome-wide association studies yielded inconsistent, underpowered and rarely replicated results such that the role of human host genetics as a contributing factor to microbiome assembly and structure remains uncertain.[5–10] In a genome-wide association analysis of 8,956 German individuals, we identified 32 genetic loci to be associated with single bacteria and overall microbiome composition. Further analyses confirm the identified associations of ABO histo-blood groups and FUT2 secretor status with *Bacteroides* and *Faecalibacterium*. Mendelian randomization analysis suggests causative and protective effects of gut microbes, with clade-specific and sometimes contrasting effects on inflammatory bowel disease. This holistic investigative approach of the host, its genetics, and its associated microbial communities as a 'metaorganism' broadens our understanding of disease aetiology and emphasizes the potential for implementing microbiota in disease treatment and management.

Main

We conducted the largest single country genome-wide association analysis of microbial traits followed by Mendelian Randomization (MR) analysis to elucidate the genetic link between humans and their associated microbiota. Our study comprised five independent cohorts from German biobanks located in Northern Germany (Kiel, Schleswig-Holstein; PopGen[11], n=724; FoCus, n=957), North-Eastern Germany (Greifswald, Mecklenburg-Western Pomerania; SHIP, n=2,029; SHIP-TREND, n=3,382),[12,13] and Southern Germany (Augsburg, Bavaria; KORA, n=1,864;[14,15] see **Methods** for details).

Baseline comparisons show similarities in anthropometric measures and microbial community compositions between cohorts (**Figure 1, Supplemental Material**). Taxonomic

groups and sequence similarity clusters included in the univariate analysis, henceforth called microbial features, covered between 98.4% and 98.7% of the whole community at the phylum level and between 77.8% (PopGen) and 82.6% (SHIP-TREND) at the genus level across cohorts. These data indicate that the cohorts share a common core microbiota (cohort-level summaries of microbial features can be found in **Supplementary Table S1**).

Association analysis was performed for 146 univariate microbial features across different levels for presence-absence patterns using logistic regression and for 225 features assessing differential abundances using (generalized) linear models (see **Methods**). Whole community multivariate association analysis of genus-level Bray-Curtis dissimilarity and weighted UniFrac distance[16] were optimized to decrease computational time (see **Supplemental Material**). We combined per-cohort results using inverse-variance weighted meta-analysis for univariate and sample-size weighted meta-analysis for multivariate analyses (see **Methods**). To ensure robustness of results, genome-wide significant results ($p_{Meta}<5\times10^{-8}$) were reported when supported by nominal significance ($p<0.05$) in at least two cohorts.

Accordingly, we reveal a total of 38 associations with microbial features and community composition involving 32 genomic loci (**Table 1, Figure 2A**), among which four associations stem from the multivariate analysis, 17 from the univariate abundance analysis, and 11 from the presence-absence patterns. The top 10,000 genetic variants for univariate and multivariate analyses are summarized in **Supplementary Tables S2-4**. All results can be queried via the mGWAS results browser (http://52.29.129.36:3838/). Of note, univariate signals with overlapping genetic loci in all cases are found from the same taxonomic group at a different taxonomic and/or clustering level.

Although not meeting the initial inclusion criteria (see **Methods**), the genus *Bifidobacterium* was included in the analysis. Its connection with the lactase gene locus (*LCT*) on chromosome 2 is important, as it is the only signal replicating across numerous previous studies.[5,9,17] The meta-analysis shows a clear association peak at *LCT* with 53 variants displaying *p*-values lower than the suggestive $p<10^{-5}$ threshold, with the lowest for rs3820794 (chr2:136505546; $p_{Meta}= 5.62\times10^{-7}$; **Figure 2B**). This is supported by nominally significant *p*-values in four of the five cohorts, with only the FoCus cohort showing a *p*-value above nominal significance ($p_{FoCus}= 0.069$), confirming the association between *LCT* and *Bifidobacterium* and the validity of the herein-used model of choice.

Our obtained genome-wide association results point to immune-mediated interactions of host and microbiota, e.g. the association detected for OTU99_55 (*Barnesiella;* OTU: operational taxonomic unit) and variants in the biliverdin reductase A (*BLVRA*; rs623108; $p_{Meta}=1.05\times10^{-8}$; **Figure 2C**) locus. Biliverdin reductase A was previously shown to inhibit Toll-like receptor 4 (TLR4) gene expression.[18] TLR4 is a pattern recognition receptor that initiates an immune response to bacterial lipopolysaccharides (LPS) present in many Gram-negative bacteria.[19] *Barnesiella*, which itself is Gram-negative, is negatively associated with LPS-induced interferon-gamma production, suggesting a contribution of this commensal to homeostasis by immune- or TLR4-signal-modulation.

We identified two independent univariate associations with a locus surrounding the histo-blood group ABO system transferase (*ABO*) gene. One *ABO* gene signal for differential abundance includes OTU99_16 belonging to *Faecalibacterium* (rs3758348; chr9:136155000; $p_{Meta}=6.16\times10^{-9}$; **Figure 2D**), which is accompanied by a second signal ~100kb downstream in the surfeit locus protein 4 (*SURF4*) gene (chr9:136239399; $p_{Meta}=4.33\times10^{-9}$). The second *ABO* association is between rs8176632 allele T and the

increased prevalence of a *Bacteroides* OTU (OTU97_27; rs8176632; chr9:136152547; $p_{Meta}=6.87\times10^{-10}$; Figure 2E). Interestingly, this same *Bacteroides* OTU is also significantly associated with variants at the *BACH2* (BTB domain and CNC homolog 2) gene locus (chr6:90978161; $p_{Meta}=4.58\times10^{-10}$). Moreover, a non-significant association between this *Bacteroides* OTU is present for the *FUT2* (Galactoside 2-alpha-L-fucosyltransferase 2) locus, whereby the strongest signal is from the missense variant rs602662 (chr19:49206985; $p_{Meta}=4.46\times10^{-7}$), which is in strong linkage disequilibrium (LD) with variant rs601338 (R^2 = 0.8898) encoding the FUT2 secretor phenotype. This variant determines whether the fucosyl-precursor for the ABO blood-group system is synthesized on mucosal surfaces in the body and secretions. Individuals homozygous for this missense variant do not have the ABO-encoded antigen on mucosal cells, independent of the ABO allele (i.e. display the non-secretor phenotype; **Figure 3B-D**). Variants at *FUT2* and *BACH2*, correlated with *Bacteroides* OTU97_27 in this study, were previously shown to be associated with inflammatory bowel disease (IBD).[20–23]

For a focused evaluation of blood-group dependent associations with microbial features, we investigated ABO histo-blood group and FUT2 secretor status (see **Methods**). The prevalence or abundance of eight taxonomic groups show at least one FDR-corrected significant association ($q<.05$) with either ABO histo-blood group alleles, secretor status, or their interaction (**Figure 3A**; **Supplementary Table S5**). These results demonstrate a positive correlation between non-O blood group and positive secretor status and the prevalence of the aforementioned *Bacteroides* OTU97_27 in four of the five cohorts ($p_{Meta}=3.65\times10^{-10}$). Intriguingly, a different *Bacteroides* branch, represented by OTU97_12, OTU99_12, and TestASV_13, exhibited significant associations with ABO histo-blood group status as well, however in this case characterized by an inverse relationship between

prevalence and non-O blood group alleles (p_{Meta}=2.1×10^{-4}). Together, these findings suggest histo-blood group dependent effects on *Bacteroides* subclades.

In addition, the model points to an association between *Faecalibacterium* OTU99_16 and the ABO histo-blood group A allele (p_{Meta}=4.7×10^{-6}). A significant association between *Holdemanella* and ABO is also identified, although the signal is exclusively driven by the SHIP-TREND cohort with only weak support from the remaining cohorts. Further, FUT2 secretor status is associated with differential abundance of *Roseburia* OTU97_30, independent of ABO blood type (p_{Meta}=4.79×10^{-6}). In conclusion, the analyses reveal a specific impact of the human ABO blood groups and secretor status on members of the intestinal community.

Mendelian randomization (MR) has recently become a popular tool to infer causal relationships of complex traits in observational data.[24] MR analysis was performed for all univariate microbial features as "exposures" and 41 selected binary traits from the MR-Base database[25] as outcomes (see **Methods**). This allows us to assess the causal effect of microbial features on disease. A total of 17 comparisons reach the per-trait suggestive threshold of p<1.22×10^{-3}, whereas three traits are below the global FDR correction threshold q<0.05 (**Table 2; Supplementary Table S6**), suggesting an influence of microbial features on either IBD or its sub-entities Crohn's disease (CD) and ulcerative colitis (UC). For example, the presence of *Prevotella* ASV (TestASV_18) appears to significantly protect against CD development (β=-0.257). Ten out of the 17 suggestive microbial effects on host traits point to either IBD, CD, or UC. Interestingly, different subgroups belonging to the genus *Bacteroides* were either protective for UC (OTU97_12 and OTU99_12, β=0.310) or causative for CD (TestASV_12, β=-1.034). Previous work has revealed *Bacteroides*-subgroups and *Prevotella* as the main determinants of gut enterotypes.[26,27] Significantly

158 more CD patients have the *Bacteroides 2*-, but not the *Bacteroides 1*-enterotype and virtually
159 none have the *Prevotella*-enterotype.[27] These descriptions support our findings of the
160 contradicting role of *Bacteroides* species in IBD subgroups. Lower levels of *Bacteroides* are
161 associated with IBD development,[28] while animal studies demonstrate a colitis induction by
162 *Bacteroides* in genetically susceptible mice compared to controls.[29] These results once
163 again emphasize the large variability of *Bacteroides* taxa in connection to genetics and
164 disease.
165
166 Further results from MR confirm host-microbiome interactions previously described in
167 observational studies. *Parabacteroides* show a protective effect on the "Obesity class 2" trait
168 (β=-0.568), supporting previous experimental observations of *Parabacteroides* species
169 alleviating obesity effects in mice.[30] Interestingly, none of the microbial traits with causal
170 effects reach genome-wide significance at any locus in the univariate analysis. In addition to
171 MR, replication of previously associated loci and gene-set enrichment and tissue specificity
172 analysis was performed using the FUMA web service[31] (see **Supplemental Material**). The
173 obtained results indicate metabolic interactions between the host and associated microbes
174 and an enrichment of genes derived from metabolic and inflammatory traits.
175
176 Our results highlight the power of combining multiple independent cohorts for genomic
177 association analyses of microbial features, as they allow for robust and replicable results.
178 Although a direct influence of ABO histo-blood group and secretor status on the microbiome
179 is debated,[32,33] our results support this interaction, potentially acting as a modulator in
180 diseases for which variants in histo-blood groups and the microbiome were independently
181 reported as risk factors,[20,34–36] The suggestive causative role of *Bacteroides* in patients
182 genetically susceptible to IBD development is notable, as multiple independent, and
183 sometimes contrasting, results were previously reported from host-microbe association and

MR analyses. The multifaceted role of *Bacteroides* in the human gut microbiome is likely insufficiently captured by 16S rRNA gene amplicon-based surveys and may therefore require future in-depth strain-level analysis. Nevertheless, our results suggest an important role of the human ABO histo-blood group antigens as candidates for direct modulation of the human metaorganism in health and disease.

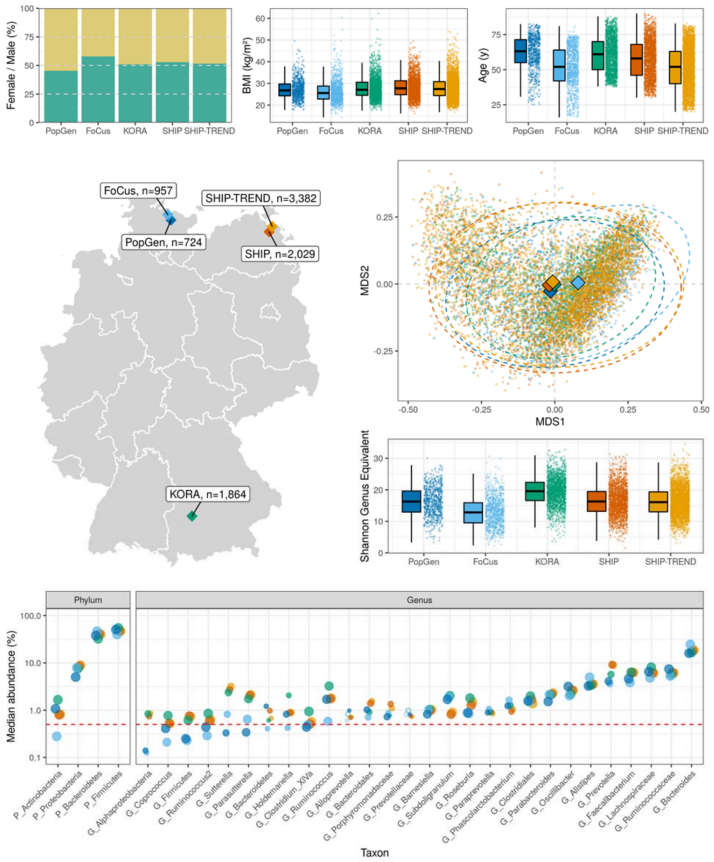


Figure 1: Summary of cohort properties. **(A)** Composition of study participants sex. **(B)** Distribution of participants BMI and **(C)** age. **(D)** Biobank/cohort locations in Germany. **(E)** Ordination of all samples based on genus-level Bray-Curtis dissimilarity. **(F)** Distribution of alpha diversities as calculated by

8

Shannon diversity genus-level equivalent and the number of observed genera. **(G)** Comparison of relative abundances of phylum- and genus-level taxonomic groups that met the inclusion criteria for the genome-wide association study in the five analysed German cohorts. Y axis represents the median abundance of the samples with non-zero abundance of the respective taxa, point size is relative to the prevalence of the respective taxon in the cohort. Taxa with cohort prevalence below the inclusion threshold of 20% are displayed as empty circles. The dashed red line represents the abundance threshold of .5% for inclusion in the analysis. Taxa are arranged from left to right by the lowest median abundance over all cohorts from high to low. Cohort-level summaries of microbial features can be found in **Supplementary Table S1**.

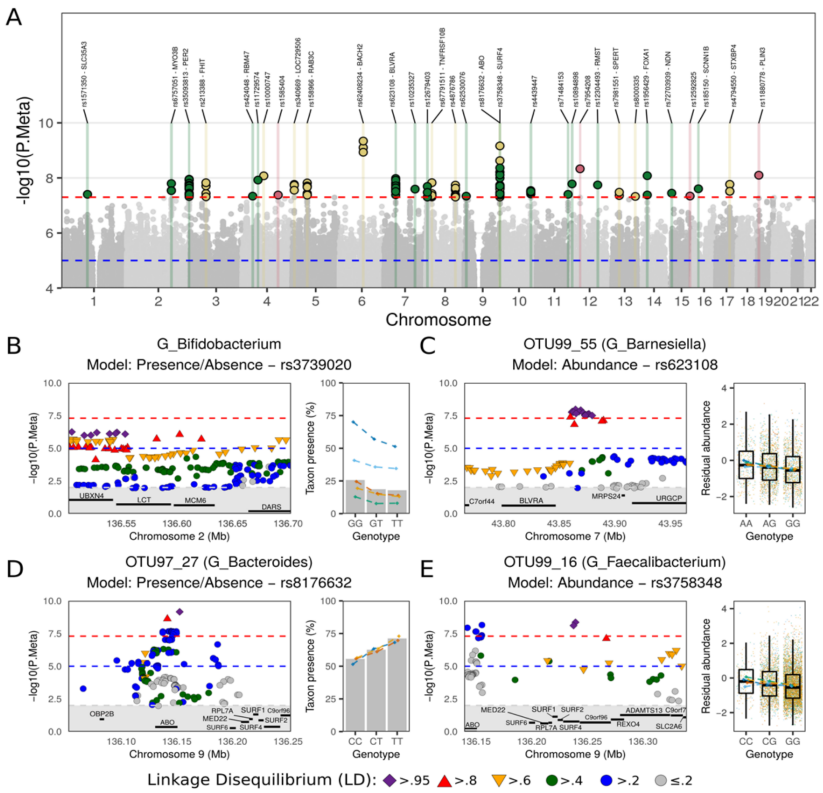


Figure 2: *Genome-wide association analysis results.* **(A)** Manhattan plot of *p*-values from the meta-analysis across all tested traits, the lowest *p*-value at each position is shown. Colour coding by analysis type. Green: abundance models; Yellow: presence-absence models (logistic regression); Red: beta diversity. Regional association plot of: **(B)** genus *Bifidobacterium* presence-absence test with variants in the *LCT* gene locus. **(C)** OTU97_55 (*Barnesiella)* abundance vs. variants at the biliverdin reductase A (*BLVRA*) gene locus. **(D)** OTU99_16 (*Faecalibacterium*) abundance vs. variants in the *ABO/SURF4* gene locus. **(E)** OTU97_27 (*Bacteroides*) presence-absence vs. *ABO* variants. The per-cohort feature abundance means and presences for each genotype are given by the diamonds in the respective colours. In panels (C) and (E) all residual abundance for the individual samples are displayed as dots in the respective colours of the cohort.

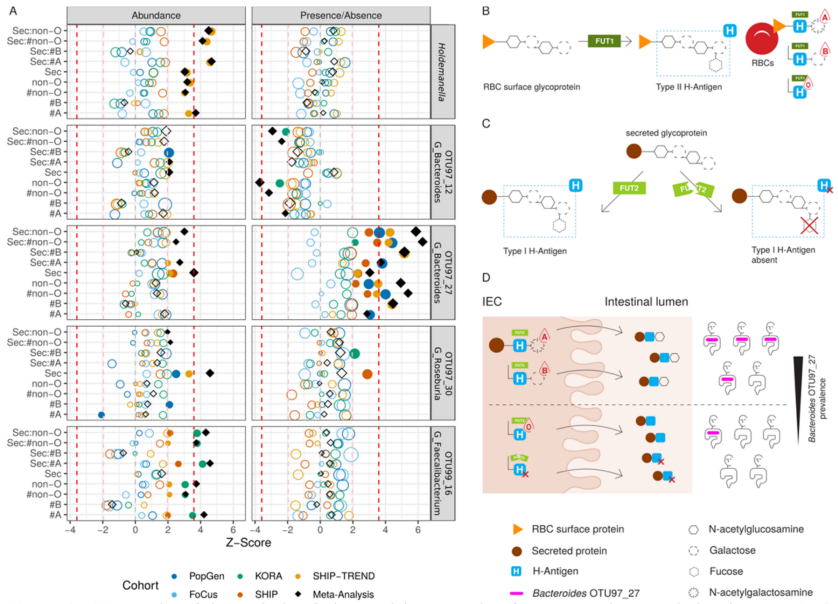


Figure 3: (A) Results of the analysis of nine models connecting feature prevalence and abundance to ABO blood group alleles and FUT2 secretor status. Shown are all univariate microbial features with at least one meta-analysis *q*-value < 0.05. *Holdemanella* is shown also representing *Holdemanella* OTU97_33, and *Bacteroides* OTU97_12 is shown representing also OTU99_12 and TestASV_13 of the same *Bacteroides* subclade with respective identical results. The Y-axes represent the nine models applied, investigating the linear effects of the number of A (#A) and B (#B) histo-blood group alleles and their sum (#non-O), as well as the effects of binary traits O vs. non-O histo-blood group (non-O) and FUT2 secretor status (Sec). The statistical interaction of Sec with all former traits is also included, indicated by the colon (:) symbol; The X-axis shows the Z-scores of the respective models. Symbols are coloured according to cohort, black diamonds represent the result of the meta-analysis of all five cohorts. Symbol size represents absolute effect size. *p*-values < 0.05 are displayed as solid shapes. Dashed vertical lines represent Z-values corresponding to nominal significance (light red line: $p<0.05$; $Z=\pm1.96$; two-sided) and adjusted significance (dark red line: $q<0.05$; $Z=\pm3.59$; two-sided). The complete results can be found in **Supplementary Table S5. (B)** The type II H-antigen on red blood cells (RBCs) is completed by addition of a fucose sugar by the enzyme Fucosyltransferase 1 (FUT1). Subsequently the A- and B-antigens are synthesized by addition of N-acetylgalactosamine or galactose, respectively. In individuals with a

O histo-blood group, no additional sugars can be added to the H-antigen. **(C)** On secreted proteins and mucosal cells, the fucosylated type I H-antigen is synthesized by the enzyme Fucosyltransferase 2 (FUT2). In individuals homozygous for the rs601338-A missense variant in *FUT2* – also known as non-secretors – there is no addition of a fucosyl-group, resulting in no H-antigen. **(D)** Consequently, no additional sugars are added to the precursor of the H-antigen in non-secretors, irrespective of the individuals' histo-blood group genotype at ABO. *Bacteroides* OTU97_27 exhibits higher prevalence in individuals with non-O ABO histo-blood groups and functioning FUT2 as compared to individuals with O histo-blood group or FUT2 non-secretors.

Table 1: Results summary of the genome-wide association analysis for beta diversity (column "Analysis": Beta), logistic regression of presence/absence patterns (LR) and analysis of abundances (NB) ordered and enumerated by genomic location of the loci. A single-variant association test was performed for each cohort and each microbial feature, adjusting the respective model for the first ten genetic principle components, age, sex and body-mass index (BMI). Results were meta-analysed weighted by inverse-variance for univariate and sample-size in multivariate non-parametric models. For univariate analysis, meta-analysis effect size (Beta) and standard error (SE) are given with respect to the effect allele, total sample numbers in the meta-analysis are given in column "N". A genome-wide significance threshold of $p_{Meta}<5\times10^{-8}$ and nominal significance ($p<0.05$) in at least two cohorts was considered to ensure robustness. Genes up to 100kb up- and downstream of the lead SNP are listed, in case multiple genes are found in the locus, the closest gene to the lead SNP is marked in bold.

Locus	Analysis	uniqID	rsID	Major	MAF	Features	chr	pos	Effect allele	*p*-value	Beta	SE	N	Genes in locus (±100kb)
1	NB	1:100446046:A:G	rs1571350	A	0.425	G_Alistipes	1	100446046	G	3.945×10^{-8}	-0.1391	0.0253	8538	AGL, **SLC35A3**, HIAT
2	NB	2:171151691:A:G	rs6751051	G	0.1078	C_Betaproteobacteria, O_Burkholderiales	2	171151691	G	1.597×10^{-8}	-0.1418	0.0252	8381	MYO3B
3	NB	2:239153765:A:T	rs35093813	T	0.0895	C_Alphaproteobacteria, G_Alphaproteobacteria	2	239153765	T	1.111×10^{-8}	-0.2607	0.0456	2903	KLHL30, FAM132B, ILKAP, LOC151174, LOC643387, HES6, **PER2**, TRAF3IP1
4	LR	3:60225409:A:C	rs213388	A	0.2411	TestASV_15 (G_Bacteroides)	3	60225409	C	1.492×10^{-8}	-0.2259	0.0399	8930	FHIT
5	NB	4:40481757:C:T	rs424048	C	0.5460	TestASV_27 (G_Ruminococcaceae)	4	40481757	T	4.588×10^{-8}	-0.2146	0.0393	1345	RBM47
6	NB	4:60918325:A:G	rs11729574	G	0.0717	OTU97_11 (G_Parabacteroides)	4	60918325	G	1.193×10^{-8}	0.2267	0.0398	4832	-
7	LR	4:82818818:A:G	rs10000747	G	0.0617	OTU97_11 (G_Parabacteroides)	4	82818818	G	8.398×10^{-9}	0.3956	0.0687	7733	-
8	Beta	4:137653726:C:T	rs1585404	C	0.4150	BrayCurtis	4	137653726	T	4.176×10^{-8}	-	-	8612	-
9	LR	5:8438531:C:T	rs340669	C	0.2910	OTU97_80 (G_Ruminococcus), OTU99_92 (G_Ruminococcus)	5	8438531	T	1.710×10^{-8}	-0.2178	0.0400	8889	LOC729506
10	LR	5:57935865:G:T	rs158966	G	0.1973	OTU97_51 (G_Barnesiella), G_Barnesiella,	5	57935865	T	1.517×10^{-8}	0.2524	0.0446	8914	RAB3C
11	LR	6:90978161:C:T	rs62408234	C	0.1411	OTU97_27 (G_Bacteroides)	6	90978161	T	4.575×10^{-10}	0.3634	0.0583	5582	BACH2
12	NB	7:43864699:A:G	rs623108	G	0.3567	OTU99_55 (G_Barnesiella)	7	43864699	G	1.045×10^{-8}	-0.1664	0.0291	2743	COA1, **BLVRA**, MRPS24, URGCP
13	NB	7:117721635:C:T	rs10235327	C	0.4526	OTU99_30 (G_Parasutterella)	7	117721635	T	2.550×10^{-8}	-0.1504	0.0270	2734	-
14	NB	8:5719816:A:G	rs12679403	G	0.3337	TestASV_26 (G_Phascolarctobacterium)	8	5719816	G	2.038×10^{-8}	0.3961	0.0706	460	-
15	LR	8:22906641:A:G	rs67791511	A	0.2907	OTU97_34 (G_Ruminococcus), OTU99_35 (G_Ruminococcus)	8	22906641	G	1.486×10^{-8}	-0.2433	0.0431	5777	RHOBTB2, **TNFRSF10B**, LOC286059, LOC254896, TNFRSF10C, TNFRSF10D
16	LR	8:112651697:A:G	rs4876786	G	0.1585	G_Sutterella	8	112651697	G	1.829×10^{-8}	0.2495	0.0443	8200	-
17	NB	9:7545825:A:G	rs62530076	A	0.1070	OTU97_15 (G_Parasutterella)	9	7545825	G	4.588×10^{-8}	0.1797	0.0329	4825	-
18	LR	9:136152547:C:T	rs8176632	C	0.1686	OTU97_27 (G_Bacteroides)	9	136152547	T	6.866×10^{-10}	0.3142	0.0509	6100	OBP2B, **ABO**, SURF6, MED22, RPL7A, SURF1, SURF2, SURF4, C9orf96
19	NB	9:136239399:C:G	rs3758348	G	0.1516	OTU99_16 (G_Faecalibacterium)	9	136239399	G	4.332×10^{-9}	-0.1434	0.0244	6559	**ABO**, SURF6, MED22, RPL7A, SURF1, SURF2, **SURF4**, C9orf96, REXO4, ADAMTS13, CACFD1, SLC2A6 †
20	NB	10:112954252:A:G	rs4439447	A	0.3334	C_Clostridia	10	112954252	G	2.861×10^{-8}	0.0878	0.0158	8821	-
21	NB	11:119792443:C:T	rs71484153	T	0.2799	TestASV_21 (G_Ruminococcaceae)	11	119792443	T	3.947×10^{-8}	0.1640	0.0299	2776	-
22	NB	11:134761316:A:G	rs10894898	A	0.3893	OTU99_4 (G_Alistipes), TestASV_4 (G_Alistipes)	11	134761316	G	1.651×10^{-8}	-0.1147	0.0203	5513	-
23	Beta	12:30561406:A:G	rs7954208	A	0.0603	BrayCurtis	12	30561406	G	4.667×10^{-9}	-	-	8903	-
24	NB	12:97768678:C:T	rs12304493	C	0.4571	G_Alloprevotella	12	97768678	T	1.798×10^{-8}	0.1922	0.0341	1738	RMST

25	LR	13:46265207:C:G	rs7981551	G	0.3511	OTU97_56 (G_Ruminococcaceae)	13	46265207	G	3.303×10^{-8}	0.1882	0.0341	8390	FAM194B, **SPERT**, SIAH3
26	LR	13:107517622:A:G	rs8000335	G	0.09352	OTU97_109 (G_Paraprevotella)	13	107517622	G	4.693×10^{-8}	-0.3318	0.0607	8640	-
27	NB	14:38073877:A:T	rs1956429	A	0.3944	F_Rikenellaceae	14	38073877	T	8.316×10^{-9}	-0.0917	0.0159	8464	MIPOL1, **FOXA1**, C14orf25
28	NB	15:23999122:C:G	rs72703939	G	0.2171	TestASV_37 (G_Ruminococcaceae)	15	23999122	G	3.567×10^{-8}	-0.3657	0.0664	615	NDN
29	Beta	15:93820994:A:G	rs12592825	A	0.3616	BrayCurtis	15	93820994	G	4.552×10^{-8}	-	-	7837	-
30	NB	16:23372110:G:T	rs185150	G	0.0673	TestASV_16 (G_Bacteroides)	16	23372110	T	2.476×10^{-8}	0.6011	0.1078	669	**SCNN1B**, COG7
31	LR	17:53069650:A:G	rs4794550	A	0.2829	TestASV_16 (G_Bacteroides)	17	53069650	G	1.707×10^{-8}	-0.2978	0.0538	8864	TOM1L1, COX11, **STXBP4**
32	Beta	19:4855248:C:T	rs11880778	C	0.2103	BrayCurtis	19	4855248	T	7.974×10^{-9}	-	-	8664	FEM1A, TICAM1, **PLIN3**, ARRDC5, C19orf31, UHRF1

† In this locus, two signals in weak LD (<.4) are found, one close to SURF4, the other close to ABO (see Figure 2E).

244 **Table 2: *Results from Mendelian Randomization (MR) analysis*.** Shown are only results with $p<1.220\times10^{-3}$ (significance threshold as determined in **Methods**)
245 and the respective FDR-adjusted *q*-values. All SNPs with $p<10^{-5}$ in the respective genome-wide association meta-analysis of presence/absence (LR) and
246 abundance (NB) patterns (exposures) were used as instrument variables and tested for their effects on 41 binary traits (see **Methods** and **Supplementary**
247 **Material**). Tests used for MR (Method) were Wald ratio (WR) in case of a single instrument variable, and inverse-variance weighted (IVW) analysis in case of two
248 and more instrument variables (#SNPs). Effect sizes (Beta) and standard errors (SE) are reported in the table.

Outcome	Exposure	Analysis	Method	#SNPs	Beta	SE	*p*-value	*q*-value
Anthropometric								
Obesity class 2 || id:91	G_Parabacteroides	NB	IVW	3	-0.568	0.166	6.14×10^{-4}	0.572
Autoimmune / inflammatory								
Asthma || id:44	OTU99_84 (G_Prevotella)	LR	WR	1	-0.726	0.211	5.82×10^{-4}	0.572
Crohn's disease || id:10	TestASV_18 (G_Prevotella)	LR	WR	1	-0.257	0.058	8.76×10^{-6}	0.034
Crohn's disease || id:10	TestASV_23 (G_Barnesiella)	LR	WR	1	0.271	0.060	6.00×10^{-6}	0.034
Crohn's disease || id:11	F_Porphyromonadaceae	NB	IVW	2	3,213	0.796	5.44×10^{-5}	0.159
Crohn's disease || id:11	TestASV_12 (G_Bacteroides)	LR	WR	1	-1,034	0.316	1.06×10^{-3}	0.716
Inflammatory bowel disease || id:293	F_Porphyromonadaceae	NB	IVW	2	2,514	0.543	3.70×10^{-6}	0.034
Inflammatory bowel disease || id:293	TestASV_12 (G_Bacteroides)	LR	WR	1	-0.999	0.265	1.68×10^{-4}	0.246
Inflammatory bowel disease || id:294	G_Clostridiales	LR	WR	1	-0.112	0.028	7.30×10^{-5}	0.171
Ulcerative colitis || id:32	OTU97_12 (G_Bacteroides)	NB	IVW	4	0.310	0.091	6.83×10^{-4}	0.572
Ulcerative colitis || id:32	OTU99_12 (G_Bacteroides)	NB	IVW	4	0.310	0.091	6.83×10^{-4}	0.572
Ulcerative colitis || id:32	OTU97_74 (G_Alistipes)	NB	IVW	7	0.104	0.031	8.50×10^{-4}	0.664
Cancer								
Gallbladder cancer || id:1057	OTU97_4 (G_Alistipes)	NB	IVW	4	5,899	1,507	9.08×10^{-5}	0.177
Gallbladder cancer || id:1057	OTU97_53 (G_Bacteroides)	NB	IVW	6	-2,904	0.771	1.66×10^{-4}	0.246
Cardiovascular								
Coronary heart disease || id:6	TestASV_23 (G_Barnesiella)	LR	IVW	4	0.150	0.043	5.10×10^{-4}	0.572
Psychiatric / neurological								
Major depressive disorder || id:804	OTU97_51 (G_Barnesiella)	NB	WR	1	0.865	0.245	4.05×10^{-4}	0.528
Schizophrenia || id:22	F_Lachnospiraceae	NB	IVW	8	0.169	0.052	1.20×10^{-3}	0.716

Online Methods

Cohort description, genotyping and imputation

PopGen: The PopGen cohort is a population-based cohort from the area around Kiel, Schleswig-Holstein, Germany.[11] From this cohort, 1,108 individuals were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 covering 906,600 genetic variants. After the initial QC, the genotyping data were prepared for imputation following the miQTL cookbook instructions (https://github.com/alexa-kur/miQTL_cookbook#chapter-2-genotype-imputation). Briefly, this Plink-based processing script includes steps to prepare variants to be in consistency with the HRC v1.1 reference panel regarding the order of reference and alternative alleles, variant naming and strand orientation. Finally, all data is converted to VCF files for imputation. Imputation of the autosomal chromosomes was performed using the Michigan Imputation Server using the Haplotype Reference Consortium (HRC) release v1.1 from 2016 as reference panel. Eagle v2.3 was chosen as phasing algorithm and EUR individuals was selected as population for quality control purposes. The process was started in "Quality Control & Imputation" mode. After downloading the final data, it was converted to binary plink files. and variants with minor allele frequency < 1% were removed. Faecal samples were available for 724 of these individuals. Faecal samples were collected by the participants themselves at their respective home in standard faecal collection tubes and mailed to the study centre where they were stored at -80°C until processing. DNA from faecal samples (approx. 200 mg) was extracted using the QIAamp DNA stool mini kit automated on the QIAcube.

Food Chain Plus (FoCus): The FoCus cohort was incepted as part of the competence network Food Chain Plus (http://www.focus.uni-kiel.de/component/content/article/88.html). This cohort consists of two parts. One part is a population-registry based cross-sectional cohort including individuals from the area around Kiel, Schleswig-Holstein, Germany. The second part is an outpatient clinic-based cohort including obese individuals (BMI > 30) with and without accompanying disease status. For our study, only the registry-based part of the cohort was included. Cohort participants were genotyped using the Infinium OmniExpressExome array. Data processing, imputation and sampling of faecal material was performed in the same way as in the PopGen cohort. Finally, out of 1,583 participants, 957 belonged

277 to the population-based part of the cohort and supplied faecal samples. DNA from faecal samples
278 (approx. 200 mg) was extracted using the QIAamp DNA stool mini kit automated on the QIAcube.

279

280 **KORA FF4:** KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a population-
281 based adult cohort study in the Region of Augsburg, Southern Germany, that was initiated in 1984
282 (https://www.helmholtz-muenchen.de/epi/research/cohorts/kora-cohort/objectives/index.html). For the
283 second follow-up study (FF4) of baseline study S4 2,279 participants were recruited and the study
284 was conducted in 2013/2014 mainly focusing on diabetes, cardiovascular disease, lung disease and
285 links to environmental factors such as the microbiome. Stool-derived DNA samples of 2,136
286 participants were obtained via the KORA Biobank. The DNA had been extracted using a
287 guanidinethiocyanat / *N*-lauroylsarcosine-based buffer[37] and subsequent clean-up with NucleoSpin
288 gDNA Clean-up (Macherey-Nagel) for further analysis. Genotyping was performed using the
289 Affymetrix Axiom array. In total, 1,864 samples with genotyping and 16S rRNA gene survey data were
290 included in the association analysis.
291

292 **SHIP and SHIP-TREND:** The Study of Health in Pomerania (SHIP) is a longitudinal population-based
293 cohort study located in the area of West Pomerania (Northeast Germany). It consists of the two
294 independent cohorts SHIP (n = 4,308; baseline examinations 1997 - 2001) and SHIP-TREND (n =
295 4,420; baseline examinations 2008 - 2012 with regular follow-up examinations every five years.[12]
296 Stool samples have been collected since the second follow-up investigation of the SHIP (SHIP-2,
297 2008 - 2012) and the baseline examination of the SHIP-TREND cohort. All faecal samples were
298 collected by the study participants in their home environment, stored in a plastic tube containing
299 stabilizing EDTA buffer and shipped to the laboratory where DNA isolation (PSP Spin Stool DNA Kit,
300 Stratec Biomedical AG, Birkenfeld, Germany) was performed as described before.[36] For a total of
301 2,029 and 3,382 samples 16S rRNA gene survey and genotype data were available and included in
302 the association analysis.
303

Written, informed consent was obtained from all study participants in all cohorts, and all protocols were approved by the institutional ethical review committee in adherence with the Declaration of Helsinki Principles.

Inference of ABO blood group and secretor status

ABO blood groups were inferred using the phased and imputed genetic data and four variants as proposed by Paré *et al.*,[38] which rs507666, rs687289, rs8176746, rs8176704 encode for the allele A1, O, B, and A2, respectively. All variants were, depending on the genotyping array used in the respective cohort, either genotyped by the array or showed very high imputation quality scores between 98.7% and 99.8%. Additionally, observed allele frequencies were manually compared to frequencies in public databases to assure highest quality blood group assignments. Secretor status was assessed by variant rs601338 on chromosome 19. Individuals homozygous for the A allele were classified as “non-secretor”. This variant was genotyped in all cohorts, except for the PopGen cohort. Here, the estimated imputation accuracy was 94.6%.

Microbial data generation and processing

Library preparation and sequencing was performed using a standardized protocol at a single wet lab in Kiel, Germany. DNA amplification by polymerase chain reaction (PCR) of the bacterial 16S rRNA gene was performed using the 27F/338R primer combination targeting the V1-V2 region of the gene employing a dual-index strategy to achieve multiplex sequencing of up to 384 samples per sequencing run. After PCR, product DNA was normalized using the SequalPrep Normalization Kit. Sequencing of the libraries was performed on an Illumina MiSeq using v3 chemistry and generating 2x300bp reads. Demultiplexing was performed allowing no mismatches in the index sequences. Data processing was performed in the R software environment (version 3.5.1)[39], using the DADA2 (v.1.10)[40] workflow for big datasets (https://benjjneb.github.io/dada2/bigdata.html) resulting in abundance tables of amplicon sequence variants (ASVs). All sequencing runs underwent quality control and error profiling separately. Briefly, forward and reverse reads were trimmed to a length of

230 and 180 bp, respectively, or at the first position with a quality score less or equal to 5. Low quality read-pairs were discarded when the estimated error in one of the reads exceeded 2 or of ambiguous bases ("N"s) were present in the base sequence. Read pairs that could not be merged due to insufficient overlap or mismatches in their nucleotide sequences were discarded. The complete workflow adjusted for the 16S rDNA V1-V2 amplicon can be found on GitHub: https://github.com/mruehlemann/german_mgwas_code/tree/master/1_preprocess. Finally, all data from the separate sequencing runs were collected in a single abundance table per dataset, followed by chimera filtering. ASVs underwent taxonomic annotation using the Bayesian classifier provided in DADA2 and using the Ribosomal Database Project (RDP) version 16 release.[41] ASV abundance tables and taxonomic annotation were passed on to the phyloseq package[42] for random subsampling to 10,000 sequences per sample (*rarefy_even_depth*()) and construction of phylum- to genus-level abundance tables (*tax_glom*()). Sequences that were not assignable to genus level were binned into the finest-possible taxonomic classification. As amplicon-based sequencing of the 16S rDNA has clade-dependent taxonomic resolution differences,[43] abundance profiles of ASVs and operation taxonomic units (OTU) based on two widely used similarity cut-offs (97% similarity for a proxy of species level 99% similarity for strain level) were included in the analysis. This enables for an unbiased assessment of genetic effects at a sub-genus taxonomic scale. Although similarity cut-offs as proxy for taxonomic resolution are element of ongoing discussion[44], clustering still allows to bundle similar sequences, and by that evolutionary closely related organisms, into units of likely also functional similarity. For this, ASV datasets were exported including their respective abundance information and combined for a dataset-spanning OTU picking at 99% and 97% identity level using the VSEARCH software.[45] ASVs and OTUs were assigned cross-dataset consistent IDs for more convenient data handling, 97%- and 99%-identity based features being named OTU97 and OTU99 throughout the article, respectively. ASVs included in the analysis were relabelled to "TestASV". OTUs on 97% identity level were aligned against the SILVA reference alignment (v132) using the SINA aligner, consistent gaps in the alignment were truncated.[46] The resulting alignment was used to construct a phylogenetic tree using the FastTree (v2.1.7)[47] software with the flags `--nt` (input is nucleotide alignment), `--gtr` (generally time-reversible model) and `--gamma` (for branch-length rescaling and calculation of gamma20-likelihood).

Statistics for cohort comparisons

Basal phenotypes of age and BMI were compared between cohorts using pairwise Wilcoxon rank sum test using the R-base function *pairwise.wilcox.test*() and the default method "holm" for *p*-value correction. Within sample diversity was assessed using the total number of observed genera and Shannon diversity index calculated on genus level using the vegan[48]::*diversity*() function in R. To generate Shannon genus level equivalents, the Shannon diversity was used as exponent in the natural exponent function *exp*(). Differences between cohorts were assessed using a pairwise Wilcoxon rank-sum test implemented in the R-base function p*airwise.wilcox.test*() and the default method "holm" for *p*-value correction. Pairwise cohort differences in between sample diversity (beta diversity) were assessed using genus-level Bray-Curtis dissimilarity and a permutational multivariate analysis of variance using distance matrices as implemented in the vegan::*adonis*() function. For each comparison, 1,000 permutations were used to assess *p*-values.

Statistical framework for genome wide association analysis

Feature filtering: All features (taxon, OTU and ASV abundances) were filtered using the same criteria. Within a cohort, a feature had to be present in at least 100 individuals and had to exceed the median abundance of 50 reads, thus .5%, in the individuals with non-zero counts. For the analysis of differential prevalence, the feature additionally had to be absent in at least 100 individuals. If these criteria were fulfilled in at least three of the cohorts, the feature was included in the analysis. Summary statistics for all cohorts and microbial features included in the analysis can be found in **Supplementary Table S1.**

Prevalence-based analysis: For the analysis of genetic effects on the prevalence of bacterial features, abundance values were recoded into 0 (absence) and 1 (presence). Genetic variants were filtered to a minor allele frequency of > 5% and coded into numeric features 0 (homozygous for reference allele), 1 (heterozygous) and 2 (homozygous for alternative allele). Taxon prevalence was submitted to a logistic regression employing a generalized linear model with binomial distribution and logit-link-function using the genotype as predictor, including age, sex, body mass index (BMI), and the ten first genetic principle components (PCs) as covariates. All tests statistical tests were performed two-sided.

Abundance-based analysis: For calculating the effects of genetic variants on the zero-truncated abundance of bacterial features, the features were first filtered for extreme outliers, deviating more than 5 interquartile ranges (IQR) from the median abundance. Using the *glm.nb*() function from the MASS package in R, count abundances were fit in a model using previously mentioned covariates age, sex, BMI and the first ten genetic PCs as covariates. Residual variation was extracted using the *residuals*() function and submitted to a linear model estimating the effect of the genetic variants on the residual abundance. Analysis of SNP vs. feature abundance directly using generalized linear models with negative binomial distribution was tested as well; however, these models' results showed highly inflated λ_{GC}-values, thus were discarded for the genome-wide association analysis. All tests statistical tests were performed two-sided.

Beta diversity analysis: In addition to the single-feature based analyses, we analyzed the effects of genetic variants on the beta-diversity. For this, the genus-level abundance tables were used to calculate the pairwise Bray-Curtis dissimilarity between the individual microbial communities. Additionally, weighted, normalized UniFrac distance was calculated based on 97% identity OTU abundances using the *UniFrac*() function in phyloseq. Distance-matrices were submitted to a distance-based redundancy analysis (dbRDA) using the vegan::*capscale*() function and the same previously mentioned covariates. The residual variance of the model was extracted using the *residuals*() function, resulting in a distance matrix adjusted for these possibly confounding factors. This distance matrix was used in a procedure to estimate the effect of genetic variants based on a distance-based F-test using moment matching[49]. The calculations were implemented to run on a GPU for further speed-up, especially in the larger cohorts (see supplemental data for benchmark). As calculations for large cohorts with n > 1,000 individuals (with tables of size n×n) still could not be finished in reasonable time, we employed a stepwise calculation of results for the cohorts (estimating from single CPU usage, processing time of 7×10 variants for the SHIP-Trend dataset would take 61 years; and even using one GPU instance, processing would take ~94 days). The stepwise calculation process was as follows: For the PopGen, FoCus and SHIP cohort, all variants were tested for an association. If a variant showed a nominal significant association ($p < 0.05$) in at least one of the cohorts, this variant was tested in the KORA cohort. If then a variant was nominal significant in at least two of these four cohorts, it was also tested in the SHIP-TREND cohort.

Meta-analysis: Genomic inflation (λ_{GC}) was assessed for all cohorts and features, and all showed values below the proposed threshold of 1.05. Results from the separate cohorts were combined using a meta-analysis framework. Prevalence- and abundance-based results were submitted to an inverse-variance based strategy, calculating effects based on effect size and variance of the respective cohorts. For the beta-diversity meta-analysis, we chose a weighing based on sample size of the respective cohorts. Both approaches were adapted from the METAL software package for GWAS meta-analysis.[50] Criteria for the reporting of a significant association were a genome-wide significant meta-analysis *p*-value < 5×10^{-8}, and nominal significance in at least two cohorts for the single-feature tests and at least three cohorts for the beta diversity analysis.

Analysis of influence of blood groups and secretor status

Hurdle models were used to investigate prevalence and abundance patterns in connection with ABO blood group and secretor status. Nine models were used for analysis. Models 1 – 4 analysed the effects of the individual's counts of A alleles, B alleles, the sum of A and B alleles and the binary status O vs. non-O, respectively. Models 5 – 8 investigated the same factors, however in interaction with FUT2 secretor status, thus only taking non-zero values when assigned as "secretor". The last model only investigated the effects of the binary secretor status. All models included the covariates age, sex, BMI and the first ten genetic principle components, in analogy to the genome-wide association analysis. Inverse-variance weighted meta-analysis was used to combine the results into a composite result per taxon and model.

Mendelian Randomization

Mendelian Randomization (MR) analysis was performed using the TwoSampleMR package (version 0.4.25)[25] for R. Using the MR-Base database (mrbase.org), 41 binary traits from the subcategories "Anthropometric", "Autoimmune / Inflammatory", "Bone", "Cancer", "Cardiovascular", "Diabetes", "Kidney", "Pediatric disease", and "Psychiatric / neurological" were selected for analysis of directional effect of microbial features on these outcomes. A full list of the selection criteria, used outcome traits and the used database IDs can be found in the Supplemental Material. For the MR, variants with *p*-value below 10^{-} were included as exposure/instrument variables in the analysis and LD clumped to include only independent signals. Using the *power_prune*() function, the best set of instrumental

variables for each trait was selected using instrument strength and sample size as selection criteria (method=2). Mendelian randomization analysis was performed for sets with multiple instrument variables and single instrument variables using the inverse variance weighted analysis and Wald Test, respectively. Per microbial trait, a suggestive threshold was defined as $p < 0.05/41 = 1.220\times10^{-3}$. For study wide significance, *p*-values were adjusted using Benjamin-Hochberg FDR correction, for the resulting *q*-value the threshold was set to 0.05. For beta diversity analysis, no MR was performed, as the non-parametric test used for analysis did not include a beta value for effect size needed for MR.

Acknowledgements

We want to thank Mr Tonio Hauptmann, Ms Ilona Urbach and Ms Ines Wulf of the IKMB Microbiome Lab for excellent technical assistance. We thank Martin Schulzky for the support in figure design. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center 1182 “Origin and Function of Metaorganisms” (DFG Grant: “SFB1182”; Project A2) and the DFG Cluster of Excellence 2167 “Precision Medicine in Chronic Inflammation (PMI)” (DFG Grant: “EXC2167”). The SHIP part of the study was supported by the PePPP-project (ESF/14-BM-A55_0045/16), and the RESPONSE-project (BMBF grant number 03ZZ0921E). SHIP is part of the Research Network Community Medicine of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg-West Pomerania.

Author contributions

A.F., J.F.B., M.M.L., and D.H. designed the experiment. G.H., M.La., W.L., U.V., H.V., A.P., performed genotype and phenotype data collection. F.D., F.F. and H.V. performed data quality control and curation. C.B., M.C.R., K.H., K.N. and F.U.W. performed microbiome sample preparation, data generation and curation. M.W. and M.C.R. implemented ABO blood-group inference. M.C.R., S.D., and J.K. implemented statistical models and performed the (meta-)analysis. M.C.R., C.B., B.M.H., L.B.T., and L.M.S. curated and interpreted results. M.C.R., B.M.H. and S.D. wrote the manuscript draft with advice from C.B., A.F. and J.F.B.. All authors reviewed, edited and approved the final manuscript.

Competing interests

All authors declare no competing interests.

Code availability

Microbiome data pre-processing, GWAS analysis and post-processing code is available via github: https://github.com/mruehlemann/german_mgwas_code.

Data availability

Cohort-level summaries of microbial feature abundances are provided in the supplemental matierial. The German mGWAS Browser application is available for local query of results from Dockerhub: https://hub.docker.com/r/mruehlemann/german_mgwas_browser_app. Due to constraints given by the written consent, participant phenotypes, genotyping and 16S rRNA gene sequencing data is available upon request from the respective biobanks:

- PopGen and Focus: https://portal.popgen.de/
- KORA FF4: https://epi.helmholtz-muenchen.de/
- SHIP and SHIP-TREND: https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php

References

1. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
2. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
3. Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol.* **0**, (2019).
4. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
5. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).

6. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
7. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
8. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
9. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
10. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
11. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
12. Völzke, H. [Study of Health in Pomerania (SHIP). Concept, design and selected results]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **55**, 790–794 (2012).
13. Völzke, H. *et al.* Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
14. Holle, R., Happich, M., Löwel, H., Wichmann, H. E. & MONICA/KORA Study Group. KORA--a research platform for population based health research. *Gesundheitswesen Bundesverb. Arzte Offentlichen Gesundheitsdienstes Ger.* **67 Suppl 1**, S19-25 (2005).
15. Reitmeier, S. *et al.* Arrhythmic gut microbiome signatures for risk profiling of Type-2 Diabetes. *bioRxiv* 2019.12.27.889865 (2019) doi:10.1101/2019.12.27.889865.
16. Lozupone, C. & Knight, R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
17. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19**, 731–743 (2016).

18. Wegiel, B. *et al.* Biliverdin inhibits Toll-like receptor-4 (TLR4) expression through nitric oxide-dependent nuclear translocation of biliverdin reductase. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18849–18854 (2011).
19. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1125-1136.e8 (2016).
20. McGovern, D. P. B. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum. Mol. Genet.* **19**, 3468–3476 (2010).
21. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
22. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
23. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
24. Smith, G. D. & Ebrahim, S. *Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies*. (National Academies Press (US), 2008).
25. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
26. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
27. Costea, P. I. *et al.* Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
28. Zhou, Y. & Zhi, F. Lower Level of Bacteroides in the Gut Microbiota Is Associated with Inflammatory Bowel Disease: A Meta-Analysis. *BioMed Res. Int.* **2016**, 5828959 (2016).

29. 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).
30. Wang, K. *et al.* Parabacteroides distasonis Alleviates Obesity and Metabolic Dysfunctions via Production of Succinate and Secondary Bile Acids. *Cell Rep.* **26**, 222-235.e5 (2019).
31. Watanabe, K., Taskesen, E., Bochoven, A. van & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).
32. Davenport, E. R. *et al.* ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genomics* **17**, 941 (2016).
33. Turpin, W. *et al.* FUT2 genotype and secretory status are not associated with fecal microbial composition and inferred function in healthy subjects. *Gut Microbes* **9**, 357–368 (2018).
34. Rausch, P. *et al.* Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 19030–19035 (2011).
35. Weiss, F. U. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. *Gut* **64**, 646–656 (2015).
36. Frost, F. *et al.* Impaired Exocrine Pancreatic Function Associates With Changes in Intestinal Microbiota Composition and Diversity. *Gastroenterology* **156**, 1010–1015 (2019).
37. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **63**, 2802–2813 (1997).
38. Paré, G. *et al.* Novel Association of ABO Histo-Blood Group Antigen with Soluble ICAM-1: Results of a Genome-Wide Association Study of 6,578 Women. *PLoS Genet.* **4**, (2008).

39. R Core Team. *R: A Language and Environment for Statistical Computing*. (2014).
40. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
41. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
42. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217 (2013).
43. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 1–11 (2019).
44. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
45. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
46. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
48. Oksanen, J. *et al.* The vegan package. *Community Ecol. Package* **10**, 631–637 (2007).
49. Rühlemann, M. C. *et al.* Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes* **9**, 68–75 (2017).
50. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

489

28

4 Disease-associated changes in the human microbiome

Publications:

Malte Christoph Rühlemann, Femke-Anouska Heinsen, Roman Zenouzi, Wolfgang Lieb, Andre Franke and Christoph Schramm: "Faecal microbiota profiles as diagnostic biomarkers in primary sclerosing cholangitis". *Gut.* 2017;66:753-754.

Malte Rühlemann*, Timur Liwinski*, Femke-Anouska Heinsen*, Corinna Bang, Roman Zenouzi, Martin Kummen, Louise Thingholm, Marie Tempel, Wolfgang Lieb, Tom Karlsen, Ansgar Lohse, Johannes Hov, Gerald Denk, Frank Lammert, Marcin Krawczyk, Christoph Schramm and Andre Franke: "Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis". *Aliment Pharmacol Ther*. 2019;50:580–589.

Malte Christoph Rühlemann, Miriam Emmy Leni Solovjeva, Roman Zenouzi, Timur Liwinski, Martin Kummen, Wolfgang Lieb, Johannes Roksund Hov, Christoph Schramm, Andre Franke and Corinna Bang: "Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species". *Gut.* 2019. [Epub ahead of print]

Hansjörg Baurecht*, **Malte C. Rühlemann***, Elke Rodrıguez, Frederieke Thielking, Inken Harder, Anna-Sophie Erkens, Dora Stölzl, Eva Ellinghaus, Melanie Hotze, Wolfgang Lieb, Sheng Wang, Femke-Anouska Heinsen-Groth, Andre Franke and Stephan Weidinger: "Epidermal lipid composition, barrier integrity, and eczematous inflammation are associated with skin microbiome configuration". *J Allergy Clin Immunol.* 2018;141(5):1668-1676.

Article D: Faecal microbiota profiles as diagnostic biomarkers in primary sclerosing cholangitis

PostScript

LETTER

Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species

LETTER TO THE EDITOR

We read with interest the recent Gut article by Lemoinne *et al* describing a dysbiosis of the fungal gut community in faeces of patients suffering from primary sclerosing cholangitis (PSC).[1] Though several reports, including our own previous data, support a functional and potentially pathogenic link between the intestinal bacteriota and liver inflammation in PSC,[2 3] the aetiology of the disease remains largely unknown.

We here report on the fungal mycobiome results of our cohort from Northern Germany approved by the local ethics committees (A148/14 and MC-111/15) comprising stool samples of 66 healthy control (HC) subjects, 65 patients with well-characterised PSC (including a subgroup with concomitant colitis (PSC-IBD), n=32) and 38 subjects with UC.[3] PCR and sequencing of the fungus-specific internal transcribed spacer 2 genomic region was performed as previously described[4] using the primer pair 5.8S-Fun and ITS4-Fun on an Illumina MiSeq machine. Sequencing data were subjected to quality control by using the open source package DADA2 (V.1.10)[5] in R (V.3.5.1; https://github.com/mrue-hlemann/ikmb_amplicon_processing). Amplicon sequence variants were taxonomically annotated using the UNITE ITS database (V.7.2).[6]

In disagreement with the findings in the French cohort,[1] overall fungal alpha diversity in the German cohort was neither altered in PSC nor in UC versus HC as calculated by Shannon species equivalent (figure 1A). None of the disease groups significantly deviated in community composition from healthy

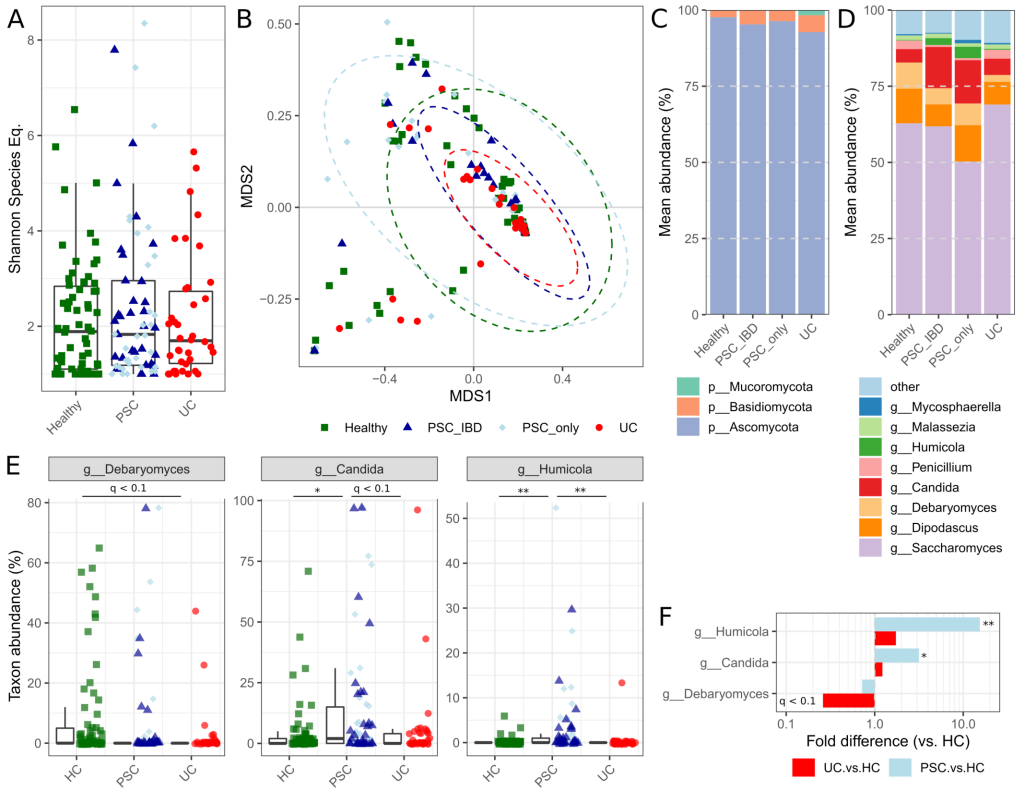


Figure 1 Mycobiome of individuals with primary sclerosing cholangitis (PSC) and UC as well as healthy controls (HC) (all of northern German origin). Rarefaction curves for Shannon diversity of sequence variants reached plateau between 50 and 100 sequences per samples, thus samples were normalised to 100 random reads per sample. (A) Alpha diversity as presented by Shannon species equivalents (all p>0.05). (B) Beta diversity ordination of the Bray-Curtis dissimilarity based on genus-level fungal abundances (all p_{adonis} >0.05). (C) Phylum-level and (D) genus-level mean abundances of all taxa with >1% mean abundance and present in at least 10 samples. (E) Group-wise box-and-whisker plots for significant genus level annotations tested for differential abundances with individual values represented as data points. (F) Differences in group-mean abundances of patients with PSC and UC, as compared with HC. *q<0.05, **q<0.01.

individuals (all p_{adonis} >0.05; figure 1B). Fungal composition on phylum level was found to be mainly dominated by Ascomycota (figure 1C), particularly by the genera *Saccharomyces*, *Candida* and *Dipodascus* (figure 1D) in relatively higher abundance of reads when compared with the findings of Lemoinne *et al*.[1] Though our results generally validate the previously described overall fungal composition in stool, we were not able to detect the genus *Exophiala*, which was found in five PSC patients from France exclusively. Whether this is due to methodological differences (choice of primer sets, data analysis tools and sampling depth) or presence of this fungus in only a subset of PSC patients not sampled in the German cohort needs to be determined.

Disease-associated differential abundance of fungal taxa was investigated by applying Student's t-test to the arcsin-squareroot-transformed relative abundances of all genera with mean abundance >1% and present in at least 10 individuals. This analysis revealed increased levels of the genera *Candida* and *Humicola* (species level annotation suggests *H. grisea*) in PSC patients with and without concomitant colitis compared with HC (all q_{BH} <0.05; figure 1E and F) and UC (all q_{BH} <0.1; figure 1E) individuals. *H. grisea*, recently reclassified as *Trichocladium griseum*,[7] belongs to the fungal class Sordariomycetes, thus our results reproduce the significant increase of this class in PSC patients, as previously described by Lemoinne and colleagues, but at increased taxonomic resolution. Previous research on *T. griseum* showed that it is most frequently isolated from soil and plants but also occasionally found in patients suffering from peritonitis.[8] In addition, the validated increase of *Candida* species in PSC patients argues for an immunogenic role of these fungi, particularly with respect to earlier findings that demonstrated their high potential to induce Th17 response in T cells.[9] Increased Th17 numbers have previously been reported in PSC patients and recently been shown to be involved in PSC pathogenesis.[10]

In summary, both the significant increase of the fungal class Sordariomycetes, likely *T. griseum*, as well as of *Candida* species in stool samples of PSC patients, now found in two independent and geographically distinct PSC patient panels that were analysed with divergent methodological approaches, strongly demands for additional analyses on these fungi and their role in PSC.

Malte Christoph Rühlemann [1], **Miriam Emmy Leni Solovjeva**,[1] **Roman Zenouzi**,[2] **Timur Liwinski** [2], **Martin Kummen** [3,4], **Wolfgang Lieb**,[5] **Johannes Roksund Hov** [3,4], **Christoph Schramm**,[2,6] **Andre Franke** [1], **Corinna Bang** [1]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
[2]Department of Internal Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[3]Norwegian PSC Research Center, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4]Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[5]Institute of Epidemiology and Biobank POPGEN, Christian-Albrechts-University of Kiel, Kiel, Germany
[6]Martin Zeitz Centre for Rare Diseases, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Correspondence to Dr Corinna Bang, Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel 24105, Germany; c.bang@ikmb.uni-kiel.de

Twitter Malte Christoph Rühlemann @mruehlemann and Johannes Roksund Hov @hov_jer

Acknowledgements We would like to thank Ms Ilona Urbach, Ms Ines Wulf and Mr Tonio Hauptmann of the IKMB microbiome laboratory for excellent technical support.

Contributors AF and CB designed the study. CS and AF obtained funding. RZ, CS and WL acquired and quality-controlled patient samples and data. CB and MELS supervised sample processing and sequencing. MCR performed statistical analyses. MCR, MELS, RZ, TL, MK, JRH, CS, AF and CB interpreted data and drafted the manuscript with input and critical revision from all authors. All authors revised and approved the final version of the manuscript.

Funding This study was supported by the Deutsche Forschungsgemeinschaft (DFG) Clinical Research Group 306' Primary sclerosing cholangitis' (no: KFO306) as well as Research Training Group 1743 and received infrastructure support from the DFG Cluster of Excellence 'Inflammation at Interfaces' (http://www.inflammation-at-interfaces.de, no: EXC306 and EXC306/2) and the German Ministry of Education and Research (BMBF) program e:Med sysINFLAME (http://www.gesundheitsforschung-bmbf.de/de/5111.php, no: 01ZX1306A).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Sequencing and clinical data of the patient samples used in this study can be applied for via the Popgen Biobank (Institute of Epidemiology, Christian-Albrechts-University of Kiel, Germany).

AF and CB contributed equally.

