TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Ernährung und Immunologie

Arrhythmic Gut Microbiome Signatures for Type 2 Diabetes Risk Profiling

The analysis of the human gut microbiome in large population-based cohorts.

Sandra Reitmeier

Vollständiger Abdruck der von der Fakultät TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Siegfried Scherer
Prüfer der Dissertation:	1. Prof. Dr. Dirk Haller
	2. Prof. Dr. Lindsay Hall
	3. Prof. Martha Merrow, Ph.D.

Die Dissertation wurde am 26.10.2020 bei der Technischen Universität München eingereicht und durch die Fakultät TUM School of Life Sciences der Technischen Universität München am 03.05.2021 angenommen.

Zusammenfassung:

Über die letzten Jahre wurde gezeigt, dass auch der Darm im Zusammenhang mit metabolischen Erkrankungen, insbesondre mit Typ 2 Diabetes (T2D), steht. Ziel dieser Arbeit war es die bakterielle Zusammensetzung des Darms in prospektiven Populationskohorten zu beschreiben und Einflussfaktoren zu bestimmen. Besonderer Fokus lag dabei auf dem Zusammenhang des Darms mit T2D, welcher in einer prospektiven Populationskohorte mit 1,976 Teilnehmern (KORA) untersucht wurde. Es konnte gezeigt werden, dass die relative Abundanz der Bakterien einer tageszeitlichen Schwankung unterliegen und die Zusammensetzung des Darms einer 24-Stunden Oszillation aufzeigt. Dabei konnte in 13 Bakterien bei T2D ein gestörter Rhythmus festgestellt werden. Diese arrhythmische bakterielle Risikosignatur wurde in kreuzvalidierte Vorhersagemodelle zur Klassifizierung und Prädiktion von T2D verwendet und zeigte eine 'Area under the Curve' (AUC) von 0.73. Die Integration von BMI als unabhängigen Risikomarker trug zusätzlicher zu einer besseren Klassifizierung bei (AUC = 0.79). In einer unabhängigen Kohorte von 1,363 Individuen (FoCus) konnte zum einen die mikrobielle Oszillation des Darms bestätigt werden und zum anderen die Übertragbarkeit der arrhythmischen Risikosignatur zur Klassifizierung von T2D (AUC = 0.76) gezeigt werden. Die Vorhersagekraft der 13 selektierten Bakterien von T2D wurde in dem prospektiven Teil der KORA Kohorte in 699 Individuen, welche eine fünf Jahres Inzidenzrate von T2D aufweisen, gezeigt (AUC = 0.78). Die Sequenzierung des Metagenoms einer Subkohorte (N = 50 Individuen, n = 100 Proben) identifizierte 26 metabolisch Stoffwechselwege, welche im Zusammenhang mit tageszeitlich oszillierenden Bakterien gebracht werden, konnte. Zusammenfassen wurde in dieser Arbeit das Zusammensetzung des Darms in großen Populationskohorten und der Einfluss von zirkadianen Rhythmen Vorkommen beschrieben. Es wurde eine Risikosignatur aus arrhythmischen Bakterien zur Klassifizierung und Prädiktion von T2D ausgearbeitet und ein Zusammenhang mit metabolischen Stoffwechselwegen erstellt.

Die Ergebnisse zeigten, dass zusätzliche Faktoren, so wie dir Zeitpunkt der Probenentnahme einen substanziellen Einfluss auf die Ergebnisse einer Studie haben. Um den Einfluss von weiteren Störfaktoren zu minimieren, wurden alle Schritte, von der Probenverarbeitung, über die bioinformatische und statistische Auswertung, analysiert validiert und optimiert.

Summary

Over the last few years, recent research has shown that the gut microbiota composition is associated with metabolic diseases, especially Type 2 Diabetes (T2D). The aim of this work was to describe the gut microbial composition in prospective population-based cohorts and to determine influencing factors. Special focus was put on the association of the gut microbiota with T2D, which was investigated in a prospective population cohort of 1,976 participants (KORA). We identified changes in the gut microbiota that are associated with a 24-hour diurnal oscillation. A disrupted rhythm of 13 bacteria was found in patients with T2D. This arrhythmic bacterial risk signature was used in cross-validated predictive models for classification and prediction of T2D. The model was able to differentiate T2D from nonT2D with an Area under the Curve (AUC) of 0.73. The integration of BMI as an independent risk marker further contributed to a better classification of the disease (AUC = 0.79). In an independent cohort of 1,363 individuals (FoCus), diurnal oscillation of the gut bacteria could be confirmed and the generalizability of the arrhythmic risk signature for the classification of T2D was shown (AUC = 0.76). The predictive power of the 13 selected bacteria of T2D was shown in the prospective part of the KORA cohort in 699 individuals with a five-year incidence rate of T2D (AUC = 0.78). Additionally, this arrhythmic risk signature was able to predict T2D in another cohort from the UK (TwinsUK) including 1,399 individuals. Metagenomic analysis of a subset of the KORA cohort (N = 50 individuals, n = 100 samples) identified 26 metabolically associated pathways could be associated with time-of-day oscillation. In summary, this work described the composition of the gut in large population cohorts and the influence of circadian rhythms. The results showed that additional factors, such as time of sample collection, have a substantial influence on shifts in gut microbiota composition. In order to minimize the influence of additional confounding factors, all steps, from sample processing to bioinformatic and statistical analysis, were validated and optimized. A risk signature from arrhythmic bacteria for the classification and prediction of type 2 diabetes was elaborated and an association with metabolic pathways was established.

Table of Content

1. Introduction

The systematic examination of the gut microbiota - the entirety of the microorganisms living in the intestine – established a close connection between changes in the intestinal microbiota and the development of various diseases. Recent findings showed that the characterization of the gut microbiota is important for the prevention, diagnosis and treatment of multiple diseases, including diabetes, cancer, inflammatory bowel diseases (IBD) and psychological disorders (Qin et al. 2012, Hsiao et al. 2013, Karlsson et al. 2013, Karlsson et al. 2013, Zhernakova et al. 2016, Pascal et al. 2017, Waldschmitt et al. 2018, Valles-Colomer et al. 2019). This work provides new insights to the role of the gut microbiota in the development of Type 2 diabetes (T2D). Assessment of DNA sequencing procedures is of great importance to obtain reliable results. This work investigate the different technical factors that could influence microbiome analysis and highlights the importance for standardization.

1.1. The human gastrointestinal tract harbors a complex and dynamic population of microorganisms

The human intestinal tract consists of two main segments. The small intestine is responsible for nutrients digestion and reabsorption. The large intestine consist of cecum, rectum and anal canal – responsible for the absorption of water. The remaining defecated material is referred to as feces. It contains bacteria which are reflecting the microbial community (microbiota) of the entire intestinal tract and is used as starting material in human population-based studies investigating the microbial composition of the gut. The analysis of the gut microbiome is the description and classification of the gut bacteria and the analysis of their function (microbiome). The entire genetic material within a microbial ecosystem is referred to as the metagenome and provides information about the bacterial species and their metabolic activity. Historically, the characterization of the human gut microbiota relied on culture-dependent methods. Especially after the ability to grow bacteria anaerobically researchers started to become interested in understanding the composition and function of the microbial communities as early as in the 1940s and 1950s (**Figure 1**). However, the inability to cultivate a large proportion of microorganisms hindered the ability to capture the complete biodiversity of human-associated microbial communities. In 1965 Schaedler et al. successfully transferred a bacterial community to germ-free mice (Wymore Brand et al. 2015). Seven years later Peppercorn and Goldman (1972) found that intestinal bacteria are influenced by the human metabolism. In 1996 Wilson and Blitchington (1996) established a culture-based identification method using sequencing of the 16S rRNA genes from each bacterium grown of a human fecal sample. In 2005 a study showed that a

change in diet alters the activity of the colonic microbiota (Toner 2005). In the same year Eckburg et al. (2005) discovered the presence of a large inter-individual variation of human gut microbiota and laid the foundations to explore its association with health and disease. Since 2006 numerous studies have highlighted the role of diet on the gut microbiota and host metabolism (Wu et al. 2011, Clarke et al. 2012, David et al. 2014, Zhernakova, Kurilshikov et al. 2016). Since the early 2000s, technologies for high-throughput sequencing and bioinformatical analysis tools improved enabling the analysis of large human populations and metagenomes (Caporaso et al. 2010, Callahan et al. 2016, Lagkouvardos et al. 2016).