To cite Rühlemann MC, Solovjeva MEL, Zenouzi R, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2019-320008

Received 4 October 2019
Revised 10 October 2019
Accepted 11 October 2019

Gut 2019;**0**:1–2. doi:10.1136/gutjnl-2019-320008

ORCID iDs
Malte Christoph Rühlemann http://orcid.org/0000-0002-0685-0052
Timur Liwinski http://orcid.org/0000-0002-1041-9142
Martin Kummen http://orcid.org/0000-0001-9660-6290
Johannes Roksund Hov http://orcid.org/0000-0002-5900-8096
Andre Franke http://orcid.org/0000-0003-1530-5811
Corinna Bang http://orcid.org/0000-0001-6814-6151

REFERENCES

1 Lemoinne S, Kemgang A, Ben Belkacem K, *et al*. Fungi participate in the dysbiosis of gut microbiota in patients with primary sclerosing cholangitis. *Gut* 2019:gutjnl-2018-317791 (published Online First: 2019/04/19).
2 Kummen M, Holm K, Anmarkrud JA, *et al*. The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. *Gut* 2017;66:611–9.
3 Rühlemann M, Liwinski T, Heinsen F-A, *et al*. Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis. *Aliment Pharmacol Ther* 2019;50:580–9.
4 Taylor DL, Walters WA, Lennon NJ, *et al*. Accurate estimation of fungal diversity and abundance through improved lineage-specific primers optimized for Illumina amplicon sequencing. *Appl Environ Microbiol* 2016;82:7217–26.
5 Callahan BJ, McMurdie PJ, Rosen MJ, *et al*. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13:581–3.
6 Nilsson RH, Larsson K-H, Taylor AFS, *et al*. The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2019;47:D259–64.
7 Wang XW, Yang FY, Meijer M, *et al*. Redefining *Humicola sensu stricto* and related genera in the *Chaetomiaceae*. *Stud Mycol* 2019;93:65–153.
8 Burns N, Arthur I, Leung M, *et al*. Humicola sp. as a cause of peritoneal dialysis-associated peritonitis. *J Clin Microbiol* 2015;53:3081–5.
9 Katt J, Schwinge D, Schoknecht T, *et al*. Increased T helper type 17 response to pathogen stimulation in patients with primary sclerosing cholangitis. *Hepatology* 2013;58:1084–93.
10 Nakamoto N, Sasaki N, Aoki R, *et al*. Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis. *Nat Microbiol* 2019;4:492–503.

Article E: Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis

Received: 1 February 2019 | First decision: 26 February 2019 | Accepted: 24 May 2019

DOI: 10.1111/apt.15375

APT Alimentary Pharmacology & Therapeutics WILEY

Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis

Malte Rühlemann[1] | Timur Liwinski[2] | Femke-Anouska Heinsen[1] | Corinna Bang[1] | Roman Zenouzi[2] | Martin Kummen[3] | Louise Thingholm[1] | Marie Tempel[1] | Wolfgang Lieb[1] | Tom Karlsen[3] | Ansgar Lohse[2] | Johannes Hov[1,3] | Gerald Denk[4] | Frank Lammert[5] | Marcin Krawczyk[5,6] | Christoph Schramm[2] | Andre Franke[1]

[1]Kiel, Germany
[2]Hamburg, Germany
[3]Oslo, Norway
[4]Munich, Germany
[5]Homburg, Germany
[6]Warsaw, Poland

Correspondence
Prof. Andre Franke, Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, University Hospital Schleswig Holstein, Campus Kiel, Rosalind-Franklin-Str. 12, 24105 Kiel, Germany.
Email: a.franke@mucosa.de

Funding information
This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Clinical Research Unit 306 ‚Primary sclerosing cholangitis' (no: KFO306) and received infrastructure support from, the DFG Cluster of Excellence 'Inflammation at Interfaces' (http://www.inflammation-at-interfaces.de, no.: EXC306 and EXC306/2), the Collaborative Research Center 1182 ‚Origin and Function of Metaorganisms' (www.metaorganism-research.com, no: SFB1182) and the German Ministry of Education and Research (BMBF) program e:Med sysINFLAME (http://www.gesundheitsforschung-bmbf.de/de/5111.php, no.: 01ZX1306A). CS receives support from the Helmut and Hannelore Greve-Foundation and the YAEL-Foundation.

Summary

Background: Single-centre studies reported alterations of faecal microbiota in patients with primary sclerosing cholangitis (PSC). As regional factors may affect microbial communities, it is unclear if a microbial signature of PSC exists across different geographical regions.

Aim: To identify a robust microbial signature of PSC independent of geography and environmental influences.

Methods: We included 388 individuals (median age, 47 years; range, 15-78) from Germany and Norway in the study, 137 patients with PSC (n = 75 with colitis), 118 with ulcerative colitis (UC) and 133 healthy controls. Faecal microbiomes were analysed by 16S rRNA gene sequencing (V1-V2). Differences in relative abundances of single taxa were subjected to a meta-analysis.

Results: In both cohorts, microbiota composition (beta-diversity) differed between PSC patients and controls ($P < 0.001$). Random forests classification discriminated PSC patients from controls in both geographical cohorts with an average area under the curve of 0.88. Compared to healthy controls, many new cohort-spanning alterations were identified in PSC, such as an increase of Proteobacteria and the bile-tolerant genus *Parabacteroides*, which were detected independent from geographical region. Associated colitis only had minor effects on microbiota composition, suggesting that PSC itself drives the faecal microbiota changes observed.

Conclusion: Compared to healthy controls, numerous microbiota alterations are reproducible in PSC patients across geographical regions, clearly pointing towards a

Malte Rühlemann, Timur Liwinski and Femke-Anouska Heinsen contributed equally to this work.

Andre Franke and Christoph Schramm jointly supervised this work.

The Handling Editor for this article was Professor Gideon Hirschfield, and it was accepted for publication after full peer-review.

The authors' complete affiliation are listed in Appendix 1.

JH is funded by the Research Council of Norway (no. 240787/F20). MKu is funded by the Regional Health Authority of South-Eastern Norway (no. 2016067).

microbiota composition that is shaped by the disease itself and not by environmental factors. These reproducibly altered microbial populations might provide future insights into the pathophysiology of PSC.

1 | INTRODUCTION

Primary sclerosing cholangitis (PSC) is a chronic progressive disease of the biliary system.[1] There is no medical treatment available and patients have an increased risk of both hepatobiliary and bowel malignancies.[2,3] As a result, most patients decease or receive liver transplantation 15-20 years after diagnosis.[4] The pathogenesis is still unclear, although a dysregulated immune reaction at the epithelial barrier in the intestine and the biliary system may play a key role. Co-occurrence of inflammatory bowel disease (IBD), typically in form of a mild pancolitis, has been reported in up to 80% of patients with PSC, and at least 2%-7.5% of IBD-patients develop PSC.[1,5-7] It has been postulated that PSC-associated colitis (PSC-IBD) represents a distinct IBD entity alongside Crohn's disease and UC.[8,9]

Several genetic risk factors have been identified that are associated with PSC, but collectively they only contribute to a small fraction of disease susceptibility.[1] Given the minor role of genetic variation, environmental factors likely play a major role in the pathogenesis of PSC. Of these, the microbiome has emerged as one of the most important potential environmental players in chronic inflammatory diseases. Gut microbiome alterations have been identified in several metabolic and inflammatory diseases,[10-12] and recent reports have demonstrated alterations in both the faecal and mucosal microbiota in patients with PSC.[13,14] The faecal microbiota of PSC patients has been characterised by a distinct profile compared to healthy controls and UC patients, including an overabundance of *Veillonella* in a Norwegian and a Czech cohort.[13,15] However, a classification between PSC and UC in a Norwegian cohort based on the abundance of *Veillonella* could not be validated in a German population.[16] In a Belgian cohort, stool samples of PSC patients showed an increased abundance of the genera *Fusobacterium*, *Enterococcus*, *Lactobacillus* and *Streptococcus* compared to healthy individuals.[17]

In summary, it is unclear whether the microbial signature so far described in PSC single-centre cohorts is centre-specific or if a PSC-specific microbial signature across different geographical regions exists. Cross-regional analysis of faecal microbiota of patients with PSC might reveal general patterns of microbial perturbation, which could elucidate the role of the microbiome in PSC and PSC-IBD and provide the basis for a better understanding of their possible pathogenetic significance in further mechanistic and clinical longitudinal studies.

2 | PATIENTS AND METHODS

Seventy-four nontransplanted German patients with PSC (n = 37 PSC only, n = 37 PSC-IBD) were recruited at the University Medical Center Hamburg-Eppendorf. In addition, a German UC study cohort (n = 88) and 95 German healthy individuals (controls) were recruited for comparison by the PopGen Biobank.[18]

Furthermore, 63 nontransplanted Norwegian patients with PSC (n = 25 PSC only, n = 38 PSC-IBD), 30 UC patients without PSC and 38 controls were recruited at the Norwegian PSC Research Center Biobank at Oslo University Hospital Rikshospitalet.[13]

PSC was diagnosed based on cholangiography and liver biopsy (if required) according to most recent guidelines.[19,20] All individuals underwent extensive screening for potential confounding (see Methods S1 for exclusion criteria). The characteristics of both study populations are summarised in Table 1.

The study was approved by the local ethics committees in Hamburg and Kiel (A148/14 and MC-111/15) and the Regional Committee for Medical and Health Research Ethics in South-Eastern Norway (reference 2012/286b). All participants gave their written informed consent.

2.1 | Assessment of dietary patterns

Dietary data were collected for 220 individuals (Table 1) of the German study cohort using standardised and validated food frequency questionnaires of the German Institute of Human Nutrition. Translation into nutrients was performed via the German Food Code and Nutrient Database (vII.3).[21]

2.2 | Stool sample processing and sequencing

Samples were collected and subjected to DNA extraction as previously described for the respective cohorts.[13,22] The amplification and library preparation of the V1-V2 region of the 16S rRNA gene using dual-indexing was performed in a single facility (Methods S1). We chose V1-V2 as target amplicon, as 2 300 bp sequencing covers the amplicon of 300-320 bp almost entirely twice, thus assuring high quality data readout. Sequencing data were subjected to quality control and data processing to obtain count-based relative abundance tables for operational taxonomic units (OTUs) and taxonomic levels from phylum to genus (Methods S1).

2.3 | Data analysis

Data analyses were performed with R statistical programming language (v3.4.3).[23]

Differences in dietary patterns were evaluated using the log-transformed average intake of the primary macronutrients protein, fat and carbohydrates (g/day), as well as fibre, water (both g/day), and total energy intake (kJ/day). To assess dietary differences of

TABLE 1 Demographic and clinical characteristics of the German and Norwegian study populations

German	Controls	PSC only	PSC-IBD	UC
Total number	n = 95	n = 37	n = 37	n = 88
General information				
Age, median years (min-max)	47 (19-64)	51 (18-73)*	46.5 (15-73)	45 (19-78)
Gender (female)	51.6% (n = 49)	32.4% (n = 12)	43.2% (n = 16)	61.4% (n = 54)
BMI, median kg/m^2 (min-max)	22.8 (20.2-24.9)	23.7 (17.9-32)**	23.6 (15.8-34.3)*	24.8 (17.0-36.5)***
Smoking (yes)	16.8% (n = 16)	8.1% (n = 3)	0%**	3.4% (n = 3)***
Dietary data				
Available	89.5% (n = 85)	83.8% (n = 31)	76.7% (n = 28)	86.4% (n = 76)
Daily intake, median (min-max)				
Energy (kJ)	9025 (4,313-23,006)	9961 (5,150-18,249)	10 153 (4,218-19,630)	9304 (5040-19 609)
Carbohydrates (g)	215.9 (93.8-772.9)	239.6 (124.7-511.0)	275.3 (91.8-588.1)	242.8 (103.3-483.9)
Fibre (g)	20.1 (9.9-25.5)	21.3 (12.9-34.1)	24.1 (10.6-46.5)	21.7 (10.4-40.7)
Fat (g)	98.0 (46.4-223.2)	105.2 (49.3-202.3)	108.2 (45.1-191.0)	95.8 (43.6-203.3)
Protein (g)	77.3 (32.4-206.0)	90.7 (45.0-182.5)	84.4 (39.7-136.1)	78.8 (41.4-156.7)
Water (L)	3.15 (1.05-7.73)	2.67 (1.47-7.15)*	2.83 (1.43-4.41)	2.79 (1.09-7.32)
Faecal Calprotectin (fCAL)				
Median (µg/g) (Q1-Q3)	27.3 (15.6-40.9)	20 (10-52.4)	29.4 (10-110)	43.3 (18.3-190.8)
fCAL low (<50 µg/g), %	80% (n = 42)	73.0% (n = 27)	56.8 (n = 21)	52.3% (n = 46)
fCAL elevated (50-200 µg/g), %	19.2% (n = 10)	13.5% (n = 5)	27.0% (n = 10)	22.7% (n = 20)
fCAL high (>200 µg/g) %	0	13.5 (n = 5)	16.2 (n = 6)	25 (n = 22)
NA	n = 43	—		—
PSC additional information				
Years since PSC diagnosis, median (min-max)	—	6.5 (0-35)	9.0 (1-28)	—
Cirrhosis (yes)	—	5.4% (n = 2)	5.4% (n = 2)	—
ALT, median U/L (min-max)	—	37 (11-165, NA = 10)	38.5 (13-286, NA = 15)	—
AP, median U/L (min-max)	—	116 (61-590, NA = 10)	125 (44-332, NA = 15)	—
Bilirubin, median U/L (min-max)	—	10.3 (5.1-35.9, NA = 11)	11.97 (3.4-34.2, NA = 16)	—
Medication (%)				
UDCA	—	97.3 (n = 36)	94.6 (n = 35)	—
5-ASA	—	2.7 (n = 1)	83.8 (n = 31)	79.5 (n = 80)
Azathioprine	—	5.4 (n = 2)	13.5 (n = 5)	30.7 (n = 27)
Budesonide	—	—	5.4 (n = 2)	31.8 (n = 28)
Biologics (Adalimumab, Infliximab)	—	—	5.4 (n = 2)	15.9 (n = 14)
PPI	—	—	—	—
Statins	—	—	—	—
Norwegian	Controls	PSC only	PSC-IBD	UC
Total number	n = 38	n = 25	n = 38	n = 30
General information				
Age, median years (min-max)	47 (35-61)	46 (31-66)	48 (21-69)	42.5 (25-69)
Gender (female) (%)	36.8 (n = 14)	36 (n = 9)	31.6 (n = 12)	53.3 (n = 16)
BMI, median kg/m^2 (min-max)	26 (19.4-39.4)	26.0 (17.8-32.2)	24.0 (17.7-34.7)	24.5 (21.4-34.3)
Smoking (yes) (%)	15.8 (n = 6)	0	2.6 (n = 1)	0*
PSC additional information				
Years since PSC diagnosis, median (min-max)	—	7.8 (2.1-31.7)	9.6 (1.4-28.8)	—

(Continues)

TABLE 1 (Continued)

German	Controls	PSC only	PSC-IBD	UC
Signs of impaired liver function (yes) (%)	—	4 (n = 1)	2.6 (n = 1)	—
ALT, median U/L (min-max)	—	65.5 (16-258), NA = 3)	54 (14-331), NA = 2)	—
AP, median U/L (min-max)	—	192 (50-548, NA = 4)	130 (30-589, NA = 2)	—
Bilirubin, median U/L (min-max)	—	13.5 (6-114, NA = 3)	13 (6-44 NA = 3)	—
Medication (%)				
UDCA	—	36 (n = 9)	26.3 (n = 10)	—
5-ASA	—	4 (n = 1)	57.9 (n = 22)	76.7 (n = 23)
Azathioprine	—	4 (n = 1)	15.8 (n = 6)	23.3 (n = 7)
Budesonide	—	4 (n = 1)	2.6 (n = 1)	6.7 (n = 2)
Biologics (Adalimumab, Infliximab)	—	—	2.6 (n = 1)	40 (n = 12)
PPI	—	—	2.6 (n = 1)	6.7 (n = 2)
Statins	—	16 (n = 4)	5.2 (n = 2)	—

Only medication taken by at least two patients is listed.
ALT, alanine aminotransferase; AP, alkaline phosphatase; ASA5, 5-aminosalicylic acid; BMI, body mass index; PPI, proton pump inhibitors; PSC, primary sclerosing cholangitis; Q1, first quartile; Q3, third quartile; UC, ulcerative colitis; NA, not available.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

diseased individuals from healthy controls, permutational analysis of variance (PERMANOVA) was performed on Euclidean distances using residuals of dietary data after regression against sex, as sex is a known major predictor of dietary behaviour.

Shannon index, as a measure of within-sample diversity (alpha-diversity), and PERMANOVA on Bray-Curtis dissimilarity, as a measure for beta-diversity, were applied to investigate differences according to disease state and other co-variables (Methods S1).

All abundances are based on a normalised number of counts, thus being relative abundances, this is always the case when the term 'abundance' is used throughout the text. To assess differential taxa abundances, we tested microbes that were identified as differentially abundant in previous studies in a first step (Methods S1), followed by a second step where we aimed to discover new associations. Details on the applied regression models and meta-analysis are provided in the Methods S1. Taxa with significant signals in both respective geographical cohorts and the meta-analysis were re-analysed with inclusion of dietary co-variables in the German cohort, for which food frequency questionnaire data was available, to correct for dietary effects.

Predictive performance of the identified taxonomic signature was evaluated using random forests classification[24] (Methods S1). To evaluate model performance, receiver operating characteristic analysis was used. Additionally, F1 score and Matthews Correlation Coefficient (MCC) as weighted measures of true and false positives rates were calculated (Methods S1).

3 | RESULTS

In total, we applied 16S rRNA gene sequencing to 257 German (n = 95 controls, n = 37 PSC only, n = 37 PSC-IBD and n = 88 UC) and 131 Norwegian samples (n = 38 controls, n = 25 PSC only, n = 38 PSC-IBD and n = 30 UC). We applied the same amplification and library preparation standard operating procedure within a single facility to all samples. For 220 of the 257 (85.6%) German study participants dietary data, assessed by standardised food frequency questionnaires, was available (Table 1).

3.1 | Healthy Norwegian and German subjects share similar core microbiota

To determine the baseline similarities and differences between the faecal microbiota of German and Norwegian healthy volunteers, we compared the healthy controls of both cohorts. We found large proportions of the core microbiota (122 of 144 taxa, 84.7%) shared by healthy controls from both populations (Figures S1-S3). The healthy Norwegian cohort displayed a significantly lower intra-individual (alpha) diversity compared to the healthy German population ($P = 0.006$). We observed slight but significant differences in between-sample diversity (beta-diversity) between German and Norwegian controls ($P = 0.01$; $R^2 = 0.021$).

3.2 | The faecal microbiota of patients with PSC is significantly different from both healthy controls and patients with UC

In both cohorts, between-sample diversity (beta-diversity) was significantly different between patients with PSC and controls ($P < 0.001$, respectively; $R^2_{GER} = 0.028$; $R^2_{NOR} = 0.042$). Differences in beta-diversity between patients with PSC and UC were also significant but less pronounced in both cohorts ($P_{NOR} = 0.016$, $R^2_{NOR} = 0.027$; $P_{GER} = 0.013$, $R^2_{GER} = 0.015$).

In the Norwegian cohort, mean within-sample diversity (alpha-diversity) of patients with PSC was reduced compared to controls ($P = 0.001$) and comparable to patients with UC ($P > 0.05$). In the

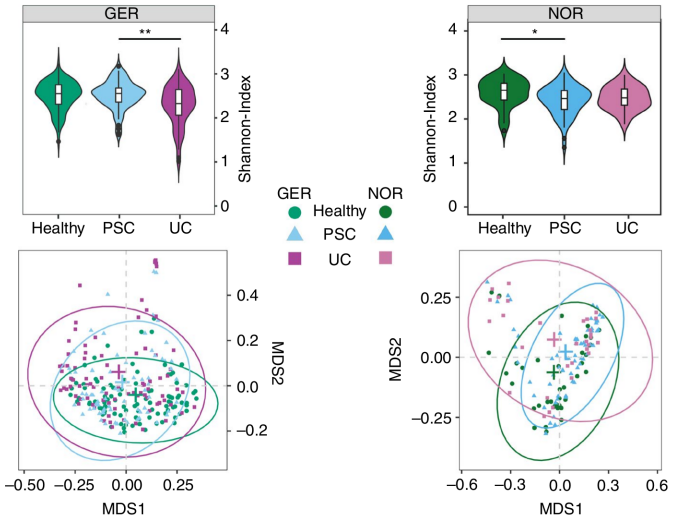


FIGURE 1 Violin plots of Shannon-Index and the unconstrained ordination plots of the Bray-Curtis dissimilarities for the German (GER) and the Norwegian (NOR) cohort. Ordination was performed on genus-level abundances subsequently plotted for each cohort separately. Centroids of ellipses are marked by crosses in the respective colours. *P < 0.05; **P < 0.01

German cohort however, alpha-diversity of patients with PSC was comparable to controls (P > 0.05) and significantly increased in contrast with patients with UC (P = 0.01) (Figure 1).

3.3 | Targeted analysis of taxa with previously reported association with PSC

A total of nine genera that were previously identified as differentially abundant[13,17] were analysed in a targeted approach aiming to reproduce the taxonomic signals. A detailed summary is provided in the Tables S1 and S2.

In both cohorts, an increased relative abundance in patients with PSC was displayed by *Veillonella* and *Streptococcus* (both P_{META} < 0.0001). In addition, an increased prevalence in patients with PSC was confirmed for *Lactobacillus* and *Enterococcus* (both P_{META} < 0.0001, respectively). Other taxa either showed inconsistent distribution patterns between cohorts or could not be recovered (with sufficient prevalence) in our samples.

3.4 | Extensive microbiota alterations in patients with PSC

In the next step, we aimed to discover new robust taxonomic distribution patterns between patients with PSC and controls across all taxonomic hierarchy levels. A detailed summary is provided in Figures 2 and 3 as well as Tables S3 and S4.

A total of 20 taxa in the German and 18 taxa in the Norwegian cohort showed altered (continuous) abundance in patients with PSC. However, only seven of these met the meta-analysis criteria (see Section 2).

Robust and cohort-spanning increased relative abundances in patients with PSC were displayed by the phylum Proteobacteria, represented by the class Gammaproteobacteria, order Lactobacillales and the class Bacilli (all Q_{META} < 0.0001, respectively). An OTU belonging to the genus *Coprococcus* was the only taxon with cohort-consistent decreased abundance in PSC (Q_{META} = 0.017). For the differentially abundant taxa, an additional analysis was performed in the German cohort, to assess whether these signals are truly driven by disease or may be influenced by dietary differences. Only for the class Bacilli and order Lactobacillales a significant influence of protein intake could be observed (both P = 0.03), this however, did not influence the strongly significant signals of disease association ($P = 1.1 \times 10^{-7}$ and $P = 8.9 \times 10^{-8}$, respectively).

Regarding microbial (binary) prevalence patterns, we observed an extensive depletion in patients with PSC compared to controls affecting 32 taxa. Among these were the genera *Holdemanella* and *Desulfovibrio* as well as OTUs classified as *Faecalibacterium* and *Clostridium IV* (both Q_{META} < 0.0001).

3.5 | Microbiota alterations in PSC are independent from associated colitis, medication or grade of colonic inflammation

For the analysis of effects of medication and calprotectin the same aforementioned statistical models were applied with inclusion of the respective data as additional independent variables. No differences in levels of faecal calprotectin could be found between PSC patients with and without IBD. Additionally, neither medical treatment with UDCA, 5-ASA or Azathioprine, nor faecal calprotectin levels exhibited any effect on or correlation with the microbiota in PSC (Supporting information).

PSC-IBD has a genetic basis and clinical phenotype different from classical UC.[9] Therefore, PSC could drive the phenotype of

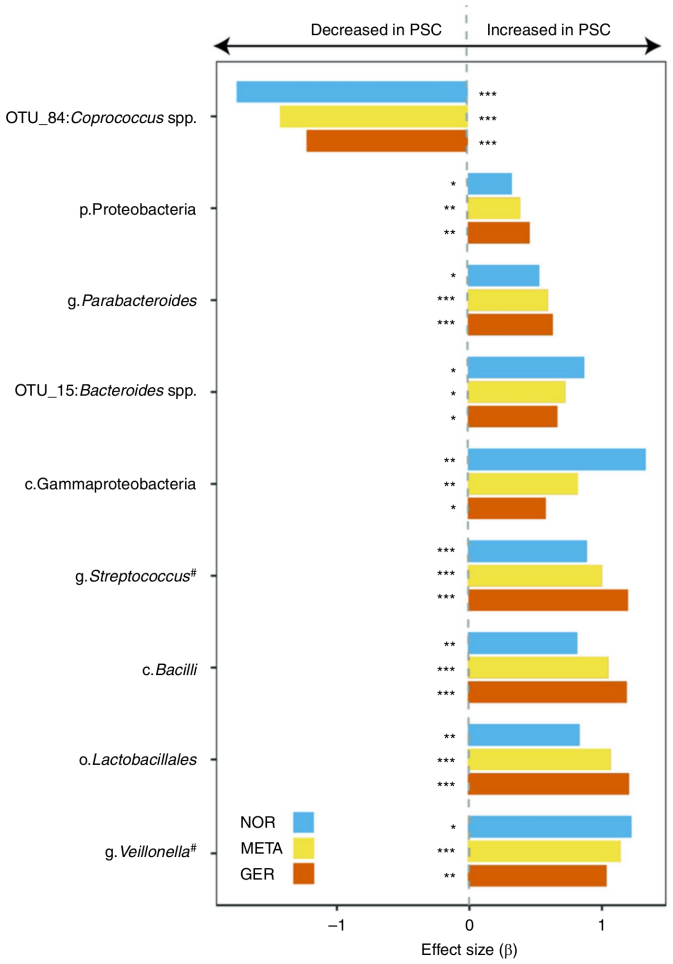


FIGURE 2 Significant and between-cohort consistent results of differentially abundant taxa in PSC patients and controls. Only taxa with $P < 0.05$ in each cohort, $Q_{META} < 0.05$ and concordant directionality are shown. Taxa from Kummen *et al* or Sabino *et al* that could be replicated in both cohorts are marked with a pound (#) symbol. Base-colours depict the respective cohort (blue: Germany; red: Norwegian) and the combined meta-analysis result (yellow). Beta-values larger than zero represent a higher abundance in PSC patients, taxa with beta-values less than zero are less abundant in PSC patients. Details on the model coefficients and the resulting P-values in the cohorts and the meta-analysis can be found in S1-S3. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$

intestinal inflammation, as well as intestinal microbiota composition. Therefore, we investigated if the microbiota signature in PSC-IBD is closer to PSC only or closer to UC.

Neither in the German nor in the Norwegian cohort there was any significant difference in beta-diversity between patients with PSC only and PSC-IBD ($P > 0.05$, respectively). Since functionally important taxa may be differentially abundant even in the absence of significant overall beta-diversity, we explored potential cohort-spanning taxonomic differences between patients with PSC only and PSC-IBD (Tables S5, S6 and S8).

Abundance-based models comparing PSC only to PSC-IBD showed no cohort-spanning signals. Additionally, there were significant differences between PSC and UC in abundance and diversity, strongly indicating that PSC drives the microbiota associations observed in patients both with PSC only as well as in those with PSC-IBD. The only robust taxonomic differences detected between PSC only and PSC-IBD were decreased prevalences of *Bilophila* ($Q_{META} = 0.017$) and an OTU assigned to *Bacteroides* (OTU_28; $Q_{META} < 0.0001$) in patients with PSC-IBD.

3.6 | PSC and UC are both characterised by altered microbiota, but cannot be differentiated by single taxa

We investigated if the observed difference of beta-diversity between patients with PSC and UC can be traced to robust differential distribution of individual taxa (Tables S9 and S10). In both cohorts, the phylum Firmicutes was significantly increased in patients with PSC compared to patients with UC ($Q_{META} = 0.011$). We found no cohort-spanning differences of lower hierarchy level taxa except for one OTU assigned to the genus *Ruminococcus* (OTU_59; $Q_{META} < 0.01$; Figure 3).

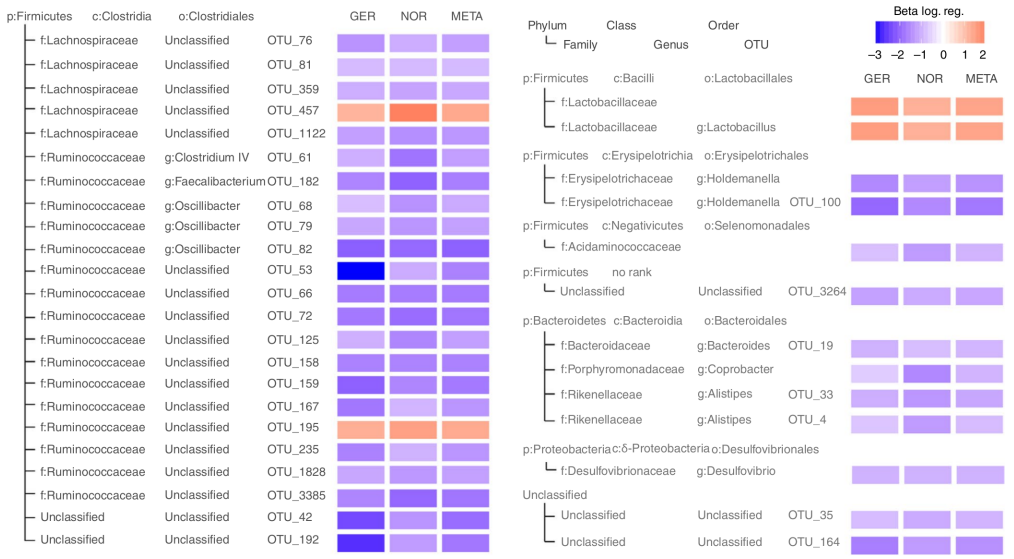


FIGURE 3 Robust results of the logistic regression within cohorts and the meta-analysis testing for differential prevalence of taxonomic groups in PSC patients and healthy controls. Only taxa with $P < 0.05$ in each cohort, $Q_{META} < 0.05$ and concordant effect direction are shown. Colour saturation expresses the effect size (Beta) of the association. Beta-values larger than zero (red boxes) represent a higher prevalence in PSC patients, taxa with values less than zero (blue) are less prevalent in PSC patients. Details on the model coefficients and the resulting *P*-values in the cohorts and the meta-analysis can be found in Table S4. p: phylum, c: class, o: order, f: family, g: genus

3.7 | The faecal microbiota profile can predict the diagnosis of PSC across different geographical regions

In order to investigate, whether faecal microbiota can be used to predict the presence of disease, we applied random forests classification to the pooled cohort of controls and patients with PSC (n = 270 subjects) using default hyperparameters. As baseline model variables all taxa with robust differential distribution were included (n = 43 features). Implementing 0.632 bootstrap resampling, a high performance with an average AUC of 0.88 was achieved (F1 = 0.83, MCC = 0.66; Figure 4A, taxon importance for the classifier evaluated by Gini index is displayed in Figure 4D). Training of the classifier on the German cohort and validation on the Norwegian subjects resulted in an AUC of 0.86 (F1 = 0.62, MCC = 0.32; Figure 4B). Classifier training with the Norwegian cohort and testing on the German population resulted in an AUC of 0.87 (F1 = 0.61, MCC = 0.51; Figure 4C). Further in-depth exploration of the classification is provided in the Supporting information.

3.8 | Diet has minor impact on microbial community alterations in PSC

Univariate comparisons of disease groups to healthy individuals showed only a minor reduction in daily water intake in PSC patients ($P_{adj} = 0.049$). Comparing multivariate differences in major dietary patterns in 220 samples of the German cohort, no significant differences were seen according to age ($P > 0.05$, $R^2 = 0.002$), BMI ($P > 0.05$, $R^2 = 0.008$) and diagnosis ($P > 0.05$, $R^2 = 0.013$). Gender, however, showed a strong impact ($P < 0.001$, $R^2 = 0.2$), as expected. Using the mentioned macronutrients and questionnaire derived intake variables as covariates in the analysis of disease-associated shifts in microbial beta-diversity yielded no significant associations ($P > 0.05$, respectively) and also revealed only minor effects on the still highly significant change in community composition associated with PSC ($P = 0.003$, $R^2 = 0.019$).

4 | DISCUSSION

It is unclear, to which extent reported single centre microbiota studies in PSC were influenced by environmental factors and whether reported associations would remain significant across different geographical regions. In this study, we analysed the faecal microbiota of patients with PSC including patients from a German and a Norwegian cohort based on 16S rRNA gene-amplicon sequencing profiles. To the best of our knowledge, this is the largest microbiota study focusing on PSC so far. The previously analysed cohorts[13,16] were reprocessed using a unified sequencing-library preparation and data analysis workflow to reduce technical and statistical disparities. Controlling for cohort-specific effects and potentially false positive results, the joint analysis of these two cohorts facilitated the identification of extensive disease-associated changes in the faecal microbiota structure in PSC independent from geographical

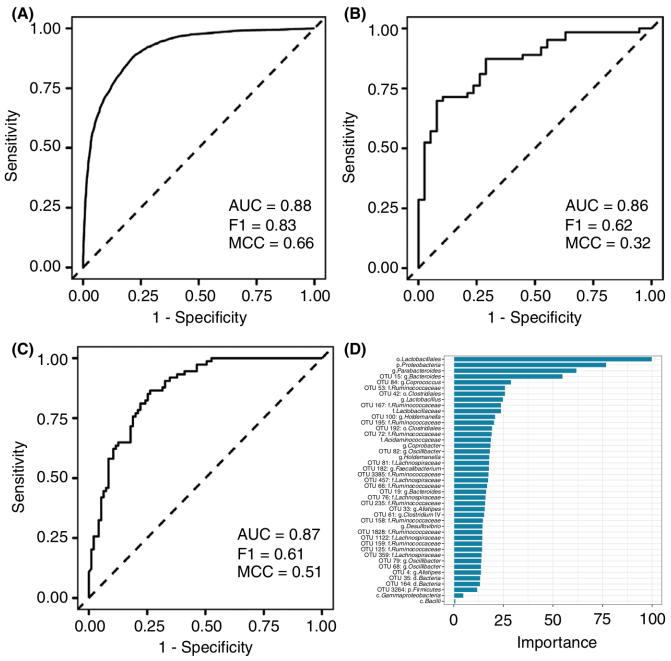


FIGURE 4 Receiver operating characteristic curve of random forest classification PSC vs controls across cohorts. Displayed are (A) the 0.632 bootstrap results from the pooled German and Norwegian cohort, (B) the classifier trained on the German cohort and validated on the Norwegian cohort and (C) vice versa. Features included in the model were the taxa with robust differential distribution between PSC and controls. (D) Feature importance of the respective taxa in the pooled classifier was ranked by Gini index

region. We identified several microbial signals previously not described in association with PSC, including an increased abundance of the phylum Proteobacteria, likely driven by the class Gammaproteobacteria, increased abundance of the genus *Parabacteroides*, and increase of one OTU belonging to the genus *Bacteroides*. Gammaproteobacteria comprise gram-negative bacteria with lipopolysaccharide (LPS) containing membranes, such as Enterobacteriaceae.[25] Multiple lines of evidence point to LPS as a common co-factor of liver injury.[26] Individual variation in Gammaproteobacteria has been directly linked with susceptibility to fatty liver disease.[27] The increase in *Parabacteroides* is intriguing, as it has been demonstrated that this bile-tolerant taxon is linked to changes in cholesterol and bile acid metabolism.[28,29]

For eight taxa across different taxonomic levels we demonstrated a consistently increased abundance in patients with PSC. These findings confirm previously found associations, eg for the genera *Veillonella*[13,16] and *Streptococcus*.[17] Increased abundance of both were previously found in patients with primary biliary cholangitis (PBC) and patients with liver cirrhosis of different origins.[12,30] This indicates that these alterations might be rather unspecific features of chronic liver disease. Additionally, we confirmed a markedly reduced abundance of one OTU belonging to the genus *Coprococcus* in patients with PSC. This taxon has previously been found to be decreased in patients with UC,[31] which is consistent with our own findings (Supporting information). This supports the notion of *Coprococcus* as an indicator of general gut integrity. Other previously described associations of bacterial taxa with PSC could not be confirmed, such as the recently described association with the genus *Klebsiella*.[32]

We could demonstrate an extensive alteration of the microbial community structure in patients with PSC, identifying 36 taxa with differential prevalence patterns associated with the disease, of which 32 were less present in PSC patients and including the genera *Faecalibacterium* and *Clostridium IV*. Both genera comprise butyrate-producing species, which provide an important energy source for intestinal epithelia and display an array of beneficial immunological properties.[33] *Faecalibacterium* was attributed with beneficial immunoregulatory properties and proposed as a pharmacobiotic agent to treat inflammatory diseases.[34] This finding further underlines its potential importance for the understanding of the pathophysiology and development of novel therapeutic approaches in inflammatory intestinal and liver diseases, where depletion of *Faecalibacterium* is a frequently observed common trait.[11,30]

The results of the machine learning classification highlight that faecal microbiota can be used to detect PSC in geographically separate cohorts. This might underline the potential pathophysiological significance of the faecal microbiota and a potential clinical value as a future biomarker. While the generalisability of the model is limited by the fact that the prevalence of PSC in the general population is significantly different from the prevalence in our cohort, the significant PSC-specific overlap of the faecal microbiota is promising. The results of the classifier additionally complement the results of the GLM-based analysis, identifying a subset of the same taxa that were found to be significantly changed in relative abundance to also be most important (scaled Gini Index > 50) for the pooled classifier.

In previous studies of the mucosal microbiota in UC patients with and without PSC, city of origin was the main determinant of gut microbiota profile despite identical handling of all samples and no consistent differences between UC and PSC with colitis were observed.[35] This is in line with the present findings, as significant differences were detected between PSC and UC in each cohort separately, however, only few of them were consistently observed across both cohorts. We can therefore not answer conclusively if microbiota profiling is also sufficient to distinguish PSC from UC in a geographically independent manner.

Our results strongly suggest that the increase of abundance or prevalence in PSC in contrast with controls is specific for certain taxa. Therefore, their potential role in the pathogenesis of PSC should be investigated. In addition, the extensive loss of many taxonomic groups is relatively unspecific and may potentially be explained by general effects of chronic inflammation rather than by disease-specific etiology.

Only marginal differences were identified in our study between patients with PSC only and PSC-IBD. This suggests that the liver disease and not the colitis is the primary driving force behind the observed gut microbial dysbiosis and that, as the differences in inflammation-localisation in the colon already suggest, PSC-IBD displays significant pathophysiological differences to UC.

Our study has several strengths and shortcomings. Notably, even though the German and Norwegian samples were processed with kits from different manufacturers, we could show extensive overlap in biological signals. Analysis of dietary patterns, that were available for the German cohort, showed no general differences between individuals based on disease status. Only minor influences on beta diversity were observed, which did not affect the clear disease-associated shift in microbial communities. Previous studies investigating the influence of dietary patterns on the microbiota could show, that indeed diet can influence microbiome compostion.[36,37] However, these were performed on larger study populations (n > 1000) and still only found small individual effects, thus to really investigate diet-disease interactions, larger, well-typed PSC cohorts are needed, which are currently not available.

Although facilitating large-scale analysis, amplicon-based marker gene surveys always come with trade-offs regarding fragment-specific biases in amplification and taxonomic assignment quality, which may in part explain differences between studies. Ultimately, the intestine must be regarded as an open, highly dynamic and spatially heterogeneous system, and faecal samples can only serve as a proxy for the actual state. Nevertheless, we here present the largest microbiota study performed in PSC to date and show robust and geography-spanning alterations in the faecal microbiota, in spite of sampling and technical differences between the centres. This strongly supports the conclusion that the observed microbiota changes are disease specific, and not primarily driven by environmental factors.

In summary, patients with PSC from different geographical regions display shared differences in gut microbial composition compared to healthy controls, independent from the presence of concurrent IBD and the use of common medication. The PSC specific gut microbiota signature might have diagnostic potential and provides a strong rationale for further microbiome meta-analyses with even larger international cohorts. These should also utilise metagenomic sequencing, profiling of microbial metabolites as well as longitudinal designs to further define the role of the gut microbiota in the pathogenesis and clinical care in PSC. Furthermore, additional patients with chronic or cholestatic liver disease should be included for validation of disease specific effects. This study focused on broad, disease-specific signals, as these are likely to be directly connected to the largely unknown etiology of PSC, independent of environmental influences. Ultimately, this may lay the foundation for new therapies targeting the gut microbiota in PSC.

ACKNOWLEDGEMENTS

We thank Ms. Ilona Urbach, Ms. Ines Wulf and Mr. Tonio Hauptmann of the IKMB microbiome laboratory and the staff of the IKMB sequencing facilities for excellent technical support.

Declaration of personal interests: None.

AUTHORSHIP

Guarantor of the article: None.

Author contributions: CS and AF designed the study and obtained funding. RZ, MKu, JH, WL, MT, TK, AWL, GD, FL and MKr acquired and quality-controlled patient samples and data. F-AH supervised sample processing and sequencing. MR and TL performed statistical analyses. MR, F-AH, TL, CB, RZ, MKu, LT, JH, CS and AF interpreted data and drafted the manuscript with input and critical revision from all authors. All authors revised and approved the final version of the manuscript.

ORCID

Malte Rühlemann https://orcid.org/0000-0002-0685-0052
Roman Zenouzi https://orcid.org/0000-0003-0136-0924
Martin Kummen https://orcid.org/0000-0001-9660-6290
Marcin Krawczyk https://orcid.org/0000-0002-0113-0777

REFERENCES

1. Karlsen TH, Folseraas T, Thorburn D, Vesterhus M. Primary sclerosing cholangitis - a comprehensive review. *J Hepatol*. 2017;67:1298-1323.
2. Boonstra K, Weersma RK, van Erpecum KJ, et al. Population-based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. *Hepatology*. 2013;58:2045-2055.
3. Tischendorf J, Hecker H, Krüger M, Manns MP, Meier PN. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: A single center study. *Am J Gastroenterol*. 2007;102:107-114.
4. Weismuller TJ, Trivedi PJ, Bergquist A, et al. Patient age, sex, and inflammatory bowel disease phenotype associate with course of primary sclerosing cholangitis. *Gastroenterology*. 2017;152:1975-1984 e8.
5. Chapman R, Fevery J, Kalloo A, et al. Diagnosis and management of primary sclerosing cholangitis. *Hepatology*. 2010;51:660-678.
6. Molodecky NA, Kareemi H, Parab R, et al. Incidence of primary sclerosing cholangitis: a systematic review and meta-analysis. *Hepatology*. 2011;53:1590-1599.
7. Lunder AK, Hov JR, Borthne A, et al. Prevalence of sclerosing cholangitis detected by magnetic resonance cholangiography in

patients with long-term inflammatory bowel disease. *Gastroenterology*. 2016;151:660-669 e4.
8. Loftus EV, Harewood GC, Loftus CG, et al. PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis. *Gut*. 2005;54:91-96.
9. de Vries AB, Janse M, Blokzijl H, et al. Distinctive inflammatory bowel disease phenotype in primary sclerosing cholangitis. *World J Gastroenterol*. 2015;21:1956-1971.
10. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457:480-484.
11. Gevers D, Kugathasan S, Denson L, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*. 2014;15:382-392.
12. Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature*. 2014;513:59-64.
13. Kummen M, Holm K, Anmarkrud JA, et al. The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. *Gut*. 2017;66:611-619.
14. Torres J, Bao X, Goel A, et al. The features of mucosa-associated microbiota in primary sclerosing cholangitis. *Aliment Pharmacol Ther*. 2016;43:790-801.
15. Bajer L, Kverka M, Kostovcik M, et al. Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. *World J Gastroenterol*. 2017;23:4548-4558.
16. Rühlemann MC, Heinsen F-A, Zenouzi R, Lieb W, Franke A, Schramm C. Faecal microbiota profiles as diagnostic biomarkers in primary sclerosing cholangitis. *Gut*. 2017;66:753-754.
17. Sabino J, Vieira-Silva S, Machiels K, et al. Primary sclerosing cholangitis is characterised by intestinal dysbiosis independent from IBD. *Gut*. 2016;65:1681-1689.
18. Krawczak M, Nikolaus S, von Eberstein H, Croucher P, El Mokhtari NE, Schreiber S. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet*. 2006;9:55-61.
19. Lindor KD, Kowdley KV, Harrison ME. Clinical guideline: primary sclerosing cholangitis. *Am J Gastroenterol*. 2015;110:646-659; quiz 60.
20. EASL Clinical Practice Guidelines: management of cholestatic liver diseases. *J Hepatol*. 2009;51:237-267.
21. Nöthlings U, Hoffmann K, Bergmann MM, Boeing H. Fitting portion sizes in a self-administered food frequency questionnaire. *J Nutr*. 2007;137:2781-2786.
22. Wang J, Thingholm LB, Skiecevičienė J, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet*. 2016;48:1396-1406.
23. Team RDC. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014.
24. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
25. Raetz C, Whitfield C. Lipopolysaccharide endotoxins. *Annu Rev Biochem*. 2002;71:635-700.
26. Su GL. Lipopolysaccharides in liver injury: molecular mechanisms of Kupffer cell activation. *Am J Physiol-Gastrointest Liver Physiol*. 2002;283:G256-G265.
27. Spencer MD, Hamp TJ, Reid RW, Fischer LM, Zeisel SH, Fodor AA. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology*. 2011;140:976-986.
28. Worthmann A, John C, Rühlemann MC, et al. Cold-induced conversion of cholesterol to bile acids in mice shapes the gut microbiome and promotes adaptive thermogenesis. *Nat Med*. 2017;23:839-849.
29. Tanaka Y, Benno Y, Ohkuma M, Sakamoto M. Parabacteroides faecis sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol*. 2015;65(Pt 4):1342-1346.
30. Tang R, Wei Y, Li Y, et al. Gut microbial profile is altered in primary biliary cholangitis and partially restored after UDCA therapy. *Gut*. 2017;67:534-541.
31. Shaw KA, Bertha M, Hofmekler T, et al. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med*. 2016;8:75.
32. Nakamoto N, Sasaki N, Aoki R, et al. Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis. *Nat Microbiol*. 2019;4:492-503.
33. Louis P, Flint HJ. Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol Lett*. 2009;294:1-8.
34. Miquel S, Martín R, Rossi O, et al. *Faecalibacterium prausnitzii* and human intestinal health. *Curr Opin Microbiol*. 2013;16:255-261.
35. Kevans D, Tyler AD, Holm K, et al. Characterization of intestinal microbiota in ulcerative colitis patients with and without primary sclerosing cholangitis. *J Crohns Colitis*. 2016;10:330-337.
36. Falony G, Joossens M, Vieira-Silva S, et al. Population-level analysis of gut microbiome variation. *Science*. 2016;352:560-564.
37. Zhernakova A, Kurilshikov A, Bonder MJ, et al. Population-based metagenomic analysis reveals markers for gut microbiome composition and diversity. *Science*. 2016;352:565-569.

SUPPORTING INFORMATION

Additional supporting information will be found online in the Supporting Information section at the end of the article.

How to cite this article: Rühlemann M, Liwinski T, Heinsen F-A, et al. Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis. *Aliment Pharmacol Ther*. 2019;50:580–589. https://doi.org/10.1111/apt.15375

APPENDIX 1

The authors' complete affiliation list

Malte Rühlemann, Femke-Anouska Heinsen, Corinna Bang, Louise Thingholm, and Andre Franke, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; Timur Liwinski, Roman Zenouzi, Ansgar Lohse, and Christoph Schramm, Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; Martin Kummen, Tom Karlsen, and Johannes Hov, Norwegian PSC Research Center, Division of Surgery, Inflammatory Medicine and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway and Research Institute of Internal Medicine, Oslo University Hospital Rikshospitalet, Oslo, Norway; Marie Tempel and Wolfgang Lieb, Institute of Epidemiology, Christian-Albrechts-University of Kiel, Kiel, Germany; Gerald Denk, Department of Medicine II, Liver Center Munich, University Hospital, LMU Munich, Munich, Germany; Frank Lammert and Marcin Krawczyk, Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany; Marcin Krawczyk, Laboratory of Metabolic Liver Diseases, Center for Preclinical Research, Department of General, Transplant and Liver Surgery, Medical University of Warsaw, Warsaw, Poland; Christoph Schramm, Martin Zeitz Centre for Rare Diseases, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Article F: Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species

PostScript

LETTER

Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species

LETTER TO THE EDITOR

We read with interest the recent Gut article by Lemoinne *et al* describing a dysbiosis of the fungal gut community in faeces of patients suffering from primary sclerosing cholangitis (PSC).[1] Though several reports, including our own previous data, support a functional and potentially pathogenic link between the intestinal bacteriota and liver inflammation in PSC,[2 3] the aetiology of the disease remains largely unknown.

We here report on the fungal mycobiome results of our cohort from Northern Germany approved by the local ethics committees (A148/14 and MC-111/15) comprising stool samples of 66 healthy control (HC) subjects, 65 patients with well-characterised PSC (including a subgroup with concomitant colitis (PSC-IBD), n=32) and 38 subjects with UC.[3] PCR and sequencing of the fungus-specific internal transcribed spacer 2 genomic region was performed as previously described[4] using the primer pair 5.8S-Fun and ITS4-Fun on an Illumina MiSeq machine. Sequencing data were subjected to quality control by using the open source package DADA2 (V.1.10)[5] in R (V.3.5.1; https://github.com/mrue-hlemann/ikmb_amplicon_processing). Amplicon sequence variants were taxonomically annotated using the UNITE ITS database (V.7.2).[6]

In disagreement with the findings in the French cohort,[1] overall fungal alpha diversity in the German cohort was neither altered in PSC nor in UC versus HC as calculated by Shannon species equivalent (figure 1A). None of the disease groups significantly deviated in community composition from healthy

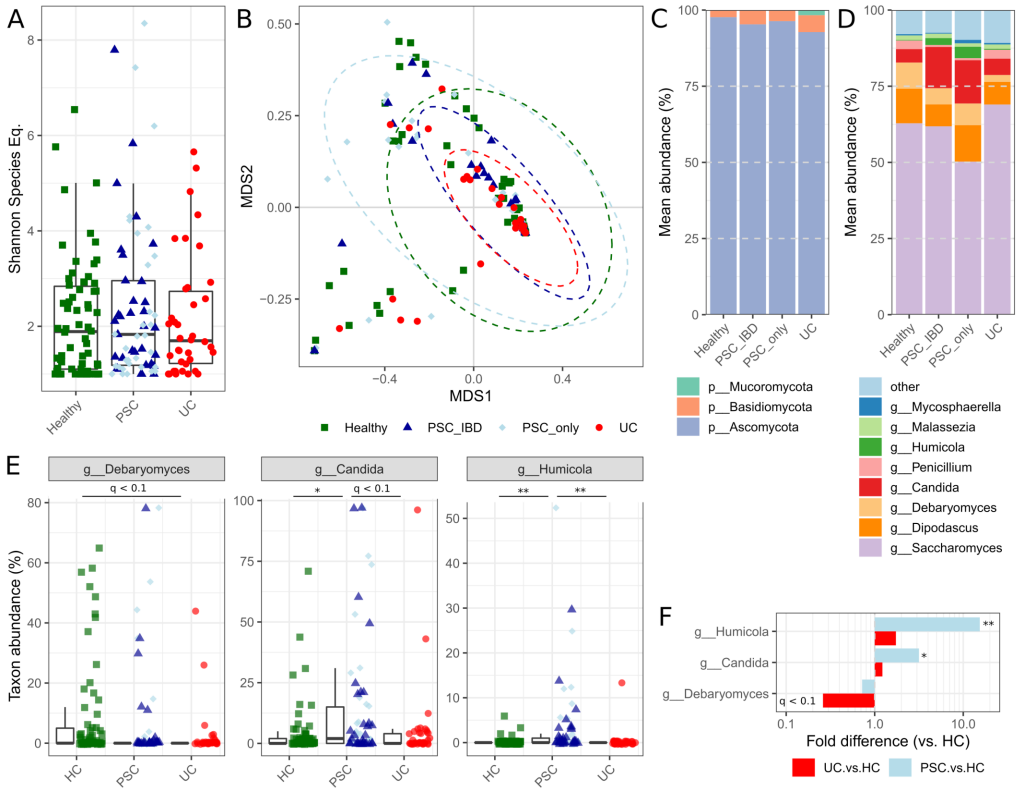


Figure 1 Mycobiome of individuals with primary sclerosing cholangitis (PSC) and UC as well as healthy controls (HC) (all of northern German origin). Rarefaction curves for Shannon diversity of sequence variants reached plateau between 50 and 100 sequences per samples, thus samples were normalised to 100 random reads per sample. (A) Alpha diversity as presented by Shannon species equivalents (all p>0.05). (B) Beta diversity ordination of the Bray-Curtis dissimilarity based on genus-level fungal abundances (all p_{adonis} >0.05). (C) Phylum-level and (D) genus-level mean abundances of all taxa with >1% mean abundance and present in at least 10 samples. (E) Group-wise box-and-whisker plots for significant genus level annotations tested for differential abundances with individual values represented as data points. (F) Differences in group-mean abundances of patients with PSC and UC, as compared with HC. *q<0.05, **q<0.01.

individuals (all p_{adonis} >0.05; figure 1B). Fungal composition on phylum level was found to be mainly dominated by Ascomycota (figure 1C), particularly by the genera *Saccharomyces*, *Candida* and *Dipodascus* (figure 1D) in relatively higher abundance of reads when compared with the findings of Lemoinne *et al*.[1] Though our results generally validate the previously described overall fungal composition in stool, we were not able to detect the genus *Exophiala*, which was found in five PSC patients from France exclusively. Whether this is due to methodological differences (choice of primer sets, data analysis tools and sampling depth) or presence of this fungus in only a subset of PSC patients not sampled in the German cohort needs to be determined.

Disease-associated differential abundance of fungal taxa was investigated by applying Student's t-test to the arcsin-squareroot-transformed relative abundances of all genera with mean abundance >1% and present in at least 10 individuals. This analysis revealed increased levels of the genera *Candida* and *Humicola* (species level annotation suggests *H. grisea*) in PSC patients with and without concomitant colitis compared with HC (all q_{BH} <0.05; figure 1E and F) and UC (all q_{BH} <0.1; figure 1E) individuals. *H. grisea*, recently reclassified as *Trichocladium griseum*,[7] belongs to the fungal class Sordariomycetes, thus our results reproduce the significant increase of this class in PSC patients, as previously described by Lemoinne and colleagues, but at increased taxonomic resolution. Previous research on *T. griseum* showed that it is most frequently isolated from soil and plants but also occasionally found in patients suffering from peritonitis.[8] In addition, the validated increase of *Candida* species in PSC patients argues for an immunogenic role of these fungi, particularly with respect to earlier findings that demonstrated their high potential to induce Th17 response in T cells.[9] Increased Th17 numbers have previously been reported in PSC patients and recently been shown to be involved in PSC pathogenesis.[10]

In summary, both the significant increase of the fungal class Sordariomycetes, likely *T. griseum*, as well as of *Candida* species in stool samples of PSC patients, now found in two independent and geographically distinct PSC patient panels that were analysed with divergent methodological approaches, strongly demands for additional analyses on these fungi and their role in PSC.

Malte Christoph Rühlemann [1], **Miriam Emmy Leni Solovjeva,**[1] **Roman Zenouzi,**[2] **Timur Liwinski** [2], **Martin Kummen** [3,4], **Wolfgang Lieb,**[5] **Johannes Roksund Hov** [3,4], **Christoph Schramm,**[2,6] **Andre Franke** [1], **Corinna Bang** [1]

[1]Institute of Clinical Molecular Biology, Christian-Albrechts-Universität zu Kiel, Kiel, Germany
[2]Department of Internal Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[3]Norwegian PSC Research Center, Oslo University Hospital, Rikshospitalet, Oslo, Norway
[4]Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[5]Institute of Epidemiology and Biobank POPGEN, Christian-Albrechts-University of Kiel, Kiel, Germany
[6]Martin Zeitz Centre for Rare Diseases, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Correspondence to Dr Corinna Bang, Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel 24105, Germany; c.bang@ikmb.uni-kiel.de

Twitter Malte Christoph Rühlemann @mruehlemann and Johannes Roksund Hov @hov_jer

Acknowledgements We would like to thank Ms Ilona Urbach, Ms Ines Wulf and Mr Tonio Hauptmann of the IKMB microbiome laboratory for excellent technical support.

Contributors AF and CB designed the study. CS and AF obtained funding. RZ, CS and WL acquired and quality-controlled patient samples and data. CB and MELS supervised sample processing and sequencing. MCR performed statistical analyses. MCR, MELS, RZ, TL, MK, JRH, CS, AF and CB interpreted data and drafted the manuscript with input and critical revision from all authors. All authors revised and approved the final version of the manuscript.

Funding This study was supported by the Deutsche Forschungsgemeinschaft (DFG) Clinical Research Group 306' Primary sclerosing cholangitis' (no: KFO306) as well as Research Training Group 1743 and received infrastructure support from the DFG Cluster of Excellence 'Inflammation at Interfaces' (http://www.inflammation-at-interfaces.de, no: EXC306 and EXC306/2) and the German Ministry of Education and Research (BMBF) program e:Med sysINFLAME (http://www.gesundheitsforschung-bmbf.de/de/5111.php, no: 01ZX1306A).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Sequencing and clinical data of the patient samples used in this study can be applied for via the Popgen Biobank (Institute of Epidemiology, Christian-Albrechts-University of Kiel, Germany).

AF and CB contributed equally.

To cite Rühlemann MC, Solovjeva MEL, Zenouzi R, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2019-320008

Received 4 October 2019
Revised 10 October 2019
Accepted 11 October 2019

Gut 2019;**0**:1–2. doi:10.1136/gutjnl-2019-320008

ORCID iDs
Malte Christoph Rühlemann http://orcid.org/0000-0002-0685-0052
Timur Liwinski http://orcid.org/0000-0002-1041-9142
Martin Kummen http://orcid.org/0000-0001-9660-6290
Johannes Roksund Hov http://orcid.org/0000-0002-5900-8096
Andre Franke http://orcid.org/0000-0003-1530-5811
Corinna Bang http://orcid.org/0000-0001-6814-6151

REFERENCES

1 Lemoinne S, Kemgang A, Ben Belkacem K, *et al*. Fungi participate in the dysbiosis of gut microbiota in patients with primary sclerosing cholangitis. *Gut* 2019:gutjnl-2018-317791 (published Online First: 2019/04/19).
2 Kummen M, Holm K, Anmarkrud JA, *et al*. The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. *Gut* 2017;66:611–9.
3 Rühlemann M, Liwinski T, Heinsen F-A, *et al*. Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis. *Aliment Pharmacol Ther* 2019;50:580–9.
4 Taylor DL, Walters WA, Lennon NJ, *et al*. Accurate estimation of fungal diversity and abundance through improved lineage-specific primers optimized for Illumina amplicon sequencing. *Appl Environ Microbiol* 2016;82:7217–26.
5 Callahan BJ, McMurdie PJ, Rosen MJ, *et al*. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13:581–3.
6 Nilsson RH, Larsson K-H, Taylor AFS, *et al*. The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2019;47:D259–64.
7 Wang XW, Yang FY, Meijer M, *et al*. Redefining *Humicola sensu stricto* and related genera in the *Chaetomiaceae*. *Stud Mycol* 2019;93:65–153.
8 Burns N, Arthur I, Leung M, *et al*. Humicola sp. as a cause of peritoneal dialysis-associated peritonitis. *J Clin Microbiol* 2015;53:3081–5.
9 Katt J, Schwinge D, Schoknecht T, *et al*. Increased T helper type 17 response to pathogen stimulation in patients with primary sclerosing cholangitis. *Hepatology* 2013;58:1084–93.
10 Nakamoto N, Sasaki N, Aoki R, *et al*. Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis. *Nat Microbiol* 2019;4:492–503.

Article G: Epidermal lipid composition, barrier integrity, and eczematous inflammation are associated with skin microbiome configuration

Atopic dermatitis and inflammatory skin disease

Epidermal lipid composition, barrier integrity, and eczematous inflammation are associated with skin microbiome configuration

Hansjörg Baurecht, PhD,[a]* Malte C. Rühlemann, MSc,[b]* Elke Rodríguez, PhD,[a] Frederieke Thielking, MD,[a] Inken Harder, MSc,[a] Anna-Sophie Erkens, MD,[a] Dora Stölzl, MD,[a] Eva Ellinghaus, PhD,[b] Melanie Hotze, MSc,[a] Wolfgang Lieb, MD, MSc,[c] Sheng Wang, PhD,[d] Femke-Anouska Heinsen-Groth, PhD,[b] Andre Franke, PhD,[b] and Stephan Weidinger, MD[a] *Kiel, Germany, and Hangzhou, China*

Background: Genomic approaches have revealed characteristic site specificities of skin bacterial community structures. In addition, in children with atopic dermatitis (AD), characteristic shifts were described at creases and, in particular, during flares, which have been postulated to mirror disturbed skin barrier function, cutaneous inflammation, or both.
Objective: We sought to comprehensively analyze microbial configurations in patients with AD across body sites and to explore the effect of distinct abnormalities of epidermal barrier function.
Methods: The skin microbiome was determined by using bacterial 16S rRNA sequencing at 4 nonlesional body sites, as well as acute and chronic lesions of 10 patients with AD and 10 healthy control subjects matched for age, sex, and filaggrin *(FLG)* mutation status. Nonlesional sampling sites were characterized for skin physiology parameters, including chromatography-based lipid profiling.
Results: Epidermal lipid composition, in particular levels of long-chain unsaturated free fatty acids, strongly correlated with bacterial composition, in particular Propionibacteria and Corynebacteria abundance. AD displayed a distinct community structure, with increased abundance and altered composition of staphylococcal species across body sites, the strongest loss of diversity and increase in *Staphylococcus aureus* seen on chronic lesions, and a progressive shift from nonlesional skin to acute and chronic lesions. *FLG*-deficient skin showed a distinct microbiome composition resembling in part the AD-related pattern.
Conclusion: Epidermal barrier integrity and function affect the skin microbiome composition. AD shows an altered microbial configuration across diverse body sites, which is most pronounced at sites of predilection and AD. Eczematous affection appears to be a more important determinant than body site. (J Allergy Clin Immunol 2018;141:1668-76.)

Key words: *Atopic dermatitis, filaggrin, lipidome, microbiome*

The skin shows strong topographic variations and spatial differences in terms of cellular makeup and molecular characteristics, which, under normal physiologic conditions, create a distinct pattern of microenvironments that differ in humidity, pH, temperature, antimicrobial peptide composition, and lipid content.[1,2] Landmark genomic studies in healthy subjects showed that the composition of microbial communities on the skin surface displays marked regional differences, which have been postulated to be driven by differences in skin physiology.[3-6] Microbiota can affect epidermal barrier function[7] and cutaneous immunity.[8] Thus the mutualistic relationship between physical, chemical, and immunologic features of the skin barrier and microbial populations can create pathogenicity islands that, under certain pathologic conditions, underlie the remarkable site specificities of many skin diseases.