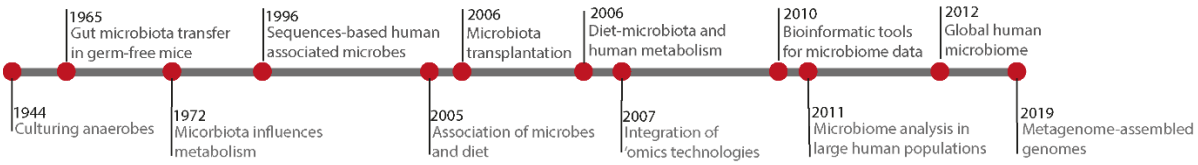


Figure 1 Milestones in the human microbiome research.

1.2. Host physiology influences the bacterial composition of the human gut

With the advent of high-throughput sequencing and the ability to analyze thousands of individuals, it quickly became clear that the composition of the human gut microbiota shows large inter-individual variation.

One of the very first large human population-based studies focusing on the bacterial composition of the gut was the Human Microbiome Project (HMP) starting in 2008. It aimed to provide a reference database of microbes and their association with health. This and further studies from all over the world provided solid evidence for a large bacterial variation found among individuals (Qin et al. 2010, Human Microbiome Project 2012, Falony et al. 2016). Partially, these differences are explained by environmental factors such as diet, geography or age. However, causes for most parts of these differences are still unclear and it was yet not possible to define a healthy reference gut microbiota for humans. The MetaHit consortium – a collaborative project exploring the genetic influence of human microbes to understand their impact on our health and wellbeing – introduced the concept of ´enterotypes´ (Arumugam et al. 2011). It was shown that humans can be divided into three clusters or ´enterotypes´ based on the dominance of certain bacterial taxa. These three clusters are characterized by an increased abundance of *Prevotella*, *Bacteroidetes* or *Ruminococcus.* Later studies tried to reproduce the concept of ´enterotypes´, however the results remain inconclusive. Some were not able to detect any (Huse et al. 2012, Koren et al. 2013), while others determined a different number of distinct clusters (Wu, Chen et al. 2011, Karlsson et al. 2014, Zhou et al. 2014, Falony, Joossens et al.

2016). A meta-analysis of human population-based cohort studies from all over the world indicated that the number of determined clusters varies according to geographical region (Costea et al. 2018). Nevertheless, most studies report that the three genera, *Prevotella*, *Bacteroides* and *Rumincoccus*, are the representative taxa of each of the cluster showing a large variation between groups of individuals. These findings can be seen as the core concept of the ´enterotypes´ (Costea, Hildebrand et al. 2018). The influence of dietary habits showed that the *Bacteroidetes* ´enterotype ´ is associated with a high-fat diets (Hildebrandt et al. 2009, Turnbaugh et al. 2009) as well as with insulin resistance, increased CRP levels and a low-grade inflammation (Bloom et al. 2011, Claesson et al. 2012, Le Chatelier et al. 2013). Nevertheless, up to now there is no causal explanation on why and how these cluster are formed.

Factors underlying the large inter-individual variation in the microbial composition are largely unknown. However, drug-related effects have been suggested as one important factor. Especially the exposure of antibiotics which is associated with a loss in microbial diversity and a depletion of specific microbial species as well as in an increase in resistance genes (De La Cochetiere et al. 2008, Dethlefsen et al. 2008, Raymond et al. 2016). Thus, due to the strong effect of antibiotics on the microbial composition this medication normally is an exclusion criterion in population-based studies. In addition, other medications e.g., proton pump inhibitors (PPI) influence the composition of the gut microbiota e.g., by an increased abundance of the phyla Firmicutes (Maier et al. 2018). It was also shown that metformin has an effect on the gut microbiome by increasing the abundance of health-related species (e.g., *Akkermansia*, *Bifidobacterium*) and resulted in an improved glucose tolerance (Wu et al. 2017).

1.3. The gut microbiota in health and disease

The bacterial composition of the gut may also affect the well-being of an individual. Since the establishment of next generation sequencing technologies in 2005, the number of publications focusing on the bacterial composition of the gut microbiota increased from 4,181 (year 2010) to 65,593 publications (year 2019) and is still rising. While many studies are analyzing the gut microbiota in a normal healthy population, some are trying to find associations of the gut microbiome with health. The majority of studies focused on the relationship between gut-related diseases, such as inflammatory bowel disease (IBD) and their association with the intestinal bacterial ecosystem. Others tried to find a link between metabolic related disorders (obesity, cardiovascular disease and T2D) and the gut microbiota.

Independent of the addressed disease, main focus of these studies was to determine bacteria which are linked to the disease and tried to define a bacterial signature related to the progression or onset of the disease. Thus, a bacterial risk signature seems to exist enabling classification of IBD or

persons at risk (Ananthakrishnan et al. 2017, Pascal, Pozuelo et al. 2017). Nevertheless, the results showed that the identification of bacterial risk signatures is only merely overlapping between studies.

1.4. Type 2 diabetes a global pandemic and one of the major challenges to human health

Globally the number of people suffering from a metabolic disorder is increasing rapidly. According to the World Health Organization (WHO), the number of T2D cases increased from 108 million in year 1980 (about 2.4% of the world's population affected) to 463 million in year 2019 (about 9% affected), which means that 1 out of 11 individuals is going to develop T2D.

The pathophysiology of T2D is typically characterized by impaired beta-cell function (ability to produce insulin) and impaired responsiveness to normal insulin levels resulting in a dysregulated glucose metabolism (Skyler et al. 2017). Most clinical treatments of T2D either target insulin resistance or aim to elevate insulin levels by increasing beta cell function. Main determinants for the development of T2D are genetics, lifestyle, exposure to medications and diet (Wu et al. 2014). A high-caloric, high-fat diet and low physical activity are key factors for the development of T2D, mostly in combination with obesity. Currently the risk assessment for the development of T2D is based on biochemical measures. The progression of T2D can be influenced by genetic markers e.g., TCF7L2 (affects insulin secretion and glucose production Huang et al. 2018), ABCC8 (regulate insulin, Andrikopoulos et al. 2016), GCGR (a glucagon hormone involved in glucose regulation, Lee et al. 2014).

Nevertheless, about 50% of people with diabetes remain undiagnosed and may have already developed complications (Gillies et al. 2008). Therefore, it is important to better understand the mechanism contributing to T2D development and to install prevention measures including early diagnoses and treatment. Based on an early diagnosis of individuals under risk dietary habits, lifestyle and behavior could be changed and thus, could delay the onset of T2D.

1.5. The association of the human gut microbiota and metabolic health

With increasing evidence the human gut microbiota seems to be linked to metabolic health (Sonnenburg and Backhed 2016) and altered microbial profiles are associated with obesity, insulin resistance and T2D (Turnbaugh et al. 2006, Qin, Li et al. 2010, Qin, Li et al. 2012, Karlsson, Tremaroli et al. 2013, Goodrich et al. 2014, Pedersen et al. 2016, Thingholm et al. 2019, Zhou et al. 2019). The identification of disease-related microbial risk profiles is complicated by the inter-individual variability (Falony, Joossens et al. 2016, Zhernakova, Kurilshikov et al. 2016), regional effects (He et al. 2018) and drug-associated changes in the gut microbiota (Forslund et al. 2015, Pryor et al. 2019). Several studies

tried to find bacterial signatures contributing to the development of T2D (Pedersen, Gudmundsdottir et al. 2016, He et al. 2018, Li et al. 2020). Despite the extensive efforts to define the role of the gut microbiota in metabolic diseases, especially obesity and T2D, there is still no consensus on disease-related bacterial taxa with diagnostic relevance. There are some overlaps between the studies, for example the genera *Akkermansia*, *Eubacterium rectale*, *Alistepes* and species *Feacalibacterium prausnitzii* are positive associated with health, i.e. absence of T2D (Turnbaugh, Ley et al. 2006, Qin, Li et al. 2012, Karlsson, Tremaroli et al. 2013, Thingholm, Ruhlemann et al. 2019), while the association with species increased in individuals with metabolic disorders is lacking. Another limitation of disease associated bacterial signatures is the lack of specificity. For instance, species of the family *Christensenellaceae*, but also *Escherichia coli* are associated with Crohn's disease and T2D (Pascal, Pozuelo et al. 2017), which complicates the identification of distinctive microbial risk factors for metabolic disorders.