A paradigmatic skin disorder with such a characteristic distribution of lesions and a postulated disturbance of the balance between host and microbes is atopic dermatitis (AD).[9,10] Despite induction of antimicrobial peptide in AD lesions,[10] colonization with *Staphylococcus aureus* is found in the vast majority of patients.[11] With the help of sequencing-based surveys, a reduced overall diversity of microbial communities in favor of *S aureus* was observed at AD predilection sites, in particular during disease flares.[12] However, it is unclear whether a shifted skin microbiome

From [a]the Department of Dermatology and Allergy, University Hospital Schleswig-Holstein, Campus Kiel; [b]the Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel; [c]the Institute of Epidemiology and Biobank PopGen, Christian-Albrechts-University of Kiel; and [d]the College of Life Information Science and Instrument Engineering, Hangzhou Dianzi University, Hangzhou.
*These authors contributed equally to this work.
The project received infrastructure support through the DFG Clusters of Excellence "Inflammation at Interfaces" (grants EXC306 and EXC306/2) and was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (sysINFLAME, grant 01ZX1306A).
Disclosure of potential conflict of interest: M. C. Rühlemann's institution received grant DFG-CRC 1182 for this work. F.-A. Heinsen-Groth's institution received grant 01ZX1306A from German Federal Ministry of Education and Research for this work. S. Weidinger's institution received DFG Clusters of Excellence "Inflammation at Interfaces" (grants EXC306 and EXC306/2) and German Federal Ministry of Education and Research (BMBF) grants within the framework of the e:Med research and funding concept (sysINFLAME, grant 01ZX1306A) for this work. The rest of the authors declare that they have no relevant conflicts of interest.
Received for publication September 15, 2017; revised December 20, 2017; accepted for publication January 3, 2018.
Available online February 5, 2018.
Corresponding author: Stephan Weidinger, MD, Department of Dermatology and Allergy, University Hospital Schleswig-Holstein, Campus Kiel, Rosalind-Franklin Str. 7, 24105 Kiel, Germany. E-mail: sweidinger@dermatology.uni-kiel.de.
The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections
0091-6749/$36.00

https://doi.org/10.1016/j.jaci.2018.01.019

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

Abbreviations used
AD: Atopic dermatitis
Af: Antecubital fossa
AS: α-Hydroxy fatty acid/sphingosine base
CoNS: Coagulase-negative staphylococci
Ef: Extensor forearm
FDR: False discovery rate
FFA: Free fatty acid
FLG: Filaggrin
GLM: Generalized linear model
OTU: Operational taxonomic unit
SFC-MS/MS: Supercritical fluid chromatography mass spectrometry
TEWL: Transepidermal water loss
Vf: Volar forearm

is a general feature in patients with AD or restricted to certain sites if the lesional microbiome signatures are dependent on the acuity of inflammation and how defined insults to epidermal barrier function and integrity affect the microbiome.

The best characterized cause for a generalized skin barrier impairment is an inherited deficiency of the epidermal structural protein filaggrin (FLG) because of a single null mutation, as seen in up to 10% of the population and up to a third of patients with AD.[9] Apart from a decreased integrity of the stratum corneum, FLG deficiency has also been suggested to affect lipid formation[13] and enhance S *aureus* growth.[14] A recent study further suggested a lower relative abundance of gram-positive anaerobe cocci in *FLG*$^{-/-}$ skin,[15] but that study examined a single location only and compared "extremes" (ie, subjects with a complete FLG deficiency with ichthyosis vulgaris). Here we set out to gain more insights into the effect of *FLG* haploinsufficiency, epidermal physiology, and incident AD on skin microbiome configuration at different body sites and through disease stages.

METHODS

Study population

For this study, 4 groups of German adults from the PopGen population-based biobank[16] were invited to participate in a follow-up examination: (1) patients with AD carrying a single *FLG* mutation (AD FLGmut), (2) patients with AD without *FLG* mutations (AD FLGwt), (3) patients with no history of chronic skin or allergic disorders carrying a single *FLG* mutation (control FLGmut), and (4) patients with no history of chronic skin or allergic disorders without *FLG* mutations (control FLGwt). Information on the *FLG* genotype was available from prior studies.[17] Proband characteristics are described in Table E1 in this article's Online Repository at www.jacionline.org. The follow-up examination included a skin examination by an experienced dermatologist blinded to the genotype. AD was diagnosed by using American Academy of Dermatology diagnostic criteria.[18] For the current analysis, 5 subjects per group were selected and matched by age and sex. Patients with AD selected had to display at least 1 unaffected antecubital crease. None of the participants had received systemic immunosuppressants or systemic antibiotics in the preceding 3 months. All participants were asked to avoid bathing and application of any topical agent 24 hours before the examination visit. The study was approved by the ethics review board of the Medical Faculty of the University Kiel, Germany, and written informed consent was obtained from all study participants.

Sampling

At sampling sites, skin pH was measured with a Skin-pH-Meter (HI-99181/HI-1414D; Hanna Instruments, Vöhringen, Germany), and transepidermal water loss (TEWL) was measured with the Tewameter TM 300 (Courage + Khazaka Electronic GmbH, Cologne, Germany), according to the manufacturer's instructions. The mean of 3 measurements was used for analysis.

Skin swabs were taken from 4 different nonlesional sites (forehead, antecubital fossa [Af], volar forearm [Vf], and extensor forearm [Ef]) and from a chronic skin lesion in patients with AD. In 6 of the patients with AD, an additional swab was taken from an acute lesion defined as a new lesion of less than 24 hours' duration without clinical signs of lichenification. Skin swabs were collected by swabbing a 4-cm^2 area with sterile Catch-All Sample Collection Swabs (Epicentre Biotechnologies, Madison, Wis) soaked in sterile specimen collection fluid (SCF-1) solution, as previously described.[19] As negative controls, we took 2 swabs exposed only to ambient air. DNA was extracted with the PowerSoil DNA Isolation Kit (MoBio Laboratories, Carlsbad, Calif), as described previously.[20]

To harvest stratum corneum lipids, 4 tape strips per site were taken by using D-Squame discs (CuDerm, Dallas, Tex), as described previously.[21] Tapes were placed in sterile 1.5-mL Eppendorf tubes (Eppendorf, Hamburg, Germany) and stored at −80°C immediately.

Processing of bacterial 16S rRNA sequenced data

Bacterial DNA was amplified, and the 16S rRNA gene was sequenced, as previously described (see the Methods section in this article's Online Repository at www.jacionline.org for details).[22] Briefly, the amplicon-specific hypervariable regions V1 and V2 of the bacterial 16S rRNA gene were sequenced on the Illumina MiSeq platform (Illumina, San Diego, Calif). After demultiplexing and quality trimming (https://github.com/najoshi/sickle), forward and reverse sequences were combined by using VSEARCH.[23] Quality control and chimera filtering were carried out with the FastX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and UCHIME algorithm.[24] Cleaned FastA sequences were combined and clustered into operational taxonomic units (OTUs) by using the UPARSE algorithm.[25] For each sample, a subset of 10,000 random reads was picked to construct the OTU abundance table. Taxonomic classification from the phylum to genus level for the same 10,000 reads was performed with the RDP classifier and the Greengenes database.[26] Classification of OTUs to *Staphylococcus* species (98% similarity threshold) was performed by comparing against the EzBioCloud database (http://www.ezbiocloud.net).[27] Taxa with more than 90% of their counts within the first 5% of the count range were excluded. From the remaining taxa, a taxon was considered a "colonist" of a body site if more than 0.1% of the reads per sample mapped to the recovered genome.

Lipid analysis

Three hundred forty-eight ceramides, 12 free fatty acids (FFAs), and cholesterol sulfate were semiquantitatively measured at 3 body sites by using a Waters UPC2/Sciex QTrap 550 mass spectrometer supercritical fluid chromatography mass spectrometry (SFC-MS/MS) system (Waters Corporation, Milford, Mass/Sciex, Framingham, Mass) with the TrueMass Stratum Corneum Lipid Panel (Metabolon, Durham, NC). Details on lipid analysis are provided in the Methods section in this article's Online Repository.

Data analysis and statistics

Details on statistical analysis are presented in the Methods section in this article's Online Repository. Briefly, all analyses were carried out with R 3.3.2 software (www.R-project.org).[28] For analyses of α- and β-diversities of the skin microbiome, the vegan (version 2.4) package was used, differences between groups were evaluated by using the Wilcoxon test, and correlation of taxa on different levels were calculated by means of Spearman correlation. Association of individual skin bacterial traits was analyzed by using generalized linear models (GLMs) with appropriate link functions and variance estimations to account for increasing numbers of zeros and overdispersion. For taxons with exhibits with more zeros than would be allowed in GLMs, we applied hurdle models implemented in the pscl (version 1.4.9) package,[29] and for multivariate taxa analysis, we used the mvabund (version 3.12.3) package. Ratios of microbial taxa abundance were first log-transformed and then carried forward to statistical analysis.

TABLE I. Mean relative abundance of genera present with at least 0.5% in healthy subjects in at least 3 sites

Phylum/genus	Af	Forehead	Ef	Vf	*P* value, Friedman test
Actinobacteria	36.7%	50.2%	44.7%	44.2%	.0327
Propionibacterium	19.0%	38.9%	26.1%	25.1%	.0062
Corynebacterium	9.8%	6.8%	10.6%	11.0%	.1968
Micrococcus	1.5%	0.9%	2.8%	2.1%	.0576
Kocuria	1.0%	0.6%	1.1%	0.9%	.2379
Actinomyces	0.7%	0.8%	0.7%	0.4%	.8964
Rothia	0.6%	1.0%	1.1%	0.3%	.0079
Firmicutes	45.0%	30.8%	36.3%	37.7%	.2933
Staphylococcus	31.3%	15.5%	16.8%	22.3%	.1524
Streptococcus	5.5%	5.6%	7.6%	3.8%	.1870
Anaerococcus	1.2%	1.6%	2.4%	2.7%	.1131
Finegoldia	1.6%	0.8%	1.8%	2.7%	.0560
Veillonella	1.1%	1.3%	1.4%	0.6%	.4618
Lactobacillus	1.0%	0.3%	1.4%	1.2%	.1576
Peptoniphilus	0.5%	0.4%	0.7%	1.1%	.7006
Proteobacteria	12.3%	14.9%	14.2%	13.4%	.5164
Acinetobacter	3.6%	2.3%	4.8%	4.5%	.5164
Haemophilus	0.7%	1.9%	1.5%	0.8%	.1604
Enhydrobacter	1.6%	0.6%	1.3%	1.4%	.2530
Neisseria	0.7%	1.0%	0.7%	0.2%	.8192
Microvirgula	1.0%	1.2%	2.5%	1.7%	.7650
Bacteriodetes	4.8%	2.7%	3.7%	4.2%	.7819
Prevotella	1.0%	0.6%	1.0%	0.9%	.3227
Chryseobacterium	1.2%	0.6%	1.1%	1.3%	.7086
Fusobacteria	0.8%	1.0%	0.8%	0.2%	.8935
Leptotrichia	0.7%	0.6%	0.5%	0.03%	.8584

Integrative analysis was applied to estimate the influence of skin lipids on microbial composition by using sparse canonical correlation analysis implemented in the PMA package.[30,31]

Gradual changes of microbial taxa from nonlesional over acute to chronic lesions are evaluated by using linear mixed models with the lme4 package (version 1.1.13).

RESULTS

Microbiome profiling

In the microbiome of healthy subjects, 5 prevalent bacterial phyla accounted for more than 99% of each subject's microbiota (Actinobacteria, Proteobacteria, Firmicutes Bacteroidetes, and Fusobacteria). Twenty-one genera were present in at least 3 sites with at least 0.5% abundance (Table I). Four of these genera (*Propionibacterium*, *Corynebacterium*, *Staphylococcus*, and *Streptococcus*) were present in all 10 healthy subjects with at least 1% abundance. Across sites, staphylococci showed the highest abundance (15.9% to 31.3%), with *Staphylococcus epidermidis* being the most prevalent *Staphylococcus* species (51.9% to 55.2%). Similarity analysis showed that Vf and Ef sites cluster together and are distinctly separated from the Af and the forehead.

FLG deficiency correlates with decreased bacterial diversity and a distinct bacterial colonization pattern

In healthy subjects, across the 4 sampling sites, FLG deficiency shows a consistent decrease of α-diversity (mean difference at Af, 0.43; Vf, 1.06; Ef, 0.39; and forehead, 0.22). Except for the forehead, we also observed a lower genera richness, which was most pronounced at the AD predilection site Af (see Fig E1 in this article's Online Repository at www.jacionline.org). The similarity analysis for each site showed a clear separation of FLG-deficient from FLG-competent subjects. Across all sites, FLG-deficient subjects displayed a decreased abundance of the Proteobacteria genera *Acinetobacter*, *Enhydrobacter*, and *Microvirgula* (Af: 1.1% vs 11.1%, $P_{\text{false discovery rate [FDR]}} = .035$; Vf: 1.8% vs 13.4%, $P_{\text{FDR}} = .002$; and Ef: 2.5% vs 14.7%, $P_{\text{FDR}} = .002$; Fig 1). At the same time, FLG-deficient subjects displayed an increased abundance of *Propionibacterium* (Af: 23.3% vs 14.7%, $P = .020$; Vf: 32.1% vs 18.0%, $P = .010$; and Ef: 34.1% vs 18.0%, $P_{\text{FDR}} = 4.0 \times 10^{-4}$). Furthermore, there was a tendency toward increased Firmicutes abundance, in particular for the genus *Staphylococcus*. However, there was no specific association with *S aureus*, *S epidermidis*, or *Staphylococcus hominis* abundance.

We also investigated whether *FLG* deficiency facilitates unique bacteria colonization. subjects carrying an *FLG* mutation showed a consistently reduced abundance of the rare genus *Chryseobacterium* (Af: 0.06% vs 2.33%, $P_{\text{FDR}} = 8.9 \times 10^{-4}$; Vf: 0.33% vs 2.20%, $P_{\text{FDR}} = .049$; and Ef: 0.51% vs 1.65%, $P = .030$). *Chryseobacterium* is found most commonly in soil, water, and contaminated medical devices,[32] is regarded an opportunistic pathogen, and has not been identified previously as a component of the skin microbiome.

The skin microbiome composition of healthy *FLG*-deficient subjects was more similar to that of patients with AD than healthy *FLG*-competent subjects, in particular at the Af (see Fig E2 in this article's Online Repository at www.jacionline.org).

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

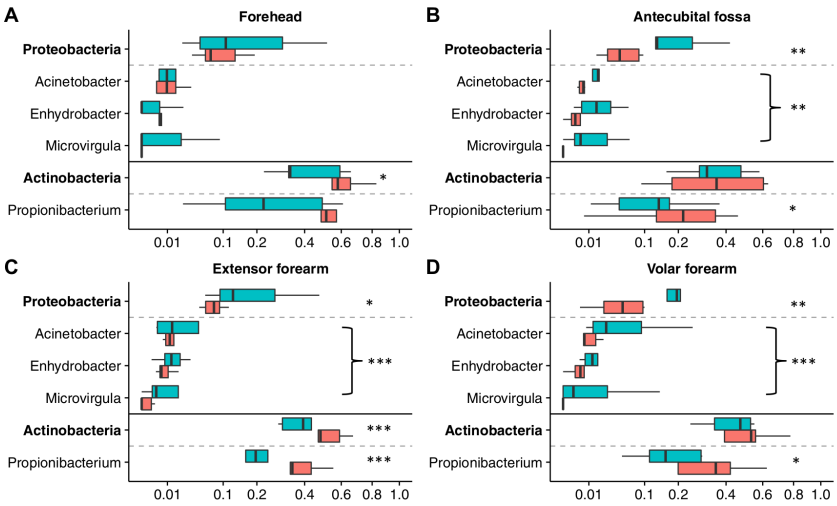


FIG 1. Effect of *FLG* deficiency on the skin microbiome. Proteobacteria and Actinobacteria genera showing differential abundance in *FLG*-competent and *FLG*-deficient subjects at the forehead **(A)**, Af **(B)**, Vf **(C)**, and Ef **(D)**. For better visualization, abundance is depicted on the square root scale. *Red bars* depict *FLG* mutation carriers, and *blue bars* depict non-*FLG* mutation carriers. *Asterisks* indicate significance obtained from the Wilcoxon test: $*P < .05$, $**P < .005$, and $***P < .0005$.

Epidermal lipid composition affects skin microbiome diversity and composition

Integrated analysis revealed a strong association between microbiome and lipidome composition ($P_{FDR} = .015$; Fig 2 and see Fig E3 in this article's Online Repository at www.jacionline.org). More specifically, the abundance of *Propionibacterium* and *Corynebacterium* showed a strongly positive correlation with levels of unsaturated long-chain FFAs, such as FA20:1, FA20:2, FA22:1, and FA24:1, and a negative correlation with saturated short-chain FFAs, such as FA16:0 and FA18:0 (Fig 2), particularly at the Af and Vf (see Fig E4 in this article's Online Repository at www.jacionline.org). Furthermore, at the AD predilection site Af, increasing levels of unsaturated long-chain FFAs correlated with decreased α-diversity (see Fig E5 in this article's Online Repository at www.jacionline.org), whereas increasing levels of α-hydroxy ceramide (AS) correlated with higher abundances of staphylococci ($r = 0.64$, $P_{FDR} = .030$; see Fig E4), in particular *S aureus* ($r = 0.57$, $P = .0088$). Interestingly, at the Af, levels of ceramides (α-hydroxy fatty acid/6-hydroxy-sphingosine base, AS, α-hydroxy fatty acid/phytosphingosine base, and nonhydroxy fatty acid/phytosphingosine base) and several FFAs (FA20:1, FA20:2, FA22:1, and FA24:1) were negatively correlated with abundance of the gram-negative Proteobacteria *Haemophilus* and *Neisseria* species and positively correlated with the abundance of the Actinobacteria and *Propionibacterium* and *Corynebacterium* species (see Fig E4). In line with previous studies,[33] levels of nonhydroxy ceramide subclasses NS (nonhydroxy fatty acid/sphingosine base) and NH (nonhydroxy fatty acid/6-hydroxy-sphingosine base) were altered in patients with AD, but no clear association between FLG deficiency and levels of a specific lipid species was observed (see Fig E6 in this article's Online Repository at www.jacionline.org). Detailed information about the investigated skin lipid profiles species are presented in Table E2 in this article's Online Repository at www.jacionline.org.

Neither TEWL nor skin pH showed clear associations with bacterial community patterns. However, skin pH, in line with previous studies,[34] showed little variation between sites and subjects.

AD displays an increased *Staphylococcus* species abundance and altered *Staphylococcus* species composition across diverse body sites

In both patients with AD and control subjects, the most prevalent *Staphylococcus* species were *S epidermidis and S hominis*, which accounted for up to 32.4% and 7.9% and up to 14.6% and 10.9 of the microbiota%, respectively (Fig 3, *B*). In line with previous studies on healthy subjects,[4,35,36] also in patients with AD, the bacterial composition was strongly dependent on the anatomic region (Fig 3, *A*).

Although diversity and genera richness were reduced at the Af but not at other sites (see Fig E7 in this article's Online Repository at www.jacionline.org), the bacterial configuration in patients with AD was clearly distinct from that of healthy subjects across body sites, including nonaffected and nonpredilection sites. In particular, across all sites, patients with AD showed decreased *Propionibacterium*, *Kocuria*, and *Chryseobacterium* and increased *Lactobacillus* species abundance (see Fig E8 in this article's Online Repository at www.jacionline.org). More specifically, in patients with AD, *Chryseobacterium* and *Kocuria* species were significantly reduced at the forehead ($P = .0210$

	Propioni-bacterium	Coryne-bacterium	Kocuria	Rothia	Staphylo-coccus	Strepto-coccus	Anaero-coccus	Fine-goldia	Acineto-bacter	Haemo-philus	Neisseria	Chryseo-bacterium
AH	0.310	0.436	0.026	-0.291	-0.081	-0.087	0.207	0.322	0.424	-0.086	-0.430	0.634
AS	0.407	0.471	-0.033	-0.154	0.060	0.068	0.391	0.416	0.389	0.118	-0.231	0.387
AP	0.337	0.468	0.019	-0.302	-0.086	-0.084	0.224	0.343	0.452	-0.082	-0.452	0.671
NP	0.179	0.287	0.081	-0.263	-0.063	-0.121	0.126	0.233	0.307	-0.127	-0.350	0.487
FA16:0	-0.436	-0.555	0.067	0.260	0.089	0.010	-0.280	-0.379	-0.498	0.000	0.444	-0.701
FA18:0	-0.429	-0.552	0.054	0.272	0.090	0.023	-0.276	-0.381	-0.500	0.014	0.454	-0.710
FA18:2	0.206	0.095	-0.339	0.320	0.032	0.325	0.101	-0.039	-0.056	0.357	0.265	-0.233
FA20:1	0.586	0.529	-0.490	0.226	-0.023	0.415	0.336	0.226	0.290	0.464	0.021	0.209
FA20:2	0.631	0.622	-0.430	0.104	-0.049	0.343	0.372	0.313	0.404	0.387	-0.128	0.403
FA22:1	0.634	0.573	-0.529	0.244	-0.026	0.449	0.364	0.246	0.315	0.501	0.021	0.228
FA24:1	0.600	0.557	-0.472	0.189	-0.031	0.393	0.347	0.253	0.325	0.441	-0.025	0.268

≥-0.8 | -0.5≤r<-0.8 | -0.3≤r<-0.5 | -0.1≤r<-0.3 | |r|<0.1 | 0.1≤r<0.3 | 0.3≤r<0.5 | 0.5≤r<0.8 | ≥0.8

FIG 2. Bacterial genera showing statistically significant correlations with skin lipid species across body sites. Overall correlation between bacterial abundance and levels of lipid species across all samples was calculated by using sparse canonical correlation analysis with feature selection. *AH,* α-Hydroxy fatty acid/6-hydroxy-sphingosine base; *AS,* α-hydroxy fatty acid/sphingosine base; *AP,* α-hydroxy fatty acid/phytosphingosine base; *NP,* nonhydroxy fatty acid/phytosphingosine base. FFAs are depicted as FAXX:Y, with *XX* indicating chain length and *Y* indicating the number of double bonds/degree of unsaturation.

and $P = .011$), Vf ($P = .0106$ and $P = .0122$), and Ef ($P_{FDR} = .049$ and $P = .0213$). In parallel, the abundance of the rare genus *Lactobacillus* was increased consistently across sites (forehead, $P_{FDR} = .034$; Af, $P = .1273$; Vf, $P = .0371$; and Ef, $P = .1077$; see Table E3 in this article's Online Repository at www.jacionline.org).

Furthermore, patients with AD exhibited a considerably greater staphylococcal abundance (forehead, 27.1% vs 15.5%; Af, 56.7% vs 31.3%; Vf, 24.5% vs 22.3%; and Ef, 21.0% vs 16.8%; see Table E3 and Fig E8). At the predilection site Af, but not at other sites, patients with AD displayed a considerably higher abundance of *S epidermidis* (32.4% vs 13.8%, $P_{FDR} = .049$), whereas the relative abundance of *S hominis* was decreased across all sites. Furthermore, *S aureus* was detected in all patients with AD and on all sites, with the greatest abundance at the Af (8.9%, $P_{FDR} = 1.1 \times 10^{-14}$), whereas it was only detectable in 4 of 10 control subjects with a very low abundance of less than 1%. Analysis of all pairwise ratios for the most prevalent taxa revealed that the AD-associated shift of microbial configuration was clearly driven by *S aureus* ($5.4 \times 10^{-6} < P < .02$), with a 31-, 85-, and 38-fold increase in the *S aureus*/*S epidermidis* ratio at the Af ($P_{FDR} = .015$), Vf ($P_{FDR} = 3.8 \times 10^{-4}$), and Ef ($P_{FDR} = .004$), respectively. Likewise, the *S aureus*/coagulase-negative staphylococci (CoNS) fold change at these 3 sites was 38, 88, and 30 times greater than that in healthy control subjects (Af, $P_{FDR} = .013$; Vf, $P_{FDR} = 3.8 \times 10^{-4}$; and Ef, $P_{FDR} = .008$; see Table E4 in this article's Online Repository at www.jacionline.org). The marked difference in the *S aureus*/CoNS ratio was independent from the *FLG* mutation status.

In a multivariate abundance analysis, AD disease severity was associated significantly with skin microbiome composition at the forehead ($P = .022$) and Ef ($P = .008$). In particular, higher disease severity was associated with increased abundance of the genus *Corynebacterium* (forehead: $r = 0.79$, $P = .0062$; Ef: $r = 0.88$, $P_{FDR} = .011$) and reduced abundance of the Proteobacteria and Acinetobacteria and *Microvirgula* species (forehead: $r = -0.73$, $P = .016$; Ef: $r = -0.81$, $P_{FDR} = .011$) but not with any prevalent *Staphylococcus* species.

A progressive shift of *Staphylococcus* species composition characterizes acute and chronic eczema

Both acute and chronic lesions showed markedly reduced bacterial diversity and genera richness that was independent of body site (see Fig E9 in this article's Online Repository at www.jacionline.org) and showed distinct clusters that were well separated from nonlesional sites in principal coordinate analysis (Fig 4, *A*). Thus the effect of inflammation appears to superimpose locoregional influences on microbiome composition. Furthermore, there was a clear distinction of the microbiome composition between acute and chronic lesions (Fig 4, *A*). Overall, the abundance of staphylococci increased from nonlesional skin samples (32.4%) over acute (46.7%) to chronic (57.3%) lesions, with a significant difference between chronic lesions and nonaffected skin ($P = .014$; Fig 4, *B*). In parallel, *S aureus* abundance showed a gradual and significant ($P_{\text{mixed model}} = .0174$) increase from nonlesional skin samples (5.3%) over acute (23.1%) to chronic lesions (37.1%), whereas the abundance of *S epidermidis* and CoNS gradually decreased (nonlesional, 17.2% and 27.1%; acute, 14.1% and 23.7%; and chronic, 12.3% and 20.2%; see Fig E10, *A-C*, in this article's Online Repository at www.jacionline.org). In particular, there was a significantly ($P_{FDR} = .027$) greater *S aureus* abundance on chronic lesions compared with nonlesional skin (37.1% vs 5.3%). *S aureus*/*S epidermidis* (and *S aureus*/CoNS) ratios increased almost exponentially from 1:5 (1:7) to 1:1 and 4:1 (2:1) at nonlesional sites and acute and chronic lesions ($P_{FDR} = .002$ and $P_{FDR} = .006$, Fig 5). We also observed a gradual decrease of the Actinobacteria and *Corynebacterium* and *Propionibacterium* species from nonlesional skin (8.7% ± 1.5% and 18.8% ± 3.1%) over acute (8.2% ± 2.1%, $P = .31$; 10.8% ± 5.3%, $P = .22$) to chronic lesions (6.3% ± 1.0%, $P = .06$; 10.4% ± 2.9%, $P = .009$; Fig 4, *B*, and see Fig E10, *D* and *E*). Chronic lesions further showed a significant reduction of the phyla Bacteroidetes ($P_{FDR} = .027$) and Fusobacteria ($P = .013$; see Fig E10, *F* and *G*).

Three of the patients with AD reported use of lower-strength to midstrength topical corticosteroids at chronic lesional skin within 5 days of sampling. To exclude a potential bias, we carried out a

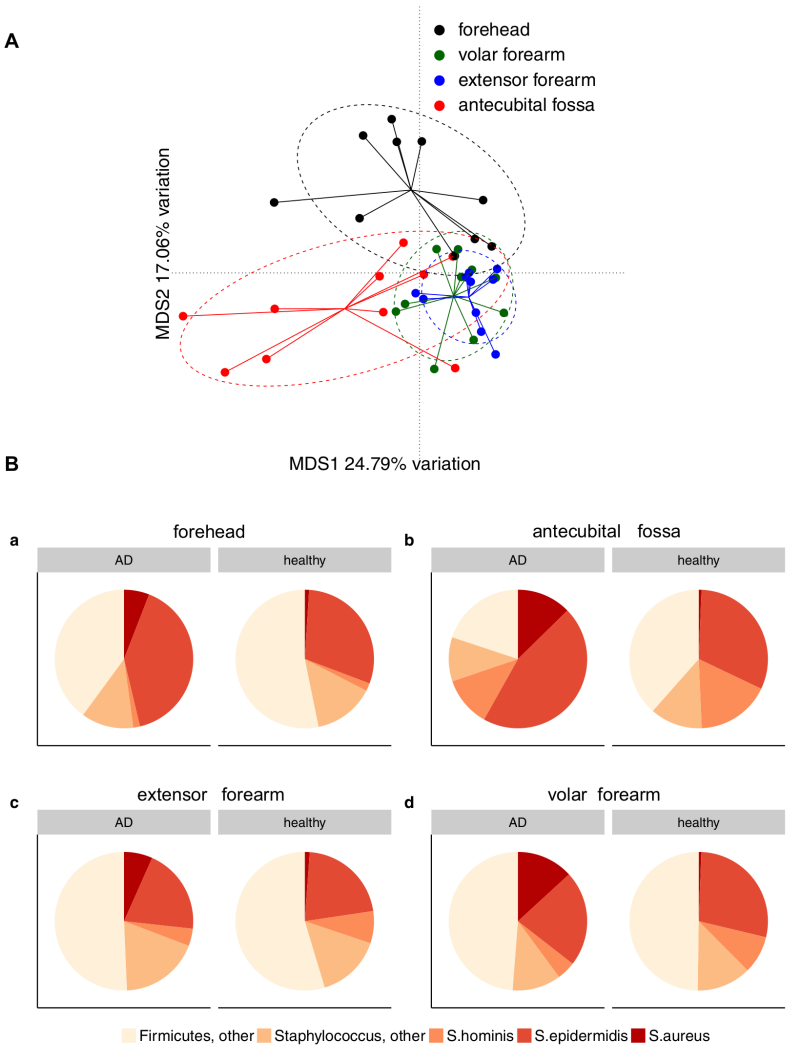


FIG 3. Site specificity of bacterial community composition in patients with AD. **A,** Based on the square root of the Bray-Curtis dissimilarities in principal coordinate analysis, the Vf and Ef clustered together but were distinct from the Af and forehead. **B,** The abundance of the most prevalent Firmicutes genera is strongly dependent on body site and shows significant differences between patients with AD and healthy control subjects.

sensitivity analysis using the Welch test, which showed no evidence of a significant effect of this treatment on diversity and microbial composition.

DISCUSSION

Although it has long been known that patients with AD show an increased cutaneous infectivity, that AD lesions are typically colonized with *S aureus*,[37] and that *S aureus*–derived toxins have the capacity to exacerbate disease activity,[38] the role of microbes at large and their interplay with skin physiology in the pathogenesis of AD is still insufficiently understood. To gain deeper insight into the composition of skin microbiome in patients with AD and its interrelation with structural and physiochemical barrier functions, we examined a cohort of carefully matched and

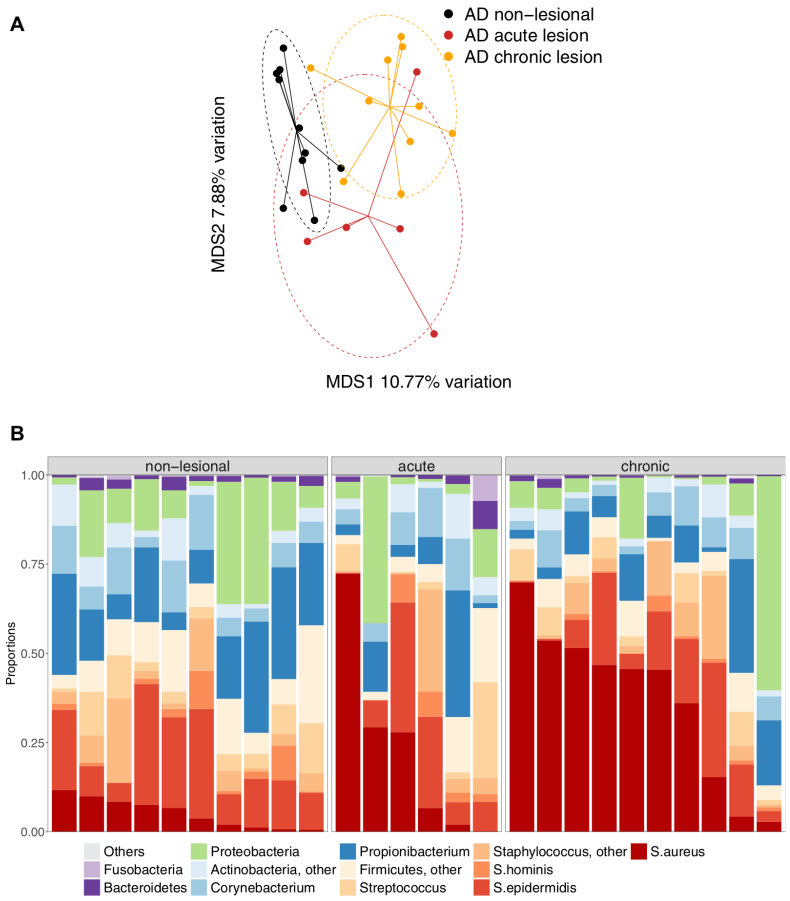


FIG 4. Affection by AD strongly affects bacterial composition. **A,** Principal component analysis showed distinct clustering of samples from nonlesional skin, acute AD lesions, and chronic AD lesions, irrespective of body site. **B,** Affection of skin by acute or chronic AD correlates with marked differences of the relative abundance of major phyla, genera, and *Staphylococcus* species, with the most pronounced shifts found on chronic lesions. The bar chart shows abundance at the individual level. *Non-lesional* depicts the average of the nonlesional samples from the 4 body sites: forehead, Af, Vf, and Ef.

comprehensively characterized patients with AD and healthy control subjects using 16S rRNA skin samples from different body sites, as well as from acute and chronic lesions.

For the first time, our study provides direct evidence for a strong effect of epidermal lipids on bacterial colonization, which presumably contributes to the marked site specificities of the skin microbiome. In particular, epidermal lipid composition correlated strongly with bacterial diversity and composition at AD predilection sites. Higher levels of long-chain unsaturated fatty acids were associated significantly with increased abundance of the lipophilic Propionibacteria and Corynebacteria, which lack a fatty acid synthase[39] and require an exogenic source of FFAs.[35,39,40] Furthermore, *S aureus* abundance was associated with higher levels of the ceramide AS, which was previously shown to be increased in AD skin, particularly in lesions.[33] In contrast, TEWL, which is often used as a proxy for epidermal barrier function, is not correlated directly with characteristics of the skin microbiome. In a previous study only correlation of total sebum from the cheeks with microbial composition and diversity was reported, which was investigated in healthy women from the Whitefield area of Bangalore city (India).[41]

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

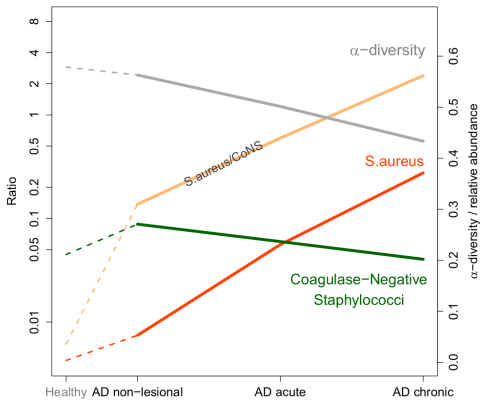


FIG 5. Hypothesized progression of bacterial composition from nonlesional to acute and chronic lesional skin in patients with AD. The *Staphylococcus* species line depicts *Staphylococcus* ratios, with the log-scaled *y-axis* on the *left-hand side*. *Red* and *green lines* indicate the *S aureus* and CoNS abundances with the corresponding *y-axis* on the *right-hand side*. The *gray line* shows the normalized α-diversity with the corresponding *y-axis* on the *right-hand side*. *Dotted lines* depict connections to the respective mean value of healthy subjects as a reference for interpretation.

Furthermore, our data show that an inherited deficiency of the epidermal structural protein FLG is associated with marked shifts in the composition of microbial communities, regardless of body site. FLG deficiency caused by inheritance of a single *FLG* mutation affects as much as 10% of the general population, which have generalized skin dryness caused by impaired formation of the stratum corneum and are at a greatly increased risk of AD.[42]

Across body sites, FLG-deficient skin showed a reduced diversity and richness, increased abundance of Firmicutes, and decreased abundance of Proteobacterium, in particular the genus *Acinetobacter*, which is capable of suppressing the cytokine response to *S aureus*,[43] the signature microbe of AD. Furthermore, it showed an increased abundance of Actinobacteria, in particular *Propionibacterium* species, which have been reported to be underrepresented in patients with psoriasis,[44] another common inflammatory skin disease with largely opposing immune deviations compared with patients with AD.[45]

Although most pronounced at disease predilection sites, in patients with AD, we observed an increased relative abundance of staphylococci, in particular *S aureus*, and an altered *Staphylococcus* species composition across nonaffected body sites, indicating that this is a general feature of AD not restricted to sites affected by eczema. AD is characterized by generalized abnormalities of skin barrier function with broad keratinocyte differentiation defects, inadequate regulation of antimicrobial peptide expression, compromised innate sensing, and considerable immune activation,[46,47] which can contribute to such alterations of cutaneous colonization patterns. However, although these AD-related changes can drive microbial alterations, altered microbes might in turn play a role in disease-related immunologic changes[48] or likely both. Future studies with larger cohorts and longitudinal collection of samples at preclinical stages of disease will be required to better understand mutual influences.

A previous study reported an increase in the proportion of *Staphylococcus* species sequences, particularly *S aureus*, during AD flares and with increased general AD disease severity.[12] However, this observation was made only for predilection sites in children with moderate-to-severe AD (ie, at sites repeatedly affected by eczema). Although we found a strongly increased *Staphylococcus* species abundance at eczematous sites and, in particular, on chronic lesions, we did not observe associations between staphylococcal abundance and disease severity. However, across sites, we observed an association between AD severity and abundance of the genus *Corynebacterium* and the phylum Proteobacteria. Considerably more striking was the observation that eczematous skin changes appear to have a dramatically stronger influence on bacterial diversity and composition than, for example, the body site, underlining the important role of cutaneous immunity on microbial communities,[49] and that the alterations of bacterial community structures are more pronounced on chronic than acute lesions.

In conclusion, our data show that the molecular composition and function of the epidermis affect skin microbiome composition; that AD shows characteristic alterations, particularly regarding staphylococcal abundance and composition at predilection sites; and that the cutaneous inflammatory milieu and its acuity in eczema lesions superimpose site specificities.

However, it is important to note that the resulting sample size was relatively small and that we conducted many analyses because of the population-based and balanced design. Therefore *P* values need to be interpreted cautiously, and it is premature to draw strong conclusions. However, even at this small sample size, we were still able to identify significant microbiome differences between groups and disease states and identify differential abundant taxa after multiple testing corrections. Furthermore, we report correlations, and cause or effect still needs to be delineated. One approach to work out the effect of specific skin pathologies on the skin microbiota and *vice versa* and to gather evidence supporting causation will be to examine other types of eczema and inflammatory skin diseases.

Key messages

- **Epidermal lipid composition affects bacterial community configuration.**
- **FLG deficiency correlates with decreased bacterial diversity and distinct bacterial colonization pattern.**
- **AD shows an altered skin microbiome configuration also at nonaffected and nonpredilection sites.**
- **The presence and chronicity of eczema appear to be more important determinants of skin microbiome configuration than body site.**

REFERENCES

1. Chuong CM, Dhouailly D, Gilmore S, Forest L, Shelley WB, Stenn KS, et al. What is the biological basis of pattern formation of skin lesions? Exp Dermatol 2006;15:547-64.
2. Kottner J, Lichterfeld A, Blume-Peytavi U. Transepidermal water loss in young and aged healthy humans: a systematic review and meta-analysis. Arch Dermatol Res 2013;305:315-23.
3. Findley K, Oh J, Yang J, Conlan S, Deming C, Meyer JA, et al. Topographic diversity of fungal and bacterial communities in human skin. Nature 2013;498:367-70.

4. Oh J, Byrd AL, Deming C, Conlan S, Program NCS, Kong HH, et al. Biogeography and individuality shape function in the human skin metagenome. Nature 2014;514:59-64.
5. 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.
6. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, et al. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A 2015;112:E2120-9.
7. Hirasawa Y, Takai T, Nakamura T, Mitsuishi K, Gunawan H, Suto H, et al. Staphylococcus aureus extracellular protease causes epidermal barrier dysfunction. J Invest Dermatol 2010;130:614-7.
8. Belkaid Y, Tamoutounour S. The influence of skin microorganisms on cutaneous immunity. Nat Rev Immunol 2016;16:353-66.
9. Weidinger S, Novak N. Atopic dermatitis. Lancet 2016;387:1109-22.
10. Harder J, Dressel S, Wittersheim M, Cordes J, Meyer-Hoffert U, Mrowietz U, et al. Enhanced expression and secretion of antimicrobial peptides in atopic dermatitis and after superficial skin injury. J Invest Dermatol 2010;130:1355-64.
11. Leyden JJ, Marples RR, Kligman AM. Staphylococcus aureus in the lesions of atopic dermatitis. Br J Dermatol 1974;90:525-30.
12. Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome Res 2012;22:850-9.
13. Vavrova K, Henkes D, Struver K, Sochorova M, Skolova B, Witting MY, et al. Filaggrin deficiency leads to impaired lipid profile and altered acidification pathways in a 3D skin construct. J Invest Dermatol 2014;134:746-53.
14. Miajlovic H, Fallon PG, Irvine AD, Foster TJ. Effect of filaggrin breakdown products on growth of and protein expression by Staphylococcus aureus. J Allergy Clin Immunol 2010;126:1184-90.e3.
15. Zeeuwen PL, Ederveen TH, van der Krieken DA, Niehues H, Boekhorst J, Kezic S, et al. Gram-positive anaerobe cocci are underrepresented in the microbiome of filaggrin-deficient human skin. J Allergy Clin Immunol 2017;139:1368-71.
16. Nothlings U, Krawczak M. [PopGen. A population-based biobank with prospective follow-up of a control group]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2012;55:831-5.
17. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat Genet 2015;47:1449-56.
18. Eichenfield LF. Consensus guidelines in diagnosis and treatment of atopic dermatitis. Allergy 2004;59(suppl 78):86-92.
19. Zeeuwen PL, Boekhorst J, van den Bogaard EH, de Koning HD, van de Kerkhof PM, Saulnier DM, et al. Microbiome dynamics of human epidermis following skin barrier disruption. Genome Biol 2012;13:R101.
20. Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, et al. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J 2013;27:1012-22.
21. Angelova-Fischer I, Becker V, Fischer TW, Zillikens D, Wigger-Alberti W, Kezic S. Tandem repeated irritation in aged skin induces distinct barrier perturbation and cytokine profile in vivo. Br J Dermatol 2012;167:787-93.
22. Wang J, Thingholm LB, Skiecevicience J, Rausch P, Kummen M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016;48:1396-406.
23. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584.
24. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;27:2194-200.
25. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013;10:996-8.
26. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72:5069-72.
27. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies. Int J Syst Evol Microbiol 2017;67:1613-7.
28. R Core Team. R: a language and environment for statistical computing. 2016. Available at: https://www.R-project.org/.
29. Zeileis A, Kleiber C, Jackman S. Regression models for count data. R J Stat Software 2008;27.
30. Witten D, Tibshirani R, Gross S, Narasimhan B. PMA: penalized multivariate analysis. 2013. R package version 1.0.9. Available at: https://CRAN.R-project.org/package=PMA.
31. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009;10:515-34.
32. Bloch KC, Nadarajah R, Jacobs R. Chryseobacterium meningosepticum: an emerging pathogen among immunocompromised adults. Report of 6 cases and literature review. Medicine (Baltimore) 1997;76:30-41.
33. van Smeden J, Bouwstra JA. Stratum Corneum lipids: their role for the skin barrier function in healthy subjects and atopic dermatitis patients. Curr Probl Dermatol 2016;49:8-26.
34. Schreml S, Zeller V, Meier RJ, Korting HC, Behm B, Landthaler M, et al. Impact of age and body site on adult female skin surface pH. Dermatology 2012;224:66-71.
35. Grice EA, Segre JA. The skin microbiome. Nat Rev Microbiol 2011;9:244-53.
36. Kennedy EA, Connolly J, Hourihane JO, Fallon PG, McLean WH, Murray D, et al. Skin microbiome before development of atopic dermatitis: early colonization with commensal staphylococci at 2 months is associated with a lower risk of atopic dermatitis at 1 year. J Allergy Clin Immunol 2017;139:166-72.
37. Williams RE, Gibson AG, Aitchison TC, Lever R, Mackie RM. Assessment of a contact-plate sampling technique and subsequent quantitative bacterial studies in atopic dermatitis. Br J Dermatol 1990;123:493-501.
38. Leung DY, Harbeck R, Bina P, Reiser RF, Yang E, Norris DA, et al. Presence of IgE antibodies to staphylococcal exotoxins on the skin of patients with atopic dermatitis. Evidence for a new group of allergens. J Clin Invest 1993;92:1374-80.
39. Bernard K. The genus Corynebacterium and other medically relevant coryneform-like bacteria. J Clin Microbiol 2012;50:3152-8.
40. Bomar L, Brugger SD, Yost BH, Davies SS, Lemon KP. Corynebacterium accolens releases antipneumococcal free fatty acids from human nostril and skin surface triacylglycerols. MBio 2016;7:e01725-35.
41. Mukherjee S, Mitra R, Maitra A, Gupta S, Kumaran S, Chakrabortty A, et al. Sebum and hydration levels in specific regions of human face significantly predict the nature and diversity of facial skin microbiome. Sci Rep 2016;6:36062.
42. Irvine AD, McLean WH, Leung DY. Filaggrin mutations associated with skin and allergic diseases. N Engl J Med 2011;365:1315-27.
43. Smeekens SP, Huttenhower C, Riza A, van de Veerdonk FL, Zeeuwen PL, Schalkwijk J, et al. Skin microbiome imbalance in patients with STAT1/STAT3 defects impairs innate host defense responses. J Innate Immun 2014;6:253-62.
44. Gao Z, Tseng CH, Strober BE, Pei Z, Blaser MJ. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS One 2008;3:e2719.
45. Eyerich S, Onken AT, Weidinger S, Franke A, Nasorri F, Pennino D, et al. Mutual antagonism of T cells causing psoriasis and atopic eczema. N Engl J Med 2011;365:231-8.
46. Suarez-Farinas M, Tintle SJ, Shemer A, Chiricozzi A, Nograles K, Cardinale I, et al. Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities. J Allergy Clin Immunol 2011;127:954-64, e1-4.
47. Nakatsuji T, Chen TH, Narala S, Chun KA, Two AM, Yun T, et al. Antimicrobials from human skin commensal bacteria protect against Staphylococcus aureus and are deficient in atopic dermatitis. Sci Transl Med 2017;9.
48. Naik S, Bouladoux N, Linehan JL, Han SJ, Harrison OJ, Wilhelm C, et al. Commensal-dendritic-cell interaction specifies a unique protective skin immune signature. Nature 2015;520:104-8.
49. Naik S, Bouladoux N, Wilhelm C, Molloy MJ, Salcedo R, Kastenmuller W, et al. Compartmentalized control of skin immunity by resident commensals. Science 2012;337:1115-9.

5 Discussion

The presented studies cover different aspects to consider in the analysis of microbiome studies. This chapter summarizes the main findings of each of the previous chapters and aims to put additional emphasis on aspects only briefly covered in the original publications. The chapter concludes with an overarching discussion of the insights gained from the individual articles, also giving an outlook on future challenges in the research of host-associated microbial communities.

For a non-redundant naming, figures and tables from the articles are referred to as the articles A-G, followed by the number of the item in the respective original publication, *i.e.* Figure 1 in article A is referred to as Figure A.1.

5.1 Extended discussion of results chapters

5.1.1 Technological influence on the survey of microbial communities

The study presented in **Article A** intended to address two main ideas with regard to the effort in the 'Collaborative Research Centre 1182': (1) How do the established lab protocols for the survey of host-associated microbial communities perform across the studied hosts? And (2), what information about the 'bigger picture' in host-microbe interactions can be generated from this unique set of samples?

The second question was discussed extensively in the article. The findings suggest that host-associated microbial communities changed drastically when hosts started to transform from aquatic to terrestrial lifestyle around 425-500 million years ago. However, this change in the microbial communities is not necessarily connected to the functional repertoire of the members of the communities, but rather happened by the introduction of more land-adapted bacterial lineages while – at community level – conserving the original functional capacity. The most notable functional changes were seen in association with carbohydrate breakdown, indicating a strong adaptation of microbes to the hosts' food sources, but could also be a by-product of evolved host-specific glycosylation patterns important for host-microbiome interactions and shaping of the associated microbiota [208], [209].

Coming back to the assessment of performance of survey strategies for microbial communities, the article led to the conclusion that the use of a one-step PCR of the V3-V4 16S rRNA gene amplicon generally is a good choice, however, host-specific and especially community-specific drawbacks apply. Focusing on human fecal samples, as they are the source material to all additional studies presented in this thesis, no systematic differences between the amplicon-based strategies

could be identified (all P>0.05; Figure A.2). Also, in beta diversity, all human samples clustered together in the ordination, indicating a clearly defined, strategy-independent community composition identified by all methods (Figure A.3). Using amplicon-data to infer the functional repertoire of microbial communities was shown to be strongly biased by the host (Figure A.5). While identification of closely related proxy organism for estimation of metabolic capacity showed to be slightly easier for V3-V4-based amplicons, none of the inferred functional profiles could satisfyingly predict the actual profiles generated by shotgun metagenomic sequencing (Figure A.5). This is possibly due to strain- and species-level functional differences in microbial taxa [169] and still unresolved strategies for the correction of 16S rRNA gene copy number variations among closely related taxa, resulting in biased estimates [210].

Taken together, the results of this study showed that 16S rRNA gene amplicon-based strategies for the analysis of host-associated microbial communities are - also in the age of larger-scale studies using shotgun metagenomics - still highly valuable, as they provide a comparably cheap and still accurate and robust estimation of microbial community structure and composition. However, amplicon-based strategies cannot be used for a reliable estimation of functional capacities of these communities. While they might serve as proxies for the generation of hypotheses to test, a clear statement about the function of a microbial community can only be guided by shotgun metagenomics, even better when accompanied by culturing of functionally important candidate taxa to also experimentally show the presence of a specific function.

The technical aspects covered in **Article A** should be considered as part of ongoing debates in the field of microbiome analysis regarding the installment of best practices for the survey and data processing of microbial communities. As already discussed in section 1.4.2, parallel efforts in the development of software suites for the analysis of amplicon-based sequencing data (see QIIME, [161], [162] mothur, [160]) have led to a landscape of tools to use in the analysis, all with advantages and drawbacks, however ultimately performing comparable [211]. Similar trends have been shown for shotgun metagenomics, however adding additional layers of complexity, including – but not limited to – sampling strategy, sample storage and choice of extraction kits (also affecting amplicon-based approaches; [212]–[214]), and finally, the choice of downstream processing of data, leaving the analyst with hundreds of possible strategies to follow, many of them targeting specific aspects, making it impossible to create a 'gold standard' or 'one-fits-all' workflows (see Sczyrba *et al.* [215]).

5.1.2 The influence of host genetics on the human intestinal microbiome

The articles **Article B** and **Article C** focused on the discovery of host-genetic associations with microbial traits. The introduction of the permutation-free distance-based F test through application of moment matching introduced in **Article B** and also applied to the larger dataset in

Article C massively reduced computation time for beta diversity-aimed GWAS. This increase in calculation speed was systematically benchmarked for the cohorts included in **Article C**, in addition also including calculations accelerated by the use of a NVIDIA Tesla P100 graphics processing unit (GPU) and is part of the articles supplementary material, reproduced in Figure 5.1. Briefly, the number of variants analyzed followed a strictly linear time complexity, independent of the cohort. Increase in cohort size (N), however, clearly led to a quadratic increase in computational time, due to the N×N-sized matrix multiplications in the approximation calculations. The speed of matrix calculations can be highly increased when using a GPU instead of CPUs, especially in the case of large matrices. This can be seen by a 10-fold increase in calculation speed for the PopGen cohort (N=724) and an up to 266-fold speed-up for the SHIP-Trend cohort (N=3,382) compared to the use of a single CPU, reducing the calculation time for a single variant in the SHIP-Trend cohort from between 44-52 minutes to only 11.6 seconds. Taken together with the speed-up compared to permutation-based approaches – which would take more time by several orders of magnitudes – the introduction of the permutation-free distance-based F test for the first time enabled genome-wide association analysis of beta diversities.

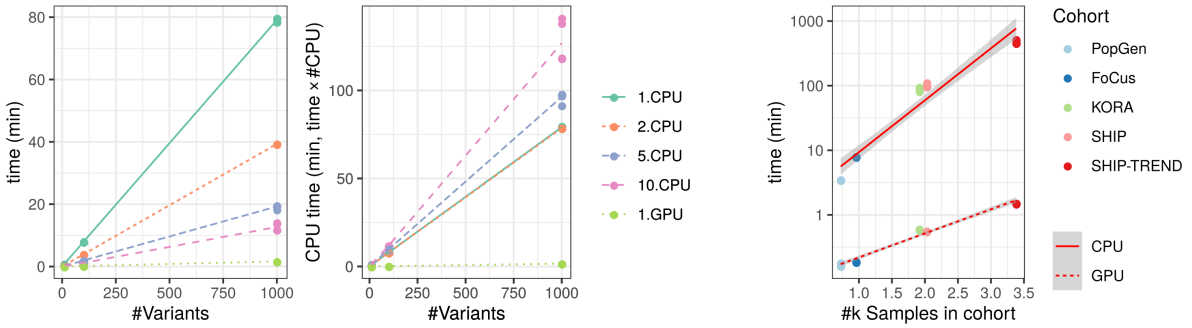


Figure 5.1: Summary of CPU- and GPU-based calculations for beta-diversity analysis. (A) Relationship of elapsed time and the number of analyzed variants in the Focus cohort depending on the number of CPUs and GPUs used in parallel. (B) Relationship of computational time in the Focus cohort depending on the number of CPUs and GPUs used in parallel. (C) Relationship of calculation time and cohort size depending on the usage of a single CPU or GPU for analysis. Reproduced from the supplementary material of **Article C**.

Among the 32 independent genetic loci identified in all three branches of the association analysis in **Article C**, focusing on beta diversity, as well as on feature abundance and presence-absence-patterns, respectively, one of the most interesting loci found to be affecting the gut microbiom is the histo-blood group ABO system transferase (*ABO*) gene locus. This result was already briefly discussed in the original manuscript, however due to space constraints of the selected publication format, in in-depth discussion of this intriguing finding was not possible, thus should now be performed here.

Depending on a set of genetic variants (see Chester *et al.* [216]), the *ABO* gene encodes for glycosyltransferases which either add N-acetylgalactosamine or galactose to a fucosylated oligosaccharide on the surface of red blood cells (RBCs; fucosylation by FUT1) or

secreted/mucosal proteins (fucosylation by FUT2), creating the *A* and *B* histo-blood group antigen, respectively. Non-functioning variants of ABO do not add any final sugar to the precursor, resulting in the *O*-antigen (see Figure C.3, panel C for reference). Multiple genetic polymorphisms encoding the production of the different antigens have been maintained in the human population – and apparently already since tens of millions of years in ancestral populations [217] – by a process called 'balancing selection'. Compared to directional selection, which favors a specific phenotypic variant in an environment, ultimately causing it to replace other variants by a so-called 'selective sweep', balancing selection leads to an active maintenance of multiple phenotypes/alleles in the gene pool of a population [218]. The multiple reasons for the maintenance of polymorphisms were reviewed by Llaurens *et al.* [218], one of them being 'heterozygote advantage', meaning that combinations of different alleles possibly lead to higher fitness than homozygotes of the respective alleles, and a second one being 'frequency-dependent selection', implying population-level fluctuations connected to an external co-fluctuating selective pressure, *e.g.* a predator-prey- or host-pathogen-relationship. A prominent example where balancing selection is thought to play a role are variations in the major histocompatibility complex (*MHC*) gene loci, as MHC heterogeneity increases antigen binding diversity, leading to broadened capacity in recognition of potential pathogens [219]. The MHC was shown to play a role in the community assembly of the microbiome in mice with respect to pathogen susceptibility [220], suggesting that adaptation to pathogens and recognition of commensals guided the diversification of the MHC. Similar effects are thought to be the basis of the persisting variation in the ABO histo-blood groups [221].

The results from the GWAS, identifying the *ABO* locus as modulator of the abundance of specific *Bacteroides* subgroups (OTU97_27) in combination with the *FUT2*-dependent secretor status, are possibly confirming this long-standing relationship of host and microbiota and once again highlight the importance of glycosylation-patterns for host-microbe interaction, as discussed in section 5.1.1. Especially for *Bacteroides* it was show that they possess large varieties of host-glycoprotein degrading enzymes [222], [223]. *Bacteroides* show to be potentially interesting for in-depth evaluation in connection to host-adaptation. As discussed in section 1.2.3, the closely related genera *Prevotella* and *Bacteroides* exhibit population-level differences in their distributions, likely connected to a westernization of lifestyle. However, also frequencies of ABO blood groups differ markedly across global populations. While in Amerindians exclusively the *O* histo-blood group is present [224], for Malawians blood groups were distributed 21.7/25.7/3.3/49.3% (A/B/AB/O) [225] and a frequencies of 39.7/10.9/4.1/45.2% were found for white non-Hispanic US americans [226]. Whether these distributions influence microbiome composition and *Prevotella/Bacteroides* abundances on a global and/or population level needs to be investigated, however the presented results suggest a contribution of the ABO histo-blood group on members of the intestinal microbial community.

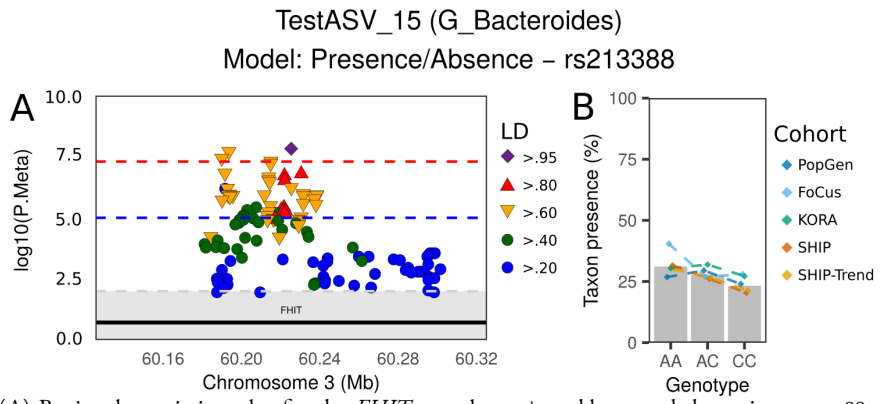


Figure 5.2 (A) Regional association plot for the *FHIT* gene locus ± 100kb around the variant rs213388 and the results of the logistic regression analysis of presence-absence-patterns of *Bacteroides* TestASV_15. (B) Differences in TestASV_15 presence in five German cohorts in dependence of the genotype at variant rs213388.

Alleles of the ABO histo-blood group system have been shown to be important for host health and variation in the *ABO* locus were shown to be associated to multiple complex traits, such as type 2 diabetes [227], hematological cell composition [228], and blood lipids [229], clearly linking this locus to metabolic syndrome and inflammation. The impact of host-microbe interactions on the genetic architecture of inflammatory diseases has been proposed previously [230], implicating possible widespread effects of microbe-associated selective pressure on host health. Intersecting the results of **Article C** (Table C.1) and a scan for signals of long-term balancing selection [231] reveals the *fragile histidine triad* (*FHIT*) gene (chr3:59,747,277-61,251,459) encoding for the Bis-(5'-adenosyl)-triphosphatase as potential target of microbe-driven selection. The GWAS identified the prevalence of again another feature belonging to *Bacteroides* (TestASV_15) to be associated with variants in the *FHIT* gene locus, a locus also described as one of the fastest evolving loci in the human genome (HAR10; human accelerated region 10)[232] since the split from the evolutionary lineage shared with chimpanzees, suggesting a connection to human speciation. Previous GWASes identified variants in *FHIT* to be connected to body mass index [233] and psychiatric traits, schizophrenia [234] and autism spectrum disorder [235].

Genome-wide association analysis of members of the microbiota still show extensive heterogeneity, with only very few loci replicating between studies. This might be explained by several influencing factors. As explained earlier, amplicon-based analyses can only serve as proxy for functional units and differences in library preparation could influence the assessment and identification of microbial features. Additionally, the presence of a microbial features - a taxon, ASV or OTU - is no proof for the presence of a function usually performed by this bacterium. Bacteria are able to quickly adapt to a changed environment [236], and each host individual and the interaction with other microorganisms in the community presents a unique surrounding, with for example ecological niches possibly taken by other bacteria or not presented by the host, due to differing dietary patterns. Further, presence and absence of microbial features can be hard to

estimate from sequencing data, as they might simply be low abundant and missed by stochastic sampling processes (random zeros) or, especially in the case of fecal samples, are not represented at all in the sampled material (structural zeros), as they colonize for example the mucus or the small intestine and are not transported in large amounts in fecal material [237]. Also the compositional nature of sequencing datasets could lead to inaccurate estimates of abundances [194]. Future analyses of host-genetic effects on the microbiome should address these issues by large-scale metagenomic sequencing to assess functional diversity within closely related organisms. Although more complicated to acquire, also mucosal biopsies or brushings from multiple parts of the intestine should be taken for an extensive analysis of spatial interactions between host-genetics and microbes. Finally, overall estimates of microbial load should be assessed and only targeted approaches for candidate taxa can ultimately demonstrate true absence or presence of features, for example by culturing and/or feature-specific quantification.

5.1.3 The human microbiota in disease

The **Articles D-G** describe disease-related changes in different sub-entities of the human-associated microbiome. **Article E** confirms and expands on the findings of **Article D**, indicating extensive changes of the intestinal bacterial composition connected to the cholestatic liver disease primary sclerosing cholangitis (PSC). The use of inverse-variance weighted meta-analysis and stringent criteria used to assess statistical significance in **Article E** enabled to find robust changes connected to PSC and ulcerative colitis, independent of the cohorts' origins. Regional variation in the microbiome can be large [238], however, the use of cross-regional approaches analyzing 16S rRNA gene-based amplicon data can exactly identify those microbial features that are directly connected to disease, and not influenced by regional differences in lifestyle, diet, and/or genetics and also technical factors (*e.g.* sampling strategies and DNA extraction kits) potentially leading to shifts in the microbiome composition. This is important as the assumption of a contribution of the microbiota to disease will likely be connected to specific taxonomic groups across all regions, where the disease is prevalent. However, the microbiota's modulation of health and disease might also be connected to microbial metabolic functions. If these functions are performed by always specific taxonomic groups, they will possibly be picked up by the analysis of amplicon sequencing data. However, as discussed already briefly in the previous sections, the use of taxonomic bins or ASVs/OTUs for the clustering of metabolically similar microbial features can only serve as a rough proxy for the bacteria's functional capacities. Analyses showed that the functional capacity of a healthy human microbiota can be very stable across taxonomically very diverse compositions [18], indicating large amounts of functional redundancy across taxonomic groups. In reverse, this might also mean that microbial metabolic functions connected to disease can be exerted by diverse taxonomic groups. Algorithms for the imputation of functional capacity from amplicon data could possibly reveal such metabolic associations, however only, when these are highly conserved in

the respective organisms. As these functions are likely strain specific [94] and horizontal gene transfer is common within host-associated microbial communities [239], these approaches are possibly over-simplistic and thus not suitable for the identification of disease associated metabolic functions, as also shown by the large divergence between real and imputed functional capacity in **Article A** and discussed in section 5.1.1. For the reliable identification of such disease causing or modulating metabolic pathways, again shotgun metagenomic sequencing data might be a valuable source for future studies.

The results of **Article F** confirm previously described disease-associated changes in the fungal microbiota in connection with PSC [240]. Even though the human fecal mycobiota is expected to be a lot less diverse than the assembly of prokaryotes [39], especially in IBD it is thought to play a pivotal role in disease pathogenesis [241]. The assessment of fungal communities has proven to be highly influenced by DNA extraction protocols [242]. In addition, only very low amounts of the total reads in shotgun metagenomic datasets can be identified as fungal sequences (1.2×10^6 out of 2.5×10^9 sequences = 0.048%; [243]). The low read count in combination with genomes larger than those of bacteria (*Candida albicans:* 14.86 Mb [244]; *Saccharomyces cerevisae:* 12.1 Mb [245]; Mean bacteria: 3.65 Mb [246]) make it challenging to gain robust estimates of fungal relative abundances [247], thus PCR-based amplification for the survey of mycobiome community compositions are needed. Especially here, long read sequencing technologies will likely prove valuable, as amplicons used in mycobiome analysis vary considerably in length across taxonomic groups [248]. Even though still little is known about stability of fungal communities in the healthy human intestine [249], the investigation of the mycobiome opens the possibility of gaining additional insights about cross-kingdom interactions within microbial communities potentially modulating health and disease. While the insights from mycobiome analysis remain descriptive (**Article F**, **[240], [241]**), experimental studies could show the influence of intestinal fungus *C. albicans* on Th17-mediated response to stimuli elsewhere in the body [250], highlighting once again the importance of the assessment of not only changes in the directly affected organs and body-sites, but rather a systemic view on the metaorganism.

The importance of this view is confirmed for the chronic inflammatory skin disease atopic dermatitis (AD). **Article G** demonstrated that not only the acutely and previously affected skin sites of AD patients exhibit compositional differences compared to healthy skin regions in the same individuals (Figure G.4), but also that bacterial communities of non-affected skin in AD patients differs from that of healthy individuals (Figure G.3). AD and Psoriasis, two common inflammatory skin diseases [251], both share genetic risk factors with inflammatory bowel disease and are known comorbidities of IBD [123], [252], and both diseases are suspected to also be associated with marked changes in the intestinal microbiota [253], [254]. Whether these changes in the intestinal microbiota are cause or consequence of potential systemic inflammatory processes

manifesting in skin lesions is yet to be clarified, however treatment options of extra-intestinal diseases via modulation of the intestinal microbiota could potentially be of interest, and in this regard, also the development of orally administered probiotics is discussed [255].

The **Articles D-G** could show that the assessment of disease-associated changes in the host's microbiota should not only rely on simple case-control comparisons within a single study population, as the analysis of high dimensional data is prone to the identification of false positive associations. The usage of an independent cohort from Oslo, Norway, in the **Articles D and E** in combination with the standardization of the sequencing library protocols lead to the identification of novel PSC-associated changes in the gut microbiota and the confirmation of previously found differential abundances of microbial features in PSC and UC. Using an approach complementary to the one used by Lemoinne *et al. [240]* for the assessment of the intestinal mycobiota, **Article F** shifts the focus away from the PSC-associated increased abundance of the genus *Exophiala* – a signal that was driven by only five samples – towards the replicating disease association of the genus *Trichocladium/Humicola,* which was also identified to be increased in abundance in the French PSC patients (only classified on class level as Sordariomycetes), however not discussed further in the original article. The inclusion of non-affected skin sites as direct controls in **Article G** should also be considered for the assessment of disease-associated changes in the intestinal microbiota. Though fecal samples are easy to obtain from large collectives as they do not require invasive sampling, using more samples acquired by biopsies or brushings of the mucosal surfaces will help to increase spatial resolution of microbial changes in inflammatory bowel diseases and are less prone to biases introduced by self-sampling procedures and transportation times. Here, also the recruitment of healthy controls for endoscopic sampling is crucial for placing the changes in the bigger picture, as **Article G** could demonstrate that also unaffected sampling sites exhibit systematic disease-associated differences in the microbial composition.

5.2 Concluding discussion and future challenges

The technical advances in recent years had great impact on the understanding of assembly and functioning of host-associated microbial communities. The **Articles A-G** presented in this thesis deliver new insights into multiple aspects of the way we see and the way we can analyze microbial communities using nucleotide sequencing approaches. Amplicon-based approaches for the survey of microbiomes have been proven to be a useful tool to investigate different parts of the human microbiota, however major challenges still remain.

Recent insights from the systematic analysis of databases of 16S rRNA gene amplicon sequences have led to a shift from 97%-identity OTUs to a focus on exact amplicon sequence variants [169] for the assessment of biologically meaningful species-level entities. However, *in silico* analysis of 16S rRNA sequences showed that also within species, and even between different gene copies within a single genome, sequence variation can be as large as 3% [256]. These properties make it complicated to find a sensible balance between covering accuracy and true biological entities. For alpha and beta diversity, measures were developed that incorporate phylogenetic trees into diversity calculations (Faith's phylogenetic diversity [185]; UniFrac [188]), however in the analysis of features differentially abundant between groups of samples, such methods have not yet reached wide acclaim (see for example the use of balance trees [257] or the PAAT (phylogeny-aware abundance testing) tool: https://github.com/mruehlemann/paat), although they might bridge the gap between coarse (and potentially inaccurate [258]) taxonomic annotations of short amplicons and fine-scale sequence variants. Inaccuracies in taxonomic assignments can be solved to large extent by the increased share of long-reads used for full-length 16S rRNA gene sequencing [258], [259], however these are only slowly reaching the cost-effectiveness and throughput of short read sequencers. In addition, long-read sequencing is gaining popularity for hybrid assemblies of shotgun metagenomic sequencing data (meaning assembly of genomes from complex communities using both short- and long-reads). Here long-read data can be highly useful for scaffolding [260], but also for binning of contigs by the help of DNA methylation patterns [261]. The acquisition of high-quality genomic assemblies from microbial communities are crucial for the determination of metabolic potentials of individual members, but most importantly of the assemblage as a whole.

An additional future challenge in microbiome analysis – part technical, part conceptual – is the incorporation of cross-kingdom interaction in the analysis. In multiple studies, the lab of Harry Sokol (see [240], [241]) could show disease-related changes in abundance correlations of bacteria and fungi. These studies were entirely based on kingdom-specific sets of target amplicon sequencing data. While such approaches enable for high-throughput analysis and the inclusion of large numbers of samples, they are likely biased by the inherent compositional nature and technical

aspects of amplicon sequencing data and need labor-intensive validation of these proposed interactions. How such cross-kingdom interactions can be validated was demonstrated for the case of the bacterium *Christensenella minuta* and the archaeon *Methanobrevibacter smithii* recently. Ruaud *et al.* [262] investigated the hypothesis that the families *Methanobacteriaceae* and *Christensenellaceae* are co-occurring, which was originally generated from 16S rRNA gene amplicon data. By using complementary approaches of *in silico* analysis of shotgun metagenomic data and *in vitro* co-culturing of the organisms, they could show co-localization and metabolic interactions of the species, leading to changed production of SCFAs, with potential connections to host health [262]. Further, the kingdom of viruses, especially bacteriophages, and their associations with host health and interacting partners have recently shifted into the focus of microbiome research. A first large-scale longitudinal study of fecal phages could show the intra-personal stability of the communities over time and the inter-personal diversity of these communities [41]. In addition, it was demonstrated that persistent members of the page community are linked to predominantly present taxa of the intestinal bacterial community [41]. These insights provide the starting point for health-related investigations of the human phageome.

The effects of host-lifestyle on the associated microbial communities remain unclear. For example the impact of physical exercise on the host are expected to be beneficial, possibly mediated via the gut microbiome [263], however intervention studies in humans showing clear effects are still lacking. One study sampled individuals participating the 2015 Boston Marathon multiple times before and after the event [264]. The study found increased relative abundances of the bacterium *Veillonella atypica* in stool after the marathon. Isolation and gavage of the bacterium to mice supposedly increased treadmill runtime and metabolic analysis pointed to a conversion of lactate produced while running to the SCFA propionate by *V. atypica,* showing a direct interaction of host-derived metabolites and specialized consumption by associated microbes [264]. However, the small sample size in this experiment, the very low abundances of *V. atypica* and high variability between individuals by orders of magnitude leave the true effects unclear and calls for the need for validation of these findings in larger studies. Although the mechanisms remain unresolved, current research mostly focusing on professional athletes and rodents as model organisms, as reviewed by Sohail *et al.* [265], suggests a reciprocal association of exercise-induced microbial changes and immune-inflammatory mediators, activating a monitored energy balance and tissue metabolism. The effects of host-lifestyle are complicated to assess, as lifestyle is complicated to assess. The accuracy of consumer-grade wearables for activity tracking vary [266], however performances are steadily increasing [267]. Nutrient intake is an important factor to consider in microbiome studies, as host-nutrition will directly influence food resources available to the members of the gut microbiome, however while changes in dietary patterns induce changes in the microbiota, intervention studies investigating long-term effect are still lacking [268]. Nutrient intake is often assessed by standardized food frequency questionnaires (FFQs) used to extract information on the

intake of micro- and macronutrients [269]. While a one-to-one correlation of nutrients and changes in individual taxa requires many individual statistical test and thus large group sizes for sufficient statistical power, recent studies suggest the use of dietary indices for the assessment of dietary patterns and their effects on the microbiota, such as the Healthy Eating Index (HEI), the Mediterranean Diet Score (MDS) and the Healthy Food Diversity index (HFD-Index) [270]. Especially the HEI was demonstrated to show high correlation with alpha diversity as well as high explained variability in beta diversity and discordance in twin pairs [270]. The use of indices or broader food groups in microbiome studies and also a longitudinal assessment of nutrition and microbiota was shown to be crucial, as changes in the microbiota were found to be connected to dietary history over multiple preceding days and personalized food choices, while profiles of micro- and macronutrients were stable and similar across individuals and over time [271].

The investigation of nutrition and lifestyle impact on the microbiome is complicated, but even though very personalized, it is a task possible to tackle in well-designed studies [271]. Even more complicated to study are early-life and life-history effects on community assembly and long-term composition. Feretti *et al.* [272] demonstrated that even in temporally and spatially dense sampled mother-infant pairs, already one day after birth, on average less than 50% of the infants fecal microbiome could be attributed to any body-associated (fecal, oral, skin, and vaginal) microbiome of the mother. However, in this study clearly demonstrated, that early members of the infants fecal microbiota are acquired from multiple known and unknown environmental sources and largely transient, with only few of them stably colonizing after one and four months after birth [272]. While robustly identified universal molecular drivers of microbial colonization of the human intestine are still lacking, for mice it could be shown that the components of the host's innate immune system (Toll-like receptors) play a pivotal role in the initial assembly of microbial communities, influencing long-term composition of the gut microbiota [273].

The vast majority of studies investigating host-associated communities are still observational and descriptive, aiming at the discovery of disease-associated microorganisms. While these discoveries are crucial for the understanding of the interaction of host and microbes, they can only serve as a first step, as they provide only a single snapshot of microbial community configuration. Although algorithmic evaluation of growth dynamics from metagenomic data has been proposed [274], longitudinal sampling should be used for a robust assessment of community composition dynamics [275]. Once candidates influencing host-health are validated, the next steps will target the directed modulation of community structure and likely specific members of the community. Targeted phage therapies for microbiome modulation are already being tested in model organisms [276] and could become a valuable tool for shaping host-associated microbiomes. An additional approach to promote host health considering the intestinal microbiota is focusing on the administration of probiotic bacteria and personalized nutrition. Zeevi *et al.* [277] could

demonstrate that the gut microbial assemblage together with additional life-style factors and anthropometric measures can predict the post-prandial glycemic response. Probiotics as supporting treatment in IBD are tested, however results remain inconclusive [278]. In addition, administration of probiotics after antibiotic treatment resulted in a prolonged reconstitution time, compared to untreated individuals and autologous fecal microbiome transplantation [279]. However, this analysis was solely based on the assumption that reconstitution to the state before treatment is the desired outcome, not assessing metabolic capacity and/or health related measurements of alternative community compositions [279].

This fact highlights that the central question remains to be answered:

What does a healthy human gut microbiome look like?

As mentioned earlier, since the split from the shared ancestor with the *Pan* lineage, the human microbiome underwent an accelerated transformation [68]. The investigation of non-human primate associated microbial communities can potentially shed light on the evolutionary history of the human microbiome, and thus, on the evolutionary history of humans. Recently, in a study covering a broad range of primates, including Lemurs, New World monkeys, Old World monkeys, and Apes, it was shown that dietary niches have an influence on the gut microbiota, however that effects of host-physiology and thus of host-phylogeny clearly outweigh those [280]. This indicates that independent of a dietary preference, the intestinal microbiota is primarily constrained by host features. But also between human populations, large variation in the microbial composition is evident [49], likely driven by a 'Western' lifestyle and accompanied by an increased prevalence of chronic inflammatory disorders [281]. The question on whether an 'ancestral' microbiota can prevent chronic disorders remains to be answered, just like the question on how this ancestral state of the human microbiota looked like. However, in any case, a less Western-centered view on the microbiota might be able to assess microbial diversity and community variability on a true global scale, as recently demonstrated by Pasolli *et al.* [282]. Also, modern, data-driven approaches for phylogenetic and taxonomic annotations of microorganism, as performed by Parks *et al.* [168], can guide the way to a better informed assessment of microbial evolution as part of a community, and thus also of human evolution. Microbiome conservatory initiatives like 'The Microbiota Vault' (http://www.microbiotavault.org/) and the 'Global Microbiome Conservancy' (http://microbiomeconservancy.org/) are crucial for tackling the challenge of further understanding host-associated community assembly and might guide future developments targeting health-related aspects of the human metaorganism.

6 References

[1] J. Wang *et al.*, "Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota," *Nat. Genet.*, vol. 48, no. 11, pp. 1396–1406, 2016, doi: 10.1038/ng.3695.

[2] J. L. Mohr, "Protozoa as Indicators of Pollution," *Sci. Mon.*, vol. 74, no. 1, pp. 7–9, 1952.

[3] J. R. Marchesi and J. Ravel, "The vocabulary of microbiome research: a proposal," *Microbiome*, vol. 3, Jul. 2015, doi: 10.1186/s40168-015-0094-5.

[4] K. Faust and J. Raes, "Microbial interactions: from networks to models," *Nat. Rev. Microbiol.*, vol. 10, no. 8, pp. 538–550, Aug. 2012, doi: 10.1038/nrmicro2832.

[5] J. A. Gilbert *et al.*, "The Earth Microbiome Project: Meeting report of the '1st EMP meeting on sample selection and acquisition' at Argonne National Laboratory October 6th 2010.," *Stand. Genomic Sci.*, vol. 3, no. 3, p. 249, Dec. 2010, doi: 10.4056/aigs.1443528.

[6] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, Oct. 2007, doi: 10.1038/nature06244.

[7] K. A. Möbius, *Die Auster und die Austernwirthschaft*. Verlag von Wiegandt, Hemple & Parey, 1877.

[8] E. Biagi, M. Candela, S. Fairweather-Tait, C. Franceschi, and P. Brigidi, "Ageing of the human metaorganism: the microbial counterpart," *Age*, vol. 34, no. 1, pp. 247–267, Feb. 2012, doi: 10.1007/s11357-011-9217-5.

[9] T. C. G. Bosch and M. J. McFall-Ngai, "Metaorganisms as the new frontier," *Zoology*, vol. 114, no. 4, pp. 185–190, Sep. 2011, doi: 10.1016/j.zool.2011.04.001.

[10] I. Zilber-Rosenberg and E. Rosenberg, "Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution," *FEMS Microbiol. Rev.*, vol. 32, no. 5, pp. 723–735, Aug. 2008, doi: 10.1111/j.1574-6976.2008.00123.x.

[11] P. R. Ehrlich and P. H. Raven, "Butterflies and Plants: A Study in Coevolution," *Evolution*, vol. 18, no. 4, pp. 586–608, 1964, doi: 10.2307/2406212.

[12] P. Baumann, L. Baumann, C.-Y. Lai, D. Rouhbakhsh, N. A. Moran, and M. A. Clark, "Genetics, Physiology and Evolutionary Relathinships of the Genus Buchnera: Intracellular Symbionts of Aphids," *Annu. Rev. Microbiol.*, vol. 49, no. 1, pp. 55–94, 1995, doi: 10.1146/annurev.mi.49.100195.000415.

[13] F. J. Pollock *et al.*, "Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny," *Nat. Commun.*, vol. 9, no. 1, pp. 1–13, Nov. 2018, doi: 10.1038/s41467-018-07275-x.

[14] R. M. Brucker and S. R. Bordenstein, "The Roles of Host Evolutionary Relationships (genus: Nasonia) and Development in Structuring Microbial Communities," *Evolution*, vol. 66, no. 2, pp. 349–362, 2012, doi: 10.1111/j.1558-5646.2011.01454.x.

[15] A. W. Brooks, K. D. Kohl, R. M. Brucker, E. J. van Opstal, and S. R. Bordenstein, "Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History," *PLOS Biol.*, vol. 14, no. 11, p. e2000225, Nov. 2016, doi: 10.1371/journal.pbio.2000225.

[16] S. J. Lim and S. R. Bordenstein, "An introduction to phylosymbiosis," PeerJ Inc., e27879v2, Dec. 2019.

[17] R. Sender, S. Fuchs, and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body," *PLoS Biol.*, vol. 14, no. 8, Aug. 2016, doi: 10.1371/journal.pbio.1002533.

[18] The Human Microbiome Project Consortium, "Structure, Function and Diversity of the Healthy Human Microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, Jun. 2012, doi: 10.1038/nature11234.

[19] J. Lloyd-Price *et al.*, "Strains, functions and dynamics in the expanded Human Microbiome Project," *Nature*, vol. 550, no. 7674, pp. 61–66, Oct. 2017, doi: 10.1038/nature23889.

[20] P. Kovatcheva-Datchary *et al.*, "Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of Prevotella," *Cell Metab.*, vol. 22, no. 6, pp. 971–982, Dec. 2015, doi: 10.1016/j.cmet.2015.10.001.

[21] H. J. Flint, K. P. Scott, S. H. Duncan, P. Louis, and E. Forano, "Microbial degradation of complex carbohydrates in the gut," *Gut Microbes*, vol. 3, no. 4, pp. 289–306, Jul. 2012, doi: 10.4161/gmic.19897.

[22] E. C. Martens *et al.*, "Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts," *PLOS Biol.*, vol. 9, no. 12, p. e1001221, Dec. 2011, doi: 10.1371/journal.pbio.1001221.

[23] D. Dodd, R. I. Mackie, and I. K. O. Cann, "Xylan degradation, a metabolic property shared by rumen and human colonic Bacteroidetes," *Mol. Microbiol.*, vol. 79, no. 2, pp. 292–304, 2011, doi: 10.1111/j.1365-2958.2010.07473.x.

[24] T. Chen, W. Long, C. Zhang, S. Liu, L. Zhao, and B. R. Hamaker, "Fiber-utilizing capacity varies in Prevotella- versus Bacteroides-dominated gut microbiota," *Sci. Rep.*, vol. 7, Jun. 2017, doi: 10.1038/s41598-017-02995-4.

[25] R. Corrêa-Oliveira, J. L. Fachi, A. Vieira, F. T. Sato, and M. A. R. Vinolo, "Regulation of immune cell function by short-chain fatty acids," *Clin. Transl. Immunol.*, vol. 5, no. 4, p. e73, Apr. 2016, doi: 10.1038/cti.2016.17.

[26] D. R. Donohoe *et al.*, "The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon," *Cell Metab.*, vol. 13, no. 5, pp. 517–526, May 2011, doi: 10.1016/j.cmet.2011.02.018.

[27] D. Parada Venegas *et al.*, "Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases," *Front. Immunol.*, vol. 10, 2019, doi: 10.3389/fimmu.2019.00277.

[28] J. M. Ridlon, D.-J. Kang, and P. B. Hylemon, "Bile salt biotransformations by human intestinal bacteria," *J. Lipid Res.*, vol. 47, no. 2, pp. 241–259, Feb. 2006, doi: 10.1194/jlr.R500013-JLR200.

[29] J. M. Ridlon, D. J. Kang, P. B. Hylemon, and J. S. Bajaj, "Bile Acids and the Gut Microbiome," *Curr. Opin. Gastroenterol.*, vol. 30, no. 3, pp. 332–338, May 2014, doi: 10.1097/MOG.0000000000000057.

[30] M. L. Jones, C. J. Martoni, and S. Prakash, "Cholesterol lowering and inhibition of sterol absorption by Lactobacillus reuteri NCIMB 30242: a randomized controlled trial," *Eur. J. Clin. Nutr.*, vol. 66, no. 11, pp. 1234–1241, Nov. 2012, doi: 10.1038/ejcn.2012.126.

[31] S. Miquel *et al.*, "Faecalibacterium prausnitzii and human intestinal health," *Curr. Opin. Microbiol.*, vol. 16, no. 3, pp. 255–261, Jun. 2013, doi: 10.1016/j.mib.2013.06.003.

[32] A. O'Callaghan and D. van Sinderen, "Bifidobacteria and Their Role as Members of the Human Gut Microbiota," *Front. Microbiol.*, vol. 7, Jun. 2016, doi: 10.3389/fmicb.2016.00925.

[33] N. Ottman, S. Y. Geerlings, S. Aalvink, W. M. de Vos, and C. Belzer, "Action and function of Akkermansia muciniphila in microbiome ecology, health and disease," *Best Pract. Res. Clin. Gastroenterol.*, vol. 31, no. 6, pp. 637–642, Dec. 2017, doi: 10.1016/j.bpg.2017.10.001.

[34] J. N. V. Martinson *et al.*, "Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults," *ISME J.*, vol. 13, no. 9, pp. 2306–2318, Sep. 2019, doi: 10.1038/s41396-019-0435-7.

[35] C. F. P. Scholz and M. Kilian, "The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. nov., Cutibacterium gen. nov. and Pseudopropionibacterium gen. nov," *Int. J. Syst. Evol. Microbiol.*, vol. 66, no. 11, pp. 4422–4432, Nov. 2016, doi: 10.1099/ijsem.0.001367.

[36] A. L. Byrd, Y. Belkaid, and J. A. Segre, "The human skin microbiome," *Nat. Rev. Microbiol.*, vol. 16, no. 3, pp. 143–155, Mar. 2018, doi: 10.1038/nrmicro.2017.157.

[37] J. Ravel *et al.*, "Vaginal microbiome of reproductive-age women," *Proc. Natl. Acad. Sci.*, vol. 108, no. Supplement 1, pp. 4680–4687, Mar. 2011, doi: 10.1073/pnas.1002611107.

[38] K. Koskinen *et al.*, "First Insights into the Diverse Human Archaeome: Specific Detection of Archaea in the Gastrointestinal Tract, Lung, and Nose and on Skin," *mBio*, vol. 8, no. 6, Dec. 2017, doi: 10.1128/mBio.00824-17.

[39] P. C. Seed, "The Human Mycobiome," *Cold Spring Harb. Perspect. Med.*, vol. 5, no. 5, May 2015, doi: 10.1101/cshperspect.a019810.

[40] P. Manrique, B. Bolduc, S. T. Walk, J. van der Oost, W. M. de Vos, and M. J. Young, "Healthy human gut phageome," *Proc. Natl. Acad. Sci.*, vol. 113, no. 37, pp. 10400–10405, Sep. 2016, doi: 10.1073/pnas.1601060113.

[41] A. N. Shkoporov *et al.*, "The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific," *Cell Host Microbe*, vol. 26, no. 4, pp. 527-541.e5, Oct. 2019, doi: 10.1016/j.chom.2019.09.009.

[42] P. J. Turnbaugh *et al.*, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, Jan. 2009, doi: 10.1038/nature07540.

[43] J. K. Goodrich *et al.*, "Genetic Determinants of the Gut Microbiome in UK Twins," *Cell Host Microbe*, vol. 19, no. 5, pp. 731–743, May 2016, doi: 10.1016/j.chom.2016.04.017.

[44] J. K. Goodrich *et al.*, "Human genetics shape the gut microbiome," *Cell*, vol. 159, no. 4, pp. 789–799, Nov. 2014, doi: 10.1016/j.cell.2014.09.053.

[45] W. Turpin *et al.*, "Association of host genome with intestinal microbial composition in a large healthy cohort," *Nat. Genet.*, vol. 48, no. 11, pp. 1413–1417, 2016, doi: 10.1038/ng.3693.

[46] M. J. Bonder *et al.*, "The effect of host genetics on the gut microbiome," *Nat. Genet.*, vol. 48, no. 11, pp. 1407–1412, Nov. 2016, doi: 10.1038/ng.3663.

[47] D. Rothschild *et al.*, "Environment dominates over host genetics in shaping human gut microbiota," *Nature*, vol. 555, no. 7695, pp. 210–215, Mar. 2018, doi: 10.1038/nature25973.

[48] J. Hellwege, J. Keaton, A. Giri, X. Gao, D. R. Velez Edwards, and T. L. Edwards, "Population Stratification in Genetic Association Studies," *Curr. Protoc. Hum. Genet.*, vol. 95, pp. 1.22.1-1.22.23, Oct. 2017, doi: 10.1002/cphg.48.

[49] T. Yatsunenko *et al.*, "Human gut microbiome viewed across age and geography," *Nature*, vol. 486, no. 7402, pp. 222–227, May 2012, doi: 10.1038/nature11053.

[50] S. L. Schnorr *et al.*, "Gut microbiome of the Hadza hunter-gatherers," *Nat. Commun.*, vol. 5, no. 1, pp. 1–12, Apr. 2014, doi: 10.1038/ncomms4654.

[51] R. Blekhman *et al.*, "Host genetic variation impacts microbiome composition across human body sites," *Genome Biol.*, vol. 16, no. 1, p. 191, Sep. 2015, doi: 10.1186/s13059-015-0759-1.

[52] E. R. Davenport, D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober, and Y. Gilad, "Genome-Wide Association Studies of the Human Gut Microbiota," *PLOS ONE*, vol. 10, no. 11, p. e0140301, Mar. 2015, doi: 10.1371/journal.pone.0140301.

[53] X. Hua *et al.*, "MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition," *bioRxiv*, p. 031187, Jan. 2015, doi: 10.1101/031187.

[54] M. Arumugam *et al.*, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, May 2011, doi: 10.1038/nature09944.

[55] G. D. Wu *et al.*, "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, vol. 334, no. 6052, pp. 105–108, Oct. 2011, doi: 10.1126/science.1208344.

[56] I. Holmes, K. Harris, and C. Quince, "Dirichlet multinomial mixtures: generative models for microbial metagenomics," *PloS One*, vol. 7, no. 2, p. e30126, 2012, doi: 10.1371/journal.pone.0030126.

[57] S. M. Huse, Y. Ye, Y. Zhou, and A. A. Fodor, "A core human microbiome as viewed through 16S rRNA sequence clusters," *PloS One*, vol. 7, no. 6, p. e34242, 2012, doi: 10.1371/journal.pone.0034242.

[58] D. Vandeputte *et al.*, "Quantitative microbiome profiling links gut community variation to microbial load," *Nature*, vol. 551, no. 7681, pp. 507–511, 23 2017, doi: 10.1038/nature24460.

[59] A. H. Moeller, P. H. Degnan, A. E. Pusey, M. L. Wilson, B. H. Hahn, and H. Ochman, "Chimpanzees and Humans Harbor Compositionally Similar Gut Enterotypes," *Nat. Commun.*, vol. 3, p. 1179, 2012, doi: 10.1038/ncomms2159.

[60] A. Bergström *et al.*, "Establishment of Intestinal Microbiota during Early Life: a Longitudinal, Explorative Study of a Large Cohort of Danish Infants," *Appl. Environ. Microbiol.*, vol. 80, no. 9, pp. 2889–2900, May 2014, doi: 10.1128/AEM.00342-14.

[61] P. I. Costea *et al.*, "Enterotypes in the landscape of gut microbial community composition," *Nat. Microbiol.*, vol. 3, no. 1, pp. 8–16, Jan. 2018, doi: 10.1038/s41564-017-0072-8.

[62] G. Falony *et al.*, "Population-level analysis of gut microbiome variation," *Science*, vol. 352, no. 6285, pp. 560–564, Apr. 2016, doi: 10.1126/science.aad3503.

[63] A. Palleja *et al.*, "Recovery of gut microbiota of healthy adults following antibiotic exposure," *Nat. Microbiol.*, vol. 3, no. 11, pp. 1255–1265, Nov. 2018, doi: 10.1038/s41564-018-0257-9.

[64] L. Maier *et al.*, "Extensive impact of non-antibiotic drugs on human gut bacteria," *Nature*, vol. 555, no. 7698, pp. 623–628, Mar. 2018, doi: 10.1038/nature25979.

[65] M. Groussin *et al.*, "Unraveling the processes shaping mammalian gut microbiomes over evolutionary time," *Nat. Commun.*, vol. 8, no. 1, pp. 1–12, Feb. 2017, doi: 10.1038/ncomms14319.

[66] H. Ochman *et al.*, "Evolutionary Relationships of Wild Hominids Recapitulated by Gut Microbial Communities," *PLOS Biol.*, vol. 8, no. 11, p. e1000546, Nov. 2010, doi: 10.1371/journal.pbio.1000546.

[67] K. E. Langergraber *et al.*, "Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 39, pp. 15716–15721, Sep. 2012, doi: 10.1073/pnas.1211740109.

[68] A. H. Moeller *et al.*, "Rapid changes in the gut microbiome during human evolution," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 46, pp. 16431–16435, Nov. 2014, doi: 10.1073/pnas.1419136111.

[69] A. H. Moeller *et al.*, "Cospeciation of gut microbiota with hominids," *Science*, vol. 353, no. 6297, pp. 380–382, Jul. 2016, doi: 10.1126/science.aaf3951.

[70] S. Zhao *et al.*, "Adaptive Evolution within Gut Microbiomes of Healthy People," *Cell Host Microbe*, vol. 25, no. 5, pp. 656-667.e8, 08 2019, doi: 10.1016/j.chom.2019.03.007.

[71] A. B. Pfisterer and B. Schmid, "Diversity-dependent production can decrease the stability of ecosystem functioning," *Nature*, vol. 416, no. 6876, pp. 84–86, Mar. 2002, doi: 10.1038/416084a.

[72] J. Tap *et al.*, "Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome," *Gastroenterology*, vol. 152, no. 1, pp. 111-123.e8, Jan. 2017, doi: 10.1053/j.gastro.2016.09.049.

[73] C. Quince *et al.*, "Extensive Modulation of the Fecal Metagenome in Children With Crohn's Disease During Exclusive Enteral Nutrition," *Am. J. Gastroenterol.*, vol. 110, no. 12, pp. 1718–1729, Dec. 2015, doi: 10.1038/ajg.2015.357.

[74] M. Kummen *et al.*, "The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls," *Gut*, vol. 66, no. 4, pp. 611–619, 2017, doi: 10.1136/gutjnl-2015-310500.

[75] S. L. Russell *et al.*, "Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma," *EMBO Rep.*, vol. 13, no. 5, pp. 440–447, May 2012, doi: 10.1038/embor.2012.32.

[76] D. Gevers, M. Pop, P. D. Schloss, and C. Huttenhower, "Bioinformatics for the Human Microbiome Project," *PLOS Comput. Biol.*, vol. 8, no. 11, p. e1002779, Nov. 2012, doi: 10.1371/journal.pcbi.1002779.

[77] H. K. Pedersen *et al.*, "Human gut microbes impact host serum metabolome and insulin sensitivity," *Nature*, vol. 535, no. 7612, pp. 376–381, Jul. 2016, doi: 10.1038/nature18646.

[78] J. Wirbel *et al.*, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nat. Med.*, vol. 25, no. 4, pp. 679–689, Apr. 2019, doi: 10.1038/s41591-019-0406-6.

[79] J. F. Cryan, K. J. O'Riordan, K. Sandhu, V. Peterson, and T. G. Dinan, "The gut microbiome in neurological disorders," *Lancet Neurol.*, vol. 0, no. 0, Nov. 2019, doi: 10.1016/S1474-4422(19)30356-4.

[80] S. R. Gill *et al.*, "Metagenomic Analysis of the Human Distal Gut Microbiome," *Science*, vol. 312, no. 5778, pp. 1355–1359, Jun. 2006, doi: 10.1126/science.1124234.

[81] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-Bacterial Mutualism in the Human Intestine," *Science*, vol. 307, no. 5717, pp. 1915–1920, Mar. 2005, doi: 10.1126/science.1104816.

[82] Q. R. Ducarmon, R. D. Zwittink, B. V. H. Hornung, W. van Schaik, V. B. Young, and E. J. Kuijper, "Gut Microbiota and Colonization Resistance against Bacterial Enteric Infection," *Microbiol. Mol. Biol. Rev.*, vol. 83, no. 3, Aug. 2019, doi: 10.1128/MMBR.00007-19.

[83] E. A. Kennedy, K. Y. King, and M. T. Baldridge, "Mouse Microbiota Models: Comparing Germ-Free Mice and Antibiotics Treatment as Tools for Modifying Gut Bacteria," *Front. Physiol.*, vol. 9, Oct. 2018, doi: 10.3389/fphys.2018.01534.

[84] Y. Belkaid and T. W. Hand, "Role of the Microbiota in Immunity and Inflammation," *Cell*, vol. 157, no. 1, pp. 121–141, Mar. 2014, doi: 10.1016/j.cell.2014.03.011.

[85] Q. Mu, J. Kirby, C. M. Reilly, and X. M. Luo, "Leaky Gut As a Danger Signal for Autoimmune Diseases," *Front. Immunol.*, vol. 8, 2017, doi: 10.3389/fimmu.2017.00598.

[86] H. Tlaskalová-Hogenová *et al.*, "The role of gut microbiota (commensal bacteria) and the mucosal barrier in the pathogenesis of inflammatory and autoimmune diseases and cancer: contribution of germ-free and gnotobiotic animal models of human diseases," *Cell. Mol. Immunol.*, vol. 8, no. 2, pp. 110–120, Mar. 2011, doi: 10.1038/cmi.2010.67.

[87] F. R. Vogenberg, C. Isaacson Barash, and M. Pursel, "Personalized Medicine," *Pharm. Ther.*, vol. 35, no. 10, pp. 560–576, Oct. 2010.

[88] N. B. La Thangue and D. J. Kerr, "Predictive biomarkers: a paradigm shift towards personalized cancer medicine," *Nat. Rev. Clin. Oncol.*, vol. 8, no. 10, pp. 587–596, Oct. 2011, doi: 10.1038/nrclinonc.2011.121.

[89] E. I. Dumbrava and F. Meric-Bernstam, "Personalized cancer therapy—leveraging a knowledge base for clinical decision-making," *Cold Spring Harb. Mol. Case Stud.*, vol. 4, no. 2, Apr. 2018, doi: 10.1101/mcs.a001578.

[90] A. G. Tamilarasan, G. Cunningham, P. M. Irving, and M. A. Samaan, "Recent advances in monoclonal antibody therapy in IBD: practical issues," *Frontline Gastroenterol.*, vol. 10, no. 4, pp. 409–416, Oct. 2019, doi: 10.1136/flgastro-2018-101054.

[91] K. Aden *et al.*, "Metabolic Functions of Gut Microbes Associate With Efficacy of Tumor Necrosis Factor Antagonists in Patients With Inflammatory Bowel Diseases," *Gastroenterology*, vol. 157, no. 5, pp. 1279-1292.e11, Nov. 2019, doi: 10.1053/j.gastro.2019.07.025.

[92] M. Zimmermann, M. Zimmermann-Kogadeeva, R. Wegmann, and A. L. Goodman, "Separating host and microbiome contributions to drug pharmacokinetics and toxicity," *Science*, vol. 363, no. 6427, Feb. 2019, doi: 10.1126/science.aat9931.

[93] J. Lloyd-Price *et al.*, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, May 2019, doi: 10.1038/s41586-019-1237-9.

[94] N. Nakamoto *et al.*, "Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis," *Nat. Microbiol.*, vol. 4, no. 3, pp. 492–503, 2019, doi: 10.1038/s41564-018-0333-1.

[95] E. B. Hollister *et al.*, "Leveraging Human Microbiome Features to Diagnose and Stratify Children with Irritable Bowel Syndrome," *J. Mol. Diagn. JMD*, vol. 21, no. 3, pp. 449–461, May 2019, doi: 10.1016/j.jmoldx.2019.01.006.

[96] A. Behrouzi, A. H. Nafari, and S. D. Siadat, "The significance of microbiome in personalized medicine," *Clin. Transl. Med.*, vol. 8, no. 1, p. 16, May 2019, doi: 10.1186/s40169-019-0232-y.

[97] C. R. Kelly *et al.*, "Update on Fecal Microbiota Transplantation 2015: Indications, Methodologies, Mechanisms, and Outlook," *Gastroenterology*, vol. 149, no. 1, pp. 223–237, Jul. 2015, doi: 10.1053/j.gastro.2015.05.008.

[98] S. J. Ott *et al.*, "Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection," *Gastroenterology*, vol. 152, no. 4, pp. 799-811.e7, 2017, doi: 10.1053/j.gastro.2016.11.010.

[99] Y. Tian *et al.*, "Fecal microbiota transplantation for ulcerative colitis: a prospective clinical study," *BMC Gastroenterol.*, vol. 19, no. 1, p. 116, Jul. 2019, doi: 10.1186/s12876-019-1010-4.

[100] R. P. Hirten *et al.*, "Microbial Engraftment and Efficacy of Fecal Microbiota Transplant for Clostridium Difficile in Patients With and Without Inflammatory Bowel Disease," *Inflamm. Bowel Dis.*, vol. 25, no. 6, pp. 969–979, 04 2019, doi: 10.1093/ibd/izy398.

[101] P. C. Kashyap, N. Chia, H. Nelson, E. Segal, and E. Elinav, "Microbiome at the Frontier of Personalized Medicine," *Mayo Clin. Proc.*, vol. 92, no. 12, pp. 1855–1864, Dec. 2017, doi: 10.1016/j.mayocp.2017.10.004.

[102] S. Alatab *et al.*, "The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 1, pp. 17–30, Jan. 2020, doi: 10.1016/S2468-1253(19)30333-4.

[103] H. Huang *et al.*, "Fine-mapping inflammatory bowel disease loci to single-variant resolution," *Nature*, vol. 547, no. 7662, pp. 173–178, 13 2017, doi: 10.1038/nature22969.

[104] D. C. Baumgart and S. R. Carding, "Inflammatory bowel disease: cause and immunobiology," *The Lancet*, vol. 369, no. 9573, pp. 1627–1640, May 2007, doi: 10.1016/S0140-6736(07)60750-8.

[105] M. Gajendran *et al.*, "A comprehensive review and update on ulcerative colitis," *Dis.--Mon. DM*, vol. 65, no. 12, p. 100851, Dec. 2019, doi: 10.1016/j.disamonth.2019.02.004.

[106] M. Gajendran, P. Loganathan, A. P. Catinella, and J. G. Hashash, "A comprehensive review and update on Crohn's disease," *Dis. Mon.*, vol. 64, no. 2, pp. 20–57, Feb. 2018, doi: 10.1016/j.disamonth.2017.07.001.

[107] M. Argollo, D. Gilardi, C. Peyrin-Biroulet, J.-F. Chabot, L. Peyrin-Biroulet, and S. Danese, "Comorbidities in inflammatory bowel disease: a call for action," *Lancet Gastroenterol. Hepatol.*, vol. 4, no. 8, pp. 643–654, Aug. 2019, doi: 10.1016/S2468-1253(19)30173-6.

[108] M. K. Washington, "Autoimmune liver disease: overlap and outliers," *Mod. Pathol.*, vol. 20, no. 1, pp. S15–S30, Feb. 2007, doi: 10.1038/modpathol.3800684.

[109] P. Udompap, D. Kim, and W. R. Kim, "Current and Future Burden of Chronic Nonmalignant Liver Disease," *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.*, vol. 13, no. 12, pp. 2031–2041, Nov. 2015, doi: 10.1016/j.cgh.2015.08.015.

[110] U. Beuers, "Hepatic overlap syndromes," *J. Hepatol.*, vol. 42, no. 1, pp. S93–S99, Apr. 2005, doi: 10.1016/j.jhep.2004.11.009.

[111] G. Paumgartner, "Ursodeoxycholic acid for primary biliary cirrhosis: treat early to slow progression," *J. Hepatol.*, vol. 39, no. 1, pp. 112–114, Jul. 2003, doi: 10.1016/S0168-8278(03)00243-5.

[112] J. E. Eaton, J. A. Talwalkar, K. N. Lazaridis, G. J. Gores, and K. D. Lindor, "Pathogenesis of Primary Sclerosing Cholangitis and Advances in Diagnosis and Management," *Gastroenterology*, vol. 145, no. 3, Sep. 2013, doi: 10.1053/j.gastro.2013.06.052.

[113] E. V. Loftus *et al.*, "PSC-IBD: a unique form of inflammatory bowel disease associated with primary sclerosing cholangitis," *Gut*, vol. 54, no. 1, pp. 91–96, Jan. 2005, doi: 10.1136/gut.2004.046615.

[114] K. N. Lazaridis and G. J. Gores, "Primary sclerosing cholangitis and cholangiocarcinoma," *Semin. Liver Dis.*, vol. 26, no. 1, pp. 42–51, Feb. 2006, doi: 10.1055/s-2006-933562.

[115] S. Weidinger and N. Novak, "Atopic dermatitis," *The Lancet*, vol. 387, no. 10023, pp. 1109–1122, Mar. 2016, doi: 10.1016/S0140-6736(15)00149-X.

[116] J. I. Silverberg and J. M. Hanifin, "Adult eczema prevalence and associations with asthma and other health and demographic factors: A US population–based study," *J. Allergy Clin. Immunol.*, vol. 132, no. 5, pp. 1132–1138, Nov. 2013, doi: 10.1016/j.jaci.2013.08.031.

[117] S. F. Thomsen, "Importance of genetic factors in the etiology of atopic dermatitis: a twin study," *Allergy Asthma Proc.*, vol. 28, no. 5, pp. 535–539, 2007, doi: 10.2500/aap2007.28.3041.

[118] A. D. Irvine, W. H. I. McLean, and D. Y. M. Leung, "Filaggrin Mutations Associated with Skin and Allergic Diseases," *N. Engl. J. Med.*, vol. 365, no. 14, pp. 1315–1327, Oct. 2011, doi: 10.1056/NEJMra1011040.

[119] K. Darabi, S. G. Hostetler, M. A. Bechtel, and M. Zirwas, "The role of Malassezia in atopic dermatitis affecting the head and neck of adults," *J. Am. Acad. Dermatol.*, vol. 60, no. 1, pp. 125–136, Jan. 2009, doi: 10.1016/j.jaad.2008.07.058.

[120] J. E. Greb *et al.*, "Psoriasis," *Nat. Rev. Dis. Primer*, vol. 2, no. 1, pp. 1–17, Nov. 2016, doi: 10.1038/nrdp.2016.82.

[121] C. E. M. Griffiths, P. van de Kerkhof, and M. Czarnecka-Operacz, "Psoriasis and Atopic Dermatitis," *Dermatol. Ther.*, vol. 7, no. Suppl 1, pp. 31–41, Jan. 2017, doi: 10.1007/s13555-016-0167-9.

[122] T. Dainichi, S. Hanakawa, and K. Kabashima, "Classification of inflammatory skin diseases: A proposal based on the disorders of the three-layered defense systems, barrier, innate immunity and acquired immunity," *J. Dermatol. Sci.*, vol. 76, no. 2, pp. 81–89, Nov. 2014, doi: 10.1016/j.jdermsci.2014.08.010.

[123] M. Cottone, C. Sapienza, F. S. Macaluso, and M. Cannizzaro, "Psoriasis and Inflammatory Bowel Disease," *Dig. Dis.*, vol. 37, no. 6, pp. 451–457, 2019, doi: 10.1159/000500116.

[124] D. Ellinghaus *et al.*, "Combined Analysis of Genome-wide Association Studies for Crohn Disease and Psoriasis Identifies Seven Shared Susceptibility Loci," *Am. J. Hum. Genet.*, vol. 90, no. 4, pp. 636–647, Apr. 2012, doi: 10.1016/j.ajhg.2012.02.020.

[125] R. W. Holley, G. A. Everett, J. T. Madison, and A. Zamir, "Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid," *J. Biol. Chem.*, vol. 240, no. 5, pp. 2122–2128, May 1965.

[126] P. G. N. Jeppesen, B. G. Barrell, F. Sanger, and A. R. Coulson, "Nucleotide sequences of two fragments from the coat-protein cistron of bacteriophage R17 ribonucleic acid," *Biochem. J.*, vol. 128, no. 5, pp. 993–1006, Aug. 1972, doi: 10.1042/bj1280993h.

[127] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *J. Mol. Biol.*, vol. 94, no. 3, pp. 441–448, May 1975, doi: 10.1016/0022-2836(75)90213-2.

[128] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 2, pp. 560–564, Feb. 1977, doi: 10.1073/pnas.74.2.560.

[129] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.

[130] A. R. Quinlan, D. A. Stewart, M. P. Strömberg, and G. T. Marth, "Pyrobayes: an improved base caller for SNP discovery in pyrosequences," *Nat. Methods*, vol. 5, no. 2, pp. 179–181, Feb. 2008, doi: 10.1038/nmeth.1172.

[131] S. Balzer, K. Malde, and I. Jonassen, "Systematic exploration of error sources in pyrosequencing flowgram data," *Bioinformatics*, vol. 27, no. 13, pp. i304–i309, Jul. 2011, doi: 10.1093/bioinformatics/btr251.

[132] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb, "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations," *Science*, vol. 299, no. 5607, pp. 682–686, Jan. 2003, doi: 10.1126/science.1079700.

[133] B. A. Flusberg *et al.*, "Direct detection of DNA methylation during single-molecule, real-time sequencing," *Nat. Methods*, vol. 7, no. 6, pp. 461–465, Jun. 2010, doi: 10.1038/nmeth.1459.

[134] D. Stoddart, A. J. Heron, E. Mikhailova, G. Maglia, and H. Bayley, "Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore," *Proc. Natl. Acad. Sci.*, vol. 106, no. 19, pp. 7702–7707, May 2009, doi: 10.1073/pnas.0901054106.

[135] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016, doi: 10.1016/j.ygeno.2015.11.003.

[136] A. Hiergeist, J. Gläsner, U. Reischl, and A. Gessner, "Analyses of Intestinal Microbiota: Culture versus Sequencing," *ILAR J.*, vol. 56, no. 2, pp. 228–240, Aug. 2015, doi: 10.1093/ilar/ilv017.

[137] J. T. Staley and A. Konopka, "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats," *Annu. Rev. Microbiol.*, vol. 39, pp. 321–346, 1985, doi: 10.1146/annurev.mi.39.100185.001541.

[138] B. G. Clement, L. E. Kehl, K. L. DeBord, and C. L. Kitts, "Terminal restriction fragment patterns (TRFPs), a rapid, PCR-based method for the comparison of complex bacterial communities," *J. Microbiol. Methods*, vol. 31, no. 3, pp. 135–142, Jan. 1998, doi: 10.1016/S0167-7012(97)00105-X.

[139] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA.," *Appl. Environ. Microbiol.*, vol. 59, no. 3, pp. 695–700, Mar. 1993.

[140] M. Tsukuda, K. Kitahara, and K. Miyazaki, "Comparative RNA function analysis reveals high functional similarity between distantly related bacterial 16 S rRNAs," *Sci. Rep.*, vol. 7, Aug. 2017, doi: 10.1038/s41598-017-10214-3.

[141] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms," *Proc. Natl. Acad. Sci.*, vol. 74, no. 11, p. 5088, Nov. 1977, doi: 10.1073/pnas.74.11.5088.

[142] S. J. Giovannoni, T. B. Britschgi, C. L. Moyer, and K. G. Field, "Genetic diversity in Sargasso Sea bacterioplankton," *Nature*, vol. 345, no. 6270, pp. 60–63, May 1990, doi: 10.1038/345060a0.

[143] K. H. Wilson and R. B. Blitchington, "Human colonic biota studied by ribosomal DNA sequence analysis.," *Appl. Environ. Microbiol.*, vol. 62, no. 7, pp. 2273–2278, Jul. 1996.

[144] M. L. Sogin *et al.*, "Microbial diversity in the deep sea and the underexplored 'rare biosphere,'" *Proc. Natl. Acad. Sci.*, vol. 103, no. 32, pp. 12115–12120, Aug. 2006, doi: 10.1073/pnas.0605127103.

[145] J. F. Siqueira, A. F. Fouad, and I. N. Rôças, "Pyrosequencing as a tool for better understanding of human microbiomes," *J. Oral Microbiol.*, vol. 4, no. 1, p. 10743, Jan. 2012, doi: 10.3402/jom.v4i0.10743.

[146] G. B. Gloor *et al.*, "Microbiome Profiling by Illumina Sequencing of Combinatorial Sequence-Tagged PCR Products," *PLOS ONE*, vol. 5, no. 10, p. e15406, Oct. 2010, doi: 10.1371/journal.pone.0015406.

[147] J. G. Caporaso *et al.*, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proc. Natl. Acad. Sci.*, vol. 108, no. Supplement 1, pp. 4516–4522, Mar. 2011, doi: 10.1073/pnas.1000080107.

[148] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform," *Appl. Environ. Microbiol.*, vol. 79, no. 17, pp. 5112–5120, Sep. 2013, doi: 10.1128/AEM.01043-13.

[149] J. J. Minich *et al.*, "High-Throughput Miniaturized 16S rRNA Amplicon Library Preparation Reduces Costs while Preserving Microbiome Integrity," *mSystems*, vol. 3, no. 6, Dec. 2018, doi: 10.1128/mSystems.00166-18.

[150] M. A. Fischer, S. Güllert, S. C. Neulinger, W. R. Streit, and R. A. Schmitz, "Evaluation of 16S rRNA Gene Primer Pairs for Monitoring Microbial Community Structures Showed High Reproducibility within and Low Comparability between Datasets Generated with Multiple Archaeal and Bacterial Primer Pairs," *Front. Microbiol.*, vol. 7, Aug. 2016, doi: 10.3389/fmicb.2016.01297.

[151] F. De Filippis, M. Laiola, G. Blaiotta, and D. Ercolini, "Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity," *Appl. Environ. Microbiol.*, vol. 83, no. 17, 01 2017, doi: 10.1128/AEM.00905-17.

[152] A. Popovic *et al.*, "Design and application of a novel two-amplicon approach for defining eukaryotic microbiota," *Microbiome*, vol. 6, no. 1, p. 228, Dec. 2018, doi: 10.1186/s40168-018-0612-3.

[153] F. Rohwer, V. Seguritan, D. h. Choi, A. m. Segall, and F. Azam, "Production of Shotgun Libraries Using Random Amplification," *BioTechniques*, vol. 31, no. 1, pp. 108–118, Jul. 2001, doi: 10.2144/01311rr02.

[154] A. I. Culley, A. S. Lang, and C. A. Suttle, "Metagenomic Analysis of Coastal RNA Virus Communities," *Science*, vol. 312, no. 5781, pp. 1795–1798, Jun. 2006, doi: 10.1126/science.1127404.

[155] F. Asnicar *et al.*, "Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling," *mSystems*, vol. 2, no. 1, Feb. 2017, doi: 10.1128/mSystems.00164-16.

[156] X. Fang *et al.*, "Metagenomics-Based, Strain-Level Analysis of Escherichia coli From a Time-Series of Microbiome Samples From a Crohn's Disease Patient," *Front. Microbiol.*, vol. 9, 2018, doi: 10.3389/fmicb.2018.02559.

[157] V. Sevim *et al.*, "Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies," *Sci. Data*, vol. 6, no. 1, pp. 1–9, Nov. 2019, doi: 10.1038/s41597-019-0287-z.

[158] H. P. Browne *et al.*, "Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation," *Nature*, vol. 533, no. 7604, pp. 543–546, May 2016, doi: 10.1038/nature17645.

[159] J.-C. Lagier *et al.*, "Culturing the human microbiota and culturomics," *Nat. Rev. Microbiol.*, vol. 16, no. 9, pp. 540–550, Sep. 2018, doi: 10.1038/s41579-018-0041-0.

[160] P. D. Schloss *et al.*, "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities," *Appl. Environ. Microbiol.*, vol. 75, no. 23, pp. 7537–7541, Dec. 2009, doi: 10.1128/AEM.01541-09.

[161] J. G. Caporaso *et al.*, "QIIME allows analysis of high-throughput community sequencing data," *Nat. Methods*, vol. 7, no. 5, pp. 335–336, May 2010, doi: 10.1038/nmeth.f.303.

[162] E. Bolyen *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nat. Biotechnol.*, vol. 37, no. 8, pp. 852–857, Aug. 2019, doi: 10.1038/s41587-019-0209-9.

[163] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010, doi: 10.1093/bioinformatics/btq461.

[164] E. Stackebrand and B. M. Goebel, "Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology," *Int. J. Syst. Evol. Microbiol.*, vol. 44, no. 4, pp. 846–849, 1994, doi: 10.1099/00207713-44-4-846.

[165] T. Z. DeSantis *et al.*, "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB," *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006, doi: 10.1128/AEM.03006-05.

[166] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D590–D596, Jan. 2013, doi: 10.1093/nar/gks1219.

[167] J. R. Cole *et al.*, "Ribosomal Database Project: data and tools for high throughput rRNA analysis," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D633–D642, Jan. 2014, doi: 10.1093/nar/gkt1244.

[168] D. H. Parks *et al.*, "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 996–1004, Nov. 2018, doi: 10.1038/nbt.4229.

[169] R. C. Edgar, "Updating the 97% identity threshold for 16S ribosomal RNA OTUs," *Bioinformatics*, vol. 34, no. 14, pp. 2371–2375, Jul. 2018, doi: 10.1093/bioinformatics/bty113.

[170] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," *Nat. Methods*, vol. 13, no. 7, pp. 581–583, Jul. 2016, doi: 10.1038/nmeth.3869.

[171] R. C. Edgar, "UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing," *bioRxiv*, p. 081257, Oct. 2016, doi: 10.1101/081257.

[172] A. Amir *et al.*, "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns," *mSystems*, vol. 2, no. 2, Apr. 2017, doi: 10.1128/mSystems.00191-16.

[173] D. T. Truong *et al.*, "MetaPhlAn2 for enhanced metagenomic taxonomic profiling," *Nat. Methods*, vol. 12, no. 10, pp. 902–903, Oct. 2015, doi: 10.1038/nmeth.3589.

[174] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014, doi: 10.1186/gb-2014-15-3-r46.

[175] E. A. Franzosa *et al.*, "Species-level functional profiling of metagenomes and metatranscriptomes," *Nat. Methods*, vol. 15, no. 11, pp. 962–968, Nov. 2018, doi: 10.1038/s41592-018-0176-y.

[176] A. Bankevich *et al.*, "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[177] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinforma. Oxf. Engl.*, vol. 31, no. 10, pp. 1674–1676, May 2015, doi: 10.1093/bioinformatics/btv033.

[178] Y.-W. Wu, B. A. Simmons, and S. W. Singer, "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets," *Bioinforma. Oxf. Engl.*, vol. 32, no. 4, pp. 605–607, Feb. 2016, doi: 10.1093/bioinformatics/btv638.

[179] D. D. Kang *et al.*, "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, p. e7359, Jul. 2019, doi: 10.7717/peerj.7359.

[180] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.

[181] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinforma. Oxf. Engl.*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.

[182] P. Legendre and L. F. J. Legendre, *Numerical Ecology*. Elsevier, 2012.

[183] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[184] E. H. Simpson, "Measurement of Diversity," *Nature*, vol. 163, no. 4148, pp. 688–688, Apr. 1949, doi: 10.1038/163688a0.

[185] D. P. Faith, "Conservation evaluation and phylogenetic diversity," *Biol. Conserv.*, vol. 61, no. 1, pp. 1–10, Jan. 1992, doi: 10.1016/0006-3207(92)91201-3.

[186] P. Jaccard, "The Distribution of the Flora in the Alpine Zone.1," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912, doi: 10.1111/j.1469-8137.1912.tb05611.x.

[187] J. R. Bray and J. T. Curtis, "An Ordination of the Upland Forest Communities of Southern Wisconsin," *Ecol. Monogr.*, vol. 27, no. 4, pp. 325–349, 1957, doi: 10.2307/1942268.

[188] C. Lozupone and R. Knight, "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities," *Appl. Environ. Microbiol.*, vol. 71, no. 12, pp. 8228–8235, Dec. 2005, doi: 10.1128/AEM.71.12.8228-8235.2005.

[189] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral Ecol.*, vol. 26, no. 1, pp. 32–46, 2001, doi: 10.1111/j.1442-9993.2001.01070.pp.x.

[190] P. Legendre and M. J. Anderson, "Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments," *Ecol. Monogr.*, vol. 69, no. 1, pp. 1–24, 1999, doi: 10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

[191] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, "Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data," *PLOS ONE*, vol. 10, no. 7, p. e0129606, Jun. 2015, doi: 10.1371/journal.pone.0129606.

[192] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.

[193] I. Tapio *et al.*, "Taxon abundance, diversity, co-occurrence and network analysis of the ruminal microbiota in response to dietary changes in dairy cows," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, doi: 10.1371/journal.pone.0180260.

[194] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Front. Microbiol.*, vol. 8, p. 2224, 2017, doi: 10.3389/fmicb.2017.02224.

[195] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[196] D. Altshuler, P. Donnelly, and The International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, Oct. 2005, doi: 10.1038/nature04226.

[197] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, Jun. 2007, doi: 10.1038/nature05911.

[198] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[199] P. M. Visscher *et al.*, "10 Years of GWAS Discovery: Biology, Function, and Translation," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, Jul. 2017, doi: 10.1016/j.ajhg.2017.06.005.

[200] T. A. Manolio, "Genomewide Association Studies and Assessment of the Risk of Disease," *N. Engl. J. Med.*, vol. 363, no. 2, pp. 166–176, Jul. 2010, doi: 10.1056/NEJMra0905980.

[201] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, Oct. 2018, doi: 10.1038/s41586-018-0579-z.

[202] R. K. Linnér *et al.*, "Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences," *Nat. Genet.*, vol. 51, no. 2, pp. 245–257, Feb. 2019, doi: 10.1038/s41588-018-0309-3.

[203] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat. Commun.*, vol. 8, no. 1, pp. 1–11, Nov. 2017, doi: 10.1038/s41467-017-01261-5.

[204] G. Hemani *et al.*, "The MR-Base platform supports systematic causal inference across the human phenome," *eLife*, vol. 7, p. e34408, May 2018, doi: 10.7554/eLife.34408.

[205] G. Hemani, J. Bowden, and G. Davey Smith, "Evaluating the potential role of pleiotropy in Mendelian randomization studies," *Hum. Mol. Genet.*, vol. 27, no. R2, pp. R195–R208, 01 2018, doi: 10.1093/hmg/ddy163.

[206] B. L. Pierce and S. Burgess, "Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators," *Am. J. Epidemiol.*, vol. 178, no. 7, pp. 1177–1184, Oct. 2013, doi: 10.1093/aje/kwt084.

[207] D. Welter *et al.*, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1001–D1006, Jan. 2014, doi: 10.1093/nar/gkt1229.

[208] J. R. Bishop and P. Gagneux, "Evolution of carbohydrate antigens--microbial forces shaping host glycomes?," *Glycobiology*, vol. 17, no. 5, pp. 23R-34R, May 2007, doi: 10.1093/glycob/cwm005.

[209] E. C. Martens, H. C. Chiang, and J. I. Gordon, "Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont," *Cell Host Microbe*, vol. 4, no. 5, pp. 447–457, Nov. 2008, doi: 10.1016/j.chom.2008.09.007.

[210] S. Louca, M. Doebeli, and L. W. Parfrey, "Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem," *Microbiome*, vol. 6, no. 1, p. 41, Feb. 2018, doi: 10.1186/s40168-018-0420-9.

[211] A. López-García *et al.*, "Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences," *Front. Microbiol.*, vol. 9, p. 3010, 2018, doi: 10.3389/fmicb.2018.03010.

[212] R. D. Bjerre, L. W. Hugerth, F. Boulund, M. Seifert, J. D. Johansen, and L. Engstrand, "Effects of sampling strategy and DNA extraction on human skin microbiome investigations," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Nov. 2019, doi: 10.1038/s41598-019-53599-z.

[213] S. von Huth, L. B. Thingholm, C. Bang, M. C. Rühlemann, A. Franke, and U. Holmskov, "Minor compositional alterations in faecal microbiota after five weeks and five months storage at room temperature on filter papers," *Sci. Rep.*, vol. 9, no. 1, pp. 1–8, Dec. 2019, doi: 10.1038/s41598-019-55469-0.

[214] K. Fiedorová *et al.*, "The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery," *Front. Microbiol.*, vol. 10, 2019, doi: 10.3389/fmicb.2019.00821.

[215] A. Sczyrba *et al.*, "Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software," *Nat. Methods*, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, doi: 10.1038/nmeth.4458.

[216] M. A. Chester and M. L. Olsson, "The ABO blood group gene: a locus of considerable genetic diversity," *Transfus. Med. Rev.*, vol. 15, no. 3, pp. 177–200, Jul. 2001, doi: 10.1053/tmrv.2001.24591.

[217] L. Ségurel *et al.*, "The ABO blood group is a trans-species polymorphism in primates," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 45, pp. 18493–18498, Nov. 2012, doi: 10.1073/pnas.1210603109.

[218] V. Llaurens, A. Whibley, and M. Joron, "Genetic architecture and balancing selection: the life and death of differentiated variants," *Mol. Ecol.*, vol. 26, no. 9, pp. 2430–2448, 2017, doi: 10.1111/mec.14051.

[219] S. Sommer, "The importance of immune gene variability (MHC) in evolutionary ecology and conservation," *Front. Zool.*, vol. 2, p. 16, Oct. 2005, doi: 10.1186/1742-9994-2-16.

[220] J. L. Kubinak *et al.*, "MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection," *Nat. Commun.*, vol. 6, Oct. 2015, doi: 10.1038/ncomms9642.

[221] L. Ségurel, Z. Gao, and M. Przeworski, "Ancestry runs deeper than blood: The evolutionary history of ABO points to cryptic variation of functional importance," *Bioessays*, vol. 35, no. 10, pp. 862–867, Oct. 2013, doi: 10.1002/bies.201300030.

[222] J. G. Ruseler-van Embden, R. van der Helm, and L. M. van Lieshout, "Degradation of intestinal glycoproteins by Bacteroides vulgatus," *FEMS Microbiol. Lett.*, vol. 49, no. 1, pp. 37–41, Mar. 1989, doi: 10.1016/0378-1097(89)90338-8.

[223] A. Ali-Ahmad *et al.*, "Structural insights into a family 39 glycoside hydrolase from the gut symbiont Bacteroides cellulosilyticus WH2," *J. Struct. Biol.*, vol. 197, no. 3, pp. 227–235, 2017, doi: 10.1016/j.jsb.2016.11.004.

[224] M. L. Olsson, S. E. Santos, J. F. Guerreiro, M. A. Zago, and M. A. Chester, "Heterogeneity of the O alleles at the blood group ABO locus in Amerindians," *Vox Sang.*, vol. 74, no. 1, pp. 46–50, 1998.

[225] E. Senga *et al.*, "ABO blood group phenotypes influence parity specific immunity to Plasmodium falciparum malaria in Malawian women," *Malar. J.*, vol. 6, no. 1, p. 102, Aug. 2007, doi: 10.1186/1475-2875-6-102.

[226] G. Garratty, S. A. Glynn, R. McEntire, and Retrovirus Epidemiology Donor Study, "ABO and Rh(D) phenotype frequencies of different racial/ethnic groups in the United States," *Transfusion (Paris)*, vol. 44, no. 5, pp. 703–706, May 2004, doi: 10.1111/j.1537-2995.2004.03338.x.

[227] A. Mahajan *et al.*, "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps," *Nat. Genet.*, vol. 50, no. 11, pp. 1505–1513, 2018, doi: 10.1038/s41588-018-0241-6.

[228] W. J. Astle *et al.*, "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease," *Cell*, vol. 167, no. 5, pp. 1415-1429.e19, 17 2016, doi: 10.1016/j.cell.2016.10.042.

[229] T. J. Hoffmann *et al.*, "A large electronic-health-record-based genome-wide study of serum lipids," *Nat. Genet.*, vol. 50, no. 3, pp. 401–413, 2018, doi: 10.1038/s41588-018-0064-5.

[230] L. Jostins *et al.*, "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease," *Nature*, vol. 491, no. 7422, pp. 119–124, Nov. 2012, doi: 10.1038/nature11582.