1.5.1. Functional shifts of the gut microbiome in metabolic disease

To better understand the relationship between bacterial composition of the gut and health it is important to examine the relationship between microbial composition and their biochemical contribution to the gut metabolic environment and interactions with host functions. Functional activity includes production of healthy metabolites, the degradation of fibers and related functions, like host immune system training. For instance, bacterial taxa, such as *Roseburia*, *Alistipes* or *Akkermansia*, were shown to be associated with the functionality of certain pathways. Especially butyrate producing bacteria seem to have a protective effect by reducing the risk to develop inflammatory and metabolic diseases (Canfora et al. 2015). Pathways associated with the metabolism of carbohydrate and amino-acids are core functional pathways expressed in all individuals. They are responsible for the biosynthesis or degradation of fructose and mannose and amino-sugars. Some pathways were found to be responsible for the degradation of N-glycan (Turnbaugh, Ley et al. 2006). Some of these pathways are enriched in obese and diabetic individuals. The under- or overrepresentation of metabolic factors (e.g., insulin level) impacts host physiology. This may result in an enrichment of xenobiotic, amino acid and glycan pathways as well as with pathway–mediated biosynthesis of vitamins and isoprenoids (Backhed et al. 2004, Gill et al. 2006). Pathways of bacterial chemotaxis and ABC transporters were reported to be associated with high-fat diet and found to be increased in individuals with increased insulin resistance (Turnbaugh, Ley et al. 2006, Karlsson, Tremaroli et al. 2013). A study from Zeevi et al. (2015) further showed that the uptake of negatively charged amino acids, such as glutamate, is associated with T2D. The functional changes are reflected by alterations in the profile of microbiota-derived metabolic byproducts.

1.5.2. Disrupted circadian rhythms influence the human metabolic health

Circadian rhythms are found in nearly every organ and tissue with similarities among mammalians. Following a day-night cycle the human body responses in mental, physiological and behavioral changes. It can also affect body function and health by controlling hormones, body temperature and other important functions. The circadian clock synchronizes daily food intake behavior and metabolism with the day and night cycle (Panda 2019) and has recently been proposed to influence microbial homeostasis (Thaiss et al. 2014). Daytime-dependent fluctuations were identified in both the oral and fecal microbiota (Thaiss, Zeevi et al. 2014, Kaczmarek et al. 2017). Circadian rhythms in gut microbiota composition and function are sensitive to diet and feeding patterns in murine models. A diet-induced obesity dampens cyclic microbial fluctuations in rodents (Zarrinpar et al. 2014) and epidemiological studies continue to show associations between circadian dysfunction due to modern lifestyle and T2D (Onaolapo and Onaolapo 2018), supporting the hypothesis that diurnal oscillations in microbiota composition and function may contribute to metabolic health. The lack of documentation of stool sampling time and in addition to the well-documented regional and individual differences in microbiota profiles may account for discrepancies between studies.

1.6. The composition of the gut microbiota in large population-based cohorts

With the development of high-throughput sequencing technologies it became possible to analyze the human gut microbiota of thousands of individuals in parallel and to find associations with the presence/absence and abundance of bacteria with host health.

1.6.1. 16S rRNA gene and metagenomic sequencing – Taxonomic and functional annotation

Next generation sequencing (NGS) allows generating thousands of sequence reads in parallel (Langille et al. 2013), while Sanger sequencing produces only one sequence at a time (Sanger et al. 1977). NGS became available in the 2000s and advanced to the most popular method for rapid analysis of diverse complex microbial communities (Hamady and Knight 2009).

Therefore, the 16S rRNA gene is used to determine the taxonomy of bacterial members in a community. This ribosomal gene has been established by Woese et al. in 1990 and was shown to be a valuable tool for bacterial species determination (Woese et al. 1990). The 16S rRNA gene codes for the 16S ribosomal RNA, which is a component of the 30S small ribosome subunit. The gene is divided into nine hypervariable regions (V1 – V9) ranging from 30 bp to 100 basepairs (bp), flanked by conserved regions and in total reaching a length of about 1,500 bp. The conserved regions of the 16S rRNA gene are very similar between almost all bacteria. In contrast, the hypervariable regions are often species-

specific and can be used for the identification of different bacterial species. The targeted variable regions are selected by using region-specific primers binding to the conserved regions. NGS amplicon sequencing using Illumina MiSeq machines enables to sequence up to two-times 300 bp (paired-end sequencing) thus, allowing a theoretical amplicon size of 600 bp at maximum. Since the 16S rRNA gene is about 1,500 bp, only part of it can be analyzed by this technique. This may cause information loss and complicates the differentiation between closely related species. Latest research is focusing on the development of full-length sequencing technologies to cover the whole 16S rRNA gene. Although methods for this are still under development, first results showed promising outcomes (Fuks et al. 2018). Overall, if taxonomic classification is the main objective short amplicon sequencing of the 16S rRNA gene sequencing provides reasonable taxonomic resolution at the genus level in a high throughput setting with low costs.