[231] B. D. Bitarello *et al.*, "Signatures of Long-Term Balancing Selection in Human Genomes," *Genome Biol. Evol.*, vol. 10, no. 3, pp. 939–955, Mar. 2018, doi: 10.1093/gbe/evy054.

[232] K. S. Pollard *et al.*, "An RNA gene expressed during cortical development evolved rapidly in humans," *Nature*, vol. 443, no. 7108, pp. 167–172, Sep. 2006, doi: 10.1038/nature05113.

[233] S. L. Pulit *et al.*, "Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry," *Hum. Mol. Genet.*, vol. 28, no. 1, pp. 166–174, 01 2019, doi: 10.1093/hmg/ddy327.

[234] M. Ikeda *et al.*, "Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect," *Schizophr. Bull.*, vol. 45, no. 4, pp. 824–834, 18 2019, doi: 10.1093/schbul/sby140.

[235] Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, "Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia," *Mol. Autism*, vol. 8, p. 21, 2017, doi: 10.1186/s13229-017-0137-9.

[236] C. Bleuven and C. R. Landry, "Molecular and cellular bases of adaptation to a changing environment in microorganisms," *Proc. R. Soc. B Biol. Sci.*, vol. 283, no. 1841, Oct. 2016, doi: 10.1098/rspb.2016.1458.

[237] R. B. Jones *et al.*, "Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab, and mucosal samples," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, Mar. 2018, doi: 10.1038/s41598-018-22408-4.

[238] Y. He *et al.*, "Regional variation limits applications of healthy gut microbiome reference ranges and disease models," *Nat. Med.*, vol. 24, no. 10, pp. 1532–1535, Oct. 2018, doi: 10.1038/s41591-018-0164-x.

[239] H. Jeong, B. Arif, G. Caetano-Anollés, K. M. Kim, and A. Nasir, "Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation," *Sci. Rep.*, vol. 9, no. 1, pp. 1–18, Apr. 2019, doi: 10.1038/s41598-019-42227-5.

[240] S. Lemoinne *et al.*, "Fungi participate in the dysbiosis of gut microbiota in patients with primary sclerosing cholangitis," *Gut*, vol. 69, no. 1, pp. 92–102, Jan. 2020, doi: 10.1136/gutjnl-2018-317791.

[241] H. Sokol *et al.*, "Fungal microbiota dysbiosis in IBD," *Gut*, vol. 66, no. 6, pp. 1039–1048, Jun. 2017, doi: 10.1136/gutjnl-2015-310746.

[242] A. Frau *et al.*, "DNA extraction and amplicon production strategies deeply inf luence the outcome of gut mycobiome studies," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17, Jun. 2019, doi: 10.1038/s41598-019-44974-x.

[243] P. D. Donovan, G. Gonzalez, D. G. Higgins, G. Butler, and K. Ito, "Identification of fungi in shotgun metagenomics datasets," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, doi: 10.1371/journal.pone.0192898.

[244] T. Jones *et al.*, "The diploid genome sequence of Candida albicans," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 19, pp. 7329–7334, May 2004, doi: 10.1073/pnas.0401648101.

[245] J. Peter *et al.*, "Genome evolution across 1,011 Saccharomyces cerevisiae isolates," *Nature*, vol. 556, no. 7701, pp. 339–344, Apr. 2018, doi: 10.1038/s41586-018-0030-5.

[246] G. C. diCenzo and T. M. Finan, "The Divided Bacterial Genome: Structure, Function, and Evolution," *Microbiol. Mol. Biol. Rev.*, vol. 81, no. 3, Sep. 2017, doi: 10.1128/MMBR.00019-17.

[247] M. Soverini, S. Turroni, E. Biagi, P. Brigidi, M. Candela, and S. Rampelli, "HumanMycobiomeScan: a new bioinformatics tool for the characterization of the fungal fraction in metagenomic samples," *BMC Genomics*, vol. 20, no. 1, p. 496, Jun. 2019, doi: 10.1186/s12864-019-5883-y.

[248] H. Toju, A. S. Tanabe, S. Yamamoto, and H. Sato, "High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples," *PLOS ONE*, vol. 7, no. 7, p. e40863, Dec. 2012, doi: 10.1371/journal.pone.0040863.

[249] A. K. Nash *et al.*, "The gut mycobiome of the Human Microbiome Project healthy cohort," *Microbiome*, vol. 5, no. 1, p. 153, Nov. 2017, doi: 10.1186/s40168-017-0373-4.

[250] P. Bacher *et al.*, "Human Anti-fungal Th17 Immunity and Pathology Rely on Cross-Reactivity against Candida albicans," *Cell*, vol. 176, no. 6, pp. 1340-1355.e15, Mar. 2019, doi: 10.1016/j.cell.2019.01.041.

[251] M.-A. Richard *et al.*, "Sex- and age-adjusted prevalence estimates of five chronic inflammatory skin diseases in France: results of the « OBJECTIFS PEAU » study," *J. Eur. Acad. Dermatol. Venereol. JEADV*, vol. 32, no. 11, pp. 1967–1971, Nov. 2018, doi: 10.1111/jdv.14959.

[252] J. Schmitt *et al.*, "Atopic dermatitis is associated with an increased risk for rheumatoid arthritis and inflammatory bowel disease, and a decreased risk for type 1 diabetes," *J. Allergy Clin. Immunol.*, vol. 137, no. 1, pp. 130–136, Jan. 2016, doi: 10.1016/j.jaci.2015.06.029.

[253] C. Hidalgo-Cantabrana *et al.*, "Gut microbiota dysbiosis in a cohort of patients with psoriasis," *Br. J. Dermatol.*, vol. 181, no. 6, pp. 1287–1295, Dec. 2019, doi: 10.1111/bjd.17931.

[254] E. B. M. Petersen, L. Skov, J. P. Thyssen, and P. Jensen, "Role of the Gut Microbiota in Atopic Dermatitis: A Systematic Review," *Acta Derm. Venereol.*, vol. 99, no. 1, pp. 5–11, 01 2019, doi: 10.2340/00015555-3008.

[255] M. Szántó, A. Dózsa, D. Antal, K. Szabó, L. Kemény, and P. Bai, "Targeting the gut-skin axis—Probiotics as new tools for skin disorder management?," *Exp. Dermatol.*, vol. 28, no. 11, pp. 1210–1218, 2019, doi: 10.1111/exd.14016.

[256] T. Větrovský and P. Baldrian, "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, doi: 10.1371/journal.pone.0057923.

[257] J. T. Morton *et al.*, "Balance Trees Reveal Microbial Niche Differentiation," *mSystems*, vol. 2, no. 1, Feb. 2017, doi: 10.1128/mSystems.00162-16.

[258] J. S. Johnson *et al.*, "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis," *Nat. Commun.*, vol. 10, no. 1, pp. 1–11, Nov. 2019, doi: 10.1038/s41467-019-13036-1.

[259] B. J. Callahan *et al.*, "High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution," *Nucleic Acids Res.*, vol. 47, no. 18, p. e103, 10 2019, doi: 10.1093/nar/gkz569.

[260] D. Bertrand *et al.*, "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," *Nat. Biotechnol.*, vol. 37, no. 8, pp. 937–944, Aug. 2019, doi: 10.1038/s41587-019-0191-2.

[261] J. Beaulaurier *et al.*, "Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation," *Nat. Biotechnol.*, vol. 36, no. 1, pp. 61–69, 2018, doi: 10.1038/nbt.4037.

[262] A. Ruaud *et al.*, "Syntrophy via Interspecies H2 Transfer between Christensenella and Methanobrevibacter Underlies Their Global Cooccurrence in the Human Gut," *mBio*, vol. 11, no. 1, Feb. 2020, doi: 10.1128/mBio.03235-19.

[263] V. Monda *et al.*, "Exercise Modifies the Gut Microbiota with Positive Health Effects," *Oxid. Med. Cell. Longev.*, vol. 2017, 2017, doi: 10.1155/2017/3831972.

[264] J. Scheiman *et al.*, "Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism," *Nat. Med.*, vol. 25, no. 7, pp. 1104–1109, Jul. 2019, doi: 10.1038/s41591-019-0485-4.

[265] M. U. Sohail, H. M. Yassine, A. Sohail, and A. A. Al Thani, "Impact of Physical Exercise on Gut Microbiome, Inflammation, and the Pathobiology of Metabolic Disorders," *Rev. Diabet. Stud. RDS*, vol. 15, pp. 35–48, 2019, doi: 10.1900/RDS.2019.15.35.

[266] T. Vetrovsky *et al.*, "Validity of six consumer-level activity monitors for measuring steps in patients with chronic heart failure," *PloS One*, vol. 14, no. 9, p. e0222569, 2019, doi: 10.1371/journal.pone.0222569.

[267] J. E. Sasaki *et al.*, "Validation of the Fitbit wireless activity tracker for prediction of energy expenditure," *J. Phys. Act. Health*, vol. 12, no. 2, pp. 149–154, Feb. 2015, doi: 10.1123/jpah.2012-0495.

[268] E. R. Leeming, A. J. Johnson, T. D. Spector, and C. I. Le Roy, "Effect of Diet on the Gut Microbiota: Rethinking Intervention Duration," *Nutrients*, vol. 11, no. 12, p. 2862, Dec. 2019, doi: 10.3390/nu11122862.

[269] U. Nöthlings, K. Hoffmann, M. M. Bergmann, and H. Boeing, "Fitting portion sizes in a self-administered food frequency questionnaire," *J. Nutr.*, vol. 137, no. 12, pp. 2781–2786, Dec. 2007, doi: 10.1093/jn/137.12.2781.

[270] R. C. E. Bowyer *et al.*, "Use of dietary indices to control for diet in human gut microbiota studies," *Microbiome*, vol. 6, Apr. 2018, doi: 10.1186/s40168-018-0455-y.

[271] A. J. Johnson *et al.*, "Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans," *Cell Host Microbe*, vol. 25, no. 6, pp. 789-802.e5, Jun. 2019, doi: 10.1016/j.chom.2019.05.005.

[272] P. Ferretti *et al.*, "Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome," *Cell Host Microbe*, vol. 24, no. 1, pp. 133-145.e5, Jul. 2018, doi: 10.1016/j.chom.2018.06.005.

[273] M. Fulde *et al.*, "Neonatal selection by Toll-like receptor 5 influences long-term gut microbiota composition," *Nature*, vol. 560, no. 7719, pp. 489–493, Aug. 2018, doi: 10.1038/s41586-018-0395-5.

[274] Y. Gao and H. Li, "Quantifying and comparing bacterial growth dynamics in multiple metagenomic samples," *Nat. Methods*, vol. 15, no. 12, pp. 1041–1044, Dec. 2018, doi: 10.1038/s41592-018-0182-0.

[275] R. Knight *et al.*, "Best practices for analysing microbiomes," *Nat. Rev. Microbiol.*, vol. 16, no. 7, pp. 410–422, Jul. 2018, doi: 10.1038/s41579-018-0029-9.

[276] B. B. Hsu *et al.*, "Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model," *Cell Host Microbe*, vol. 25, no. 6, pp. 803-814.e5, Jun. 2019, doi: 10.1016/j.chom.2019.05.001.

[277] D. Zeevi *et al.*, "Personalized Nutrition by Prediction of Glycemic Responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015, doi: 10.1016/j.cell.2015.11.001.

[278] B. P. Abraham and E. M. M. Quigley, "Probiotics in Inflammatory Bowel Disease," *Gastroenterol. Clin. North Am.*, vol. 46, no. 4, pp. 769–782, 2017, doi: 10.1016/j.gtc.2017.08.003.

[279] J. Suez *et al.*, "Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT," *Cell*, vol. 174, no. 6, pp. 1406-1423.e16, Sep. 2018, doi: 10.1016/j.cell.2018.08.047.

[280] K. R. Amato *et al.*, "Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes," *ISME J.*, vol. 13, no. 3, pp. 576–587, Mar. 2019, doi: 10.1038/s41396-018-0175-0.

[281] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nat. Rev. Genet.*, vol. 13, no. 4, pp. 260–270, Apr. 2012, doi: 10.1038/nrg3182.

[282] E. Pasolli *et al.*, "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle," *Cell*, vol. 176, no. 3, p. 649, Jan. 2019, doi: 10.1016/j.cell.2019.01.001.

Acknowledgments

First and foremost, I want to thank Andre Franke for his guidance and his trust in my work, ideas and opinions. You recruited me to the IKMB already as a Master student and gave me the opportunity to work on exciting projects in an inspiring environment. I am truly grateful for this!

A very special thanks I want to express to Corinna Bang. Your supervision, organization skills and support made this thesis possible and most importantly thank you for all the discussions, chats and jokes, it is a pleasure to be working with you!

Many thanks also to the great people of the growing "Team Microbiome": Louise Thingholm, Lucas Moitinho e Silva, Alba Troci and Philipp Rausch. The discussions with and input from you help me a lot on a daily basis and throughout writing of this thesis. And thanks to Femke Heinsen-Groth, who supported me a lot in the early time of my PhD.

Working at the IKMB means working with a lot of amazing and fun people. I want to thank all of you for the great scientific discussions and informal chats. There are a few additional former and present colleagues I want to thank in particular: Sören Franzenburg, Matthias Hübenthal and the "cool corner" - Elisa Rosati, Montse Torres and Simonas Juzenas.

Also, a big thank you to the technical staff of the microbiome lab, Ilona Urbach, Tonio Hauptmann and Ines Wulf, and the NGS lab, Yewgenia Dolshanskaya, Sandra Greve, Melanie Schlapkohl, Melanie Vollstedt und Catharina von der Lancken. Also, I want to thank everyone working in the IKMB administration and IT department. The reliability of the work all of you do is the cornerstone to all my output.

I want to thank John Baines for his support and valuable input for my work, and his group, especially Katja Cloppenborg, Shauni Doms and Britt Hermes. I really enjoy seeing you at my weekly visits to your lab!

Thanks to Thomas Bosch in representation for the CRC1182 and the possibility to conduct my exciting projects in this amazing framework. Thanks to Hinrich Schulenburg for being part of my thesis advisory committee and thanks to Cleo Pietschke for keeping the bunch of us PhD students in check.

I had the pleasure to work and collaborate with many nice people in the preparation of this work. I want to thank in particular Christoph Schramm, Timur Liwinski, Roman Zenouzi, Jörg Heeren, Anna Worthmann, Clara John, and the people of the CFU306 in Hamburg, Martin Kummen and Johannes Hov from Oslo and Markus Lerch and Fabian Frost from Greifswald.

Zuletzt möchte ich meinen Eltern Andrea und Frank, meiner Familie und meiner Verlobten Johanna aus tiefstem Herzen danken. Eure Liebe, Unterstützung und Vertrauen geben mir die Kraft für meine Arbeit. Danke, dass ihr immer für mich da seid!

Curriculum Vitae

Personal information:

Name:	Malte Christoph Rühlemann
Date and place of birth:	28.05.1988, in Detmold, Germany
Current residence:	Hanssenstr. 24, 24106 Kiel

Education:

08/1998 - 06/2007	Gymnasium Leopoldinum, Detmold
10/2008 - 09/2011	Bachelor of Science in Life Science, Leibniz University Hannover
	Grade: 1.2
	Thesis: "Reinigung und Charakterisierung rekombinant produzierter Fragmente des phage shock protein A in *Escherichia coli*" (Grade: 1.0) Institute of Microbiology, Prof. Dr. Thomas Brüser
10/2011 - 03/2012	Master of Science Biology, Kiel University (left after 1 semester)
04/2012 - 06/2014	Master of Science in Medical Life Sciences, Kiel University
	Grade: 1.4
	Thesis: "The Intersection Between Natural Selection, Susceptibility to Chronic Inflammatory Disorders and the Human Microbiome" (Grade: 1.0) Institute of Experimental Medicine, Prof. Dr. John F. Baines
07/2014 - current	PhD student, Kiel University
	Supervisor: Prof. Dr. Andre Franke

Awards and Grants:

11/2017	ZMB Young Scientist Grant for Doctoral Students (10,000 €)

List of Publications

Published in peer-reviewed journals

1. Forster M, Szymczak S, Ellinghaus D, Hemmrich G, **Rühlemann M**, Kraemer L, Mucha S, Wienbrandt L, Stanulla M, UFO Sequencing Consortium within I-BFM Study Group, Franke A: "Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data". *Sci Rep*. 2015;5:11534.

2. Tschurtschenthaler M, Kachroo P, Heinsen FA, Adolph TE, **Rühlemann MC**, Klughammer J, Offner FA, Ammerpohl O, Krueger F, Smallwood S, Szymczak S, Kaser A, Franke A: "Paternal chronic colitis causes epigenetic inheritance of susceptibility to colitis". *Sci Rep*. 2016;6:31640.

3. Heinsen FA, Fangmann D, Müller N, Schulte DM, **Rühlemann MC**, Türk K, Settgast U, Lieb W, Baines JF, Schreiber S, Franke A, Laudes M: "Beneficial Effects of a Dietary Weight Loss Intervention on Human Gut Microbiome Diversity and Metabolism Are Not Sustained during Weight Maintenance". *Obes Facts*. 2016;9(6):379-391.

4. Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen FA, **Rühlemann MC**, Szymczak S, Holm K, Esko T, Sun J, Pricop-Jeckstadt M, Al-Dury S, Bohov P, Bethune J, Sommer F, Ellinghaus D, Berge RK, Hübenthal M, Koch M, Schwarz K, Rimbach G, Hübbe P, Pan WH, Sheibani-Tezerji R, Häsler R, Rosenstiel P, D'Amato M, Cloppenborg-Schmidt K, Künzel S, Laudes M, Marschall HU, Lieb W, Nöthlings U, Karlsen TH, Baines JF, Franke A: "Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota". *Nat Genet*. 2016;48(11):1396-1406

5. **Rühlemann MC**, Heinsen FA, Zenouzi R, Lieb W, Franke A, Schramm C: "Faecal microbiota profiles as diagnostic biomarkers in primary sclerosing cholangitis". *Gut*. 2017;66(4):753-754.

6. Bauer CR, Knecht C, Fretter C, Baum B, Jendrossek S, **Rühlemann M**, Heinsen FA, Umbach N, Grimbacher B, Franke A, Lieb W, Krawczak M, Hütt MT, Sax U: "Interdisciplinary approach towards a systems medicine toolbox using the example of inflammatory diseases". *Brief Bioinform*. 2017;18(3):479-487.

7. Worthmann A, John C, **Rühlemann MC**, Baguhl M, Heinsen FA, Schaltenberg N, Heine M, Schlein C, Evangelakos I, Mineo C, Fischer M, Dandri M, Kremoser C, Scheja L, Franke A, Shaul PW, Heeren J: "Cold-induced conversion of cholesterol to bile acids in mice shapes the gut microbiome and promotes adaptive thermogenesis". *Nat Med*. 2017;23(7):839-849.

8. Sommer F*, **Rühlemann MC***, Bang C, Höppner M, Rehman A, Kaleta C, Schmitt-Kopplin P, Dempfle A, Weidinger S, Ellinghaus E, Krauss-Etschmann S, Schmidt-Arras D, Aden K, Schulte D, Ellinghaus D, Schreiber S, Tholey A, Rupp J, Laudes M, Baines JF, Rosenstiel P, Franke A: "Microbiomarkers in inflammatory bowel diseases: caveats come with caviar". *Gut*. 2017;66(10):1734-1738.

9. **Rühlemann MC**, Degenhardt F, Thingholm LB, Wang J, Skiecevičienė J, Rausch P, Hov JR, Lieb W, Karlsen TH, Laudes M, Baines JF, Heinsen FA, Franke A: "Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci". *Gut Microbes*. 2018;9(1):68-75.

10. Zhu C, Miller M, Marpaka S, Vaysberg P, **Rühlemann MC**, Wu G, Heinsen FA, Tempel M, Zhao L, Lieb W, Franke A, Bromberg Y: "Functional sequencing read annotation for high precision microbiome analysis". *Nucleic Acids Res*. 2018;46(4):e23.

11. Baurecht H*, **Rühlemann MC***, Rodríguez E, Thielking F, Harder I, Erkens AS, Stölzl D, Ellinghaus E, Hotze M, Lieb W, Wang S, Heinsen-Groth FA, Franke A, Weidinger S: "Epidermal lipid composition, barrier integrity, and eczematous inflammation are associated with skin microbiome configuration". *J Allergy Clin Immunol*. 2018;141(5):1668-1676.e16.

12. Crusell MKW, Hansen TH, Nielsen T, Allin KH, **Rühlemann MC**, Damm P, Vestergaard H, Rørbye C, Jørgensen NR, Christiansen OB, Heinsen FA, Franke A, Hansen T, Lauenborg J, Pedersen O: "Gestational diabetes is associated with change in the gut microbiota composition in third trimester of pregnancy and postpartum". *Microbiome*. 2018;6(1):89.

13. Wang J, Kurilshikov A, Radjabzadeh D, Turpin W, Croitoru K, Bonder MJ, Jackson MA, Medina-Gomez C, Frost F, Homuth G, **Rühlemann M**, Hughes D, Kim HN; MiBioGen Consortium Initiative, Spector TD, Bell JT, Steves CJ, Timpson N, Franke A, Wijmenga C, Meyer K, Kacprowski T, Franke L, Paterson AD, Raes J, Kraaij R, Zhernakova A: "Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative". *Microbiome*. 2018;6(1):101.

14. Luedde M*, Winkler T*, Heinsen FA*, **Rühlemann MC***, Spehlmann ME, Bajrovic A, Lieb W, Franke A, Ott SJ, Frey N: "Heart failure is associated with depletion of core intestinal microbiota". *ESC Heart Fail*. 2017;4(3):282-290.

15. Thingholm L, **Rühlemann M,** Wang J, Hübenthal M, Lieb W, Laudes M, Franke A, D'Amato M: "Sucrase-isomaltase 15Phe IBS risk variant in relation to dietary carbohydrates and faecal microbiota composition". *Gut*. 2019;68(1):177-178.

16. Frost F, Kacprowski T, **Rühlemann M**, Bülow R, Kühn JP, Franke A, Heinsen FA, Pietzner M, Nauck M, Völker U, Völzke H, Aghdassi AA, Sendler M, Mayerle J, Weiss FU, Homuth G, Lerch MM: "Impaired Exocrine Pancreatic Function Associates With Changes in Intestinal Microbiota Composition and Diversity". *Gastroenterology*. 2019;156(4):1010-1015.

17. Frost F, Kacprowski T, **Rühlemann MC**, Franke A, Heinsen FA, Völker U, Völzke H, Aghdassi AA, Mayerle J, Weiss FU, Homuth G, Lerch MM: "Functional abdominal pain and discomfort (IBS) is not associated with faecal microbiota composition in the general population". *Gut*. 2019;68(6):1131-1133.

18. Liwinski T, Zenouzi R, John C, Ehlken H, **Rühlemann MC**, Bang C, Groth S, Lieb W, Kantowski M, Andersen N, Schachschal G, Karlsen TH, Hov JR, Rösch T, Lohse AW, Heeren J, Franke A, Schramm C: "Alterations of the bile microbiome in primary sclerosing cholangitis". *Gut*. 2019. [Epub ahead of print]

19. Frost F, Storck LJ, Kacprowski T, Gärtner S, **Rühlemann M**, Bang C, Franke A, Völker U, Aghdassi AA, Steveling A, Mayerle J, Weiss FU, Homuth G, Lerch MM: "A structured weight loss program increases gut microbiota phylogenetic diversity and reduces levels of Collinsella in obese type 2 diabetics: A pilot study". *PLoS One*. 2019;14(7):e0219489.

20. Thingholm LB, **Rühlemann MC**, Koch M, Fuqua B, Laucke G, Boehm R, Bang C, Franzosa EA, Hübenthal M, Rahnavard A, Frost F, Lloyd-Price J, Schirmer M, Lusis AJ, Vulpe CD, Lerch MM, Homuth G, Kacprowski T, Schmidt CO, Nöthlings U, Karlsen TH, Lieb W, Laudes M, Franke A, Huttenhower C: "Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition". *Cell Host Microbe*. 2019;26(2):252-264.e10.

21. Rausch P*, **Rühlemann M***, Hermes BM, Doms S, Dagan T, Dierking K, Domin H, Fraune S, von Frieling J, Hentschel U, Heinsen FA, Höppner M, Jahn MT, Jaspers C, Kissoyan KAB, Langfeldt D, Rehman A, Reusch TBH, Roeder T, Schmitz RA, Schulenburg H, Soluch R, Sommer F, Stukenbrock E, Weiland-Bräuer N, Rosenstiel P, Franke A, Bosch T, Baines JF: "Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms". *Microbiome*. 2019;7(1):133.

22. **Rühlemann M***, Liwinski T*, Heinsen FA*, Bang C, Zenouzi R, Kummen M, Thingholm L, Tempel M, Lieb W, Karlsen T, Lohse A, Hov J, Denk G, Lammert F, Krawczyk M, Schramm C, Franke A: "Consistent alterations in faecal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis". *Aliment Pharmacol Ther*. 2019;50(5):580-589.

23. **Rühlemann M**, Franke A: "Editorial: gut microbial profile associated with primary sclerosing cholangitis-what is new and how do we progress from here? Authors' reply". *Aliment Pharmacol Ther*. 2019;50(5):606-607.

24. Esser D, Lange J, Marinos G, Sieber M, Best L, Prasse D, Bathia J, **Rühlemann MC**, Boersch K, Jaspers C, Sommer F: "Functions of the Microbiota for the Physiology of Animal Metaorganisms". *J Innate Immun*. 2019;11(5):393-404.

25. **Rühlemann MC**, Solovjeva MEL, Zenouzi R, Liwinski T, Kummen M, Lieb W, Hov JR, Schramm C, Franke A, Bang C: "Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species". *Gut*. 2019. [Epub ahead of print]

26. Herken J, Bang C, **Rühlemann MC**, Finke C, Klag J, Franke A, Prüss H: "Normal gut microbiome in NMDA receptor encephalitis". *Neurol Neuroimmunol Neuroinflamm*. 2019;6(6).

27. Tueffers L, Barbosa C, Bobis I, Schubert S, Höppner M, **Rühlemann M**, Franke A, Rosenstiel P, Friedrichs A, Krenz-Weinreich A, Fickenscher H, Bewig B, Schreiber S, Schulenburg H: "Pseudomonas aeruginosa populations in the cystic fibrosis lung lose susceptibility to newly applied β-lactams within 3 days". *J Antimicrob Chemother*. 2019;74(10):2916-2925.

28. von Huth S, Thingholm LB, Bang C, **Rühlemann MC**, Franke A, Holmskov U: "Minor compositional alterations in faecal microbiota after five weeks and five months storage at room temperature on filter papers". *Sci Rep*. 2019;9(1):19008.

29. Frost F, Kacprowski T, **Rühlemann M**, Bang C, Franke A, Zimmermann K, Nauck M, Völker U, Völzke H, Biffar R, Schulz C, Mayerle J, Weiss FU, Homuth G, Lerch MM: "Helicobacter pylori infection associates with fecal microbiota composition and diversity". *Sci Rep*. 2019;9(1):20100.

30. Seybold H, Demetrowitsch T, Hassani MA, Szymczak S, Reim E, Haueisen J, Lübbers L, **Rühlemann M**, Franke A, Schwarz K, Stukenbrock E: "Fungal pathogen induces systemic susceptibility and systemic shifts in wheat metabolome and microbiome composition". *Nature Communications. 2020. Manuscript accepted.*

Manuscripts submitted to peer-reviewed journals

1. **Rühlemann MC**, Hermes BM, Bang C, Doms S, Moitinho-Silva L, Thingholm LB, Frost F, Degenhardt F, Wittig M, Kässens J, Weiss FU, Peters A, Neuhaus K, Völker U, Völzke H, Homuth G, Laudes M, Lieb W, Haller D, Lerch MM, Baines J, Franke A: "ABO histo-blood groups influence gut microbiome, with causal relationship between *Bacteroides* and inflammatory bowel disease". *Manuscript under review in Nature Genetics.*

2. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, Le Roy C, Raygoza Garay JA, Finnicum C, Liu X, Zhernakova D, Bonder MJ, Hansen T, Frost F, **Rühlemann M**, Turpin W, Young Moon J-Y, Kim H-N, Lüll K, Barkan E, Shah S, Fornage M, Szopinska-Tokov J, Wallen Z, Borisevich D, Agreus L, Andreasson A, Bang C, Bedrani L, Bell J, Bisgaard H, Boehnke M, Boomsma D, Burk R, Claringbould A, Croitoru K, Davies G, van Duijn C, Duijts L, Falony G, Fu J, van der Graaf A, Hansen T, Homuth G, Hughes D, Ijzerman R, Jackson M, Jaddoe V, Joossens M, Jørgensen T, Keszthelyi D, Knight R, Laakso M, Laudes M, Launer L, Lieb W, Lusis A, Masclee A, Moll H, Mujagic Z, Qi Q, Rothschild D, Shin H, Sørensen S, Steves C, Thorsen J, Timpson N, Tito R, Vieira-Silva S, Voelker U, Völzke H, Võsa U, Wade K, Walter S, Watanabe K, Weiss S, Weiss F, Weissbrod O, Westra H-J, Willemsen G, Payami H, Jonkers D, Arias Vasquez A, de Geus E, Meyer K, Stokholm J, Segal E, Org E, Wijmenga C, Kim H-L, Kaplan R, Spector T, Uitterlinden A, Rivadeneira F, Franke A, Lerch M, BIOS Consortium , Sann S, D'Amato M, Pedersen O, Paterson A, Kraaij R, Raes J, Zhernakova S: "Genetics of human gut microbiome composition". *Manuscript under review in Nature Genetics*.

3. Hughes D, Bacigalupe R, Wang J, **Rühlemann M**, Tito R, Falony G, Joossens M, Vieira-Silva S, Henckaerts L, Rymenans L, Verspecht C, Ring S, Franke A, Wade K, Raes J, Timpson N: "Genome-wide associations of human gut microbiome variation and implications for causal inference analyses". *Manuscript under review in Nature Microbiology.*

4. Crusell MKW, Hansen TH, Nielsen T, Allin KH, **Rühlemann MC**, Damm P, Vestergaard H, Rørbye C, Jørgensen NR, Christiansen OB, Heinsen F-A, Franke A, Hansen T, Lauenborg J, Pedersen O: "Offspring of women with gestational diabetes have an aberrant gut microbiota". *Manuscript under consideration in Frontiers in Cellular and Infection Microbiology.*

5. Kaleviste E, **Rühlemann M**, Kärner J, Haljasmägi L, Tserel L, Org E, Trebušak Podkrajšek K, Battelino T, Bang C, Franke A, Peterson P, Kiesand K: "IL-22 deficiency in APECED is associated with impaired barrier function and microbial alterations in oral cavity". *Manuscript under review in Frontiers in Immunology.*

6. Jaspers C, Weiland-Bräuer N, **Rühlemann MC**, Baines JF, Schmitz RA, Reusch TBH: "Microbiota differences of native and invasive gelatinous zooplankton organisms in a low saline environment". *Manuscript under review in Science of the Total Environment.*

7. Liwinski T, Casar C, **Rühlemann M**, Bang C, Sebode M, Hohenester, S, Denk G, Lieb W, Lohse A, Franke A, Schramm C: "Disease specific enteric microbial alterations with depletion of Bifidobacterium in autoimmune hepatitis". *Manuscript under review in Alimentary Pharmacology & Therapeutics*.

Declaration

I hereby declare,

I. that apart from my supervisor's guidance, the content and design of this thesis is completely my own work. Contributions of other authors are listed in the following section.

II. this thesis has not been submitted either partially or completely as part of a doctoral degree to another examining institution. No other materials are published or submitted for publication than indicated in this thesis.

III. this thesis was prepared in compliance with the "Rules of Good Scientific Practice" of the German Research Foundation (DFG).

IV. that I did not have an academic degree revoked.

Kiel, _____

Malte Christoph Rühlemann

Author Contributions:

Chapter 1: Introduction

Review of literature and introduction of concepts relevant concepts fully written by doctoral candidate Malte Christoph Rühlemann.

Chapter 2: Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms

Article A: Rausch*, Rühlemann* *et al.* (2019): "Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms"

Conception and design of the study: Philipp Rausch, Philip Rosenstiel, Thomas Bosch, John F. Baines

Data analysis: Philipp Rausch, Malte Christoph Rühlemann

Interpretation of results and writing of the manuscript: Philipp Rausch, Malte Christoph Rühlemann, Britt Hermes, Shauni Doms, John F. Baines

Generation and interpretation of host-specific data and intellectual input: Philipp Rausch, Malte Christoph Rühlemann, Tal Dagan, Katja Dierking, Hanna Domin, Shauni Doms, Sebastian Fraune, Jakob von Frieling, Ute Hentschel, Femke-Anouska Heinsen, Britt Hermes, Marc Höppner, Martin Jahn, Cornelia Jaspers, Kohar Annie B. Kissoyan, Daniela Langfeldt, Ateequr Rehman, Thorsten B. H. Reusch, Thomas Roeder, Ruth A. Schmitz, Hinrich Schulenburg, Ryszard Soluch, Felix Sommer, Eva Stukenbrock, Nancy Weiland-Bräuer, Philip Rosenstiel, Andre Franke, Thomas Bosch, John F. Baines

All authors read and approved the final manuscript.

Chapter 3: Host-genetic influence on the human intestinal microbiome

Article B: Rühlemann *et al.* (2018): "Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci".

Conception and design of the study: Andre Franke, John F. Baines, Tom H. Karlsen, Malte Christoph Rühlemann

Data analysis: Malte Christoph Rühlemann

Interpretation of results and writing of the manuscript: Malte Christoph Rühlemann, Frauke Degenhardt, Louise B. Thingholm, Philipp Rausch

All authors read and approved the final manuscript.

Article C: Rühlemann *et al.* (under review): "ABO histo-blood groups influence gut microbiome, with causal relationship between *Bacteroides* and inflammatory bowel disease"

Conception and design of the study: Andre Franke, John F. Baines, Markus Maximilian Lerch, Dirk Haller

Genotype and phenotype data collection: Georg Homuth, Matthias Laudes, Wolfgang Lieb, Uwe Völker, Henry Völzke, Annette Peters

Data quality control and curation: Frauke Degenhardt, Fabian Frost, Henry Völzke

Microbiome sample preparation, data generation and curation: Corinna Bang, Malte Christoph Rühlemann, Klaus Neuhaus, Frank Ulrich Weiß

ABO blood group inference: Michael Wittig, Malte Christoph Rühlemann

Implementation of statistical models and (meta-)analysis: Malte Christoph Rühlemann, Shauni Doms, Jan Kässens

Curation and interpretation of results: Malte Christoph Rühlemann, Corinna Bang, Britt Marie Hermes, Lousie Bruun Thingholm, Lucas Moitinho-Silva

Writing of manuscript draft: Malte Christoph Rühlemann, Britt Marie Hermes, Shauni Doms, Corinna Bang, Andre Franke, John F Baines

All authors read and approved the final manuscript.

Chapter 4: Disease-associated changes in the human microbiome

Article D: Rühlemann *et al.* (2017): "Faecal microbiota profiles as diagnostic biomarkers in primary sclerosing cholangitis".

Conception and design of the study: Andre Franke, Christoph Schramm

Patient data and material: Christoph Schramm, Roman Zenouzi, Wolfgnag Lieb

Data analysis: Malte Christoph Rühlemann, Femke-Anouska Heinsen, Andre Franke

Writing of the manuscript: Andre Franke, Chistoph Schramm, Malte Christoph Rühlemann, Femke-Anouska Heinsen

All authors read and approved the final manuscript.

Article E: Rühlemann *et al.* (2019): "Consistent alterations in feacal microbiomes of patients with primary sclerosing cholangitis independent of associated colitis".

Conception and design of the study: Andre Franke, Christoph Schramm

Acquisition and curation of patient samples and data: Roman Zenouzi, Martin Kummen, Johannes Hov, Wolfgang Lieb, Marie Tempel, Tom Karlsen, Ansgar W. Lohse, Gerald Denk, Frank Lammert, Marcin Krawczyk

Sample processing and sequencing: Femke-Anouska Heinsen

Data analysis: Malte Christoph Rühlemann, Timur Liwinski

Interpretation of results and writing of the manuscript: Malte Christoph Rühlemann, Femke-Anouska Heinsen, Timur Liwinski, Corinna Bang, Roman Zenouzi, Martin Kummen, Louise Thingholm, Johannes Hov, Christoph Schramm, Andre Franke

All authors read and approved the final manuscript.

Article F: Rühlemann *et al.* (2019): "Gut mycobiome of primary sclerosing cholangitis patients is characterised by an increase of *Trichocladium griseum* and *Candida* species".

Conception and design of the study: Andre Franke, Corinna Bang

Funding of the study: Andre Franke, Christoph Schramm

Acquisition and curation of patient samples and data: Roman Zenouzi, Christoph Schramm, Wolfgang Lieb

Sample processing and sequencing: Corinna Bang, Miriam Emmy Leni Solovjeva

Data analysis: Malte Christoph Rühlemann

Interpretation of results and writing of the manuscript: Malte Christoph Rühlemann, Miriam Emmy Leni Solovjeva, Johannes R. Hov, Christoph Schramm, Andre Franke, Corinna Bang

All authors read and approved the final manuscript.

Article G: Baurecht*, Rühlemann* *et al.* (2018): "Epidermal lipid composition, barrier integrity, and eczematous inflammation are associated with skin microbiome configuration".

Conception and design of the study: Stephan Weidinger

Sequencing data processing: Malte Christoph Rühlemann

Statistical analysis: Hansjörg Baurecht, Malte Christoph Rühlemann

Interpretation of results and writing of the manuscript: Hansjörg Baurecht, Malte Christoph Rühlemann, Elke Rodriguez

All authors gave additional input for interpretation of the findings, read and approved the final manuscript.

Chapter 5: Discussion

Discussion of results and outlook on future challenges fully written by doctoral candidate Malte Christoph Rühlemann.

Appendix

Appendix A: Article supplements

This section contains supplemental methods and results of the **Articles A**, **C**, **E**, and **G**. Large tables not printable in A4 format are excluded from the section. All articles, including the complete supplements, can be found on the CD attached to this thesis.

Appendix A.1: Supplement of Article A

Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms

Philipp Rausch[1,2,3,†,*], Malte Rühlemann[4,†], Britt M. Hermes[1,2,5], Shauni Doms[1,2], Tal Dagan[6], Katja Dierking[7], Hanna Domin[8], Sebastian Fraune[8], Jakob von Frieling[9], Ute Hentschel[10,11], Femke-Anouska Heinsen[4], Marc Höppner[4], Martin T. Jahn[10], Cornelia Jaspers[11,12], Kohar Annie B. Kissoyan[7], Daniela Langfeldt[6], Ateeqr Rehman[4], Thorsten B. H. Reusch[11,12], Thomas Roeder[9], Ruth A. Schmitz[6], Hinrich Schulenburg[7], Ryszard Soluch[6], Felix Sommer[4], Eva Stukenbrock[13,14], Nancy Weiland-Bräuer[6], Philip Rosenstiel[4], Andre Franke[4], Thomas Bosch[8], John F. Baines[1,2,*]

Affiliations:

[1] Evolutionary Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany

[2] Institute for Experimental Medicine, Kiel University, Kiel, Germany

[3] Laboratory of Genomics and Molecular Biomedicine, Department of Biology University of Copenhagen, Copenhagen Ø, Denmark

[4] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

[5] Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany

[6] Institute of General Microbiology, Kiel University, Kiel, Germany

[7] Department of Evolutionary Ecology and Genetics, Zoological Institute, Kiel University, Kiel, Germany

[8] Zoological Institute, Kiel University, Kiel, Germany

[9] Molecular Physiology, Zoological Institute, Kiel University, Kiel, Germany

[10] Marine Ecology, Research Unit Marine Symbioses, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

[11] Kiel University, Kiel, Germany

[12] Marine Ecology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

[13] Environmental Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany

[14] Environmental Genomics, Botanical Institute, Kiel University, Kiel, Germany

† Authors contributed equally

* Corresponding authors: Philipp Rausch (philipp.rausch@bio.ku.dk), John F. Baines (baines@evolbio.mpg.de)

Supplementary Material:

MEGAN [1] uses the information from all reads generated in the sequencing. Kraken uses a user-specified library of genomes as a database, where the records consist of a *k*-mer and the lowest common ancestor (LCA) whose genomes contain the k-mer. By querying the database for each *k*-mer in a sequence, and then using the resulting set of LCA taxa an appropriate label for the sequence can be determined. This drastically speeds up the classification process, but requires absolute matches which sacrifices sensitivity. MEGAN first needs a preprocessing step where reads are compared against a database (usually the non-redundant NCBI GeneBank database) using a BLAST algorithm or another comparison tools to search the top hits for each short read (here DIAMOND [2]). MEGAN then uses the NCBI taxonomy of the matched database to classify the sequences using a LCA algorithm, where reads are assigned to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. MEGAN has the advantage that it simultaneously performs a taxonomical and functional classification of the reads simultaneously due to the inherent information of the database hits.

Supplementary Figures:

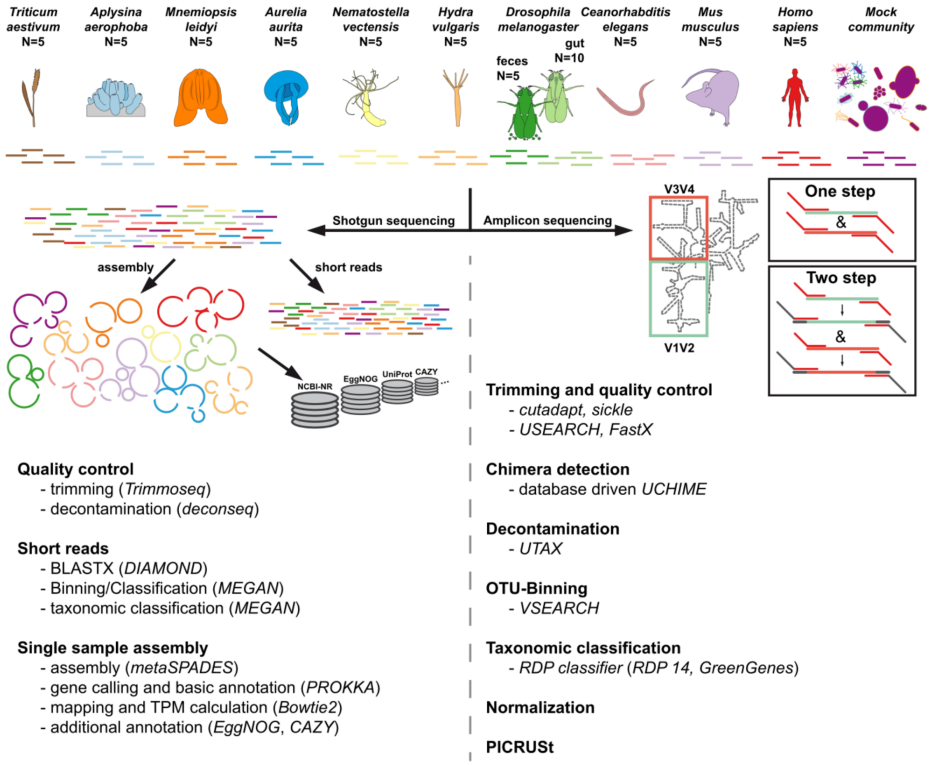


Figure S1: Visual abstract of the experiment, comparing different host organisms of the CRC 1182 investigated via different shotgun und 16S rRNA amplicon based methods (for final sample sizes see Table S1).

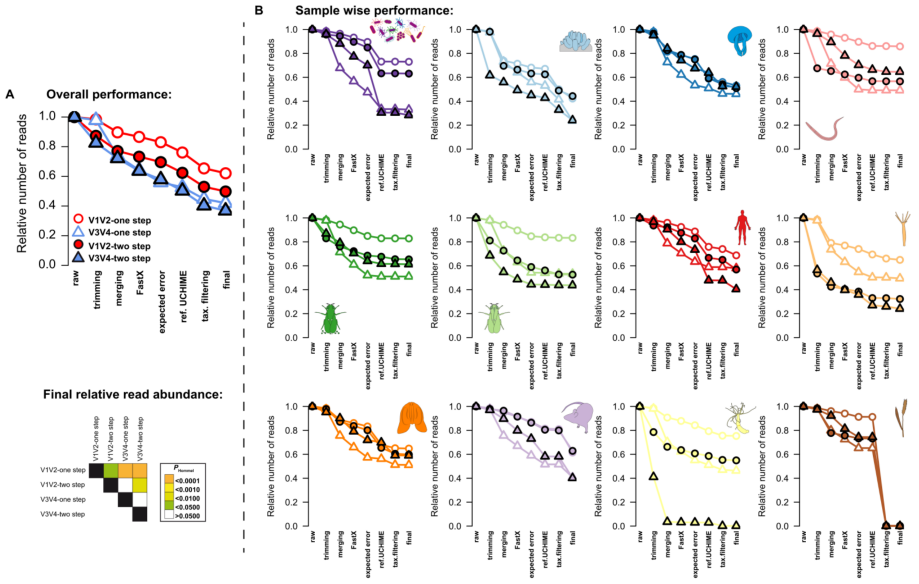


Figure S2: Average relative read count of the different 16S rRNA gene amplicon techniques (one step/two step, V1V2/V3V4) after each QC step employed in this study (for details please refer to the Methods section). (**A**) Average relative abundances of reads throughout the QC pipeline showing the significantly highest number of preserved reads in the V1V2-one step protocol (pairwise comparisons via pairwise *t*-Tests and Hommel *P*-value adjustment, significance levels are indicated by color). (**B**) Single plots display the average relative read counts across the different QC steps for each sample/host type (mock community, *A. aerophoba*, *A. aurita*, *C. elegans*, *D. melanogaster* feces, *D. melanogaster* gut tissue, *H. sapiens*, *H. vulgaris*, *M. leidyi*, *M. musculus*, *N. vectensis*, *T. aestivum*).

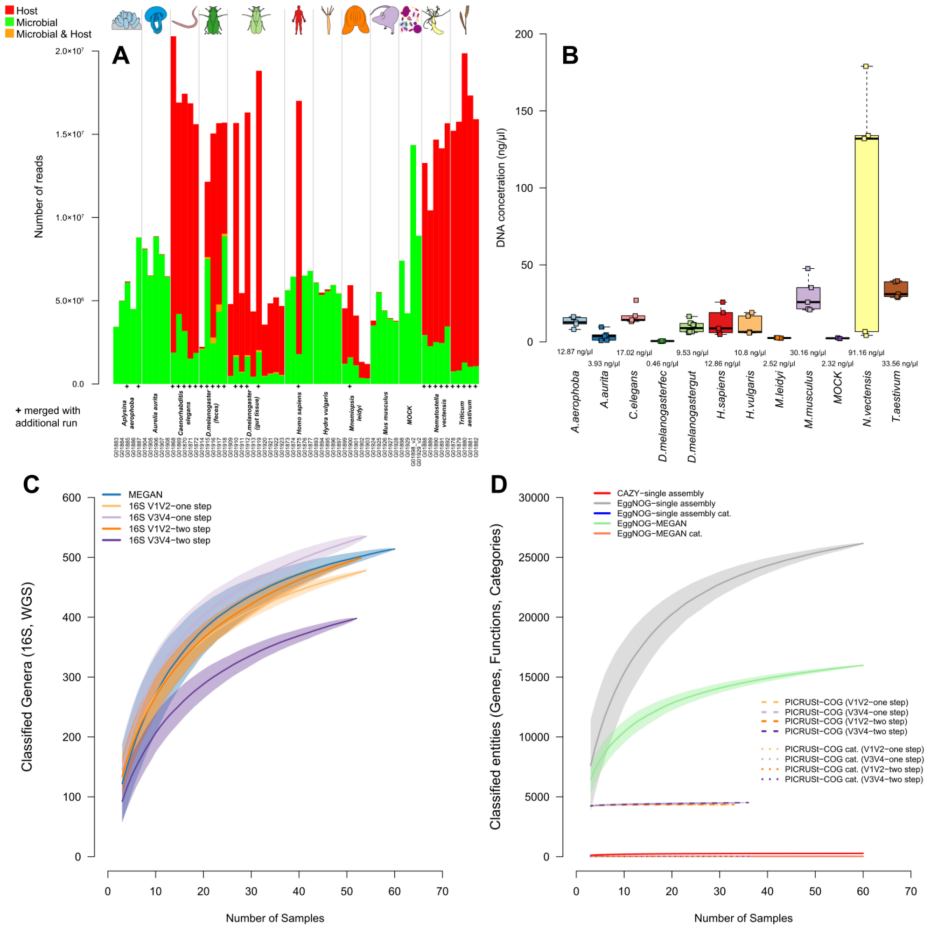


Figure S3: (A) Barplot displays the number of reads detected to be of host- or microbial origin for each sample. Resequenced samples are marked with a "+". **(B)** Average DNA concentration of samples (Qubit measurements). Collectors curves based on 1000 random re-samplings samples to see saturation of the number of **(C)** genera derived from shotgun and 16S rRNA gene amplicon techniques, as well as the number genes and functions **(D)** as derived from different shotgun based annotations and imputed functions (PICRUSt). Shading indicates standard deviations of the random samplings at each step (for sample sizes see Table S1).

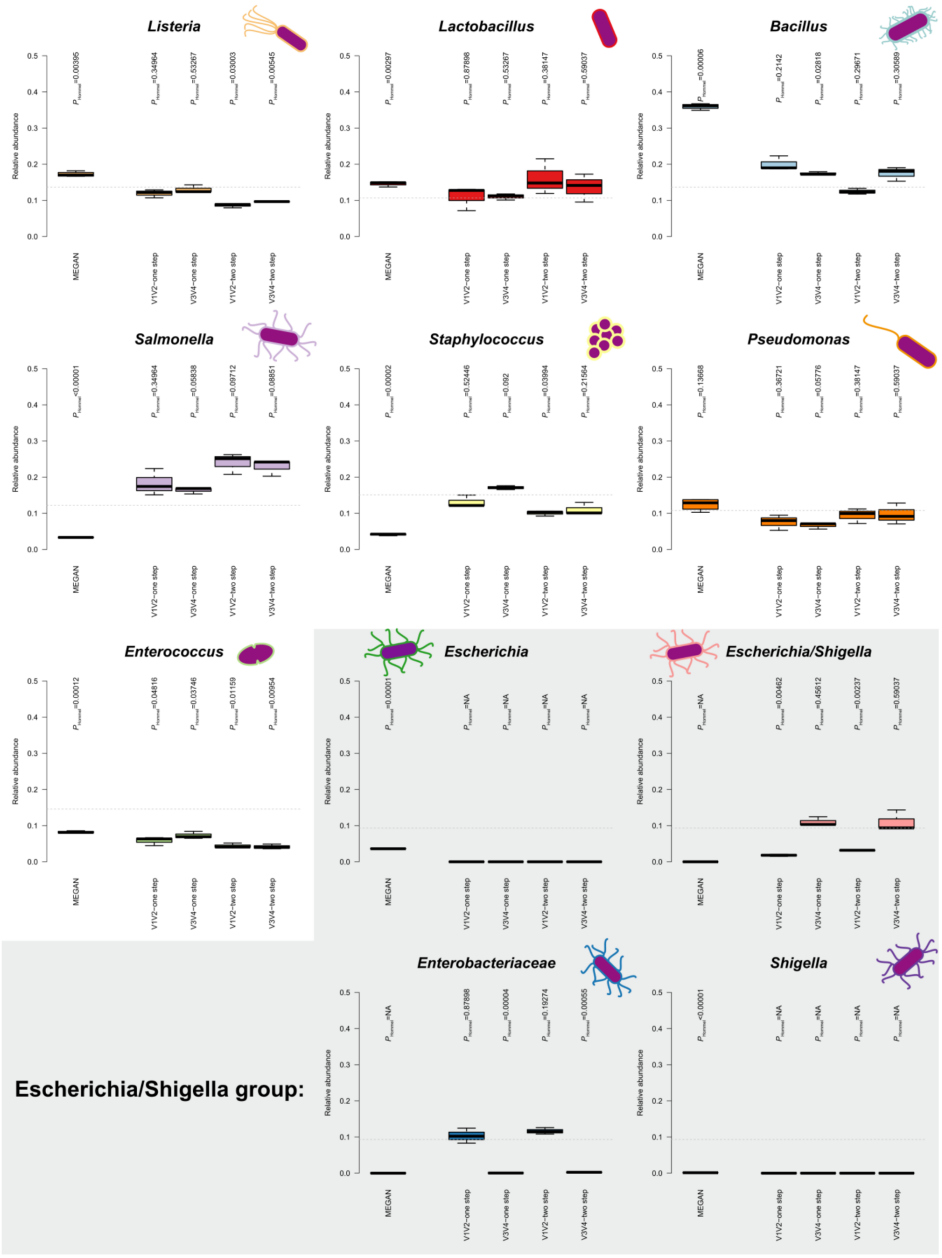

Listeria
Lactobacillus
Bacillus
Salmonella
Staphylococcus
Pseudomonas
Enterococcus
Escherichia
Escherichia/Shigella
Escherichia/Shigella group:
Enterobacteriaceae
Shigella
Relative abundance
MEGAN
V1V2-one step
V3V4-one step
V1V2-two step
V3V4-two step

Figure S4: Comparison of relative bacterial abundances in mock community samples to the expected relative abundances (dashed line) via one-sample Wilcoxon test (two-sided). Abundances are derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=4, $N_{V1V2\text{-}one\ step}$=3, $N_{V1V2\text{-}two\ step}$=3, $N_{V3V4\text{-}one\ step}$=3, and $N_{V3V4\text{-}two\ step}$=3. *P*-values are corrected for multiple testing by Hommel *P*-value adjustment for each technique.

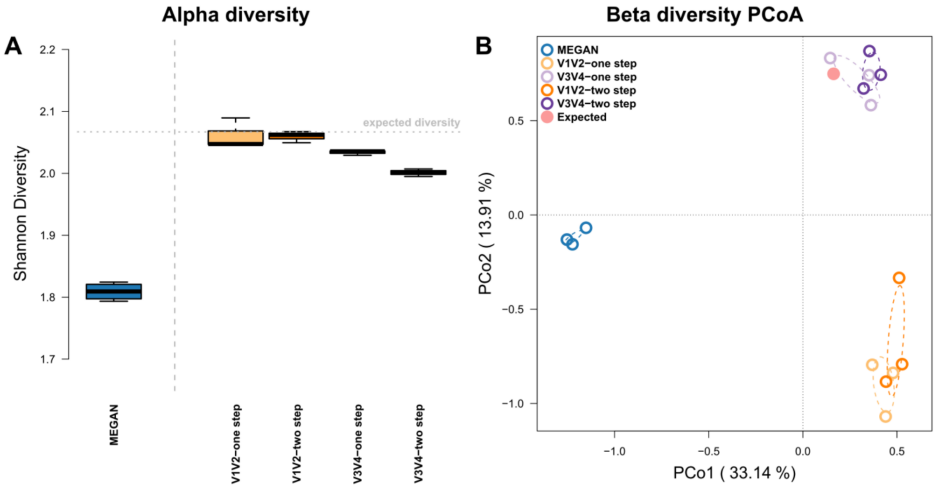


Figure S5: (A) Bacterial alpha diversity (Shannon H) derived via different techniques in comparison to the expected community diversity (dotted line). **(B)** Principle Coordinate Analyses (PCoA) of mock community compositions based on the different shotgun and 16S rRNA gene amplicon techniques focusing on shared occurrences (Jaccard). Ellipses represent standard deviations of points within the respective groups. Sample sizes for the different approaches are $N_{shotgun}$=4, $N_{V1V2\text{-}one\ step}$=3, $N_{V1V2\text{-}two\ step}$=3, $N_{V3V4\text{-}one\ step}$=3, and $N_{V3V4\text{-}two\ step}$=3.

Supplemental single host analyses:

Summary *Aplysina aerophoba* (Nardo, 1843):

Aplysina aerophoba is a Mediterranean-Atlantic member of the *Verongida*, which lives at light-exposed sites at depths from 5 to 15 m. *A. aerophoba*, as many demosponges, hosts a dense microbial community localized extracellularly in the mesohyl matrix which contributes to around one third of the sponge biomass [3, 4]. This sponge species serves as a model system to study basal host-microbial interactions and is further known for its biotechnological potential [5].

***Methods**:* Mediterranean *A. aerophoba* specimens were sampled offshore Girona, Spain, by scuba diving. Sponge specimens were rinsed with sterile sea water, fixed in RNAlater, and stored at -80°C. DNA was extracted from tissue via the Fast DNA Spin Kit for Soil (MP) with an additional ethanol precipitation step. No enrichment for bacterial DNA was used.

***Results**:* We see the highest alpha diversity in the shotgun profiles derived by MEGAN (Figure S6, left panel). The principle coordinate analysis of the beta diversities show a clear separation of shotgun vs amplicon-sequenced samples based on abundance and co-occurrence of genera (Figure S6, middle panel) and displays clustering of MEGAN based community profiles. Among the genera profiles derived from the 16S rRNA gene, there is a clear separation by variable region sequenced (*i.e.* V1V2 or V3V4) but less by amplification method. However, some samples based on the V3V4-two step method are clustering closely with MEGAN derived profiles (Figure S6, middle panel; Table S4), which also display a noticeable higher community variation (Figure S6, right panel).

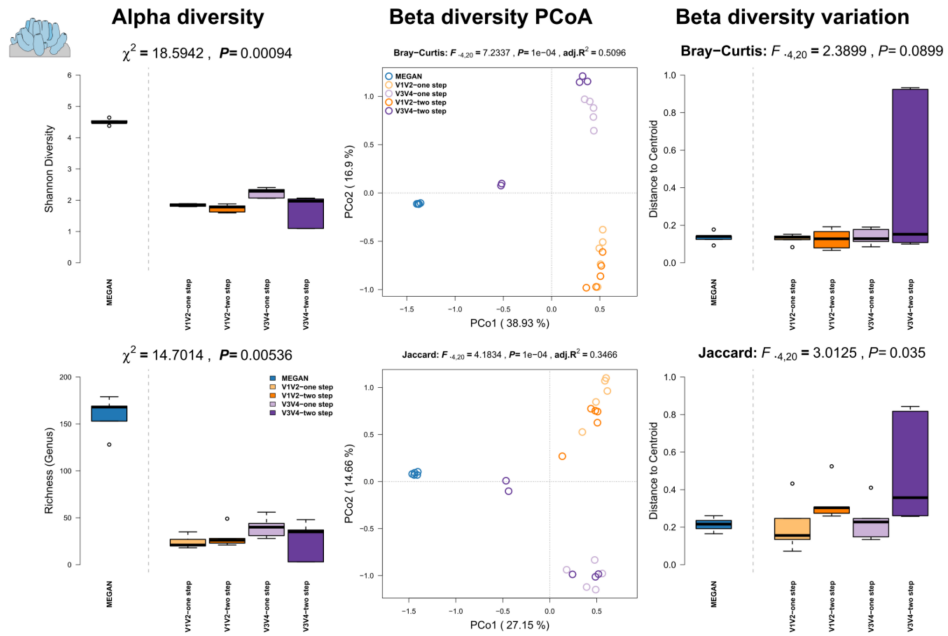


Figure S6: Comparison of community alpha diversity, composition, and variability of *Aplysina aerophoba* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test [6] (alpha diversity), PERMANOVA [7, 8] (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step) and MEGAN based classification (short reads). Sample sizes for the

different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Aurelia aurita* (Linnaeus, 1758):

The moon jelly (*Aurelia aurita*) is a widely distributed pelagic scyphozoan found in almost all warm and temperate waters of coastal zones worldwide and is recognized as a key player in marine ecosystems [9]. Due to its ability to tolerate a wide range of environmental conditions, especially temperature and salinity, and its highly diverse food spectrum, it successfully colonizes different environments and often causes jellyfish blooms around the world [10].

***Methods**:* Individual *A. aurita* medusae (mean umbrella diameter 23 cm, N=5) were sampled from one location in the Eckernförder Bight, Baltic Sea (54.462654 N, 9.842743 E) in June 2016 by using a dip net. The animals were transported immediately to the laboratory, washed thoroughly with sterile filtered artificial seawater (ASW) to remove non-associated microbes. Pieces (2 × 2 cm) of the umbrella were cut out with a sterile scalpel. Dissociation of tissues was performed overnight at 4°C with 1 mg/mL collagenase (Sigma-Aldrich, St. Louis/USA). Homogenates were filtered through 10 µm Nylon gaze, followed by adding 0.1% IGEPAL CA-630 (Sigma-Aldrich, St. Louis/USA) and centrifugation of samples for 25 min at 300 × g at 4 °C. Supernatant including the prokaryotic fraction was centrifuged 5 min at 7,500 × g. DNA was extracted using the Wizard genomic purification kit (Promega, Madison, WI, USA). Pellets from eukaryotic/prokaryotic cell separation were homogenized in 480 µl 50mM EDTA and incubated at 37°C for 30 min after the addition of 10 mg/mL lysozyme (Carl Roth, Karlsruhe/Germany) and 60 Units Proteinase K (Life Technologies, Darmstadt/Germany). The remaining preparation steps were performed according to the manufacturer's protocol.

***Results**:* Alpha diversity estimates are mostly comparable between shotgun- and 16S rRNA amplicon -based community profiles (Figure S7, left panel). MEGAN however, detects a lower number of genera compared to amplicon techniques. The 16S rRNA gene amplicon profiles are homogeneous and only show a slightly higher diversity in the V3V4 based community profiles. The principle coordinate analyses show a relatively close clustering of MEGAN and the V1V2 profiles which together are separated from the V3V4 profiles based on shared abundances of genera. However, there is a clear separation among the shotgun- and two amplicon-derived profiles when we only consider the presence of bacterial genera among samples (Figure S7, middle panel; Table S4). Within the 16S rRNA gene amplicon profiles, there is a separation by variable region based on abundance differences between samples (*i.e.* V1V2 or V3V4). When we focus on community variation we see a slightly higher variation in beta diversity in one-step PCR protocols compared to the two-step PCR protocols and MEGAN, although not significant (Figure S7, right panel).

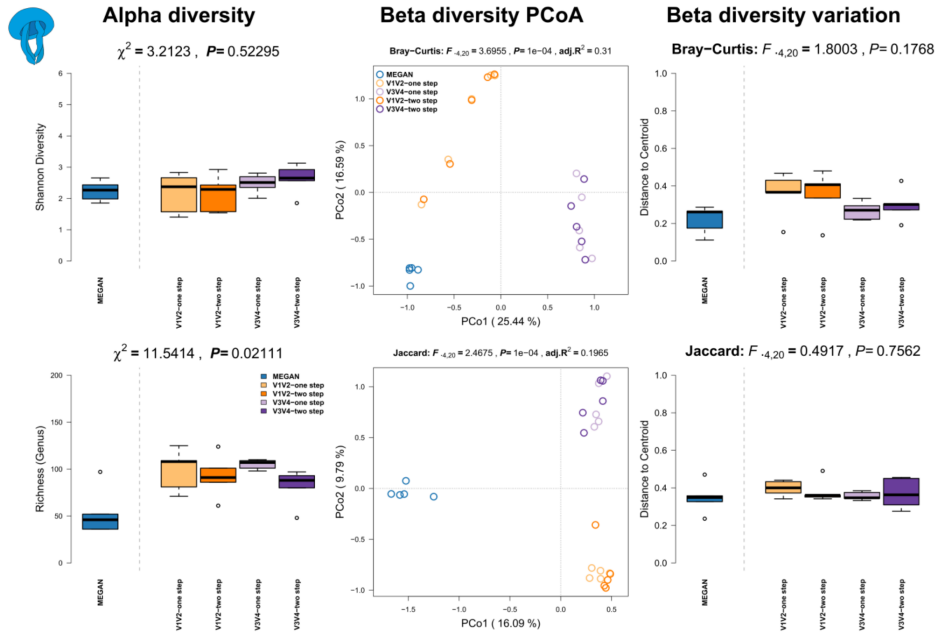


Figure S7: Comparison of community alpha diversity, composition, and variability of *Aurelia aurita* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Caenorhabditis elegans* (Maupas, 1900):

The nematode *Caenorhabditis elegans* provides a multitude of experimental advantages, such as small size, large-scale culturing, short generation time, transparency, genetic tractability, and thus, it has become one of the most widely used model organisms in biological research. In the laboratory *C. elegans* is almost exclusively cultivated mono-axenically on its food bacterium *E. coli* OP50. Therefore, almost all of the numerous studies with this nematode ignore a potential influence of the microbiome. In contrast, natural *C. elegans* are associated with a wide range of different microorganisms. Hence, wild caught and naturally colonized worms are of great interest to study the influence of a natural bacterial flora on nematode biology and fundamental developmental and genetic pathways.

***Methods**: C. elegans* were isolated directly from compost (n=2) and slugs found on the same compost (n=3) in the botanical garden in Kiel, Northern Germany (54°200N and 10°060E) in 2012, as previously described [11-13]. The samples were frozen in 30% glycerol-TSB and stored at -80°C. Worm samples were thawed and washed three times with M9 buffer with 0.05% Triton X-100 prior to DNA isolation [14]. DNA extractions from worm samples and a negative control (nuclease free water) were performed using the CTAB (Cetyl Trimethyl Ammonium Bromide) protocol as previously described [11, 13, 15]. No enrichment procedure for bacterial DNA was used.

***Results**:* We can observe no differences in alpha diversity between the shotgun- and 16S rRNA amplicon based community profiles (Figure S8, left panel). Communities further do not differ significantly between methods according to the bacterial abundances, but do slightly between MEGAN and 16S rRNA based techniques when we only consider the shared presence of genera (Figure S8, middle panel; Table S4). Furthermore, the different techniques do not differ in the amount of community variation.

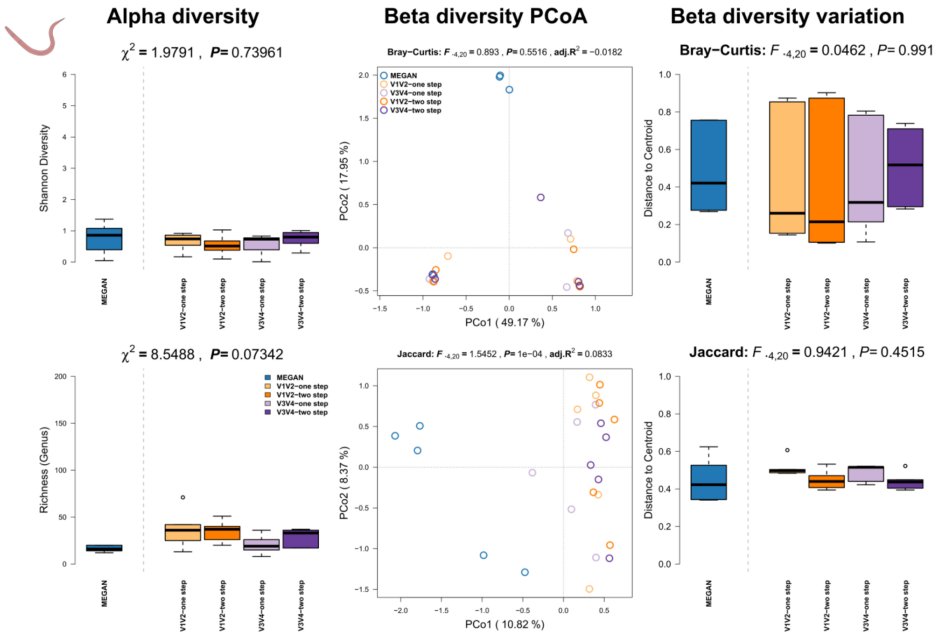


Figure S8: Comparison of community alpha diversity, composition, and variability of *Caenorhabditis elegans* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one

step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Drosophila melanogaster* (Meigen, 1830):

Drosophila melanogaster is one of the best established model organisms for genetic and developmental investigations with a plethora of tools and protocols available for genetic, physiological and neurological manipulations and was the first metazoan genome to be sequenced [16]. Due to its rather simple metagenome, its easy husbandry, and well established tools, *D. melanogaster* is becoming a model system for experimental microbial community analyses as well [17].

***Methods**: D. melanogaster* (strain: w^{1118}) were sampled in different ways. Five samples of 5 and 10 female flies each were dissected as previously described and fecal spots of five independent culture vials were sampled and extracted with the MOBIO Soil Kit [18] to contrast the tissue associated and fecal microbial communities of *D. melanogaster*. No enrichment for bacterial DNA was used.

***Results**:* Among *D. melanogaster* samples we observe a lower alpha diversity in the shotgun profiles compared to amplicon based estimates in fecal and gut tissue samples (Figure S9, Figure S10, left panels). Community complexity, as measured by Shannon H index, is slightly higher for V1V2-two step compared to the other methods for amplicon generation. Genus richness however is highest in V3V4- one step derived samples. Analysis of beta diversities shows only little differentiation between 16S rRNA gene amplicon- and shotgun-derived profiles based on differences in abundance and more pronounced when only the presence of genera is considered (Figure S9, Figure S10, middle panels; Table S4). However, community variability is only slightly influenced by sequencing technique and shows a higher community variability of bacterial abundances in gut samples analyzed via MEGAN (Figure S9, Figure S10, right panels).

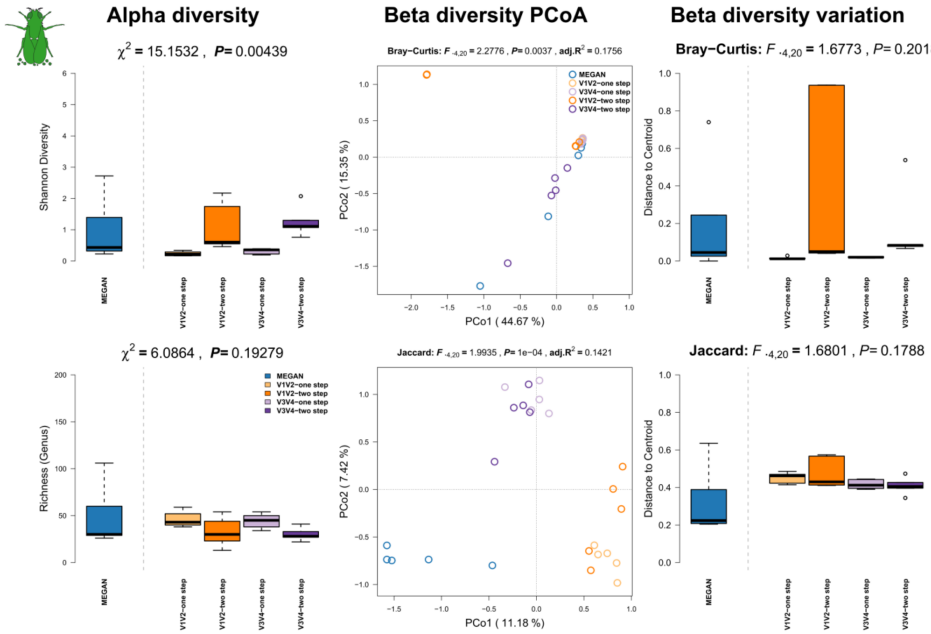


Figure S9: Comparison of community alpha diversity, composition, and variability of *Drosophila melanogaster* fecal microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

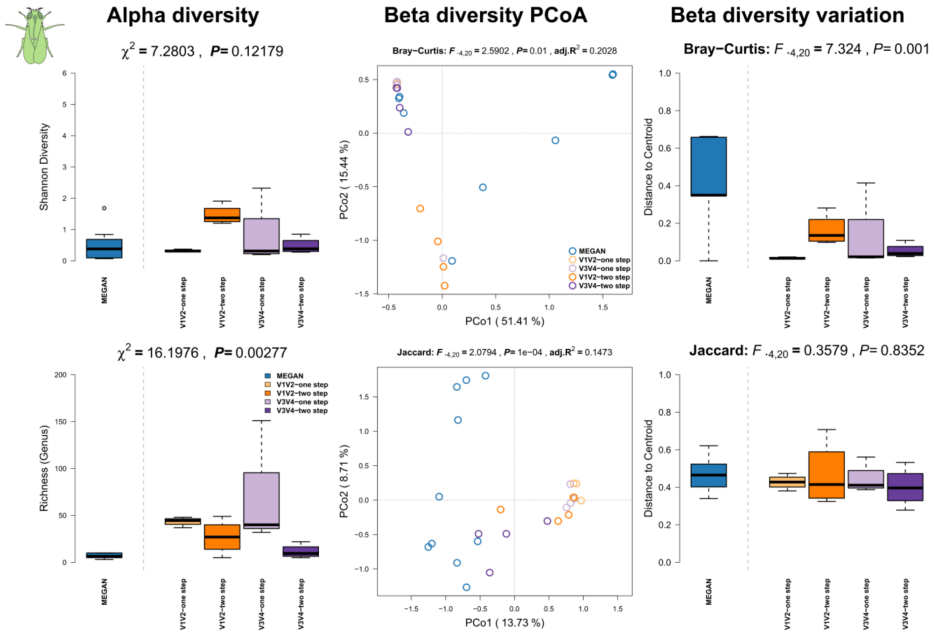


Figure S10: Comparison of community alpha diversity, composition, and variability of *Drosophila melanogaster* gut tissue associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=10, $N_{V1V2\text{-}one\ step}$=4, $N_{V1V2\text{-}two\ step}$=4, $N_{V3V4\text{-}one\ step}$=4, and $N_{V3V4\text{-}two\ step}$=4.

Summary *Homo sapiens* (Linnaeus, 1758):

The human (*Homo sapiens*) microbiome has been studied intensively and is presumed to be involved in wide array of diseases. Several microbiome studies have shown the influence of host genetics, environmental/life style variables, or diseases themselves [19, 20].

***Methods**:* Human feces (N=4) and biopsy samples (N=1) were sampled and extracted following the procedures as described in Wang *et al.* 2016 [19]. No enrichment for bacterial DNA was used.

***Results**:* We see again the lowest alpha diversity in the shotgun profiles using MEGAN, while all other 16S rRNA gene amplicon profiles show very similar levels of complexity and richness. Interestingly, principle coordinate analyses show no clear separation between shotgun- and

amplicon-derived profiles, except when we only consider the occurrence of genera, which separates mainly amplicon and shotgun based community profiles (Figure S11, middle panel; Table S4). There is also no clear difference in community variation noticeable among methods (Figure S11, right panel).

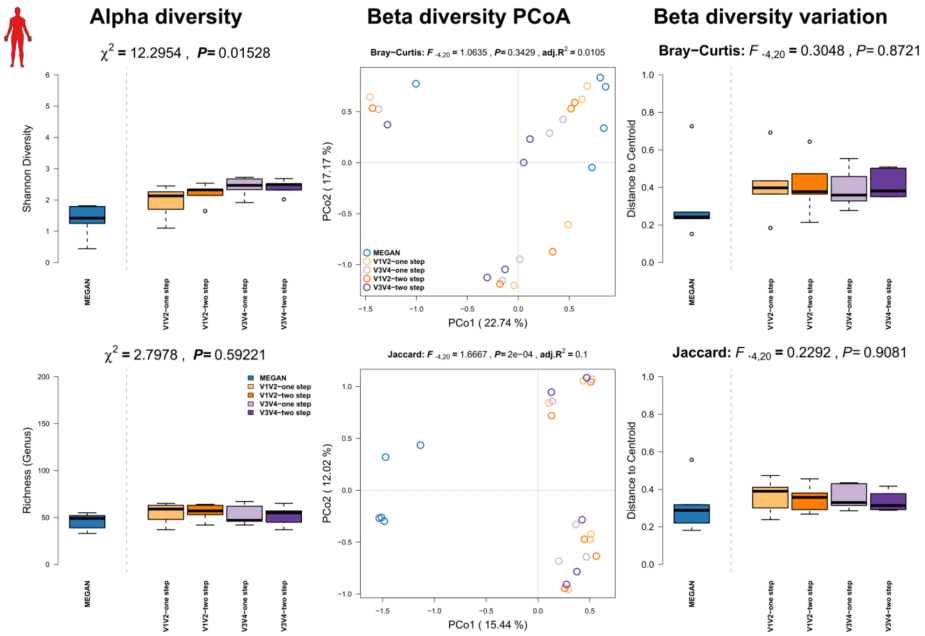


Figure S11: Comparison of community alpha diversity, composition, and variability of *Homo sapiens* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Hydra vulgaris* (Pallas, 1766)*:*

The fresh-water cnidarian *Hydra vulgaris* is an established model organism in evolutionary developmental biology belonging to the Hydrozoa. Since *Hydra* is colonized by a stable and species specific bacterial community, it is also used for the study of host-microbe interactions [21]. Its body and tissue structure resembles in some ways vertebrate intestine with the

endodermal epithelium and can be used to emulate dynamics as seen in vertebrates due to its ancestral relationship [22]. It also recently got a complete genome and has established methods for genetic manipulations [23, 24].

***Methods**:* Two hundred adult polyps of lab raised *H. vulgaris* were used for each sample and stored at -20°C until extraction via the QiaAMP stool kit following the manufacturer's instructions. All animals were cultured under constant, identical environmental conditions including culture medium, food (first-instar larvae of *Artemia nauplii*, fed three times per week) and temperature according to standard procedures. *H. vulgaris* polyps were cultured following standard operating procedure as described in Lenhoff & Brown (1970) [25]. No enrichment for bacterial DNA was used.

***Results**:* We see the highest alpha diversity (richness) in the shotgun profiles based on MEGAN (Figure S12, left panel), while community complexity (Shannon index) is fairly similar among sequencing techniques. The analysis of beta diversities via principle coordinate analysis shows clear separation between shotgun- and amplicon-based profiles, in particular based on differential occurrence of genera and far less based on differential abundances of taxa (Figure S12, middle panel; Table S4). V1V2- one step derived profiles are least differentiated to other methods. Variation of communities is quite comparable among the different profiling methods, however we can observe a higher variation in community membership in the 16S rRNA-two step derived community profiles (Figure S12, right panel).

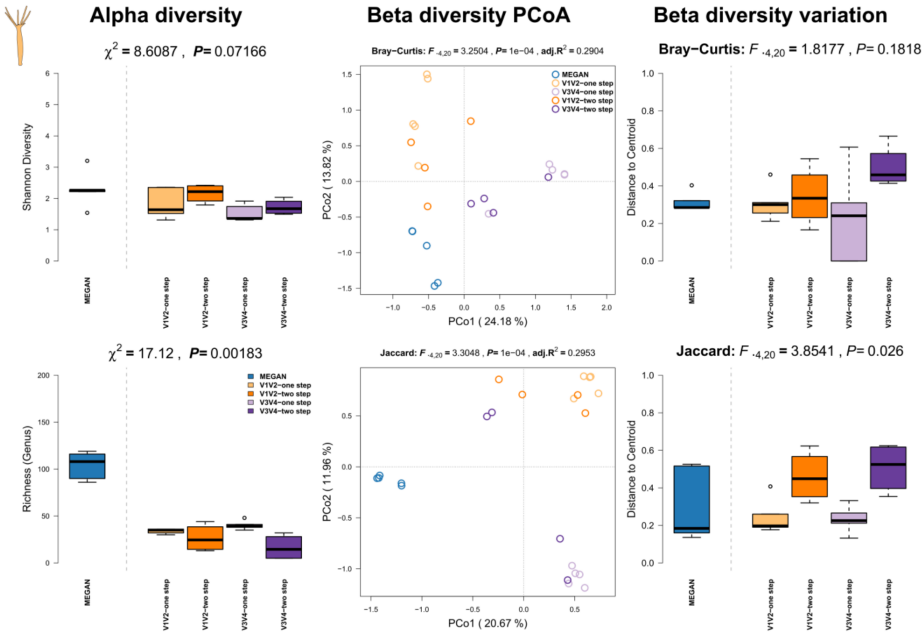


Figure S12: Comparison of community alpha diversity, composition, and variability of *Hydra vulgaris* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=4, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=4.

Summary *Mnemiopsis leidyi* (Agassiz, 1865):

Mnemiopsis leidyi is a widely distributed lobate comb jelly (Ctenophora) native to the western Atlantic coasts. *M. leidyi* has become an ecologically and economically important invasive species as it recently expanded its range to the Black, Caspian, North- and Baltic Sea through human influence. *M. leidyi* was also established as model system to study regeneration, patterning, and luminescence, as well as the evolution of the metazoans. A recently finished genome is a recent addition to the set of resources available for *M. leidyi* [26].

***Methods**:* Individual *M. leidyi* (mean size 4 cm, N= 5) were sampled from one location in the Kiel Bight, Baltic Sea (54.330107 N, 10.149735 E) in September 2016 by using a dip net. The animals were transported immediately to the laboratory, washed thoroughly with sterile filtered

artificial seawater (ASW) to remove non-associated microbes. Separation of eukaryotic and prokaryotic cells from whole animals as well as DNA extraction were performed as described in the *A. aurita* section.

***Results**:* We can observe the highest complexity but lowest richness in the shotgun profiles using MEGAN (Figure S13, left panel). In the 16S rRNA gene amplicon profiles we see a slightly higher diversity if V3V4 amplicons are analyzed as compared to V1V2. Principle coordinate analyses show a strong separation between methods, with a strong differentiation between MEGAN based profiles and the different variable regions, particularly to V1V2 derived profiles (Figure S13, middle panel; Table S4). However, variation in community composition varied the least in the profiles based on V1V2 amplicon methods. This pattern is not observed when we only consider genus occurrences in the community profiles (Figure S13, right panel).

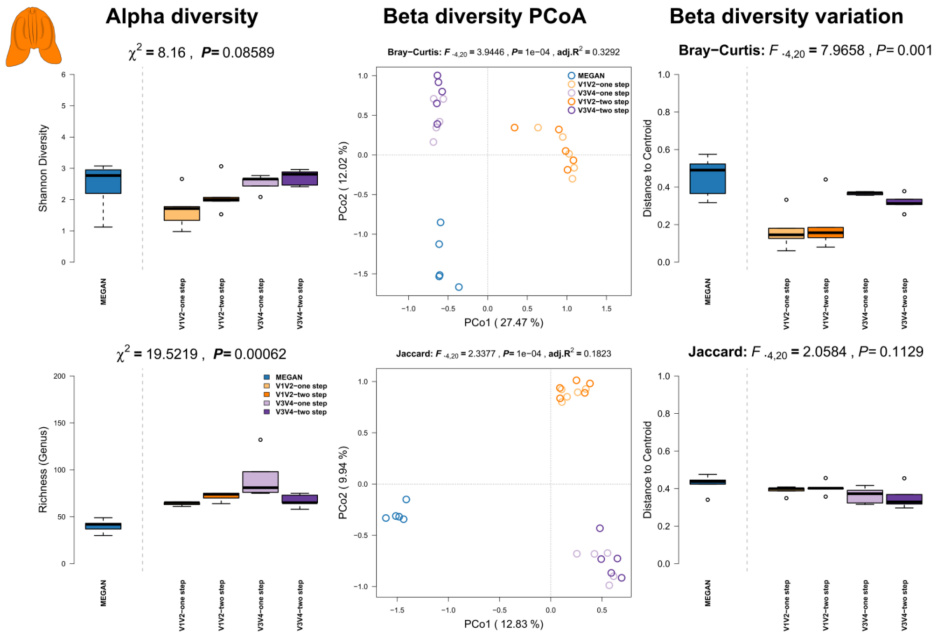


Figure S13: Comparison of community alpha diversity, composition, and variability of *Mnemiopsis leidyi* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Mus musculus* (Linnaeus, 1758):

The house mouse (*Mus musculus*) is a widely used model organism in biology and medicine. As the mouse shares many similarities to humans in terms of anatomy, physiology and genetics, it has become the preferred mammalian model for genetic research, with a wide array of genetic and physiological tools available. It has also become a widely used model for microbiome studies.

***Methods**:* Five male hybrid mice were used for sampling. The mice originate from hybrid house mouse breeding stocks of wild derived *M. m. musculus* × *M. m. domesticus* hybrids captured in 2008, kept at the Max Planck Institute in Plön (11^{th} lab generation at time of sampling). The stocks are derived from the Bavarian hybrid zone around Munich (Germany), in particular the locations FS (N=2), HA (N=1), and TU (N=2) [27]. Handling and killing of the mice was conducted according to the German animal welfare law and Federation of European Laboratory Animal Science Associations guidelines. All mice were sacrificed with CO_2 followed by cervical dislocation. Cecal content was preserved in 1ml of RNAlater and stored at -20°C after 12h at 5°C, following centrifugation and removal of supernatant (N=5). This was later extracted using the DNA/RNA AllPrep kit (Qiagen) according to the manufacture's protocol as described previously [28, 29], with an initial bead-beating step using Lysing Matrix E tubes (MP Biomedicals). No enrichment for bacterial DNA was used.

***Results**: Mouse samples show the* the highest alpha diversity (Shannon H, richness) when community profiles are based on shotgun profiles (MEGAN) compared to the amplicon based approaches (Figure S14, left panel). We can observe further observe a clear separation between shotgun- and amplicon-derived community profiles in the principle coordinate analyses. However, we see a slightly smaller distance between shotgun- and V1V2 derived communities (Figure S14, middle panel; Table S4) and among the 16S rRNA gene amplicon profiles we observe separation between variable regions but not between amplification methods (*i.e.* V1V2 or V3V4). Furthermore, V1V2 based profiles are most variable, in particular compared to MEGAN based profiles (Figure S14, right panel).

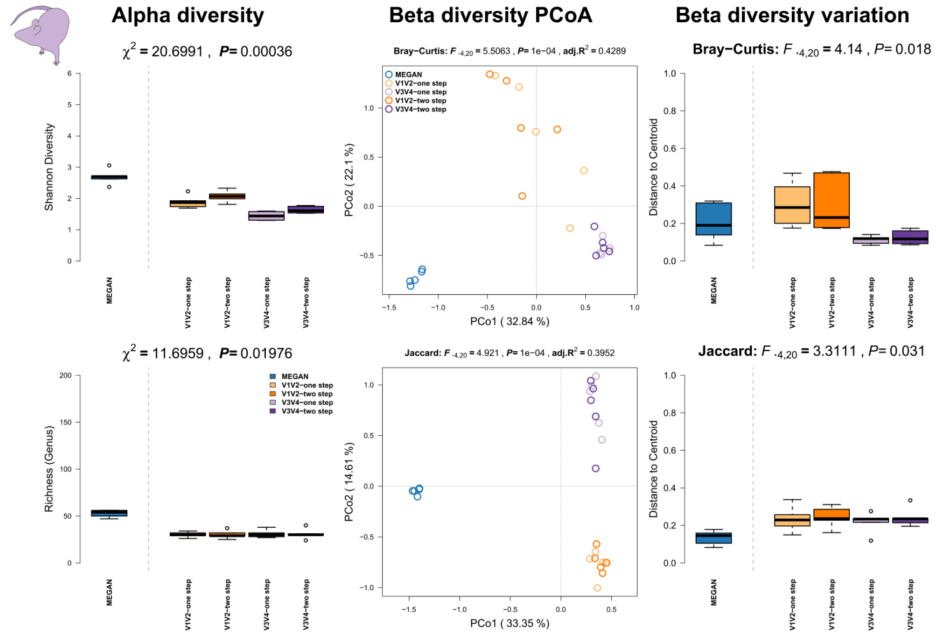


Figure S14: Comparison of community alpha diversity, composition, and variability of *Mus musculus* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

Summary *Nematostella vectensis* (Stephenson, 1935):

The marine Starlet Sea Anemone *Nematostella vectensis* is a marine cnidarian (Anthozoa) living in the shallow coastal waters and marshes of Canada, the United States, and England. *N. vectensis* is a model organism for embryonic development and has an enormous adaptive potential to variable environmental factors. The animals can be cultured in the lab and reproduce sexually and asexually throughout the year.

***Methods**:* Five solitary adult lab raised polyps of *N. vectensis* were sampled and stored at -20°C until extraction via the QIAamp DNA Microbiome Kit following the manufacturer's instructions. All animals were cultured under constant, identical environmental conditions including culture medium (Artificial sea water, 16‰), food (first-instar larvae of *Artemia nauplii*, fed two times per

week) and temperature according to standard procedures (18°C, in complete darkness). No enrichment for bacterial DNA was used.

***Results**:* In contrast to the other host organisms we can see the highest alpha diversity in the one step amplicon methods, followed by MEGAN (Figure S15, left panel). In the amplicon profiles V3V4-two step method display the lowest alpha diversity. Principle coordinate analyses show differentiation between shotgun- and amplicon-sequenced samples with the lowest distance to V3V4 based community profiles (Figure S15, middle panel; Table S4). Based on the co-occurrence of genera we can observe more differentiation between amplicon- and shotgun based profiles, however V3V4 communities still cluster closest to the shotgun based communities. We can further observe a also a higher variation in the V3V4-two step PCR as well as MEGAN based community profiles (Figure S15, right panel).

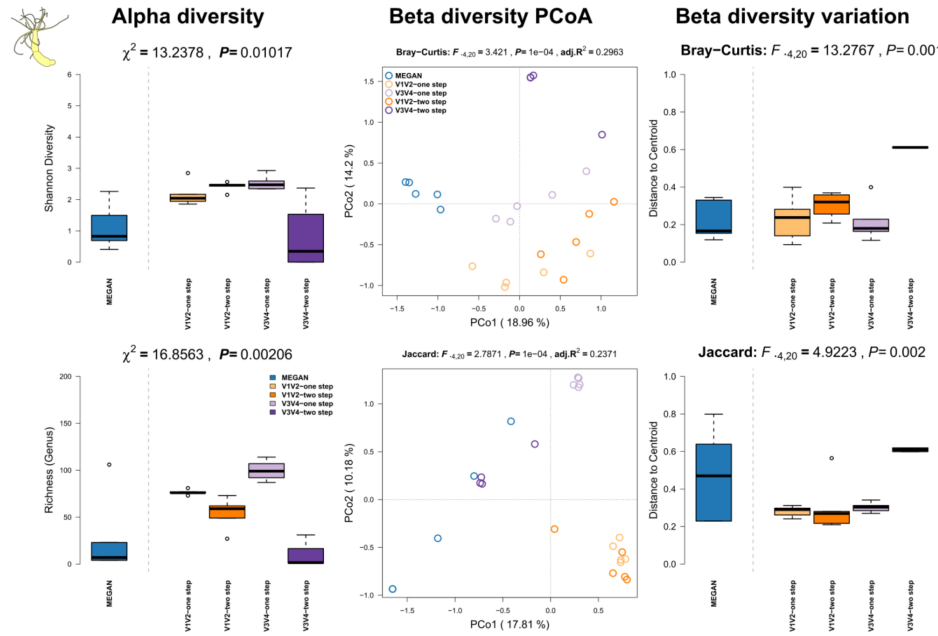


Figure S15: Comparison of community alpha diversity, composition, and variability of *Nematostella vectensis* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=4.

Summary *Triticum aestivum* (Linnaeus, 1758):

Triticum aestivum a member of the *Poaceae* is the ancestral hybrid of *Triticum dicoccum* and *Aegilops tauschii* and, potentially first cultured 7000 yr. a.d. in the fertile crescent. It is one of the most widely distributed crops, used for bread, cereal and starch. Due to its industrial use it is of high genetic homogeneity. Grains and their associated microbial communities might be beneficial and interact during germination, as well as be a direct route of heritability from a fully grown plant to the next generation via the seed [30].

***Methods**:* DNA was extracted from grains (0.1 gram) of commercially purchased *Triticum aestivum* (Davert GmbH, brand name: Bioland Davert) (N=2) and farm raised *T. aestivum* (N=3, University farm CAU Kiel, Runal cultivar). 0.1 gram of seeds was washed in 70% ethanol and surface sterilized in 3% Sodium hypochlorite for 10 minutes. Afterwards grains were washed with sterile water. Surface sterilized grains were homogenized using Precellys (SK38; according to manufacturer's protocol). Homogenate was further processed using GeneJet Plant Genomic DNA Purification Kit (ThermoFisher Scientific). Quality of extracted DNA was tested on agarose gel. Presence of the bacterial 16S rRNA gene was confirmed by PCR using bacteria specific primers with less affinity to chloroplast signatures (799F-1391R). No enrichment for bacterial DNA was used.

***Results**:* Although a high percentage of sequences has been lost due to host sequence contamination, we could still perform the majority of analyses. The highest alpha diversity has been recovered in the 16S rRNA gene amplicon based samples, particularly in the V1V2-two step (also highest richness), while the shotgun based community profiles show the lowest diversity (Figure S16, left panel). Principle coordinate analyses show a clear separation between shotgun- and amplicon-derived profiles (Figure S16, middle panel; Table S4), but also between 16S rRNA gene amplicon profiles by variable region (*i.e.* V1V2 or V3V4), which is further reduced when we only consider the presence of taxa between samples (Jaccard). On average, the variability of communities (genus abundance or occurrence) is highest in the 16S rRNA gene amplicon derived community profiles; in particular in the one step V1V2 based samples (Figure S16, right panel).

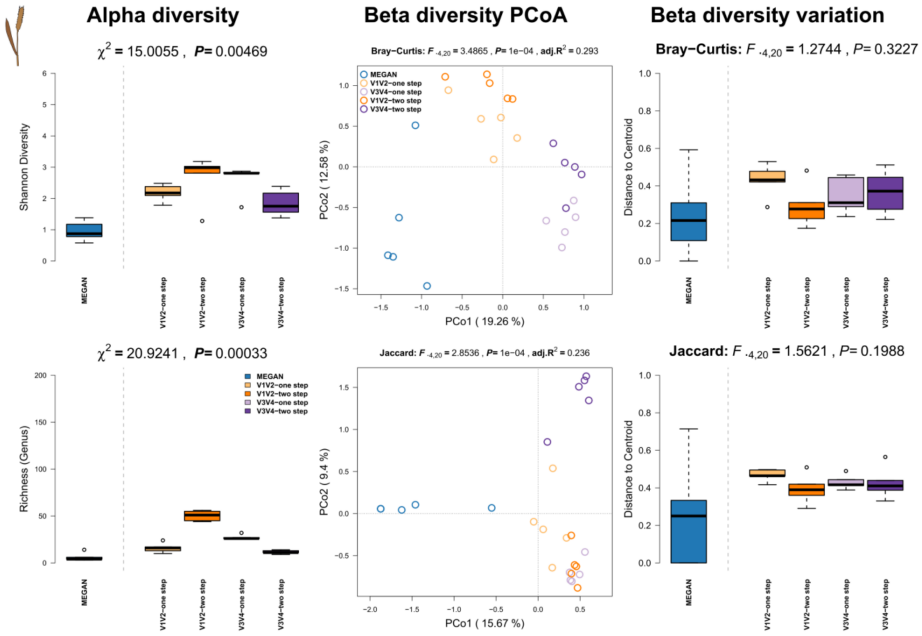


Figure S16: Comparison of community alpha diversity, composition, and variability of *Triticum aestivum* associated microbial communities among 16S rRNA gene amplicon and shotgun based pipelines. Comparisons are tested via approximate Kruskal-Wallis test (alpha diversity), PERMANOVA (beta diversity), and permutative anova (variation of beta diversity). Community profiles were derived from 16S rRNA gene amplicon sequencing (V1V2, V3V4, one step, two step), MEGAN based classification (short reads). Sample sizes for the different approaches are $N_{shotgun}$=5, $N_{V1V2\text{-}one\ step}$=5, $N_{V1V2\text{-}two\ step}$=5, $N_{V3V4\text{-}one\ step}$=5, and $N_{V3V4\text{-}two\ step}$=5.

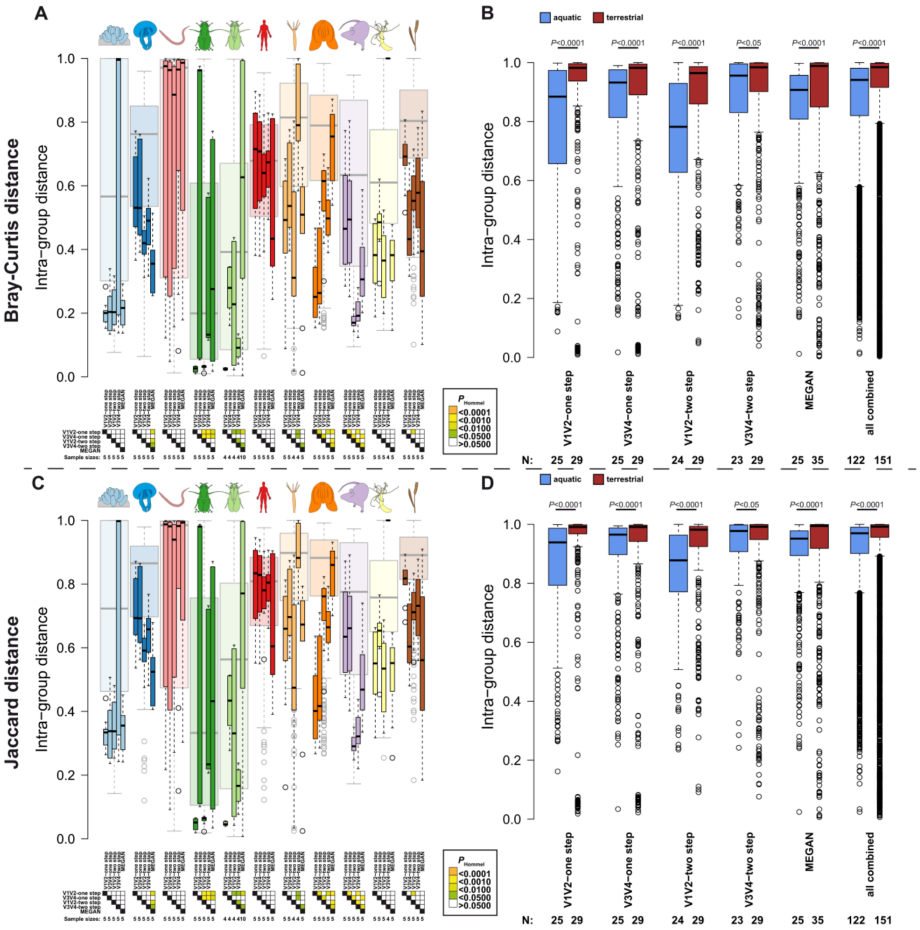


Figure S17: Boxplots visualize the pairwise distances based on shared abundance (**A, B**) and shared presence (**C, D**) between samples belonging to the respective host groups and methods (**A, C**), and host environments (**B, D**) among the different sequencing methods. Pairwise comparisons were computed via pairwise Wilcoxon tests (Hommel *P*-value adjustment) within host groups, and approximate Wilcoxon tests [6] between host environments. Significance levels after correction for multiple testing are indicated by color (for sample sizes see Table S1).

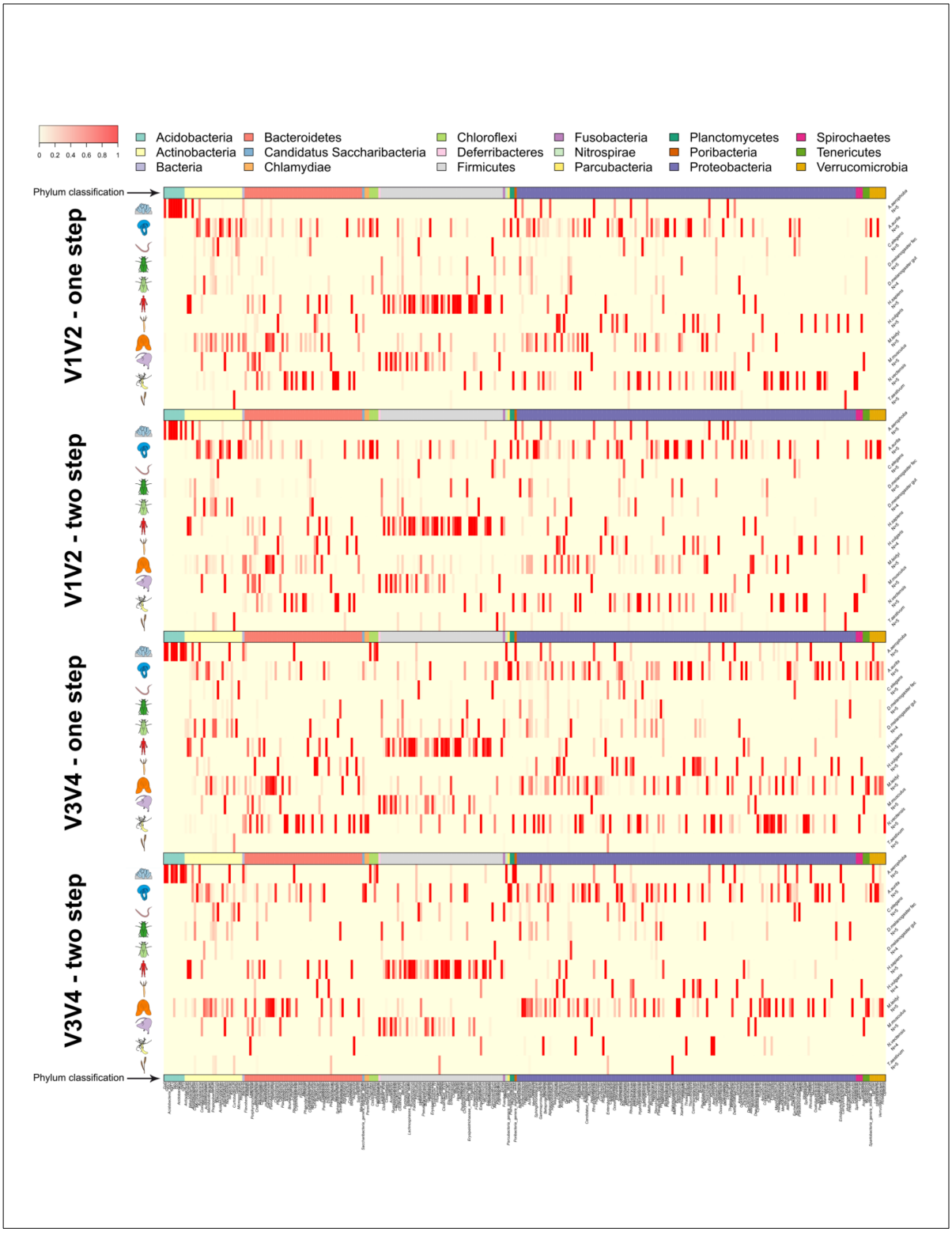

0 0.2 0.4 0.6 0.8 1
Acidobacteria
Actinobacteria
Bacteria
Bacteroidetes
Candidatus Saccharibacteria
Chlamydiae
Chloroflexi
Deferribacteres
Firmicutes
Fusobacteria
Nitrospirae
Parcubacteria
Planctomycetes
Poribacteria
Proteobacteria
Spirochaetes
Tenericutes
Verrucomicrobia
Phylum classification
V1V2 - one step
V1V2 - two step
V3V4 - one step
V3V4 - two step
Phylum classification

Figure S18: Heatmaps of indicator genera derived from the different 16S rRNA gene amplicon strategies. Heat color indicates relative abundance across sample groups and the color bar indicates phylum affiliation of the different indicators (**see Table S6**).

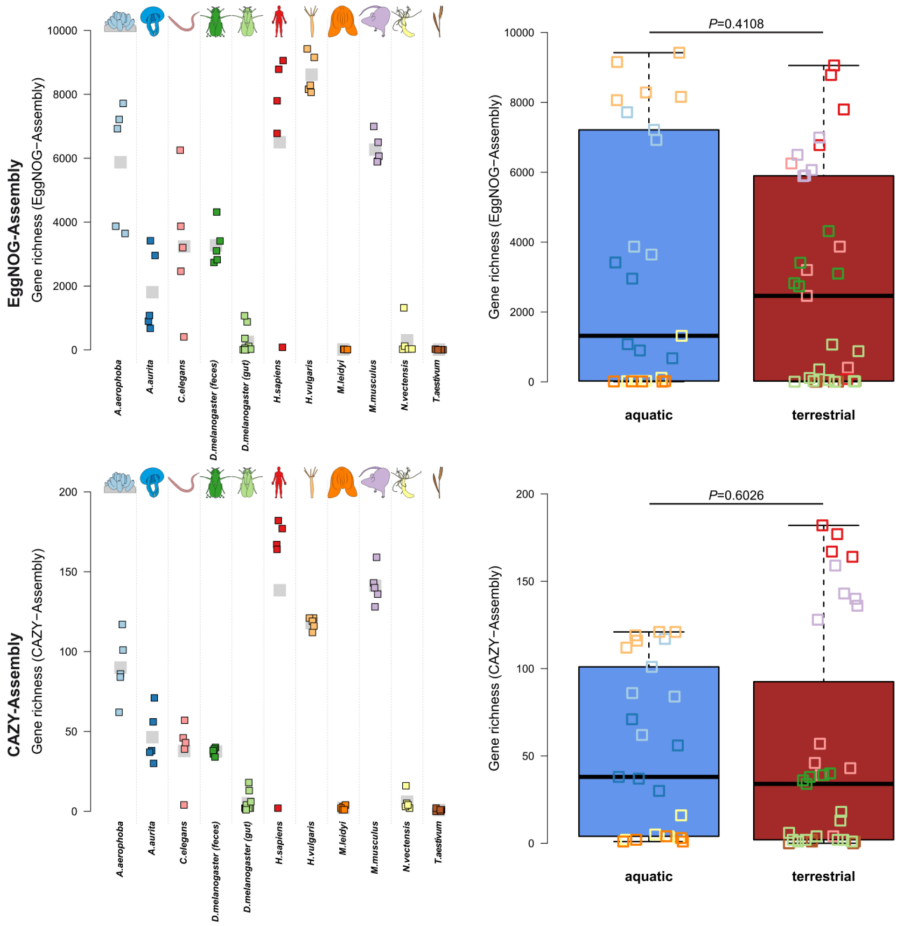


Figure S19: Functional diversity of EggNOG derived genes and CAZY predictions in the different organisms and host environments. Pairwise differences of functional diversity between hosts were tested via pairwise *t*-Tests with pooled standard variations. Approximate Wilcoxon tests were employed to compare the host environmental groups. Sample size for the host taxa is N=5, except for D. melanogaster gut tissue (N=10). The sample size of aquatic hosts is N=25, while terrestrial hosts have a sample size of N=35 (see Table S1).

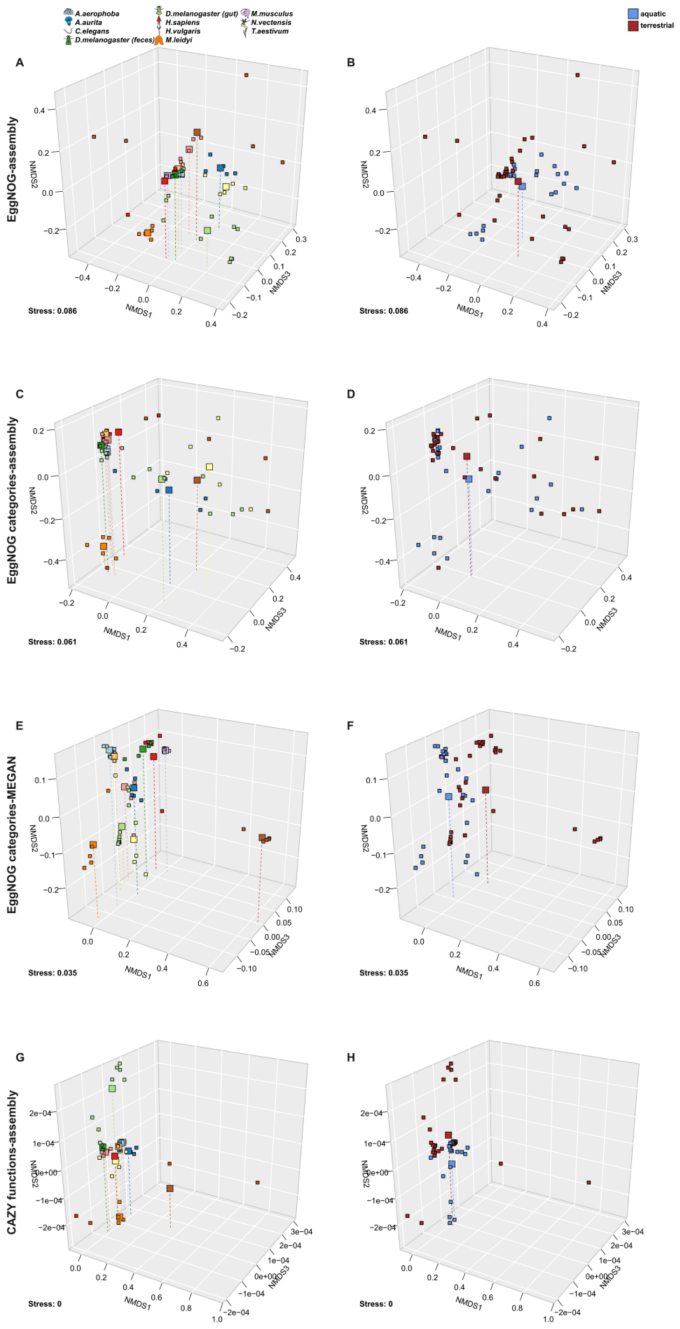

A.aerophoba
A.aurita
C.elegans
D.melanogaster (feces)
D.melanogaster (gut)
H.sapiens
H.vulgaris
M.leidyi
M.musculus
N.vectensis
T.aestivum
aquatic
terrestrial
A
B
C
D
E
F
G
H
EggNOG-assembly
EggNOG categories-assembly
EggNOG categories-MEGAN
CAZY functions-assembly
NMDS1
NMDS2
NMDS3
Stress: 0.086
Stress: 0.061
Stress: 0.035
Stress: 0

Figure S20: Non-metric Multidimensional Scaling of Bray-Curtis distances derived from abundance differences of functional components highlighting functional differences between the host organisms (**A, C, E, G; see Table 3**) and host environments (**B, D, F, H; see Table 3**). Large symbols indicate the centroid of the respective host groups and vertical lines help to determine their position in space. Functional variation of communities based on pairwise Bray-Curtis distances within host organism groups and environmental groups. Sample size for the host taxa is N=5, except for *D. melanogaster* gut tissue (N=10). The sample size of aquatic hosts is N=25, while terrestrial hosts have a sample size of N=35 **(**see Table S1**)**.

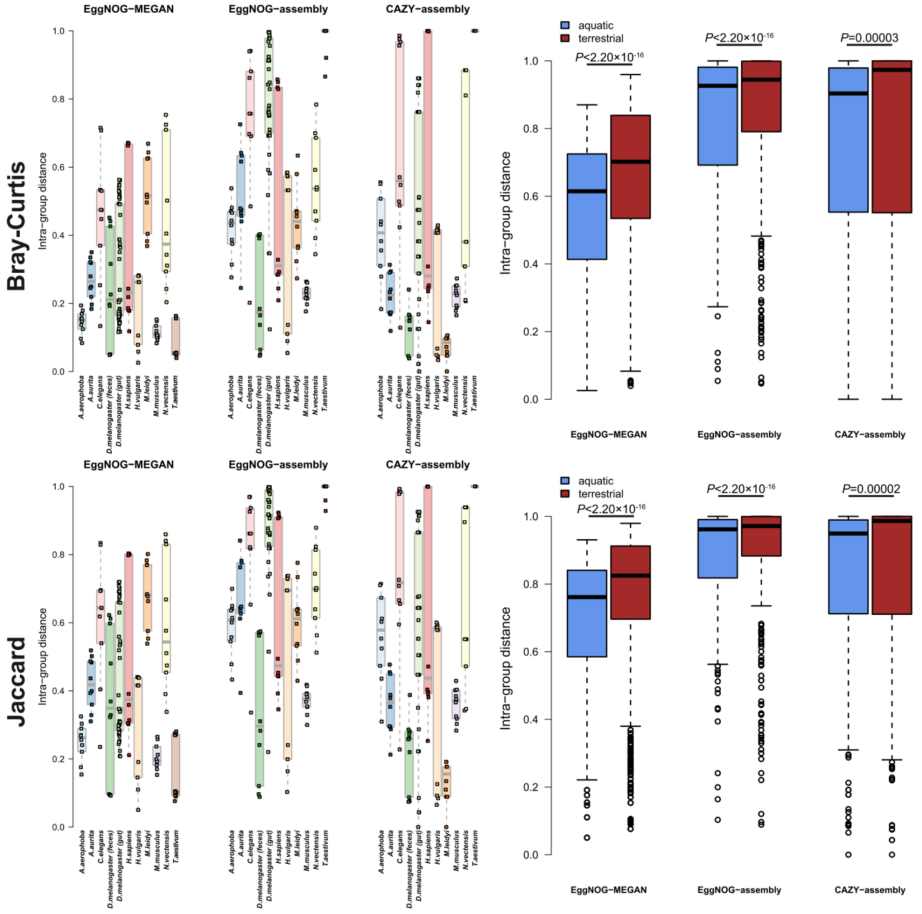


Figure S21: Functional variation of communities based on EggNOG derived genes and CAZY predictions based on pairwise Bray-Curtis distances within host organism groups and environmental groups. Terrestrial hosts show a significantly higher Functional variation than aquatic hosts. Sample size for the host taxa is N=5, except for *D. melanogaster* gut tissue (N=10). The sample size of aquatic hosts is N=25, while terrestrial hosts have a sample size of N=35 (see Table S1).

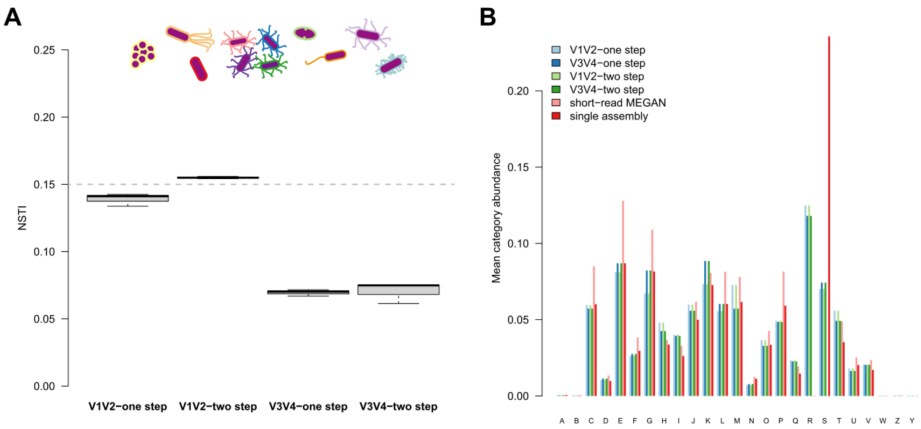


Figure S22: (A) NSTI/imputation success differences between variable regions in the mock community samples. **(B)** Barplot displays average abundances of the single functional categories derived by different 16S rRNA gene amplicon based and shotgun based techniques for mock community samples. Sample sizes for the different approaches are $N_{shotgun}$=4, $N_{V1V2\text{-}one\ step}$=3, $N_{V1V2\text{-}two\ step}$=3, $N_{V3V4\text{-}one\ step}$=3, and $N_{V3V4\text{-}two\ step}$=3.

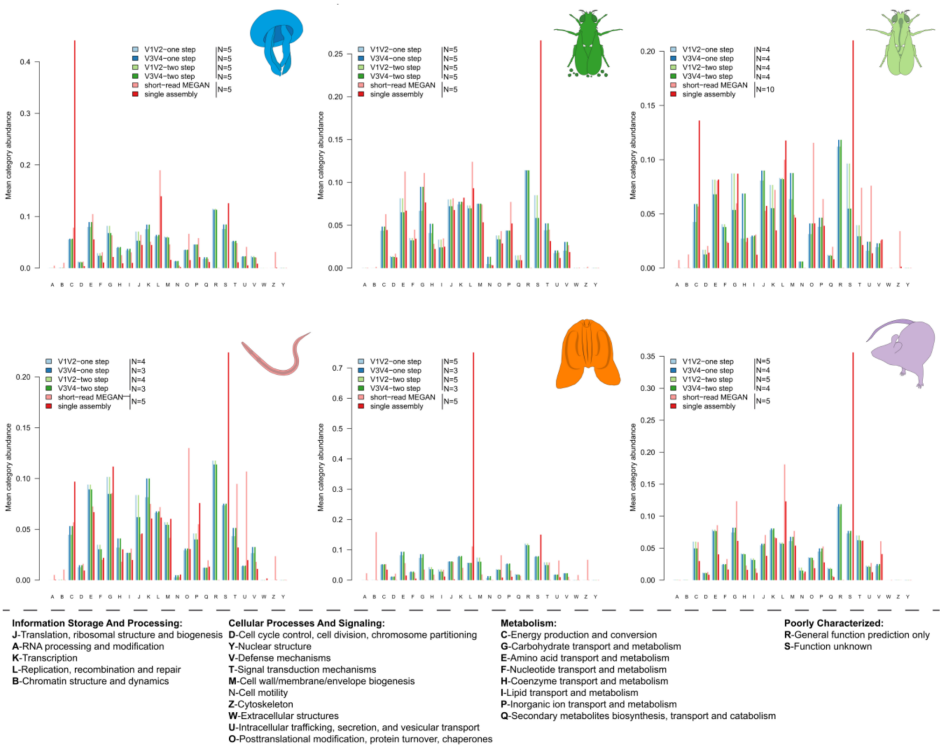


Figure S23: Barplots display average abundances of the single functional categories derived by PICRUSt from the different 16S rRNA amplicon techniques and shotgun based annotations for specific host community samples with sufficient sample coverage.

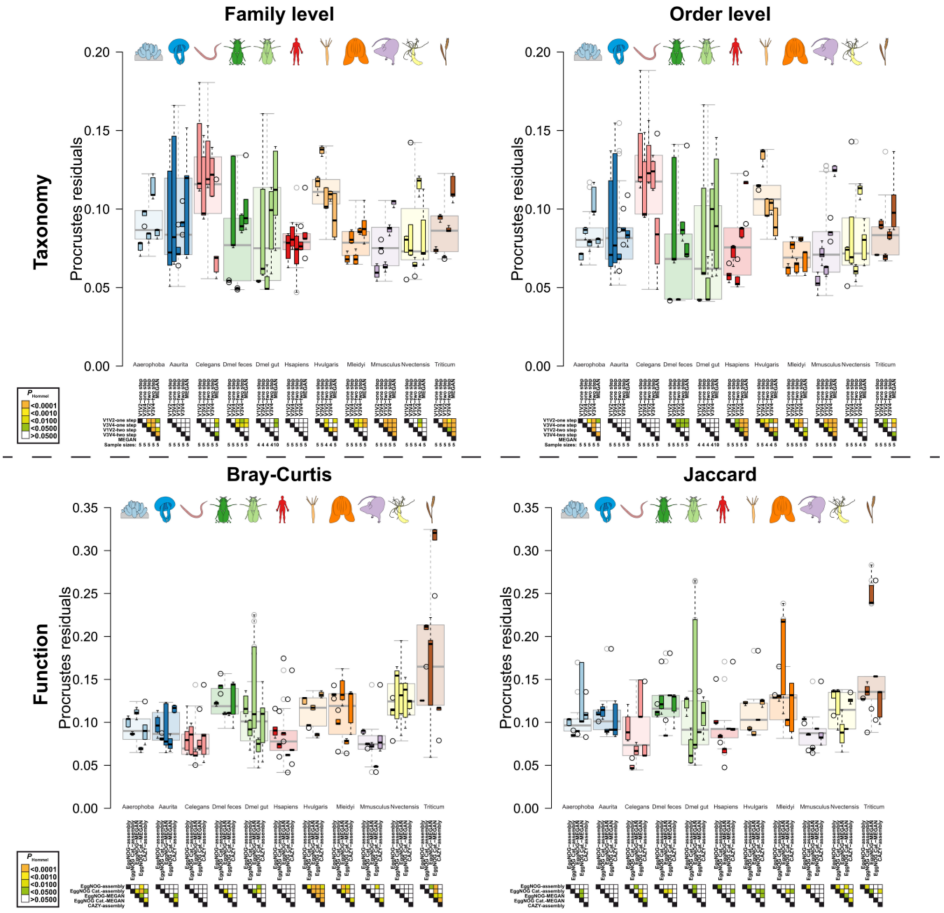


Figure S24: Upper panels visualize the residuals derived from Procrustes correlations between phylogenetic and taxonomic microbial community distances among host samples. Sample sizes are indicated below samples. Lower panels show residuals of the correlation between phylogenetic and community distances derived from functional profiles. Sample size for the host taxa is N=5, except for *D. melanogaster* gut tissue (N=10, see Table S1). Pairwise comparisons were computed via pairwise *t*-Tests (Hommel *P*-value adjustment) within host groups. Significance levels after correction for multiple testing are indicated by color.

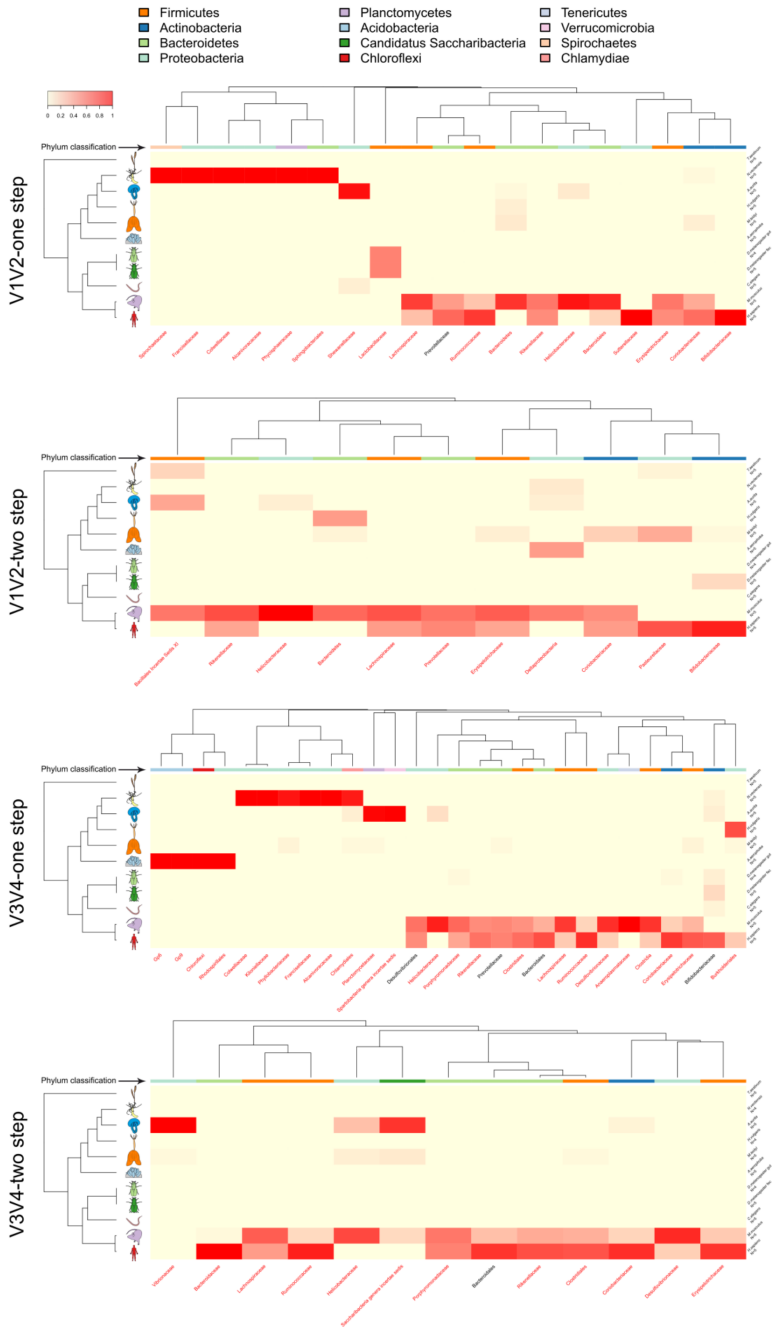

Firmicutes
Actinobacteria
Bacteroidetes
Proteobacteria
Planctomycetes
Acidobacteria
Candidatus Saccharibacteria
Chloroflexi
Tenericutes
Verrucomicrobia
Spirochaetes
Chlamydiae
Phylum classification
V1V2-one step
V1V2-two step
V3V4-one step
V3V4-two step

Figure S25: Heatmaps of family abundances derived from the different 16S rRNA gene amplicon strategies which associate significantly to phylogenetic distance, as determined by Moran's I test [31]. Heat colors indicate relative abundances across sample groups and the color bar indicates phylum affiliation of the different taxa (see Table S15). Row dendrogram depicts the 18S rRNA gene-based phylogeny of host species. The column dendrogram highlights the similarity of genera distributions as determined by Bray-Curtis distance and average neighbor clustering. Taxon names in red indicate a significant association as an indicator taxon (for sample sizes see Table S1).

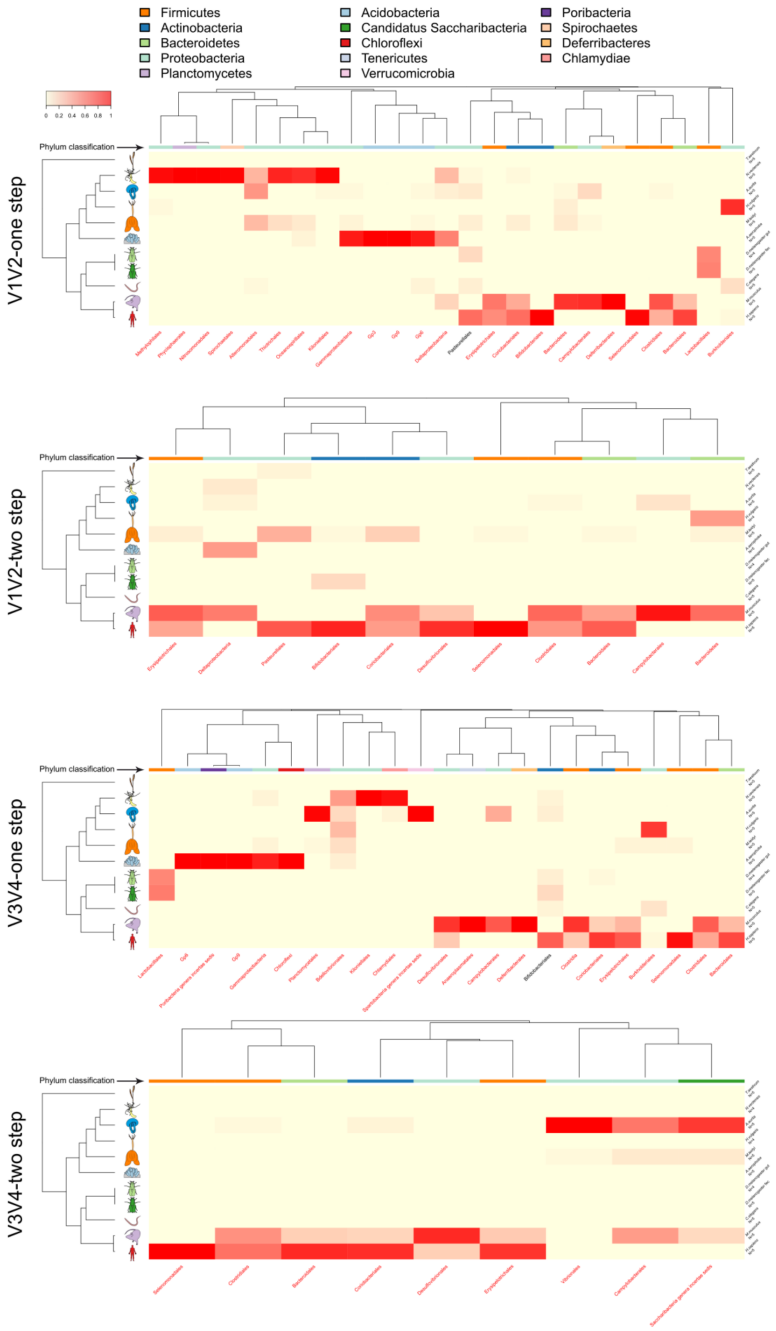

Firmicutes
Actinobacteria
Bacteroidetes
Proteobacteria
Planctomycetes
Acidobacteria
Candidatus Saccharibacteria
Chloroflexi
Tenericutes
Verrucomicrobia
Poribacteria
Spirochaetes
Deferribacteres
Chlamydiae
Phylum classification
V1V2-one step
V1V2-two step
V3V4-one step
V3V4-two step

Figure S26: Heatmaps of order abundances derived from the different 16S rRNA gene amplicon strategies which associate significantly to phylogenetic distance, as determined by Moran's I test [31]. Heat colors indicate relative abundances across sample groups and the color bar indicates phylum affiliation of the different taxa (see Table S15). Row dendrogram depicts the 18S rRNA gene-based phylogeny of host species. The column dendrogram highlights the similarity of genera distributions as determined by Bray-Curtis distance and average neighbor clustering. Taxon names in red indicate a significant association as an indicator taxon (for sample sizes see Table S1).