Metagenomic shotgun sequencing is another approach to identify bacterial community members. Here the complete DNA of a sample is sequenced. Depending on sequencing depth, species can be determined even on strain level. Metagenomes also allow connecting strains with functional pathways. Nevertheless, not all genes present in an organism are from the environment the sample was taken from (Bleicher et al. 2010). Furthermore, some genes are not functionally annotated which limits the functional analysis. For instance, the number of unknown genes varies between 15% for *E. coli* (e.g. Escherichia coli Ghatak et al. 2019) and 65% for *Bifidobacterium infantis* (e.g. Bifidobacterium infantis Albert et al. 2019). The short fragments generated by shotgun NGS are assembled into contigs and mapped against databases to analyze the gene content for each community or environment (Segata et al. 2012, Sunagawa et al. 2013, Wood et al. 2019).

1.6.2. Technical factors affecting the outcome of gut microbiota profiling

Sample preprocessing for 16S rRNA gene sequencing starts with the isolation of the microbial DNA followed by the amplification of the targeted region of the 16S rRNA gene (Klindworth et al. 2013, Kozich et al. 2013) via the Polymerase-chain reaction (PCR) (Bartram et al. 2011). Each of these steps influence the outcome and is a potential source of bias. While in the DNA extraction, the bacterial DNA yield and composition may vary according to the method (Claassen et al. 2013, Wesolowska-Andersen et al. 2014, Wagner Mackenzie et al. 2015). The PCR is influenced by the primer selection (Berry et al. 2011, Klindworth, Pruesse et al. 2013) and the number of cycles for the amplification resulting in over- or underestimation of certain bacteria (Bartram, Lynch et al. 2011, Salter et al. 2014).

Fecal sample collection protocols and storage conditions (storage time, temperature etc.) were shown to influence the composition of the bacteria. Short- and long-term storage have an effect on microbial DNA stability (Carroll et al. 2012, Dominianni et al. 2014) with some bacteria being more

sensitive than others (Cuthbertson et al. 2014, Shaw et al. 2016). Recommended preservation method is to immediately freeze samples at -20°C and, for long-term storage, keep samples at -80 °C (Goodrich et al. 2014). It was further shown that a DNA stabilization liquid has advantages for preservation of the DNA and facilitates the process of sample collection and storage in studies (Ilett et al. 2019). After sample collection and possible storage DNA is isolated. There is a variety of different protocols, ranging from automatized commercial approaches to in-house developed protocols (Claassen, du Toit et al. 2013, Schirmer et al. 2015). However, there is no extraction method suitable for every sample type and protocols are not standardized, which introduces variability between sequencing facilities and between studies (Hiergeist et al. 2016), showing differences observed in the amount of isolated DNA as well as in the bacterial species found within a sample (Claassen, du Toit et al. 2013, Salter, Cox et al. 2014, Wagner Mackenzie, Waite et al. 2015). Amplification of the 16S rRNA genes from the extracted DNA is conducted by using universal primers targeting one to three hypervariable regions of the 16S rRNA gene (Klindworth, Pruesse et al. 2013), while again the selection of chemicals and methods (e.g. primer selection, polymerase reagents, number of cycles) are influencing the results (Berry, Ben Mahfoudh et al. 2011, Klindworth, Pruesse et al. 2013). Thus, despite its popularity and usefulness sample preparation for 16S rRNA gene sequencing is prone to technical artifacts and biases at various levels of the workflow.

1.6.3. Bioinformatical methods and tools for preprocessing of sequencing data

For the processing of raw sequencing data (i.e., FASTQ files), a common approach is the clustering of similar nucleotide sequences into operational taxonomic units (OTUs) (Edgar 2013, Sunagawa, Mende et al. 2013, Rideout et al. 2014). This approach is integrated in many metagenomic pipelines (Caporaso et al. 2012, Lagkouvardos, Joseph et al. 2016). A 97% identity cutoff is used as proxy for species-level diversity as it was proposed by several clustering approaches (Schloss et al. 2009, Edgar 2010, Rideout, He et al. 2014).