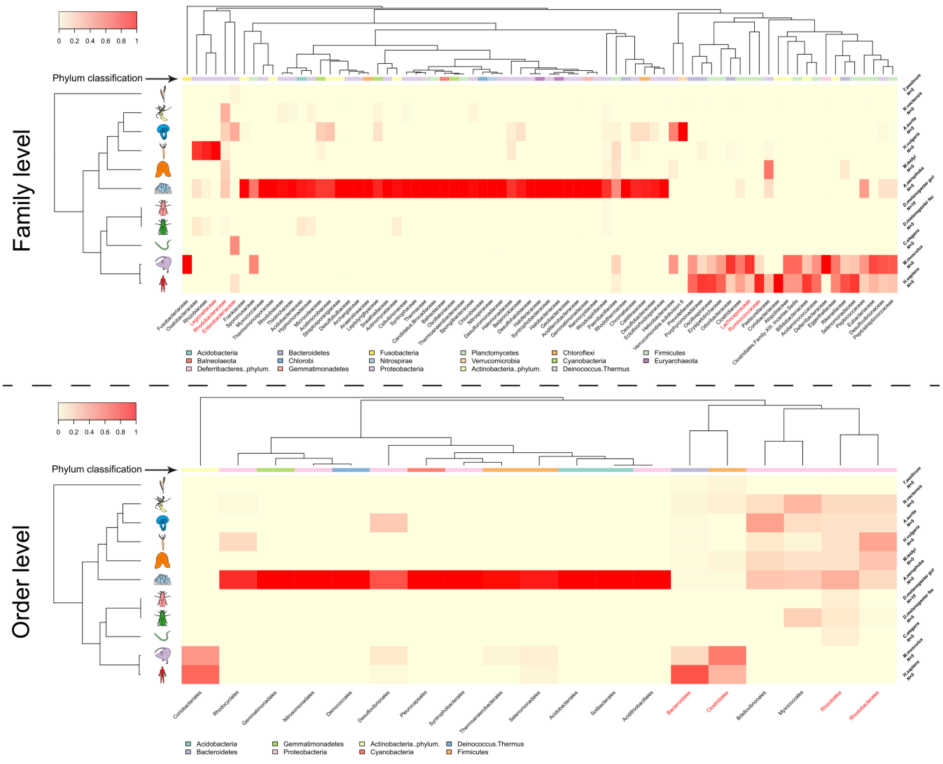


Figure S27: Heatmaps of family abundances derived from the different shotgun strategies which associate significantly to phylogenetic distance, as determined by Moran's I test [31]. Heat color

indicates relative abundance across sample groups and the color bar indicates phylum affiliation of the different indicators (see Table S16). Row dendrogram shows the 18S rRNA gene-based tree of host species (ML tree) and the column dendrogram highlights the similarity of genera distributions as determined by Bray-Curtis distance and average neighbor clustering. Taxon names in red indicate a significant association as an indicator taxon. Sample size for the host taxa is N=5, except for *D. melanogaster* gut tissue (N=10, see Table S1).

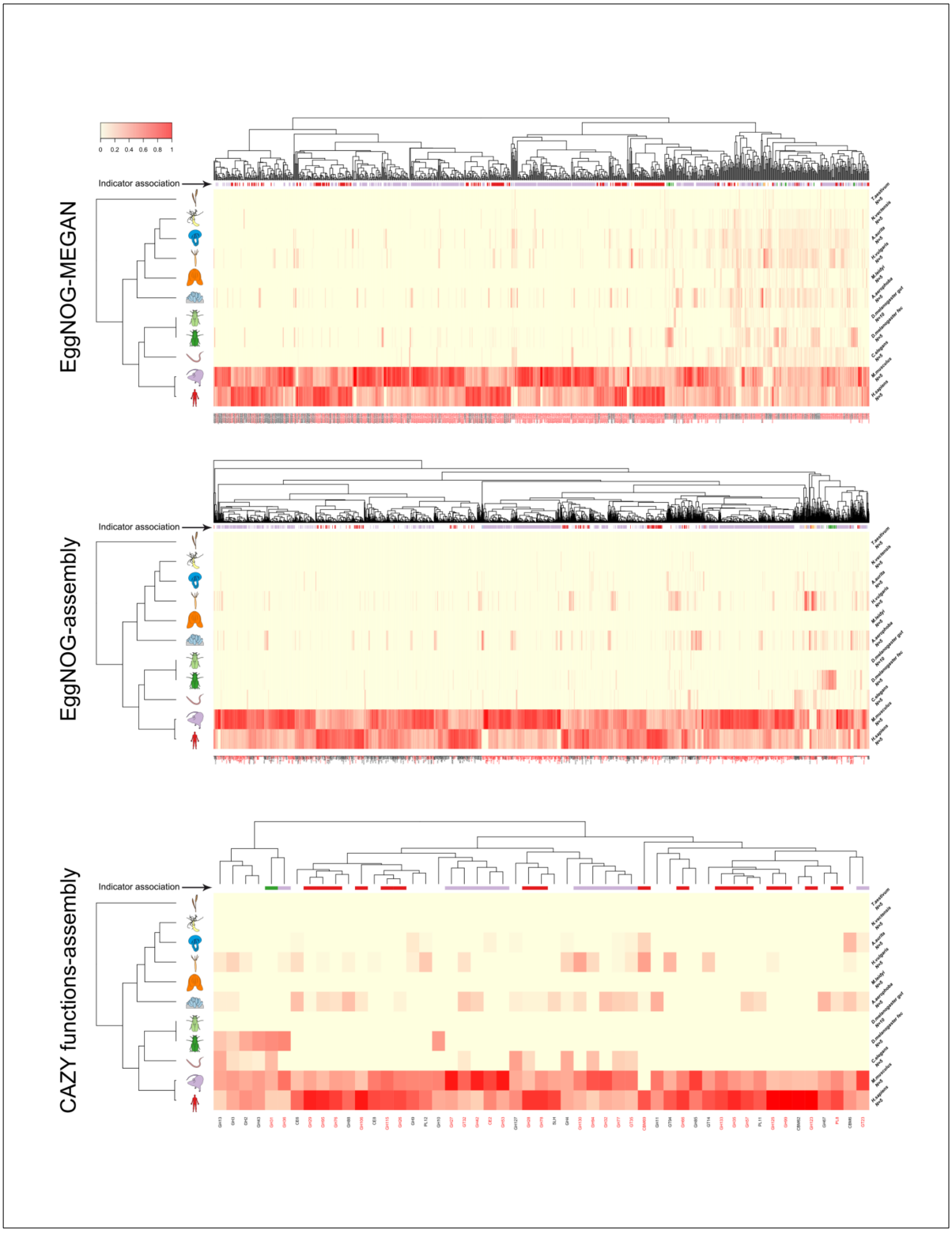

EggNOG-MEGAN
EggNOG-assembly
CAZY functions-assembly
Indicator association
0 0.2 0.4 0.6 0.8 1

Figure S28: Heatmaps of function abundances derived from the different metagenomic functional annotation strategies associate significantly to phylogenetic distance, as determined by Moran's I test [31]. Heat colors indicate relative abundance across sample groups and the color bar indicates single host indicator associations (see Table S17-S20). Function names in red indicate a significant association as an indicator function. Row dendrogram shows the 18S rRNA gene-based tree of host species (ML tree) and the column dendrogram highlights the similarity of genera distributions as determined by Bray-Curtis distance and average neighbor clustering. Sample size for the host taxa is N=5, except for *D. melanogaster* gut tissue (N=10, see Table S1).

Supplementary Tables:

Table S1: Sample sizes of the 16S rRNA amplicon based- and shotgun based approaches for single host taxa and environments.

Table S2: Differences between expected and observed alpha diversity of the mock community, based on different shotgun- and 16S rRNA gene amplicon based community compositions (bacterial genus level), tested via a one-sample *t*-test (two-sided, *P*-values adjusted via Hommel procedure).

Table S3: Pairwise differences in mock community composition between the 16S rRNA gene amplicon and shotgun based techniques focusing on shared abundance (Bray-Curtis) and shared presence (Jaccard) of bacterial genera (PERMANOVA *P*-values based on 10'000 permutations, *P*-values adjusted via FDR procedure).

Table S4: Differences in community composition based on genus abundances and occurrences between the 16S rRNA gene amplicon and shotgun based techniques in different hosts (PERMANOVA *P*-values based on 10'000 permutations, *P*-values adjusted via Hommel procedure).

Table S5: Differences in alpha diversity estimates based on genus abundances between different hosts based on 16S rRNA amplicon and shotgun based techniques (pairwise *t*-test with Hommel adjusted *P*-values).

Table S6: Pairwise community compositional differences between the host species based on shared abundance of genera in the different amplicon techniques (Bray-Curtis). *P*-values are derived by PERMANOVA based on 10'000 permutations (*P*-values adjusted via Hommel procedure for each amplicon technique).

Table S7: Indicator genera for host taxa using different 16S rRNA gene amplicon techniques. Analyses are based on 10'000 permutations and *P*-value correction for each amplicon technique separately. Host taxa in grey highlight disagreement in their association and overlaps with shotgun techniques are listed in a separate column.

Table S8: Indicator genera for host taxa using the MEGAN based shotgun technique. Analyses are based on 10'000 permutations and *P*-value correction for each technique separately. Host taxa in grey highlight disagreement in their association overlaps with the results from the 16S rRNA gene amplicon analyses are listed in a separate column.

Table S9: Indicator genera for host environments using different 16S rRNA gene amplicon techniques. Analyses are based on 10'000 permutations and *P*-value correction for each 16S rRNA gene amplicon technique separately. Overlapping associations with the MEGAN based shotgun technique are listed in a separate column.

Table S10: Indicator genera for host environments using the different shotgun techniques. Analyses are based on 10'000 permutations and *P*-value correction for each technique separately. Overlapping associations with amplicon techniques are listed in a separate column.

Table S11: Indicator functions for host taxa using the EggNOG classification based on MEGAN. Analyses are based on 10'000 permutations with Benjamini-Yeketuli *P*-value correction.

Table S12: Indicator functions for host taxa using the EggNOG classification based on single sample assemblies and *emapper* annotation. Analyses are based on 10'000 permutations with Benjamini-Yeketuli *P*-value correction.

Table S13: Indicator functional categories for host taxa using the EggNOG classification based on single sample assemblies and *emapper* annotation and MEGAN. Analyses are based on 10'000 permutations with Benjamini-Yeketuli *P*-value correction.

Table S14: Indicator functions for host taxa using the CAZY classification based on single sample assemblies. Analyses are based on 10'000 permutations with Benjamini-Yeketuli *P*-value correction.

Table S15: Pairwise PERMANOVA comparisons between functional repertoires (EggNOG, COG) derived from single assemblies, MEGAN, and PICRUSt imputations (V1V2, V3V4, one step, two step).

Table S16: Moran's I eigenvector analyses for family and order abundances in combination with indicator analysis results for single and multiple hosts (maximum 3) using the 16S rRNA gene amplicon and metagenomic shotgun (MEGAN) based techniques (10'000 permutations). Repeated associations for single bacterial taxa, including their indicator associations are highlighted.

Table S17: Moran's I eigenvector analyses for CAZY functional abundances.

Table S18: Moran's I eigenvector analyses for EggNOG functional abundances derived from single sample assemblies.

Table S19: Moran's I eigenvector analyses for EggNOG functional abundances derived from MEGAN.

Table S20: Moran's I eigenvector analyses for EggNOG functional category abundances derived from MEGAN and single sample assemblies.

Supplementary References:

1. Huson D, Auch A, Qi J, Schuster S: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**(3):377 - 386.
2. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Meth* 2015, **12**(1):59-60.
3. Vacelet J: **Etude en microscopic électronique de l'association entre bacteries et spongiaires du genre Verongia (Dictyoceratida)**. *J Microsc Biol Cell* 1975, **23**:271-288.
4. Hentschel U, Hopke J, Horn M, Friedrich AB, Wagner M, Hacker J, Moore BS: **Molecular Evidence for a Uniform Microbial Community in Sponges from Different Oceans**. *Applied and Environmental Microbiology* 2002, **68**(9):4431-4440.
5. Hentschel U, Piel J, Degnan SM, Taylor MW: **Genomic insights into the marine sponge microbiome**. *Nat Rev Micro* 2012, **10**(9):641-654.
6. Hothorn T, Hornik K, Van de Wiel MA, Zeileis A: **A Lego system for conditional inference**. *American Statistician* 2006, **60**(3):257-263.
7. Legendre P, Anderson MJ: **Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments**. *Ecological Monographs* 1999, **69**(1):1-24.
8. Anderson MJ: **A new method for non-parametric multivariate analysis of variance**. *Austral Ecology* 2001, **26**(1):32-46.
9. Dawson MN, Jacobs DK: **Molecular Evidence for Cryptic Species of Aurelia aurita (Cnidaria, Scyphozoa)**. *The Biological Bulletin* 2001, **200**(1):92-96.
10. Lucas CH: **Reproduction and life history strategies of the common jellyfish, Aurelia aurita, in relation to its ambient environment**. In*: 2001; Dordrecht*. Springer Netherlands: 229-246.
11. Haber M, Schungel M, Putz A, Muller S, Hasert B, Schulenburg H: **Evolutionary history of Caenorhabditis elegans inferred from microsatellites: evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding**. *Mol Biol Evol* 2005, **22**(1):160-173.
12. Petersen C, Dirksen P, Prahl S, Strathmann EA, Schulenburg H: **The prevalence of Caenorhabditis elegans across 1.5 years in selected North German locations: the importance of substrate type, abiotic parameters, and Caenorhabditis competitors**. *BMC ecology* 2014, **14**:4-4.
13. Petersen C, Saebelfeld M, Barbosa C, Pees B, Hermann RJ, Schalkowski R, Strathmann EA, Dirksen P, Schulenburg H: **Ten years of life in compost: temporal and spatial variation of North German Caenorhabditis elegans populations**. *Ecology and evolution* 2015, **5**(16):3250-3263.
14. Dirksen P, Marsh SA, Braker I, Heitland N, Wagner S, Nakad R, Mader S, Petersen C, Kowallik V, Rosenstiel P *et al*: **The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model**. *BMC Biology* 2016, **14**(1):38.
15. von der Schulenburg JH, Hancock JM, Pagnamenta A, Sloggett JJ, Majerus ME, Hurst GD: **Extreme length and length variation in the first ribosomal internal transcribed spacer of ladybird beetles (Coleoptera: Coccinellidae)**. *Mol Biol Evol* 2001, **18**(4):648-660.

16. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al*: **The Genome Sequence of Drosophila melanogaster**. *Science* 2000, **287**(5461):2185-2195.
17. Fink C, Staubach F, Kuenzel S, Baines JF, Roeder T: **Noninvasive Analysis of Microbiome Dynamics in the Fruit Fly Drosophila melanogaster**. *Applied and Environmental Microbiology* 2013, **79**(22):6984-6988.
18. Fink C, von Frieling J, Knop M, Roeder T: **Drosophila Fecal Sampling**. *Bio-protocol* 2017, **7**(18):e2547.
19. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen F-A, Ruhlemann MC, Szymczak S *et al*: **Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota**. *Nat Genet* 2016, **48**(11):1396-1406.
20. Rehman A, Rausch P, Wang J, Skieceviciene J, Kiudelis G, Bhagalia K, Amarapurkar D, Kupcinskas L, Schreiber S, Rosenstiel P *et al*: **Geographical patterns of the standing and active human gut microbiome in health and IBD**. *Gut* 2015, **65**(2):238-248.
21. Fraune S, Bosch TCG: **Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra**. *Proceedings of the National Academy of Sciences* 2007, **104**(32):13146-13151.
22. Franzenburg S, Fraune S, Künzel S, Baines JF, Domazet-Lošo T, Bosch TCG: **MyD88-deficient Hydra reveal an ancient function of TLR signaling in sensing bacterial colonizers**. *Proceedings of the National Academy of Sciences* 2012, **109**(47):19374-19379.
23. Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D *et al*: **The dynamic genome of Hydra**. *Nature* 2010, **464**:592.
24. Wittlieb J, Khalturin K, Lohmann JU, Anton-Erxleben F, Bosch TCG: **Transgenic *Hydra* allow *in vivo* tracking of individual stem cells during morphogenesis**. *Proceedings of the National Academy of Sciences* 2006, **103**(16):6208-6211.
25. Lenhoff HM, Brown RD: **Mass culture of hydra: an improved method and its application to other aquatic invertebrates**. *Laboratory Animals* 1970, **4**(1):139-154.
26. Ryan JF, Pang K, Schnitzler CE, Nguyen A-D, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, Smith SA *et al*: **The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution**. *Science* 2013, **342**(6164):1242592.
27. Turner LM, Harr B: **Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions**. *eLife* 2014, **3**:e02504.
28. Staubach F, Künzel S, Baines AC, Yee A, McGee BM, Bäckhed F, Baines JF, Johnsen JM: **Expression of the blood-group-related glycosyltransferase B4galnt2 influences the intestinal microbiota in mice**. *ISME J* 2012, **6**(7):1345-1355.
29. Rausch P, Basic M, Batra A, Bischoff SC, Blaut M, Clavel T, Gläsner J, Gopalakrishnan S, Grassl GA, Günther C *et al*: **Analysis of factors contributing to variation in the C57BL/6J fecal microbiota across German animal facilities**. *International Journal of Medical Microbiology* 2016, **306**(5):343-355.
30. Shade A, Jacques M-A, Barret M: **Ecological patterns of seed microbiome diversity, transmission, and assembly**. *Curr Opin Microbiol* 2017, **37**:15-22.
31. Gittleman JL, Kot M: **Adaptation: Statistics and a Null Model for Estimating Phylogenetic Effects**. *Systematic Zoology* 1990, **39**(3):227-241.

Appendix A.2: Supplement of Article C

Supplemental material: ABO histo-blood groups influence gut microbiome, with causal relationship between *Bacteroides* and inflammatory bowel disease

Authors and affiliations:
Malte Christoph Rühlemann[1], Britt Marie Hermes[2,3,4], Corinna Bang[1], Shauni Doms[2,3], Lucas Moitinho-Silva[1,5], Louise Bruun Thingholm[1], Fabian Frost[6], Frauke Degenhardt[1], Michael Wittig[1], Jan Kässens[1], Frank Ulrich Weiss[6], Annette Peters[7,8], Klaus Neuhaus[9], Uwe Völker[10], Henry Völzke[10], Georg Homuth[10], Matthias Laudes[11], Wolfgang Lieb[12], Dirk Haller[9,13], Markus Maximilian Lerch[6], John Baines[2,3], Andre Franke[1]

[1] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany
[2] Evolutionary Genomics, Max-Planck-Institute for Evolutionary Biology, Plön, Germany
[3] Institute of Experimental Medicine, Kiel University, Kiel, Germany
[4] Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany
[5] Department of Dermatology, Kiel University, Kiel, Germany
[6] Department of Medicine A, University Medicine Greifswald, Greifswald, Germany
[7] Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany
[8] German Center for Diabetes Research (DZD), Neuherberg, Germany
[9] ZIEL - Institute for Food & Health, Technical University of Munich, Freising, Germany
[10] Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
[11] Department of Internal Medicine 1, Kiel University, Kiel, Germany
[12] Institute of Epidemiology, Kiel University, Kiel, Germany
[13] Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany

Correspondence should be addressed to: Prof. Dr. Andre Franke, Institute of Clinical Molecular Biology (IKMB), Kiel University, Rosalind-Franklin-Str. 12, 24106 Kiel, Germany. Email: a.franke@mucosa.de

Supplementary Tables

Supplementary Table S1: Cohort level summaries of all taxa included in the mGWAS analysis, including feature label used in the analysis, taxonomic/clustering level, amplicon nucleotide sequence of the V1-V2 region of the 16S rRNA gene, presence in cohort (%), mean/median abundance in all individuals in the cohort; mean/median abundance of all individuals with the respective feature present.

Supplementary Table S2: The top 10,000 variants across all features from the univariate logistic regression analysis for presence/absence patterns of microbial features.

Supplementary Table S3: The top 10,000 variants across all features from the univariate linear regression analysis of residual abundance of microbial features.

Supplementary Table S4: The top 10,000 variants across all features from the multivariate analysis of Bray-Curtis and 97%-identity OTU based weighted UniFrac distance

Supplementary Table S5: Results of all univariate analyses of ABO hist-blood groups on microbial features using hurdle models.

Supplementary Table S6: Results of all Mendelian Randomization analyses of univariate traits.

Supplementary Table S7: Replication of results from previous host-microbiome GWAS analyses.

Supplementary Table S8: Results of the gene-set and tissue enrichment analysis using the FUMA web application.

Supplementary Methods

Traits used for Mendelian Randomization

The traits used for Mendelian Randomization were selected to be binary with “log odds” as effect units. Excluding traits with sample size < 500 and from the categories “Behavioural” and “Education”. Additionally excluded were subtypes of ovarian and lung cancer, melanoma, neuroblastoma, extreme height, “Top 1% surival”, “Diabetic nephropathy” and “oligoclonal band status”.

Subcategories, traits and respective MR-Base IDs used in the Mendelian Randomization analysis:

Subcategory	Trait	MR-Base IDs
Anthropometric	Extreme body mass index	85
Anthropometric	Extreme waist-to-hip ratio	87
Anthropometric	Obesity class 1	90
Anthropometric	Obesity class 2	91
Anthropometric	Obesity class 3	92
Anthropometric	Overweight	93
Autoimmune / inflammatory	Asthma	44
Autoimmune / inflammatory	Celiac disease	1058, 1059, 1060, 276, 278
Autoimmune / inflammatory	Crohn's disease	10, 11, 12, 13, 14, 15, 30
Autoimmune / inflammatory	Eczema	996
Autoimmune / inflammatory	Gout	1054
Autoimmune / inflammatory	Inflammatory bowel disease	292, 293, 294, 295, 296, 31, 819
Autoimmune / inflammatory	Multiple sclerosis	1024, 1025, 280, 286, 820, 821
Autoimmune / inflammatory	Rheumatoid arthritis	283, 831, 832, 833, 834
Autoimmune / inflammatory	Sarcoidosis	981
Autoimmune / inflammatory	Systemic lupus erythematosus	288, 815

Autoimmune / inflammatory	Ulcerative colitis	32, 968, 969, 970, 971, 972, 973
Bone	Paget's disease	975
Cancer	Gallbladder cancer	1057
Cancer	Ovarian cancer	1120
Cancer	Pancreatic cancer	822
Cancer	Prostate cancer	823
Cancer	Prostate cancer (overall)	1174
Cancer	Upper gastrointestinal cancers	825
Cardiovascular	Coronary heart disease	6, 7, 8, 9
Diabetes	Type 2 diabetes	1090, 23, 24, 25, 26, 976
Kidney	Chronic kidney disease	1102, 17
Kidney	IgA nephropathy	1081
Kidney	Microalbuminuria	1097, 20
Paediatric disease	Hirschsprung's disease	983
Psychiatric / neurological	Alzheimer's disease	297, 298, 824
Psychiatric / neurological	Amyotrophic lateral sclerosis	1085, 1086
Psychiatric / neurological	Anorexia nervosa	45
Psychiatric / neurological	Attention deficit hyperactivity disorder	799
Psychiatric / neurological	Autism	802, 806
Psychiatric / neurological	Bipolar disorder	800, 801, 808
Psychiatric / neurological	Bulimia nervosa	990
Psychiatric / neurological	Major depressive disorder	804, 805
Psychiatric / neurological	Parkinson's disease	811, 812, 818
Psychiatric / neurological	Schizophrenia	22, 810

Replication of signals from previous studies

We included a total of 179 independent loci with previously identified associations from four genome-wide association studies.[1–4] Of these 40, 51 and 9 were identified to be associated with taxon abundances by Wang et al., Turpin et al., and Bonder et al., respectively. Another 42 and

4 were associated with beta-diversity by Wang et al. and Rühlemann et al. The remaining 13 and 20 were found to be associated with GO2000 terms and MetaCyc pathways by Bonder et al. The threshold for replication was adjusted to the total number of taxa involved in the taxon-based test: p_{thresh} = 0.05/(146 + 225) = 0.05/371 = 1.35 x 10^{-4}. We marked a locus as replicated, if we found an association surpassing this threshold in a window of +/- 10kb surrounding the variant, independent of the original trait the signal was associated with. Loci surpassing this threshold were annotated with the protein coding gene overlapping or closest to the locus, with a maximum distance of 100kb up- or downstream.

Gene set enrichment and tissue specificity analysis

Genes overlapping with genome-wide significant risk loci and closest to loci replicated from previous studies were subjected to gene set enrichment analysis using the GENE2FUNC module of the FUMA GWAS webservice (https://fuma.ctglab.nl/gene2func). All parameters were kept in their default state (Ensembl v92, GTEx v6) and “All genes” were selected as background for enrichment analysis.

Supplementary Results

Cohorts baseline comparison

The proportion of female study participants was between 45.3% (PopGen) and 57.8% (FoCus; see Figure 1A). The lowest mean BMI was observed in the FoCus cohort (26.4 kg/m²), the highest in the SHIP cohort (28.3 kg/m²; Figure 1B), pairwise comparisons (Wilcoxon rank sum test) of the cohorts showed significant values (q_{Holm} <.05), except for KORA vs. PopGen (q_{Holm}=.0611) and KORA vs. SHIP-Trend (q_{Holm} =.451). The highest average age was observed in the PopGen and KORA cohorts (61.5 and 60.6 years, respectively), the lowest average age was seen in the FoCus and SHIP-TREND cohorts (51.4 and 51.3 years, respectively; Figure 1, Panel C). Here, also all pairwise comparisons were significant, except for SHIP-TREND vs. FoCus (q_{Holm} =.4) On a community scale, between sample diversity (beta diversity) as calculated by genus-level Bray-Curtis dissimilarity clearly differed between cohorts (all pairwise p_{adonis}<.001), while effect sizes generally were low, with pairwise variance between cohorts ranging from R^2=0.076% between SHIP and SHIP-Trend and R^2=5.85% between FoCus and KORA, median pairwise difference being 2.49%. The three highest values of the pairwise comparisons are all seen with the FoCus cohort. This is also evident when looking at the

ordination (Figure 1E), where the centroid of the FoCus cohort separates from the other centroids. Alpha diversities as assessed by Shannon genus-equivalent and the total number of observed genera showed to be generally in the same range between cohorts, however clearly significantly different, with $q_{Holm}<10^{-5}$ in all comparisons, except for SHIP vs. SHIP-TREND in both diversity indices ($q_{Holm}=.057$ and $q_{Holm}=.25$ for observed genera and Shannon genus-equivalent, respectively) and PopGen vs. SHIP and SHIP-TREND in Shannon genus-equivalent (both $q_{Holm}=.93$; Figure 1F). Though we see clear statistical differences in the overall comparison between cohorts, we can observe highly similar patterns in most of the highly abundant and prevalent taxa on phylum and genus level that met the inclusion criteria for genome-wide association analysis (Figure 1G), and it is only around the taxon median abundances in single digit percentage range and below where larger between-cohort variations are seen.

Benchmarking of beta diversity calculations on CPU and GPU

To assess speedup of beta diversity calculations implemented on GPU hardware, we set up a benchmark. For all cohorts 10, 100 and 1000 SNPs with minor allele frequency of > 10% were randomly drawn from chromosome 1. Calculations were performed using 1, 2, 5 and 10 CPU threads on our local HPC infrastructure, always requesting complete nodes (Intel Xeon E5-2670, 8 CPU, 16 threads, 128Gb RAM) to ensure undisturbed performance measures. For GPU-based calculations, we performed the same set of calculations on a NVIDIA Tesla P100 with 16Gb RAM and a single CPU core for pre-processing of the data. Assessed calculation times only include the actual calculation steps and do not included data loading and preprocessing, maximum runtime per job was set to 48 hours. Each calculation was run in 5 replicates to get confident estimates of average performance.

For the two smaller cohorts (Focus and PopGen, n < 1000) all benchmarking jobs finished. For the intermediate sized cohorts (SHIP and KORA, n ~ 2000) the jobs running on 10 parallel CPUs and calculation batches of 1,000 SNPs failed due to exceeding maximum available memory of 128Gb. For the largest cohort with n = 3382 (SHIP-Trend), jobs with 1,000 variants on a single CPU failed du to exceeding the time limit, for 5 and 10 CPUs, all jobs with batches of 100 and 1,000 variants failed due to exceeding memory. Only for 2 parallel CPUs all jobs completed. All GPU-based calculations completed without failure.

Based on the Focus cohort, the results show the expected linear dependence between calculation time and the number of variants analyzed (Figure S1A). The number of utilized CPU

decreases calculation time, however CPU time does not scale linearly with CPU time being 24.0% and 50.3% higher when using 5 and 10 CPUs, respectively, compared to a single CPU (Figure S1, A and B,). Utilization of a GPU massively surpasses CPU speed, showing a 20 to 30-fold speedup compared to 1 and 10 CPUs in computation time, respectively (Figure SX, B). This trend is visible and even more pronounced in the other cohorts, especially with increasing sample size, peaking in the largest cohort with n = 3382 (SHIP-Trend) with a 226-fold speedup of computation time comparing GPU with 10 parallel CPUs.

The results show, that while an increase of parallel CPUs decreases total time, though computational time increases for the single variant. The bottleneck in parallel computation is the amount available RAM, as increased number of parallel processes increases memory usage and may lead to failure, as seen in the multi-CPU calculations for the larger cohorts with n > 1000. GPUs are known for their good performance in multiplication of large matrices and it shows in this application, with a single GPU massively surpassing CPU performance.

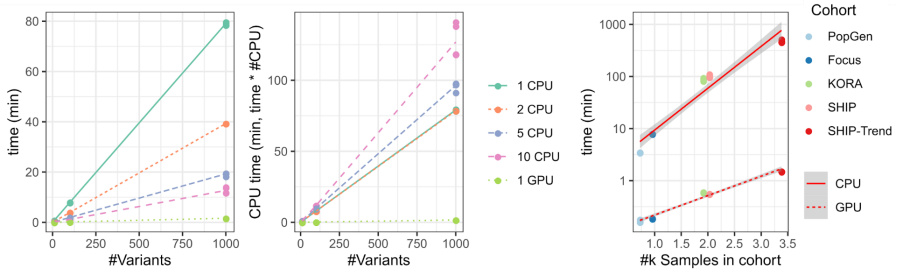


Figure S1: Summary of CPU- and GPU-based calculations for beta-diversity analysis. (A) Relationship of elapsed time and the number of analyzed variants in the Focus cohort depending on the number of CPUs and GPUs used in parallel. (B) Relationship of computational time in the Focus cohort depending on the number of CPUs and GPUs used in parallel. (C) Relationship of calculation time and cohort size depending on the usage of a single CPU or GPU for analysis.

Average CPU/GPU time in seconds for 10 random variants on chromosome 1 based on 5 replicates per calculation

	Samples	1 CPU	2 CPU	5 CPU	10 CPU	1 GPU
PopGen	724	21.0	20.9	21.8	26.6	2.2

Focus	957	47.9	47.5	59.4	72.0	2.4
KORA	1915	491.9	474.6	559.8	565.2	4.7
SHIP	2029	586.2	567.3	662.5	670.0	5.1
SHIP-Trend	3382	2749.4	2644.3	3116.9	3091.9	11.6

Replication of previous loci from mGWAS analyses

A total of 179 study-wise independent loci were included in the replication analysis, of which 88 (=49.16%) met the replication criteria. The highest replication rate was achieved for signals from Bonder et al. in association with taxonomic groups (6 out of 9, 66.7%). The lowest replication rate was found for GO2000 terms from the same study (5 out of 13, 38.5%). For for 73 of the 88 loci, a close-by gene wascould be identified. Among the replicated loci we found the SLC9A8 gene locus (chr20:48,429,250-48,508,779) encoding for NHE8, a sodium/hydrogen exchanger, which has been identified as associated locus twice before in the German cohorts [Wang, Rühlemann]. SLC9A8 is expressed in goblet cells in the intestine has been shown to be essential for mucosal integrity. Loss of expression has been shown to be connected to increased bacterial adhesion and inflammation in mice after DSS treatment. One locus replicated from Bonder et al. overlapped with the c*ontactin 6* (CNTN6; chr3:1,134,620-1,445,278) gene, which is also the closest gene to one locus replicated from Wang et al. (~ 50kb upstream). This locus showed the second lowest replication P-value (P_{META}=9.98 x 10^{-7}). The lowest P-value for a replicating locus was found in the ALDH1A3 gene locus on chromosome 15 (rs8040493), first found by Wang et al. in association with beta diversity, now showing association with OTU97_11 belonging to the genus *Parabacteroides* (P_{META}=8.47 x 10^{-7}). A complete list of loci and their replication results is provided as **supplementary table S6.**

Gene set enrichment and tissue specificity analysis

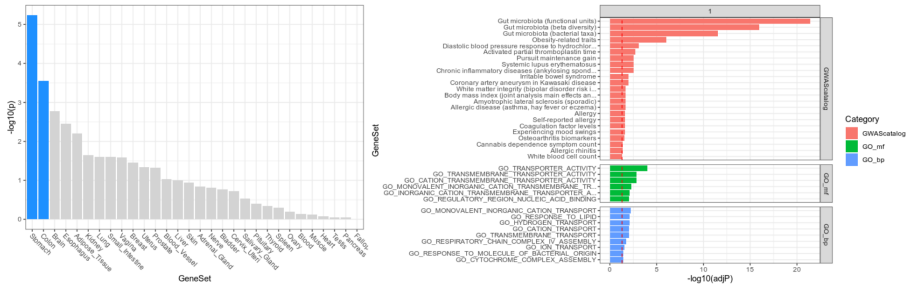


Figure S2: Summary of the tissues-specificity and gene-set enrichment analysis. All Resuls can be found in supplementary table S7.

The genome-wide association analysis and replication of previously reported results identified a total of 82 unique genes in loci with associations to microbial traits. These genes were subjected to enrichment and tissues specificity analysis using the FUMA webservice (see **Methods**). The tissue specificity analyses targeting 30 general tissues and 53 tissue types identified both an enrichment of genes differentially expressed in stomach tissue. The analysis of general tissue types additionally identifies an enrichment of genes differentially expressed in Colon (both directions) and genes up-regulated in esophagus. Gene-set enrichment identified a total of 103 gene sets to be significantly enriched (q < .05). The largest groups consisted of transcription factor targets (n=33) and enrichments in loci from traits in the GWAS catalogue (n=27). As the enrichment analysis included replications of previously identified loci associated with microbial traits, these categories show to be highly enriched. Among the remaining sets we find connections to obesity (*Obesity-related traits*, q=9.26 x 10^{-7}; *Body-mass-index*, q=2.15 x 10^{-2}) and chronic inflammation (*systemic lupus erythematosus* and *Chronic inflammatory diseases (pleiotropy)*, both q=2.95 x 10^{-3}). Enrichments in set from *Gene Ontology (GO) Terms* (GO molecular functions, n=6; GO biological processes, n=5) suggest metabolic interactions between host and microorganisms, indicated by the enrichment of different terms involving transport, e.g. "GO:*Transporter Activity*" (q=9.17 x 10^{-5}) and "GO:*Transmembrane Transporter Activity*" (q=1.36 x 10^{-3}). Additionally, we find enrichments possibly suggesting response to dietary intake (GO:*Response to lipids*, q=7.85 x 10^{-3}) and direct response to bacteria derived molecules (GO:*Response to molecule of bacterial origin*, q=3.73 x 10^{-3}). Further categories with enriched gene sets were Positional gene sets (n=4), Immunological signatures (n=1), canonical pathways (n=1), curated gene sets (n=7), chemical and genetic perturbation (n=16) and microRNA targets (n=3). The complete lists can be found in **Supplementary table S7**.

Supplemental Literature

1. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
2. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
3. Rühlemann, M. C. *et al.* Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes* **9**, 68–75 (2017).
4. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).

Appendix A.3: Supplement of Article E

SUPPLEMENTARY MATERIAL: Consistent Alterations in Fecal Microbiomes of Patients With Primary Sclerosing Cholangitis

Malte Rühlemann[1,#], Timur Liwinski[2,#], Femke-Anouska Heinsen[1,#], Corinna Bang[1], Roman Zenouzi[2], Martin Kummen[3,4], Louise Thingholm[1], Marie Tempel[5], Wolfgang Lieb[5], Tom Karlsen[3,4,6,7], Ansgar Lohse[2], Johannes Hov[3,4,6,7], Gerald Denk[8], Frank Lammert[9], Marcin Krawczyk[9,10], Christoph Schramm[2,11,†], Andre Franke[1,†]

[1] Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany

[2] I. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[3] Norwegian PSC Research Center, Division of Surgery, Inflammatory Medicine and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway

[4] Research Institute of Internal Medicine, Oslo University Hospital Rikshospitalet, Oslo, Norway

[5] Institute of Epidemiology, Christian-Albrechts-University of Kiel, Kiel, Germany

[6] Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[7] Section of Gastroenterology, Department of Transplantation Medicine, Division of Surgery, Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway

[8] Department of Medicine II, Liver Center Munich, University Hospital, LMU Munich, Munich, Germany

[9] Department of Medicine II, Saarland University Medical Center, Saarland University, Homburg, Germany

[10] Laboratory of Metabolic Liver Diseases, Center for Preclinical Research, Department of General, Transplant and Liver Surgery, Medical University of Warsaw, Warsaw, Poland

[11] Martin Zeitz Centre for Rare Diseases, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[#] MCR, TL and F-AH contributed equally to this work; [†] AF and CS jointly supervised this work

TABLE OF CONTENTS

SUPPLEMENTARY METHODS

Sample selection and exclusion

All individuals underwent extensive screening for confounding factors. Any signs of systemic disease assessed by routine laboratory parameters deviating from healthy reference ranges, namely CRP ≤5 mg/l, glucose between 55 and 115 mg/dl, and insulin between 2.6 and 24.9 mU/l, and a comprehensive questionnaire covering self-reported status on disease (diabetes, cancer, respiratory-, liver, coronary heart disease, heart attack, neuropathy, IBD, IBS, chronic diarrhea, asthma, organ transplantation, phlebitis, varices, venous insufficiency) were criteria for study exclusion. All study participants were unrelated individuals.

To avoid confounding by unspecific microbiota changes in chronic liver disease, PSC patients with concurrent signs of autoimmune hepatitis were excluded from the analysis. The vast majority of PSC samples did not exhibit any signs of liver cirrhosis or impaired liver function. UC was diagnosed based on medical history and colonoscopy with biopsy according to recent guidelines and excluding infectious colitis.[S1]

Exclusion criteria for all groups were treatment with antibiotics within six weeks before stool collection, missing data on age, gender or BMI, other chronic liver diseases and prior colectomy. Individuals with Crohn's disease-like inflammatory patterns affecting the small bowel or undetermined type of colitis were excluded from the analysis due to small sample numbers in the German cohort (n=7 and n=2, respectively).

Stool Biomarkers

Calprotectin (as an indicator of intestinal inflammation) was measured in all German fecal samples using the Bühlmann fCAL™ ELISA kit (BÜHLMANN LABORATORIES AG) and analyzed using the SoftMax Pro Software (Molecular Devices).

Sequencing and bioinformatics processing of 16S rRNA gene libraries

The 16S rRNA gene library generation and sequencing was performed as previously described.[S2] In short, the V1-V2 region of the 16S rRNA gene was sequenced on the MiSeq platform, using the 27F-338R primer pair and a dual indexing approach. Obtained sequences data were trimmed in paired-end mode using sickle and default parameters.[S3] Forward and reverse reads were merged using VSEARCH v2.5 and quality controlled using FastX-Toolkit. Sequences with more than 5% per-base-quality below 30 were discarded. Chimeric reads were identified by both a reference-based (utilizing the UCHIME 'gold.fa' database obtained from https://drive5.com/uchime/uchime_download.html) and a de-novo approach based on

the UCHIME algorithm implemented in VSEARCH.[S4] Sequences were taxonomically annotated using the SINTAX algorithm in USEARCH (v9) and the RDP database version 14.[S5] Sequences that could not be assigned to the domain 'Bacteria' were discarded. Clustering into Operational Taxonomic Units (OTUs) was performed using the VSEARCH. Sequencing-depth was normalized by randomly picking 10.000 reads per sample and abundance tables for all taxonomic levels and OTUs were generated. The number of 10,000 sequences was choses, as this was defined as the minimum sequence count threshold after QC and a widely used choice for human gut microbiome analysis.

Analysis of differences in alpha- and beta-diversity

Differences in alpha-diversity were assessed using the non-parametric Wilcoxon rank-sum test. Permutational MANOVA (*adonis* function from the *vegan* package[S6]) using Bray-Curtis dissimilarity was applied to assess differences in correcting for the covariates age, gender and BMI. Unconstrained ordination plots were created based on Bray-Curtis dissimilarity (*vegdist* function in the *vegan* package). Square-root transformation of the dissimilarity was used to eliminate negative eigenvalues of ordination axes.[S7]

Analysis of healthy core microbiota

To explore the overlap and differences in healthy microbiota between the Norwegian and the German cohorts, core microbiota analysis was performed. A Venn diagram was drawn to illustrate common taxa prevalence on the genus taxonomic level ("vennDiagram" R package[S8]). To avoid bias by rare taxa, only taxa present in at least 5% prevalence in each respective cohort were considered. Core microbiota heatmaps were drawn separately and compared for the Norwegian and German healthy cohorts as described previously,[S9] considering genera with minimum prevalence of 10% in the respective cohorts.

Selection of genera from Kummen *et al.* and Sabino *et al.* for replication

In the article by *Kummen et al.*[S10], Figure 3 summarizes 10 taxonomic groups as differentially abundant in PSC compared to HC. Of these, five were classified on genus level. The genus *Desulfovibrio* was not present in any of the two cohorts, thus discarded from the analysis. All members of the family Christensenellaceae were classified down to genus level *Christensenella,* perfectly correlating to the genus level abundance, thus being included in the analysis. Sabino *et al*.[S11] identified four genera as "PSC signature" (*Fusobacterium, Enterococcus, Streptococcus* and *Lactobacillus*), as summarized in their Figure 3, all of which could be recovered in both cohorts presented in this study.

Analysis for differential prevalence and abundance patterns and meta-analysis

All taxa with at least 20% presence were transformed to dichotomous features (0 for absence, 1 for presence) and subjected to logistic regression (R base function *glm)*, adjusting for the covariates age, gender and BMI. If the abundance in the samples where the taxon was present exceeded a mean abundance of 0.5% in one of the groups being compared, the taxa were additionally subjected to abundance-based analysis using hurdle models as implemented in the *pscl* R package [S12] if the respective taxa contained zero values or otherwise generalized linear models (GLM) as provided in the *MASS* R package.[S13] Both models were applied assuming a negative binomial distribution for the count part of the abundances and including the covariates age, gender and BMI analogously to the logistic regression. Extreme abundance values deviating more than five interquartile ranges from the group median were excluded to minimize outlier-driven false positive results. Count models give more statistical power to detect differential expression than approximate normal models in high-throughput sequencing data.[14] Therefore, high-throughput sequencing datasets were advised to be treated as count data.[15] Moreover, count models have the advantage for properly separating biological from technical variation.[16] The meta-analysis was performed separately for logistic regression and abundance-based models, however following the same approach. Briefly, this approach combines the effect sizes from the models combined weighted by a termed based on the inverted variance as calculated from the standard errors taken from the respective models. The Z-score resulting from this model can be used to calculate the meta-analysis P-value. These P-values were subsequently corrected for multiple testing using the Benjamini-Hochberg-correction. Only taxa with a resulting $P<0.05$ in each cohort and $Q_{META}<0.05$, after correction across all taxonomic levels used in the analysis to ensure a minimum of false postitves, were regarded as robustly differentially distributed.

Analysis of effects of medical treatment

To assess the effects of PSC medication on the microbiota, all medications received by at least 10% of PSC patients were tested. These included ursodeoxycholic acid (UDCA), mesalazine (5ASA) and azathioprine (AZA). Analyses of effects on beta-diversity and single taxonomic groups were performed as described earlier, using intake of the respective drugs as additional covariates.

Machine learning classification

Random forest classification was implemented using the package *ranger* [S17] via the *caret* interface with default hyperparameters. The number of variables at each node (mtry) was

held constant according to the square root of included model features as suggested by Breiman.

Taxa with robust differential distribution were included either as continuous variables or dichotomous features, according to the prespecified abundance threshold.

To estimate the generalization error the 0.632 bootstrap estimator was applied with 999 iterations. [S18] For Evaluation of model performance, the area under the receiver operating characteristic (ROC) curve (AUC) was used. In addition, contingency table related measures are reported: The F1 score represents the harmonic average between the recall (sensitivity) and precision (true positive rate) of a model. The optimal value is reached at F1=1. Matthews Correlation coefficient (MCC) reflects the improvement of agreement between predicted and actual values over random prediction with respect to frequency of each class. [S19] Therefore, the MCC represent a performance measure which is robust to class imbalance in a data set. Perfect agreement is achieved at MCC=1.

Feature importance of the pooled random forest classifier was calculated using the Gini index (scaled 0 to 100).

Learning curves were drawn to explore the relationship between the size of the training cohort and classifier performance.

To validate the resampling results of the pooled random forest classification, additional resampling methods were implemented: repeated cross validation (3 folds, 10 repeats) and leave one out cross-validation. [S20]

To proof that the classification performance is better explained by the data structure than the chosen model, we implemented further independent machine learning algorithms: (guided) regularized random forest, radial kernel support vector machine and extreme gradient boosting. [S21-S23] Generalization error of each model was estimated using 0.632 bootstrapping.

To control for potential overfitting due to many features, we used two different feature reduction strategies: First, we implemented recursive feature elimination using repeated cross-validation (3 folds, 10 repeats). The dependence of ROC and MCC on the number of features was assessed visually. Second, we applied the 'Boruta' feature selection algorithm with default parameters.[S24] Within this framework, so called 'shadow variables' are created via permutation of the original variables. The most important variables are selected via comparison of the importance of each real variable with the maximum value of all shadow variables. The pooled random forest classification was repeated with a reduced feature number selected by the Boruta algorithm using 0.632 bootstrapping.

Analysis of calprotectin measures

Differences in fecal calprotectin were assessed using the non-parametric Wilcoxon rank-sum test (*wilcox.test* from the R-base package).

SUPPLEMENTARY RESULTS

Healthy Norwegian and German cohorts share similar core microbiota

Of the 144 genera present in at least 5% of either the Norwegian or the German cohort, 122 (84.72%) were prevalent in both cohorts (**Figure S1**). Only six genera (4.16%) were found exclusively in Norwegian patients. Exclusive taxa were confined to low abundance genera (3rd quartile of mean relative abundance 3.49×10^{-5}). Core microbiota heatmaps showed similar taxonomic abundance/prevalence patterns in both cohorts (**Figure S2**).

Furthermore, applying principal coordinates analysis to the healthy cohorts, the first principal coordinate, which showed the most pronounced difference between Norwegian and German healthy controls (Wilcoxon rank-sum test; $p=1.89\times10^{-4}$) was highly correlated with the Shannon index (Spearman's rank correlation test; $r=0.53$, $p=6.23\times10^{-11}$; **Figure S3**).

To summarize, Norwegian and German control cohorts share similar healthy core microbiota, differences in beta-diversity are mainly explained by the difference in alpha-diversity, and to a lesser extent by disparities of the taxonomic composition.

Minor differences between German and Norwegian healthy controls

Four genera (*Blautia* ($P_{adj}=4x10^{-4}$, $\beta=-1.04$), *Escherichia/Shigella* ($P_{adj}=9.5x10^{-3}$, $\beta=-4.27$), *Akkermansia* ($P_{adj}=8x10^{-4}$, $\beta=-3.06$) and *Haemophilus* ($P_{adj}=1.3x10^{-2}$, $\beta=-2.8$)) as well as the family *Enterobacteriaceae* ($P_{adj}=3x10^{-4}$, $\beta=-3.45$) were significantly higher abundant in the German samples compared to the Norwegian samples, whereas the genus *Clostridium Cluster IV* ($P_{adj}=8.9x10^{-3}$, $\beta=1.18$) and the class Alphaproteobacteria ($P_{adj}=1.3x10^{-3}$, $\beta=2.53$) showed higher abundances in the Norwegian cohort. Additionally, the *Escherichia/Shigella* cluster was more prevalent in the German cohort ($P_{adj}=1.21x10^{-2}$, $\beta=-1.89$).

Medical treatment of PSC has no influence on the fecal microbiome.

Initiation of UDCA therapy has previously been reported to change the intestinal microbiome of patients with primary biliary cirrhosis (PBC).[19] In our cross-sectional cohort of PSC patients, neither UDCA, nor 5ASA and AZA had any significant influence on the overall beta-diversity ($P>0.05$) or individual taxa (all $P_{META}>0.05$).

Calprotectin not significantly elevated in PSC

Pairwise Wilcoxon-rank-sum-test of fecal Calprotectin measurements did not show any differences between PSC patients and healthy controls ($P_{Wilcox,controls\text{-}PSC}$=0.97), however UC patients differed from both PSC and controls ($P_{Wilcox,\ controls\text{-}UC}$=0.0077, $P_{Wilcox,PSC\text{-}UC}$=0.02). Splitting PSC patients into PSC-only and PSC-IBD, the test is significant between PSC-only and UC ($P_{Wilcox,PSConly\text{-}UC}$=0.013), however neither between PSC-only and PSC-IBD ($P_{Wilcox,PSConly\text{-}PSC\text{-}IBD}$=0.48) and PSC-IBD and UC ($P_{Wilcox,PSC\text{-}IBD,UC}$=0.12).

Known patterns in the microbiota of UC compared to healthy controls.

We observed major differences in the microbiota of patients with UC and controls and were able to reproduce results of former studies.

Patients with UC displayed a cohort-spanning increase in bacteria belonging to the phylum of Proteobacteria and more specifically the class Gammaprotebacteria (Q_{META}=1.3x10^{-7}, P_{GER}=4.7x10^{-6}, P_{NOR}=5.7x10^{-5} and Q_{META}=7.7x10^{-3}, P_{GER}=8.3x10^{-3}, P_{NOR}=3.6x10^{-2}). Additionally, we could observe an increase in *Pasteurellacae* (Q_{META}=1.8x10^{-2}, P_{GER}=3.8x10^{-2}, P_{NOR}=3.3x10^{-2}) and OTU_13 belonging to the genus *Bacteroides* (Q_{META}=1.0x10^{-4}, P_{GER}=2.9x10^{-4}, P_{NOR}=5.8x10^{-3}). A decreased abundance associated with UC was found in the phylum Firmicutes (Q_{META}=1.0x10^{-4}, P_{GER}=2.9x10^{-4}, P_{NOR}=5.8x10^{-3}) and in sub-clades thereof. The strongest effects were seen in two OTUs, OTU_84: *Coprococcus spp.* (Q_{META}=1.0x10^{-4}, P_{GER}=4.8x10^{-5}, P_{NOR}=1.8x10^{-2}) and OTU_3385 belonging to the family of *Ruminococcaceae* (Q_{META}=1.0x10^{-4}, P_{GER}=9.8x10^{-3}, P_{NOR}=6.4x10^{-3}, **Figure S4**, **Supplementary Table S7**).

Furthermore, we found a higher prevalence of known mediators like *Akkermansia*, *Bilophila*, *Coprococcus* and *Ruminococcus* in healthy controls, which were among a total of 47 taxonomic groups more likely to be absent in UC cases (**Supplementary Table S8**, **Figure S5**), also including four OTUs belonging to the genus *Faecalibacterium*. *Bilophila* was also found to be reduced in PSC patients, indicating an association to overall gut health. Only three taxa were found to be more prevalent in UC patients, three of which were unspecified members of the family *Lachnospiraceae* (OTU_119, OTU_376 and OTU_457) and one classified as *Clostridium Cluster XVIII spp.* (OTU_109).

Machine learning results

A detailed summary of the cross-cohort classification is provided in **Table S13**. Ranked Gini variable importance of the pooled random forest classifier is illustrated in **Figure S6**.

Learning curves of the random forest classifier (**Figure S6**) indicate a constantly perfect classification of the training set (regardless of the size). The baseline test performance is

high already at small training set sizes (AUC>0.8) and converges with the training performance with increasing training sample sizes (AUC>0.9).

Repeating the pooled random forest classification over more conservative resampling methods resulted in a less optimistic but still high classifier performance (**Table S14**).

Repeating the pooled classification (n=270 subjects, n=43 features) with different models resulted in a comparably high performance (**Table S15**).

Implementing recursive feature elimination, random forest classifier performance steeply increased with increasing feature number and plateaued at ca. 20 features (**Figure S7**).

Boruta feature selection algorithm identified 23 variables which significantly contributed to loss random forest classifier performance (**Figure S8**). Repeating the classification with the reduced variables set did not result in a loss of classifier performance (AUC=0.89, F1=0.83, MCC=0.67; **Figure S9**). The selected taxa are shown in **Figure S8** ranked by Gini importance (scaled 0 to 100).

SUPPLEMENTARY TABLES

Supplementary Table S1: Summary of the targeted analysis of previously identified genera associated with PSC. Only genera that could be recovered in the dataset are presented. Mean abundance is calculated based on non-zero values. Details of models and P-values in the cohorts and the meta-analysis can be found in Supplementary Table S2.

Differential prevalence

	Prevalence (%)						
	GER		NOR		Results Logistic Regression		
Genus	Controls	PSC	Controls	PSC	GER	NOR	META
Kummen et al.							
Christensenella	18.9	18.9	18.4	9.5	-	-	-
Coprococcus	88.4	67.6	97.4	85.7	*	n.s.	n.s.
Phascolarctobacterium	46.3	32.4	68.4	41.3	n.s.	**	n.s.
Succinivibrio	0	0	2.6	1.5	-	-	-
Veillonella	70.5	75.7	60.5	73	n.s.	n.s.	n.s.
Sabino et al.							
Enterococcus	13.7	33.8	2.6	19	***	*	***
Fusobacterium	10.5	20.3	7.9	17.5	n.s.	n.s.	n.s.
Lactobacillus	37.9	77	26.3	55.6	***	**	***
Streptococcus	91.6	97.3	92.1	96.8	n.s.	n.s.	n.s.

Differential abundance

	Mean Abundance (%)						
	GER		NOR		Results GLM		
Genus	Controls	PSC	Controls	PSC	GER	NOR	META
Kummen et al.							
Christensenella	0.02	0.02	0.01	0.01	-	-	-
Coprococcus	0.43	0.34	1.1	0.54	n.s.	n.s.	n.s.
Phascolarctobacterium	3.26	2.95	1.91	1.46	n.s.	n.s.	n.s.
Succinivibrio	0	0	7.99	0.06	-	-	-
Veillonella	0.16	1.8	0.12	0.42	*	**	***
Sabino et al.							
Enterococcus	0.04	0.51	0.01	0.31	-	-	-
Fusobacterium	0.03	0.15	0.01	0.13	-	-	-
Lactobacillus	0.16	1	0.11	1.01	*	n.s.	n.s.
Streptococcus	0.3	0.95	0.17	0.46	***	***	***

n.s.: not significant; * P<0.05; ** P<0.01; *** P<0.001; - : not tested due to low prevalence or abundance

Supplementary Table S2: Targeted analysis of previously identified genera associated with PSC. Differential-abundance and differential-prevalence analysis was performed using generalized linear models on the non-zero count values and presence/absence status, respectively. Meta-analysis was performed using the inverse-variance weighted Z-score.

Differential abundance

GER	BETA	SE	Z	P-value	Excl. Thresh	n Excluded	n Count	n Zero
g.Veillonella	1,0478887	0,520233	2,0142679	0,04398141	106	19	104	46
g.Phascolarctobacterium	0,2685215	0,4539669	0,5915002	0,554185327	2600	1	67	101
g.Coprococcus	-0,3091227	0,2698947	-1,1453453	0,252066139	220,25	4	130	35
g.Lactobacillus	1,2129193	0,5449075	2,2259177	0,026019692	181	11	82	76
g.Streptococcus	0,8866685	0,235146	3,7707147	0,000162781	293	4	155	10

NOR	BETA	SE	Z	P-value	Excl. Thresh	n Excluded	n Count	n Zero
g.Veillonella	1,23446171	0,4553572	2,71097452	0,006708578	74	4	65	32
g.Phascolarctobacterium	-0,25765376	0,4210821	-0,6118848	0,540614	1686,25	0	52	49
g.Coprococcus	-0,18317647	0,2755271	-0,66482204	0,5061643	330,5	8	83	10
g.Lactobacillus	-0,08117386	0,9242498	-0,08782675	0,9300144	151	3	42	56
g.Streptococcus	1,14899675	0,2450015	4,68975331	2,74E-06	125	6	90	5

META	w GER	w NOR	SE	BETA	Z	P-value
g.Veillonella	3,694913	4,82276	0,3426412	1,15352751	3,36657598	0,000761076
g.Phascolarctobacterium	4,852345	5,639835	0,3087217	-0,01431215	-0,04635941	0,9630238
g.Coprococcus	13,728123	13,172598	0,1928049	-0,24745003	-1,28342196	0,1993443
g.Lactobacillus	3,367863	1,170634	0,469401	0,87912823	1,8728727	0,06108597
g.Streptococcus	18,08526	16,659517	0,1696505	1,01245035	5,96785826	2,40E-04

Differential prevalence

GER	BETA	SE	Z	P-value	Excl. Thresh	n Excluded	n Count	n Zero
g.Veillonella	0,08013801	0,3845355	0,2084021	0,834915	106	19	104	46
g.Phascolarctobacterium	-0,56652463	0,3402311	-1,6651172	0,09588939	2600	1	67	101
g.Coprococcus	-1,40763613	0,4256146	-3,3073021	0,000941992	220,25	4	130	35
g.Lactobacillus	1,55340248	0,3757886	4,1337134	0,0000357	181	11	82	76
g.Enterococcus	1,11313454	0,4070803	2,7344347	0,006248749	NA	0	38	131
g.Fusobacterium	0.514840	0.476090	1.555	0.1198	19	1	25	144
g.Streptococcus	1,22858584	0,8627039	1,4241107	0,1544144	293	4	155	10

NOR	BETA	SE	Z	P-value	Excl. Thresh	n Excluded	n Count	n Zero
g.Veillonella	0,4775925	0,4539256	1,052138	0,292736093	74	4	65	32
g.Phascolarctobacterium	-1,2079426	0,443242	-2,725244	0,006425392	1686,25	0	52	49
g.Coprococcus	-1,6575232	1,0862435	-1,525922	0,127029198	330,5	8	83	10
g.Lactobacillus	1,2504417	0,4637517	2,69636	0,007010181	151	3	42	56
g.Enterococcus	2,114603	1,0680297	1,97991	0,047713615	NA	0	101	88
g.Fusobacterium	1,0566854	0,7143608	1,479204	0,13908576	NA	0	101	87
g.Streptococcus	2,5288593	1,6838979	1,501789	0,133151621	125	6	90	5

META	w GER	w NOR	SE	BETA	Z	P-value
g.Veillonella	6,76281	4,8532282	0,2934073	0,2461961	0,8390932	0,401417
g.Phascolarctobacterium	8,638772	5,0900057	0,2698883	-0,8043332	-2,9802448	0,002880181
g.Coprococcus	5,520354	0,8475116	0,3962808	-1,4408941	-3,6360429	0,000276858
g.Lactobacillus	7,081296	4,6497429	0,2919657	1,4333202	4,9092083	0,000000914
g.Enterococcus	6,034479	0,8766643	0,3803865	1,2401688	3,2602863	0,001112998
g.Fusobacterium	4,411861	1,9595882	0,396169	0,6814887	1,720197	0,9146034
g.Streptococcus	1,34362	0,35267	0,7678032	1,4989213	1,9522206	0,05091201

Supplementary Tables S3-S12 can be found in the attached 'Supplementary_Tables_S3-S12.xlsx' file

Table S13: Summary of cross-cohort random forest classification.

Cohort [a]		Performance measures					
Training	Testing	AUC [b]	Sensitivity	Specificity	Accuracy	F1 [c]	MCC [d]
German	Norwegian	0.86	1	0.25	0.63	0.62	0.32
Norwegian	German	0.87	0.44	1	0.72	0.61	0.51

[a] N=169 German and n=101 Norwegian individuals, n=43 features; [b] area under the receiver operating curve; [c] F1 score is the harmonic average between precision (true positive rate) and recall (sensitivity), the best performance is reached at F1=1; [d] Matthews Correlation Coefficient, the MCC reflects the improvement of agreement between predicted and actual values over random prediction with respect to frequency of each class, perfect agreement is achieved at MCC=1.

Table S14: Performance diagnostics of the random forest classification of the pooled cohort over different resampling methods.

	Performance measures					
Method	AUC [d]	Sensitivity	Specificity	Accuracy	F1 [e]	MCC [f]
Boot632 [a]	0.88	0.84	0.82	0.83	0.83	0.66
CV [b]	0.82	0.75	0.71	0.73	0.73	0.47
LOOCV [c]	0.81	0.77	0.71	0.74	0.75	0.48

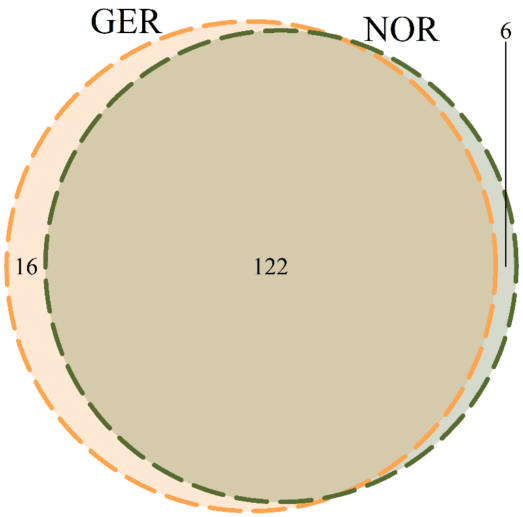
[a] Bootstrap 0.632; [b] Repeated three-fold cross-validation (10 repeats); [c] Leave one out cross-validation; [d] area under the receiver operating characteristic curve; [e] F1 score is the harmonic average between precision (true positive rate) and recall (sensitivity), the best performance is reached at F1=1; [f] Matthews Correlation Coefficient, the MCC reflects the improvement of agreement between predicted and actual values over random prediction with respect to frequency of each class, perfect agreement is achieved at MCC=1.

Table S15: Performance diagnostics of the classification of the pooled cohort using different models.

	Model performance measures [a]					
Model	AUC [e]	Sensitivity	Specificity	Accuracy	F1 [f]	MCC [g]
RRF [b]	0.87	0.82	0.82	0.82	0.82	0.64
SVM [c]	0.85	0.79	0.74	0.78	0.78	0.56
Xgboost [d]	0.83	0.78	0.76	0.77	0.77	0.54

[a] Bootstrap 0.632 was used as resampling method; [b] (guided) regularized random forest; [c] radial kernel support vector machine; [d] extreme gradient boosting; [e] area under the receiver operating curve; [f] F1 score is the harmonic average between precision (true positive rate) and recall (sensitivity), the best performance is reached at F1=1; [f] Matthews Correlation Coefficient, the MCC reflects the improvement of agreement between predicted and actual values over random prediction with respect to frequency of each class, perfect agreement is achieved at MCC=1.

SUPPLEMENTARY FIGURES



295

Figure S1: Venn diagram of co-prevalence of bacterial genera in the Norwegian (NOR) and German (GER) cohorts.

13

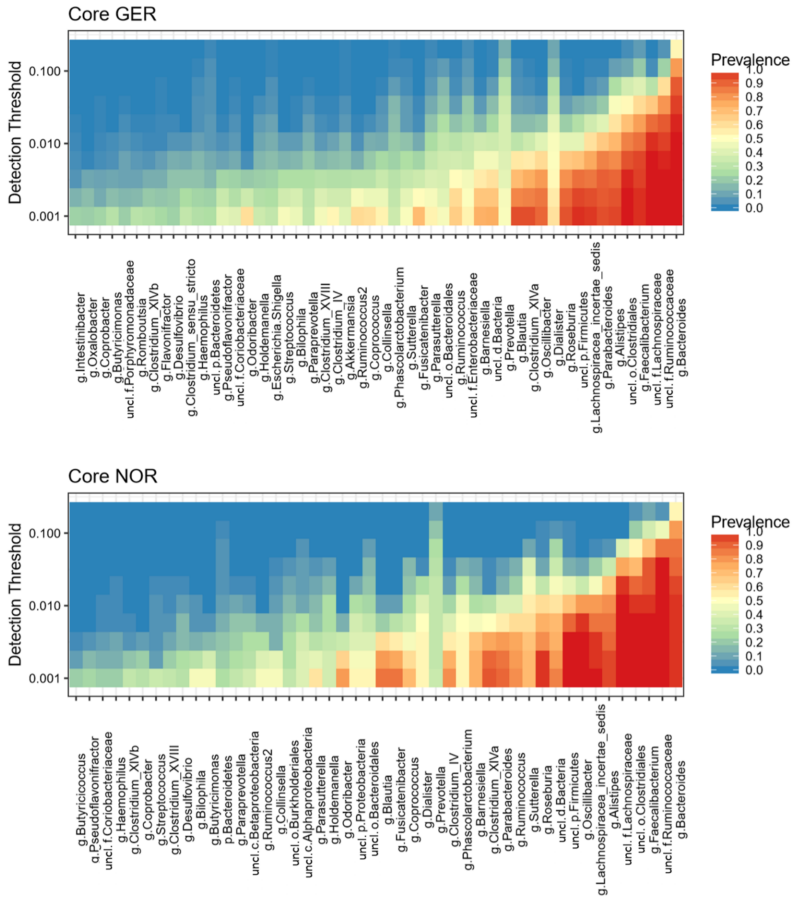


Figure S2: Core microbiota heatmaps illustrating prevalence rates (color gradient, red indicates ubiquitous prevalence [i.e. 1], blue indicates zero prevalence) over different detection thresholds (relative abundance) in the German (Core GER) and Norwegian (Core NOR) cohorts. Non-genus level taxa reflect community members that could only be classified to the respective taxonomic level.

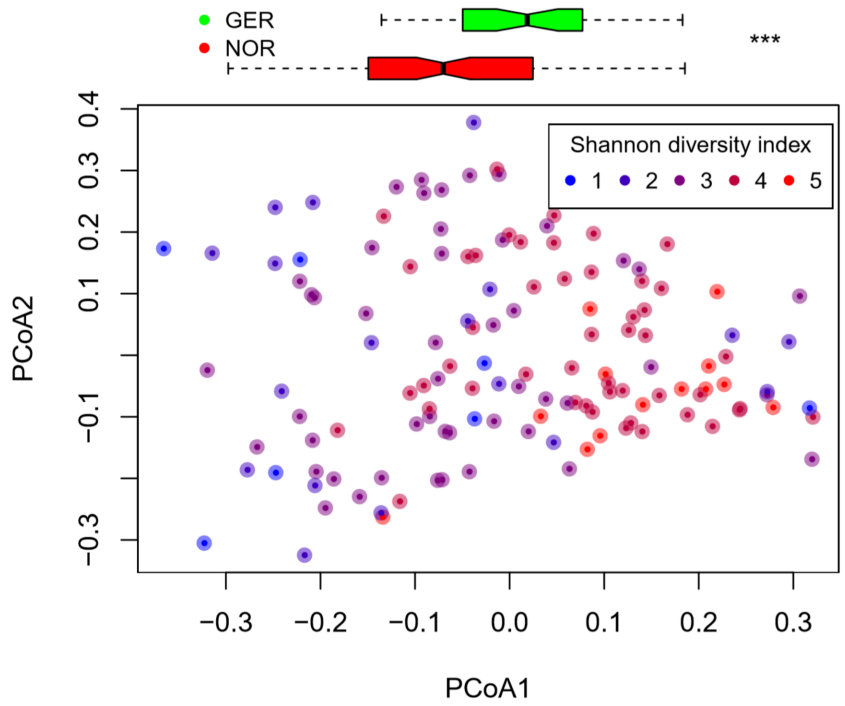


Figure S3: Principal coordinates analysis of fecal microbiomes of Norwegian and German healthy cohorts. The most pronounced difference between Norwegian and German healthy cohorts (boxplots) was seen on the first principal coordinate (PCoA1; Wilcoxon rank-sum test; ***: p<0.001), whereby the first principal coordinate was highly correlated with alpha diversity (color gradient according to Shannon diversity index from left to right).

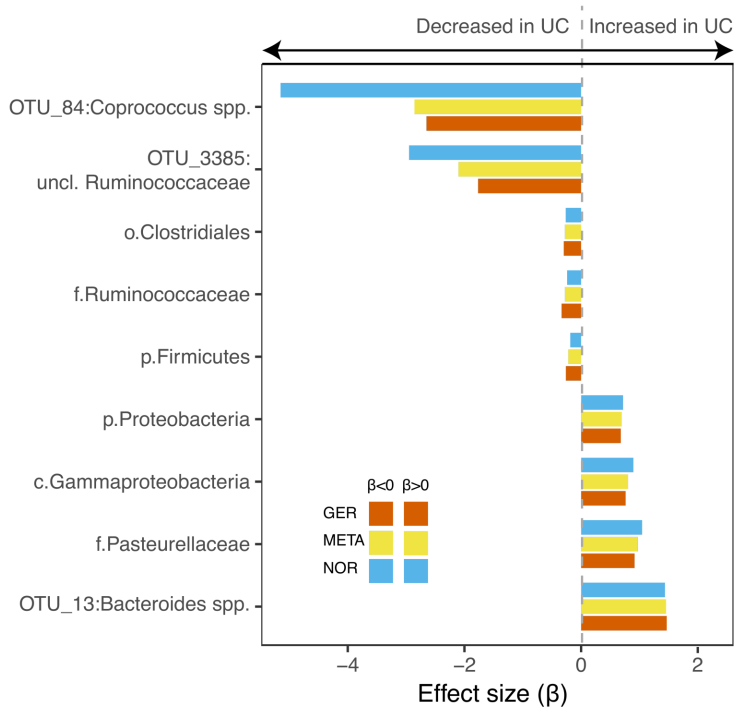


Figure S4: Significant and robust results of differentially abundant taxa in UC patients and healthy controls from the analysis using generalized linear models and hurdle models. Only taxa with P<0.05 in each cohort, Q_{META}< 0.05 and concordant directionality are shown. Base-colors depict the effect direction (Beta) of the association. Beta-values larger than zero (red) represent a higher abundance in UC patients, taxa with values less than zero are less abundant in PSC patients. Details on the model coefficients and the resulting P-values in the cohorts and the meta-analysis can be found in **Supplementary Table S7**.

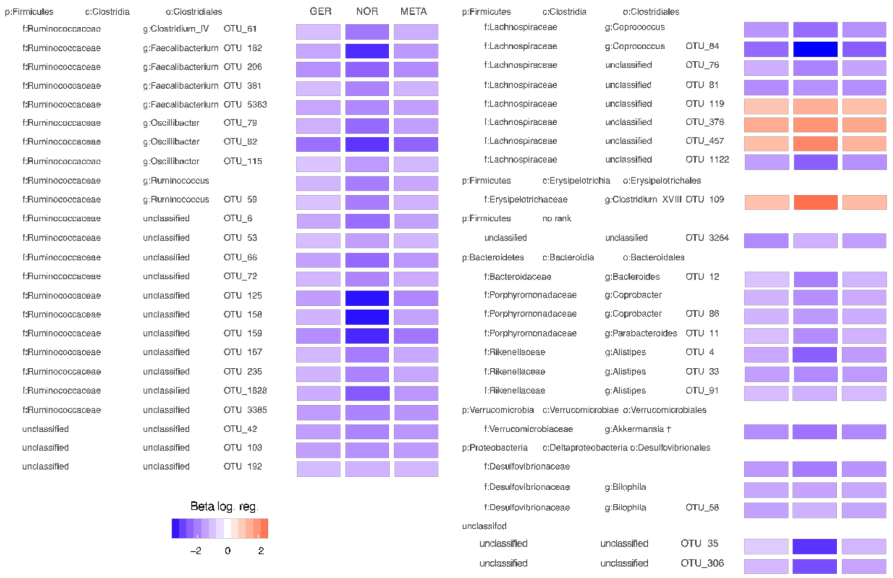


Figure S5: Significant and robust results of the logistic regression within cohorts and the inverse-variance weighted meta-analysis testing for differential prevalence of taxonomic groups depending on UC disease status. Only taxa with P<0.05 in each cohort, Q_{META}< 0.05 and concordant directionality are shown. Colors depict the effect size (Beta) of the association. Beta-values larger than zero (red) represent a higher prevalence in UC patients, taxa with values less than zero are less prevalent in UC patients. †: Akkermansia signal could be found from phylum (Verrucomicrobia) to genus level. Details on the model coefficients and the resulting P-values in the cohorts and the meta-analysis can be found in **Supplementary Table S8**.

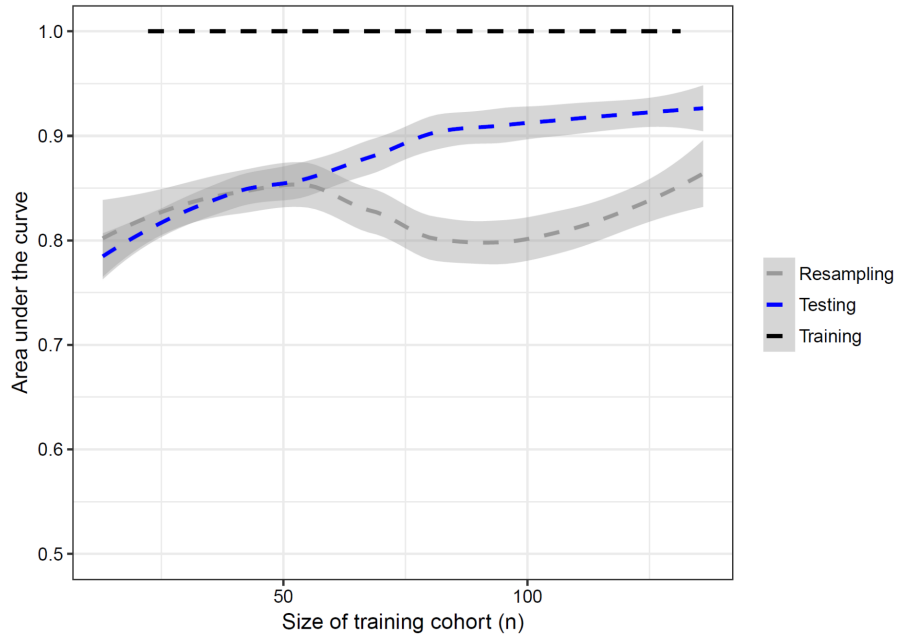


Figure S6: Learning curves of the pooled random forest classification illustrate the area under the curve of the training and test set classification performance as a function of the training set size.

18

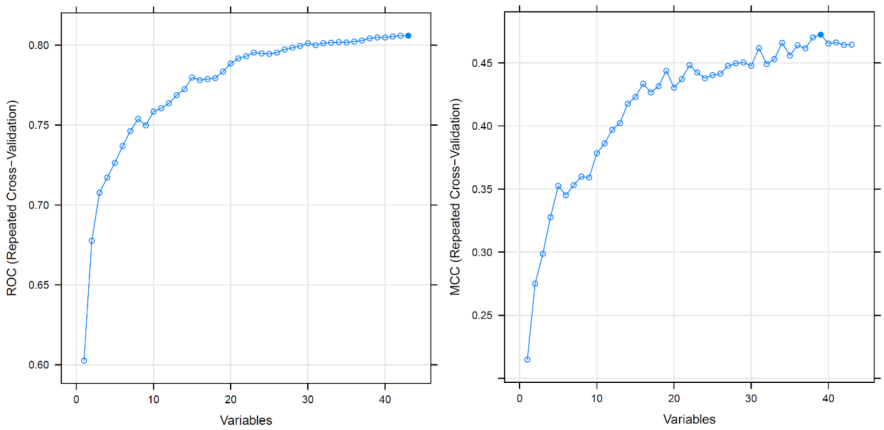


Figure S7: Random forest classifier performance measured by are under the receiver operating characteristic curve (AUC; left) and Matthews Correlation Coefficient (MCC; right) displayed as function of the number of independent variables included. The performance plateaus at approximately n=20 features.

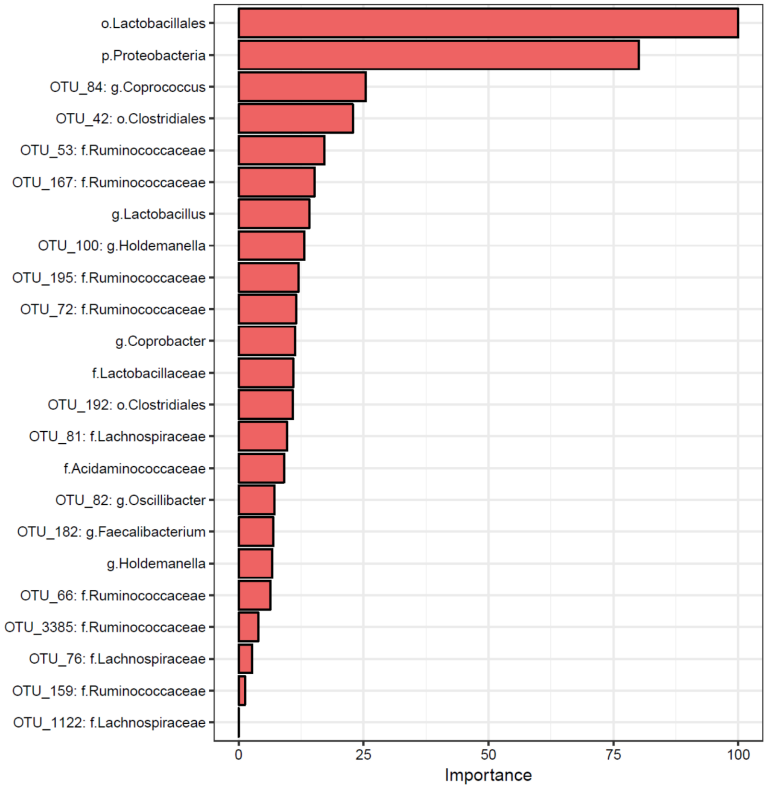


Figure S8: Important taxa identified by Boruta algorithm for pooled random forest classification

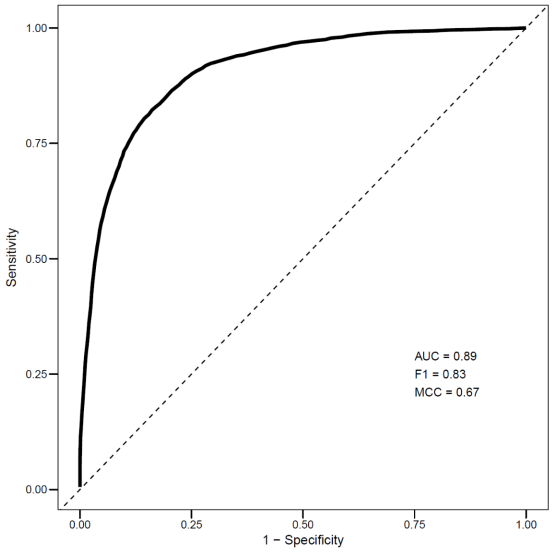


Figure S9: Area under the receiver operating characteristic curve for random forest classifier performance with reduced variable set size (n=23). The 0.632 bootstrap estimator was used to estimate the generalization error.

SUPPLEMENTARY REFERENCES

[S1] Magro F, Gionchetti P, Eliakim R, et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. Journal of Crohn's and Colitis. 2017;11(6):649-70.

[S2] Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 2016;48:1396-1406.

[S3] Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.; 2011.

[S4] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;27:2194-2200.

[S5] Edgar R. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv 2016.

[S6] Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.4–3. https://CRAN.R-project.org/package=vegan. 2017

[S7]: Legendre P and Anderson MJ. Distance-based redundancy analysis: Testing multispecies responses in mulfactorial ecological experiments. Ecological Monographs 1999;69:1-24.

[S8] Chen H. 2018; VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.6.20. Available from: https://CRAN.R-project.org/package=VennDiagram

[S9] Shetty A, Hugenholz F, Lathi L, Smidt H, de Vos WM. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. FEMS Microbiology Reviews 2017; 41, 182-199.

[S10] Kummen M, Holm K, Anmarkrud JA, et al. The gut microbial profile in patients with primary sclerosing cholangitis is distinct from patients with ulcerative colitis without biliary disease and healthy controls. Gut. 2017;66(4):611-9. Epub 2016/02/17.

[S11] Sabino J, Vieira-Silva S, Machiels K, et al. Primary sclerosing cholangitis is characterised by intestinal dysbiosis independent from IBD. Gut. 2016;65(10):1681-9. Epub 2016/05/20.

[S12] Jackman S. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.2. URL https://github.com/atahk/pscl/. 2017 [cited; Available from:

[S13] Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0 (http://www.stats.ox.ac.uk/pub/MASS4).

[S14] Robinson, M.D., and A. Oshlack. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11 (3): R25–R25.

[S15] Kuczynski, J., C.L. Lauber, et al. 2011. Experimental and analytical tools for studying the human microbiome. Nature Reviews Genetics 13 (1): 47–58.

[S16] Robinson, M.D., and G.K. Smyth. 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23 (21): 2881–2887.

[S17] Wright MN , Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 2017;77:1-17.

[S18] Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 1983;78:316-331.

[S19] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 1975;405:442-451.

[S20] Kuhn M and Johnson K. Applied Predictive Modeling. Springer, New York, 2013.

[S21] Deng H. Guided Random Forest in the RRF Package. 2013;arXiv:1306.0237.

[S22] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien. 2017; Available from: https://cran.r-project.org/web/packages/e1071/index.html

[S23] Chen T, He T. xgboost: eXtreme Gradient Boosting. 2018;R package version 0.6.4.1; Available from: https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

[S24] Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software 2010;36:1-13.

[S25] Tang R, Wei Y, Li Y, et al. Gut microbial profile is altered in primary biliary cholangitis and partially restored after UDCA therapy. Gut. 2017.

Appendix A.4: Supplement of Article G

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

METHODS

Study population and sample collection

In addition to skin examination by an experienced dermatologist, the follow-up examination included a questionnaire on medical history. At examination, disease severity of the selected patients with AD was assessed quantitatively by using the SCORAD score.

All skin biophysical measurements, such as pH and TEWL, were done in standardized environmental conditions (room temperature, 22°C to 25°C; humidity levels, 30% to 35%) by the same investigator, according to international guidelines,[E1] to minimize the influence of exogenous, environmental, and instrument-related factors.

Skin microbiota was collected by firmly swabbing a 4-cm^2 skin area for at least 30 seconds with sterile Catch-All Sample Collection Swabs (Epicentre Biotechnologies, Madison, Wis) soaked in sterile SCF-1 solution (50 mmol/L Tris buffer, 1 mmol/L EDTA, and 0.5% Tween-20 [pH 8.0]).

Stratum corneum lipids were obtained by attaching circular adhesive tape strips (3.8 cm^2; D-Squame) to each sampled skin and pressed for 5 seconds of constant pressure (225 g/cm^2) by using a D-Squame Pressure Instrument D500 (CuDerm).

Processing of bacterial 16S rRNA sequenced data

16S rRNA gene libraries were generated by using PCR from purified genomic DNA with primers 27F and 338R, targeting the hypervariable regions V1 and V2 of the 16S rRNA gene. Amplification and sequencing was performed by using a dual-indexing approach (8-nt on forward and reverse primers), as described by Kozich et al,[E2] on the Illumina MiSeq platform generating 2× 300-bp reads.

After demultiplexing, which is based on zero mismatches to the barcode sequence, reads were quality trimmed toward the 3′- and 5′-ends by using a sliding window of length of 0.1 times read length and cuts when the quality decreases to less than an average quality within the window of 20 (https://github.com/najoshi/sickle). Then forward and reverse sequences were combined into a single sequence by using overlap between the paired-end reads (minimum length, 280; maximum length, 350) with VSEARCH (https://github.com/torognes/vsearch/).[E3] Quality control was performed with the FastX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and chimera filtering was done with the UCHIME algorithm (http://drive5.com/usearch/manual/uchime_algo.html) implemented in USEARCH software.[E4] The cleaned FASTA sequences from all samples were combined, and sequences were clustered into OTUs at a similarity level of 97% by using the UPARSE algorithm implemented in the USEARCH software (http://www.drive5.com/usearch/manual/cmd_cluster_otus.html).[E5] For each sample, a subset of 10,000 random reads was picked to construct the OTU abundance table. Taxonomic classification from the phylum to genus level for the same 10,000 reads was performed with the RDP classifier and the Greengenes database.[E6] Classification of OTUs to *Staphylococcus* species (98% similarity threshold) was performed by comparing against the EzBioCloud taxonomic database (http://www.ezbiocloud.net).[E7]

Lipid analysis

At 3 body sites (Af, forehead, and Vf), levels of 348 ceramides, 12 FFAs, and cholesterol sulfate were determined by using SFC-MS/MS with the TrueMass Stratum Corneum Lipid Panel (Metabolon). To this end, all D-Squame disks were extracted together with 4 quality control samples and using a polar and nonpolar organic solvent after addition of a known amount of surrogate standard solution: C16 ceramide-d31 ([S{C18}16:0]-d31), cis-10-heptadecenoic acid (FA17:1n7), and cholesteryl sulfate-d7. The organic extracts are combined and evaporated to dryness. The dried extract is reconstituted, and an aliquot is analyzed on a Waters UPC2/Sciex QTrap 5500 mass spectrometer SFC-MS/MS system in MRM mode by using characteristic parent-fragment mass transitions for each analyte trace.

The semiquantitative determination of individual analytes is based on their peak area comparison with the peak areas of their corresponding surrogate standards for which concentrations are known. C16 ceramide-d31 is used as a surrogate standard for all ceramides, all fatty acids are referenced to cis-10-heptadecenoic acid, and cholesteryl sulfate is referenced to cholesteryl sulfate d7. Concentrations are given in picomoles per disk for individual analytes, as well as each lipid class. Additionally, the percentage composition of individual ceramide subtypes of the ceramide fraction is listed for each sample.

Data analysis and statistics

All data are presented as means ± SEMs, unless otherwise indicated. β-Diversity based on genus-level composition was computed by using the square root–transformed Bray-Curtis dissimilarity. Constrained principal coordinate analysis (nonmetric multidimensional scaling) on community ordination was performed by using the capscale function in vegan (version 2.4). Thereby either categorical data or continuous variables are linearly correlated with the coordinates of the microbial communities, and model selection is based on the Akaike information criterion provided by the standard R step function. Significance testing is carried out by means of permutation.

The Wilcoxon test for independent and dependent samples was applied to evaluate differences in α-diversities and relative abundance of individual skin bacterial traits on different taxa levels between groups at the same site or between sites within groups. Correlations of taxa of different levels (eg, phylum and genus) were calculated by using the Spearman correlation coefficient.

Additionally, multivariate and univariate association testing of the abundance of individual skin bacterial traits on different taxa levels with AD, *FLG* halploinsufficiency, and biophysical parameters was carried out by using GLMs. For prevalent taxa in human skin with almost no zeros (eg, on the phylum level), we applied GLM with a Poisson distribution and a log link with correction for overdispersion by using robust sandwich covariance estimation with the sandwich package.[E8,E9] For bacteria of low abundance, which are characterized by an increasing number of zeros at lower taxonomic levels and a right-skewed distribution, often with a long tail, we selected a GLM with a negative binomial (negbin) distribution and a log link for the statistical analysis (glm.nb function of the MASS version 7.3 package). For taxa that exhibit more zero observations than would be allowed for the negative binomial model (especially on the genus or species level), we applied the 2-component hurdle model with a negbin distribution using the pscl (version 1.4.9) package.[E10] Therefore the GLM framework was applied consistently from the phylum to genus level, whereas analysis of rare genera and species was supported with the hurdle model. Multivariate taxa analysis was carried out with the mvabund (version 3.12.3) package (function manyglm). Significance testing was carried out by using nonparametric bootstrapping in which bootstrap samples were obtained by residual resampling of the probability integral transformation residuals.[E11]

Because the use of ratios can lead to a strong reduction in overall variance, we computed and tested all possible pair microbial taxa abundance ratios of the most prevalent microbes. Microbial taxa abundance ratios were log-transformed before computing the statistics, and their back-transformation on the original scale can be interpreted as a ratio fold change. Note that testing ratios between 2 microbial taxa A and B are independent of their order, as follows:

$$\log(A/B) = -\log(B/A),$$

which halves multiple testing burden.

For estimating the influence of skin lipids on microbiome composition, we applied the integrative analysis method of sparse canonical correlation analysis using the R package PMA.[E12,E13] To this end, we standardized lipid, TEWL, and pH measurements by using

$$y_i = (x_i - \bar{x})/std(x),$$

whereas taxa counts were transformed by using the inverse hyperbolic sine transformation as follows:

$$y_i = \log\left(x_i + \left(x_i^2 + 1\right)^{0.5}\right).$$[E14]

Significance was evaluated by using a permutation test.[E15]

Gradual changes of microbial taxa from nonlesional over acute to chronic lesions are evaluated by using linear mixed models with random intercept for each subject by using the lme4 package (version 1.1.13).

All analyses have to be interpreted as exploratory, and the presented *P* values are unadjusted, if not otherwise stated. If adjustment for multiple testing was applied, the FDR for univariate analysis and the stepdown resampling procedure (mvabund package) for multivariate analysis were used, and respective *P* values are indicated.

REFERENCES

E1. du Plessis J, Stefaniak A, Eloff F, John S, Agner T, Chou TC, et al. International guidelines for the in vivo assessment of skin properties in non-clinical settings: part 2. transepidermal water loss and skin hydration. Skin Res Technol 2013; 19:265-78.

E2. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 2013;79:5112-20.

E3. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584.

E4. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011;27:2194-200.

E5. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 2013;10:996-8.

E6. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72:5069-72.

E7. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: s taxonomically united database of 16S rRNA and whole genome assemblies. Int J Syst Evol Microbiol 2017;67:1613-7.

E8. Zeileis A. Econometric computing with HC and HAC covariance matrix estimators. J Stat Software 2004;11:1-17.

E9. Zeileis A. Object-oriented computation of sandwich estimators. J Stat Software 2006;16:1-16.

E10. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Software 2008;27.

E11. Dunn PK, Smith GK. Randomized quantile residuals. Journal of Computational and Graphical Statistics 1996;5:236-44.

E12. Witten D, Tibshirani R, Gross S, Narasimhan B. PMA: penalized multivariate analysis. 2013.

E13. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009;10:515-34.

E14. Burbidge JB, Magee L, Robb AL. Alternative transformation to handle extreme values of the dependent variable. J Am Stat Assoc 1988;83:123-7.

E15. Yamada T, Sugiyama T. On the permutation test in canonical correlation analysis. Comp Stat Data Analysis 2006;50:2111-23.

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

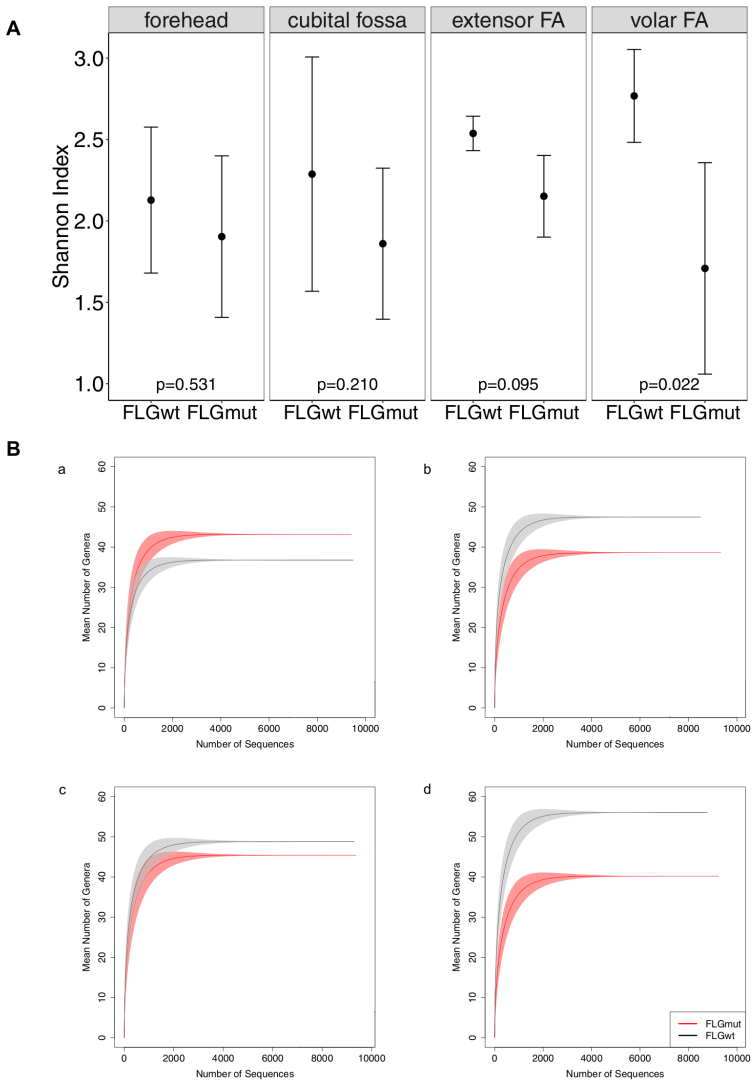


FIG E1. Reduced α-diversity and species richness in *FLG*-deficient subjects across body sites. **A,** α-Diversity (Shannon index) related to *FLG* deficiency in healthy subjects at the forehead, Af, Ef, and Vf. **B,** Species richness at the genus taxonomic level in healthy subjects divided by *FLG* haploinsufficiency. *a*, Forehead; *b*, Af; *c*, Ef; and *d*, Vf. Rarefaction curves are depicted with 2 times the sampling error.

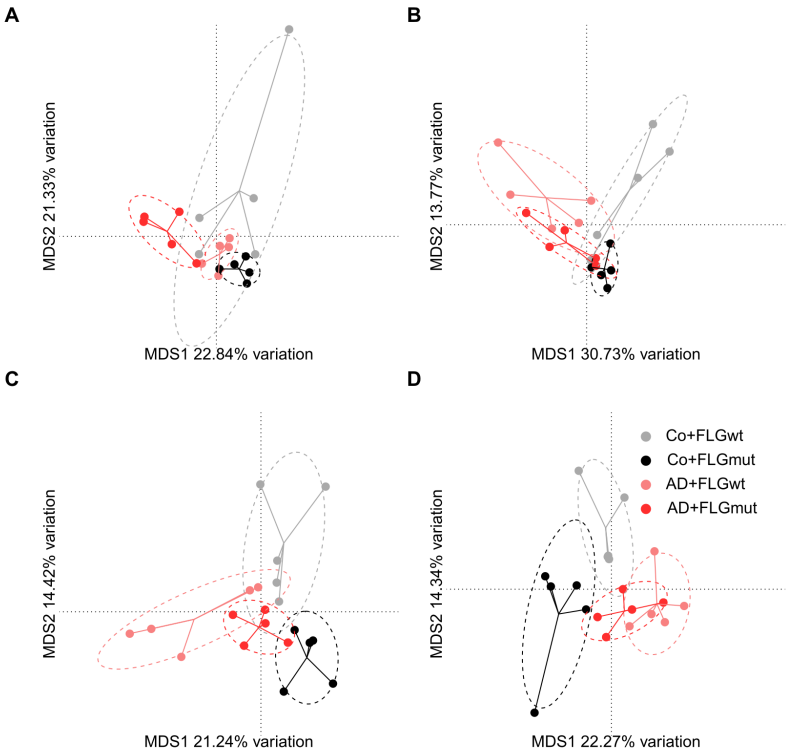


FIG E2. Bacterial community composition of *FLG*-deficient subjects is shifted toward the composition of patients with AD. Principal coordinates analysis was calculated based on the square root of the Bray-Curtis distance. **A,** Forehead; **B,** Af; **C,** Ef; and **D,** Vf.

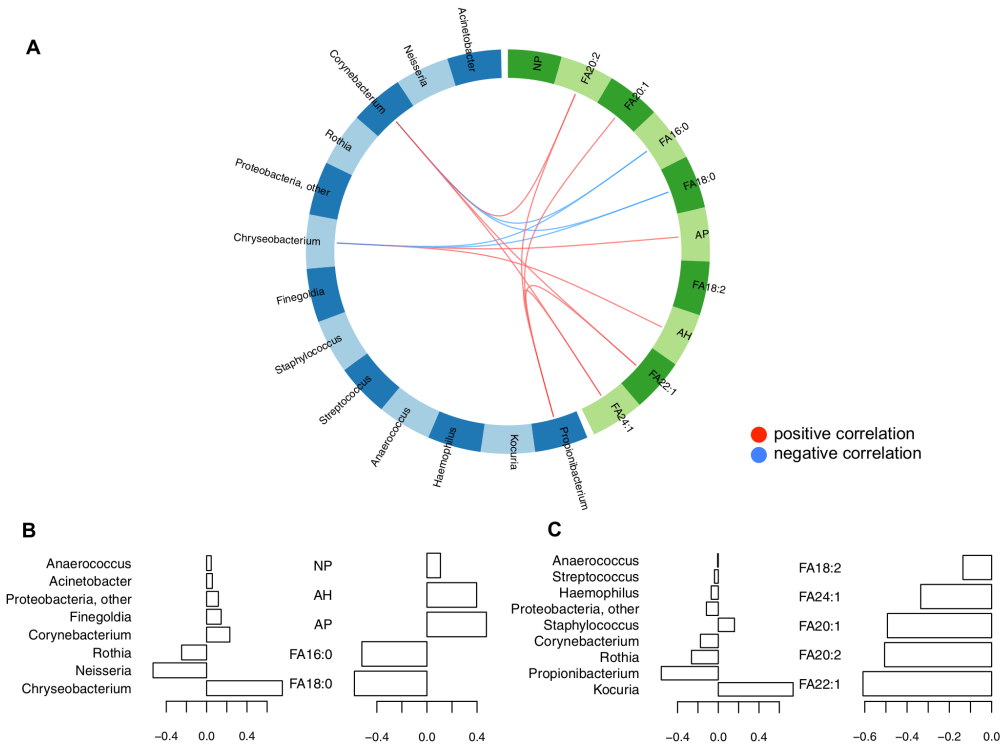


FIG E3. Effect of epidermal lipid composition on bacterial community composition. **A,** The Circos plot shows genera and epidermal lipids contributing to the overall correlation between bacterial abundance and levels of lipid species identified by using sparse canonical correlation analysis. The strongest pairwise positive (negative) correlations ($r > 0.6$) between selected genera and epidermal lipids are depicted. **B** and **C,** Coefficient weights of the selected genera and lipids on the first (Fig E3, *B*) and second (Fig E3, *C*) components of each omics level. *AH*, α-Hydroxy fatty acid/6-hydroxy-sphingosine base; *AP*, α-hydroxy fatty acid/phytosphingosine base; *NP*, nonhydroxy fatty acid/phytosphingosine base.

	Propioni-bacterium	Coryne-bacterium	Kocuria	Rothia	Staphylo-coccus	Strepto-coccus	Anaero-coccus	Finegoldia	Acineto-bacter	Haemo-philus	Neisseria	Chryseo-bacterium
Antecubital fossa												
AH	0.37	0.39	-0.14	-0.41	0.31	-0.27	0.45	0.43	0.18	-0.38	-0.43	0.33
AS	0.28	0.24	-0.20	-0.47	0.37	-0.45	0.47	0.47	-0.08	-0.50	-0.58	0.14
AP	0.36	0.28	-0.13	-0.67	0.17	-0.46	0.39	0.54	0.07	-0.56	-0.62	0.27
NP	0.46	0.32	-0.25	-0.66	0.27	-0.47	0.38	0.51	-0.09	-0.54	-0.65	0.12
FA16:0	-0.33	-0.27	-0.06	0.47	-0.13	0.45	-0.20	-0.44	-0.14	0.51	0.68	-0.53
FA18:0	-0.33	-0.30	-0.13	0.43	-0.14	0.46	-0.30	-0.54	-0.16	0.46	0.69	-0.50
FA18:2	0.65	0.46	-0.14	-0.08	0.06	0.16	0.00	0.06	0.05	0.12	-0.01	0.03
FA20:1	0.57	0.61	-0.20	-0.29	0.23	-0.33	0.42	0.46	0.00	-0.18	-0.56	0.10
FA20:2	0.52	0.41	-0.03	-0.45	0.36	-0.37	0.43	0.54	0.10	-0.44	-0.61	0.32
FA22:1	0.46	0.60	-0.31	-0.26	0.07	-0.31	0.30	0.41	-0.10	-0.09	-0.55	0.05
FA24:1	0.15	0.56	-0.49	-0.25	0.11	-0.36	0.30	0.35	-0.13	-0.11	-0.58	0.01
Forehead												
AH	-0.35	0.06	0.05	-0.30	-0.24	-0.18	-0.27	-0.12	0.46	-0.31	-0.13	0.51
AS	-0.63	-0.27	0.34	0.22	-0.41	0.02	-0.38	-0.34	0.13	-0.18	0.27	0.26
AP	-0.50	-0.24	-0.10	-0.04	-0.42	-0.19	-0.45	-0.35	0.20	-0.34	-0.13	0.48
NP	-0.43	-0.28	-0.38	0.12	-0.44	-0.24	-0.40	-0.46	-0.05	-0.31	-0.11	0.25
FA16:0	0.00	-0.10	0.11	0.39	-0.07	0.16	0.11	-0.25	-0.23	0.12	0.31	-0.49
FA18:0	0.01	-0.13	0.01	0.33	-0.15	0.07	0.02	-0.33	-0.24	0.02	0.24	-0.44
FA18:2	0.20	-0.09	-0.41	0.05	-0.19	-0.06	-0.16	-0.37	-0.30	-0.09	0.04	-0.33
FA20:1	-0.47	-0.26	-0.80	0.16	-0.71	-0.58	-0.61	-0.80	-0.40	-0.61	-0.28	0.06
FA20:2	0.17	0.11	-0.69	0.13	-0.01	0.13	-0.04	-0.09	-0.07	0.06	0.03	0.28
FA22:1	0.24	0.22	-0.62	0.36	0.10	0.38	0.13	-0.08	-0.15	0.27	0.32	0.04
FA24:1	0.16	0.22	-0.40	0.45	0.15	0.52	0.16	-0.02	-0.10	0.37	0.46	0.09
Volar forearm												
AH	0.21	0.32	-0.30	0.14	-0.13	0.40	0.43	0.16	0.46	0.40	-0.27	0.55
AS	0.19	0.24	-0.42	0.06	0.04	0.36	0.42	0.17	0.33	0.39	-0.33	0.37
AP	0.30	0.44	0.07	0.14	-0.52	0.32	0.33	-0.07	0.49	0.25	-0.27	0.66
NP	0.50	0.47	-0.23	0.04	-0.41	0.35	0.36	0.10	0.42	0.39	-0.29	0.62
FA16:0	-0.12	-0.25	-0.05	0.34	0.24	0.17	-0.12	-0.04	-0.04	0.16	0.59	-0.60
FA18:0	-0.19	-0.35	0.09	0.41	0.25	0.12	-0.26	-0.17	-0.13	0.02	0.48	-0.64
FA18:2	0.12	-0.28	0.00	0.45	-0.03	0.21	-0.37	-0.35	-0.16	0.05	0.16	-0.59
FA20:1	0.59	0.24	-0.11	0.53	-0.44	0.53	0.02	-0.12	0.19	0.39	0.05	-0.07
FA20:2	0.68	0.57	-0.03	0.32	-0.56	0.57	0.39	0.04	0.39	0.45	-0.27	0.34
FA22:1	0.48	0.23	-0.26	0.39	-0.43	0.40	0.02	0.01	0.16	0.29	-0.04	0.15
FA24:1	0.35	0.13	-0.22	0.22	-0.38	0.19	-0.09	0.02	0.06	0.08	-0.21	0.24

≥-0.8	-0.5≤r<-0.8	-0.3≤r<-0.5	-0.1≤r<-0.3	|r|<0.1	0.1≤r<0.3	0.3≤r<0.5	0.5≤r<0.8	≥0.8

FIG E4. Correlation of bacterial genera with skin lipids at different body sites. *AH,* α-Hydroxy fatty acid/6-hydroxy-sphingosine base; *AS,* α-hydroxy fatty acid/sphingosine base; *AP,* α-hydroxy fatty acid/phytosphingosine base; *NP,* nonhydroxy fatty acid/phytosphingosine base. FFAs are depicted as FAXX:Y, with *XX* indicating chain length and *Y* indicating number of double bonds/degree of unsaturation.

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

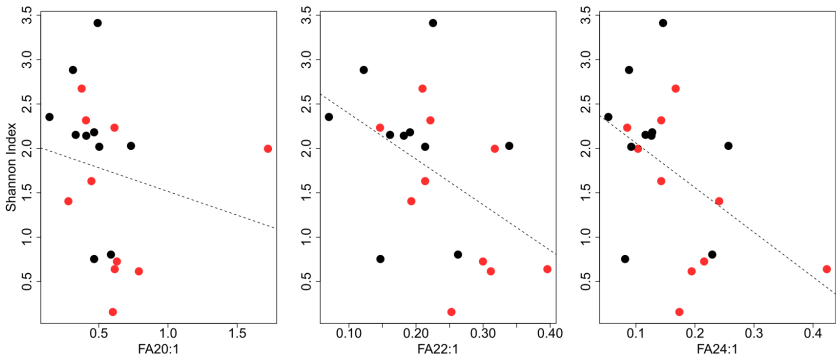


FIG E5. Reduced α-diversity is correlated with increased levels of unsaturated long-chain FFAs at the AD predilection site. Spearman correlation of α-diversity (Shannon index) with levels of 3 unsaturated long-chain FFAs at the Af are calculated. *Red dots* depict patients with AD, and *black dots* depict healthy subjects.

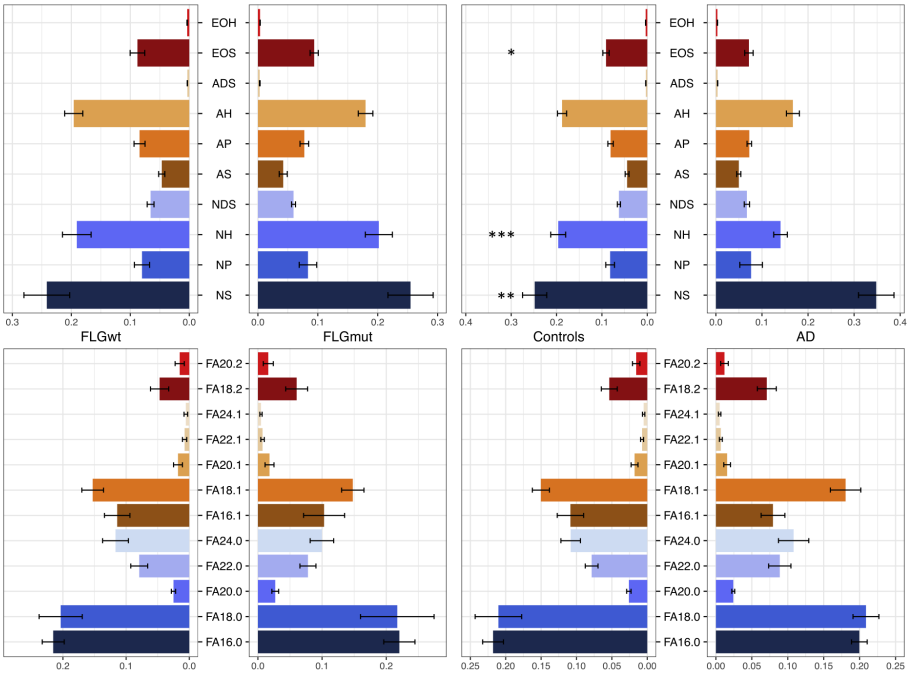


FIG E6. Association of the investigated ceramides and FFAs with FLG deficiency *(left)* and AD status *(right)*. *Asterisks* indicate significance as follow: $*P < .05$, $**P < .005$, and $***P < .0005$. For ceramides, the first letter depicts the fatty acid, and the remaining letters depict the respective sphingoid base. *AX*, α-Hydroxy fatty acid; *DS*, dihydrosphingosine base; *EX*, esterfied ω-hydroxy fatty acid; *NX*, nonhydroxy fatty acid; *XH*, 6-hydroxy-sphingosine base; *XP*, phytosphingosine base; *XS*, sphingosine base. FFAs are depicted as FAXX:Y, with XX indicating chain length and Y indicating number of double bonds/degree of unsaturation.

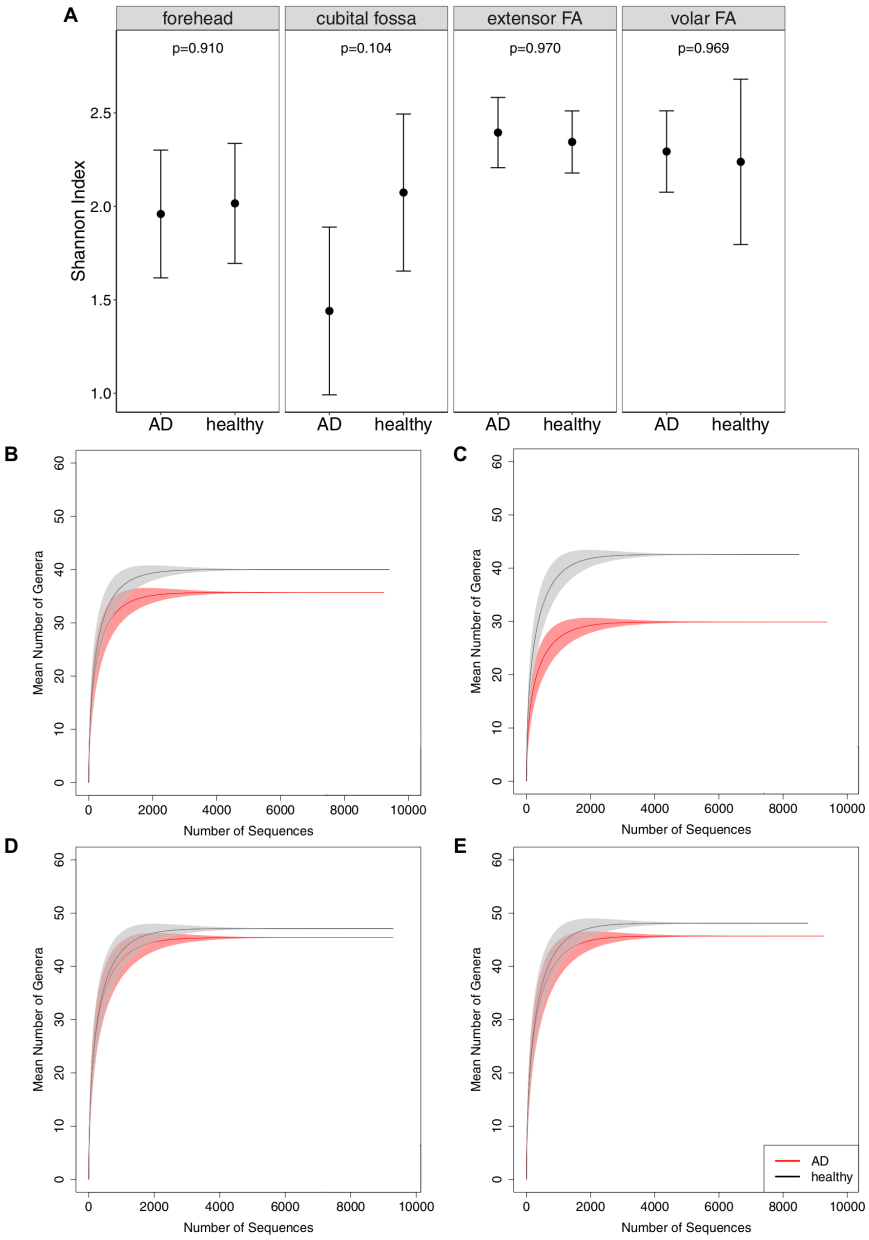


FIG E7. Reduced α-diversity and species richness in patients with AD at the AD predilection site. α-Diversity (Shannon index) and richness are calculated at the genus taxonomic level for patients with AD and healthy subjects. **A,** α-Diversity at all 4 sites. **B-E,** Species richness at the forehead (Fig E7, *B*), Af (Fig E7, *C*), Ef (Fig E7, *D*), and Vf (Fig E7, *E*). Rarefaction curves are depicted with 2 times the sampling error.

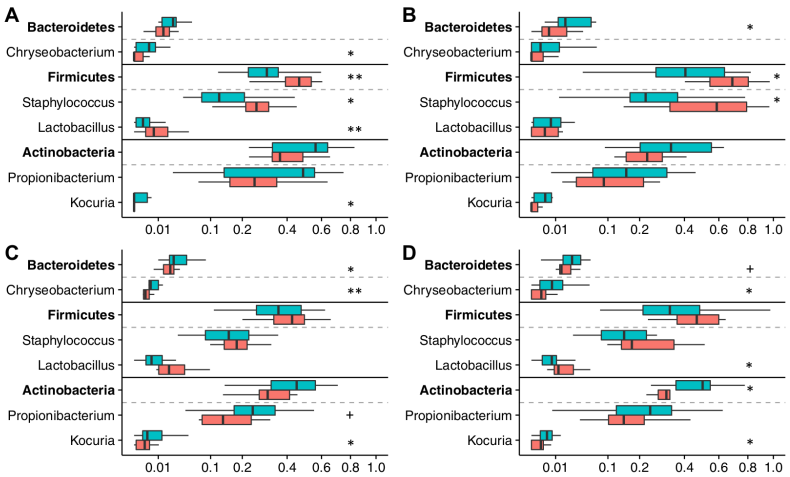


FIG E8. Effect of AD on the skin microbiome. Bacteroidetes, Firmicutes, and Actinobacteria genera showing differential abundance in patients with AD and healthy control subjects. For better visualization, abundance is depicted on the square root scale: **A,** forehead; **B,** Af; **C,** Vf; and **D,** Ef. *Red bars* depict patients with AD, and *blue bars* depict healthy control subjects. *Asterisks* indicate significance obtained from the Wilcoxon test as follows: $+P < .1$, $*P < .05$, and $**P < .005$.

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

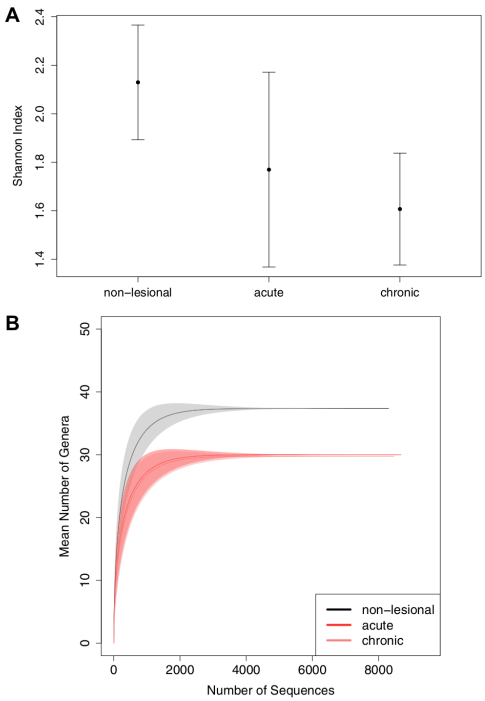


FIG E9. Reduced α-diversity and species richness in acute and chronic lesions. **A,** For patients with AD, α-diversity (Shannon index) is calculated at both nonlesional skin and acute and chronic lesions. *Non-lesional* reflects the mean diversity of the 4 body sites. **B,** Species richness at the genus taxonomic level for patients with AD at nonlesional skin, as well as at acute and chronic lesions. *Non-lesional* reflects the mean richness of the 4 body sites. Rarefaction curves are depicted with 2 times the sampling error.

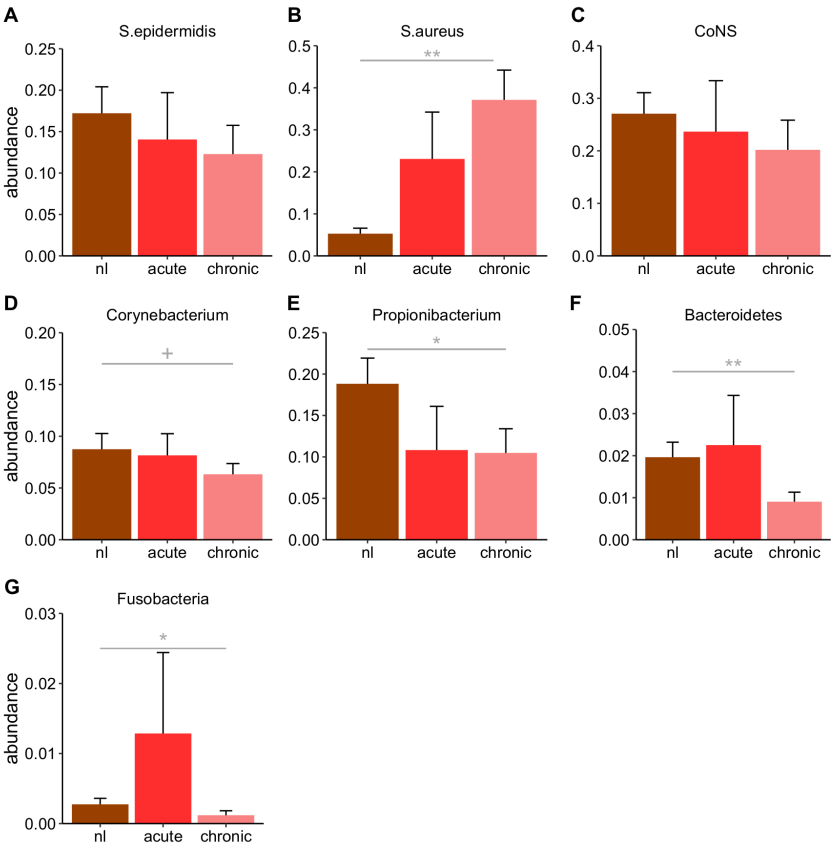


FIG E10. A-G, Differential bacterial abundance at nonlesional *(nl)* and acute and chronic lesional skin in patients with AD. Depicted is the mean ± SE abundance of the prevalent taxa with an abundance of at least 5%. *Asterisks* indicate significance obtained from the Wilcoxon test as follows: $+P < .1$, $*P < .05$, and $**P < .005$.

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

TABLE E1. Proband characteristics

ID	FLG	Sex	Age (y)	SCORAD	Acute site	Chronic site
Co09	FLGwt	Male	29			
Co08	FLGwt	Male	41			
Co01	FLGwt	Female	27			
Co04	FLGwt	Female	27			
Co07	FLGwt	Female	54			
Co04	FLGmut	Male	36			
Co08	FLGmut	Female	39			
Co09	FLGmut	Female	25			
Co01	FLGmut	Female	32			
Co11	FLGmut	Female	44			
AD04	FLGwt	Male	26	29	Af (left)	wrist (left)
AD05	FLGwt	Male	48	33		Af (left)
AD10	FLGwt	Female	20	22		neck
AD03	FLGwt	Female	25	34	Vf (right)	Vf (left)
AD01	FLGwt	Female	46	31	Lower back	Lower back
AD09	FLGmut	Male	30	42	Chest	Af (left)
AD06	FLGmut	Male	45	46	Wrist (right)	Wrist (right)
AD08	FLGmut	Female	21	37	Upper arm (left)	Af (left)
AD02	FLGmut	Female	32	34		Wrist (right)
AD05	FLGmut	Female	51	29		Wrist (left)

TABLE E2. TrueMass stratum corneum lipid panel

FFAs			
Chain length	**Saturated**	**Monounsaturated**	**Diunsaturated**
16	16:0	16:1	
18	18:0	18:1	18:2
20	20:0	20:1	20:2
22	22:0	22:1	
24	24:0	24:1	

Ceramides			
Sphingoid base	**Nonhydroxy**	**α-Hydroxy**	**Esterified ω-hydroxy**
Ceramides	NS (65)	AS (37)	EOS (21)
Dihydroceramides	NDS (54)	ADS (7)	
6-Hydroxyceramides	NH (39)	AH (31)	EOH (12)
Phytoceramides	NP (41)	AP (31)	

Twelve FFA and 10 ceramide classes were measured. For the ceramide classes, a different number of lipid species is measured and depicted in parentheses.
ADS, α-Hydroxy fatty acid/dihydrosphingosine base; *AH*, α-hydroxy fatty acid/6-hydroxy-sphingosine base; *AP*, α-hydroxy fatty acid/phytosphingosine base; *AS*, α-hydroxy fatty acid/sphingosine base; *NDS*, nonhydroxy fatty acid/ dihydrosphingosine base; *NH*, nonhydroxy fatty acid/6-hydroxy-sphingosine base; *NP*, nonhydroxy fatty acid/phytosphingosine base; *NS*, nonhydroxy fatty acid/ sphingosine base.

J ALLERGY CLIN IMMUNOL
VOLUME 141, NUMBER 5

TABLE E3. Mean bacterial abundance of Bacteroidetes, Firmicutes, and Actinobacteria genera in accounting for the abundance difference at the phylum level with regard to AD status

	Forehead				Af				Vf				Ef			
	Patients with AD	Healthy subjects	*P* value		Patients with AD	Healthy subjects	*P* value		Patients with AD	Healthy subjects	*P* value		Patients with AD	Healthy subjects	*P* value	
Bacteriodetes	2.4% ± 1.0%	2.7% ± 0.5%	.9824		1.3% ± 0.5%	4.8% ± 2.1%	.0231		2.0% ± 0.3%	4.2% ± 1.5%	.0845		2.2% ± 0.3%	3.7% ± 0.8%	.0292	
Chryseobacterium	0.1% ± 0.1%	0.6% ± 0.2%	.0210		0.3% ± 0.2%	1.2% ± 0.7%	.1381		0.3% ± 0.1%	1.3% ± 0.6%	.0106		0.3% ± 0.1%	1.1% ± 0.5%	.0054	
Firmicutes	44.9% ± 4.1%	30.8% ± 4.3%	.0066		67.9% ± 5.9%	45.0% ± 8.2%	.0281		45.9% ± 4.9%	37.7% ± 8.1%	.2753		42.1% ± 4.5%	36.3% ± 5.1%	.3348	
Staphylococcus	27.1% ± 4.7%	15.5% ± 4.1%	.0313		56.7% ± 8.9%	31.3% ± 8.4%	.0333		24.5% ± 5.0%	22.3% ± 8.6%	.7360		21.0% ± 3.9%	16.8% ± 3.3%	.3195	
S aureus	2.8% ± 1.3%	0.3% ± 0.1%	5.4E-04		8.9% ± 3.9%	0.2% ± 0.1%	5.1E-17		6.5% ± 2.1%	0.4% ± 0.3%	3.3E-04		2.9% ± 0.7%	0.5% ± 0.4%	.0165	
S epidermidis	17.8% ± 4.3%	10.2% ± 2.9%	.0782		32.4% ± 8.4%	13.8% ± 2.9%	.0049		10.6% ± 2.4%	14.6% ± 7.8%	.6165		8.0% ± 1.7%	8.0% ± 1.2%	.8875	
S hominis	0.6% ± 0.2%	0.6% ± 0.2%	.9930		7.9% ± 4.4%	10.9% ± 5.7%	.7063		2.0% ± 0.7%	3.7% ± 0.9%	.2137		1.4% ± 0.3%	2.8% ± 0.6%	.0425	
Lactobacillus	3.0% ± 1.9%	0.3% ± 0.2%	.0027		2.5% ± 2.0%	1.0% ± 0.4%	.1273		3.8% ± 2.0%	1.1% ± 0.4%	.0371		3.2% ± 0.9%	1.4% ± 0.7%	.1077	
Actinobacteria	40.2% ± 4.3%	50.2% ± 6.3%	.4172		22.5% ± 3.6%	36.7% ± 6.2%	.2673		34.9% ± 4.6%	44.2% ± 6.5%	.3451		34.5% ± 5.1%	44.7% ± 5.7%	.2133	
Propionibacterium	28.5% ± 5.8%	38.9% ± 7.9%	.3440		12.4% ± 3.3%	19.0% ± 5.0%	.6501		18.0% ± 3.8%	25.1% ± 5.9%	.3665		16.4% ± 3.0%	26.1% ± 5.0%	.0852	
Kocuria	0.0% ± 0.0%	0.6% ± 0.4%	.0115		0.1% ± 0.1%	1.0% ± 0.5%	.1304		0.2% ± 0.1%	0.9% ± 0.4%	.0122		0.3% ± 0.1%	1.1% ± 0.5%	.0213	

TABLE E4. Ratio of *Staphylococcus* species

	Patients with AD		Healthy subjects				
	Median	Q1-Q3	Median	Q1-Q3	RFC	95% CI	*P* value
S aureus/*S epidermidis*							
Forehead	0.056	0.029-0.151	0.034	0.003-0.076	5.9	0.8-42.2	.0937
Vf	0.613	0.241-0.960	0.003	0.001-0.016	85.6	22.9-320.2	6.9×10^{-6}
Ef	0.298	0.148-0.979	0.006	0.002-0.040	38.3	9.4-155.3	2.0×10^{-4}
Af	0.080	0.035-0.499	0.003	0.001-0.023	31.2	5.7-172.0	.0010
S aureus/CoNS							
Forehead	0.042	0.020-0.981	0.027	0.001-0.050	5.4	0.8-36.2	.1047
Vf	0.322	0.112-0.652	0.002	0.001-0.009	88.4	25.1-311.4	7.0×10^{-6}
Ef	0.142	0.092-0.173	0.003	0.001-0.021	30.8	8.0-117.8	4.0×10^{-4}
Af	0.077	0.029-0.142	0.002	0.001-0.008	38.3	6.5-225.7	8.0×10^{-4}

Ratios are obtained by calculating the proportion of species count A divided by species log(A/B). To avoid dividing by zero, we added 1 to each count. For hypothesis testing, ratios were log-transformed to obtain normality, and the *t* test was applied.
Q1, Lower quartile; *Q3*, upper quartile; *RFC*, ratio fold change.