Another strategy was implemented in recent years, which tries to build a model on the sequencing errors found and to correct for those. This concept was introduced as the ‘exact sequence variant’ (ESV) or meanwhile commonly named ‘amplicon sequence variant’ (ASV) (Callahan et al. 2016). While the OTU approach integrates different sequences of similarity ≥ 97%, ASVs consist of identical but perhaps corrected sequences and as such, each ASV is a different variant. The increase in resolution (more ‘species’ present) brings along the major disadvantage of this approach resulting in a reduced overlap in taxonomically similar entities between samples and thus, increases the complexity to find overlaps in ‘species’ (i.e., ASVs) common between samples (Callahan et al. 2017).

Mapping algorithms (Altschul et al. 1990, Edgar et al. 2011, Katoh and Standley 2013) allow the assignment of nucleotide sequences to the corresponding bacteria. 16S rRNA amplicon sequences are aligned against reference sequences from databases (e.g. RDP (Wang et al. 2007), Greengenes (Yokono et al. 2018) and Silva (Quast et al. 2013)). Depending on the settings of the algorithm, e.g. allowed mismatches, the best sequence match is determined, and the considered taxonomy is assigned. The taxonomic classification is influenced by the database used to build the alignment, which may differ in nomenclature or may be incomplete (Edgar 2018, Park and Won 2018).

In microbiology it is assumed that more closely related species (e.g., similar 16S rRNA genes) have similar ecological functions. Based on phylogenetic trees is possible to determine closely related sequences and to define microbial differences between groups of individuals (Chen et al. 2012). In order to build a tree a multiple sequence alignment (Madeira et al. 2019) is generated from the nucleotide sequences of OTUs (or ASVs) (Price et al. 2009). The tree is either built using greedy algorithms, e.g. using the Neighbor-Joining approach, which are fast in computational time but prone to error (e.g. no global solution) and with limitations for large datasets. Another building approach is the bootstrapping method, e.g. Maximum Likelihood approach (Price, Dehal et al. 2009), which works better for big data sets but performs with poor computational power.

Downstream analysis of OTU (Schloss, Westcott et al. 2009, Lagkouvardos et al. 2017) or ASV tables (Callahan, McMurdie et al. 2016) is carried out for the calculation and visualization of diversity and composition of complex microbial communities, extract information about taxonomic composition of the environment and address differences between groups and microbial diversity. The integration of adequate statistical methods allows the identification of OTUs/ASVs, or taxonomic groups that are differentially abundant between individuals (or treatments) and statistical methods allow inferring correlations between microbial taxa and metadata (like health status) characterizing the environment of interest.

1.7. Limitations of 16S rRNA gene amplicon analysis

While 16S rRNA gene sequencing has the potential for great insights in microbiota composition, this technology bears some challenges. These include contaminations either from the sample (e.g., DNA derived from ingested bacteria of the food and not native to the gut), from the processing environment (e.g., aerosols carry DNA from one sample to another), chemicals for amplification and sequencing (e.g., some bacterial genera are known to grow in production pipelines for kit components (Klindworth, Pruesse et al. 2013), or from the sequencing itself (e.g., overlapping cluster causing misreading, index cross-talk, (MacConaill et al. 2018), and others. Filtering methods have been developed to determine and remove such contaminants from the data set. One of the most

commonly used methods, removes all sequence reads which appeared only once in the data over all samples (referred to as singleton removal) (Schloss, Westcott et al. 2009, Caporaso, Kuczynski et al. 2010, Callahan, McMurdie et al. 2016). Nevertheless, this method highly depends on sequencing depth and on the number of samples included within one sequencing run as well as the used sequencing technology. As a consequence there has been a lot of confusion in the field about how many bacterial species can be detected in the human intestine based on sequencing, with values ranging from a few hundred to several thousand (Avershina and Rudi 2015, Clavel et al. 2016). Reference studies based on low-error amplicon analysis protocols or shotgun metagenomics suggested the detection of 150 to 200 species in one individual sample, though this was based on sample size of <200 individuals (Qin et al. 2010, Faith et al. 2013). Therefore, a proper filtering must be considered for meaningful and trustworthy analysis of high throughput 16S rRNA gene amplicon datasets.

Moreover, falsely detected OTUs (referred to as spurious OTUs) should be removed in order to exclude spurious taxa which are not part of the microbial ecosystem. Despite the widespread use of 16S rRNA gene amplicon sequencing approaches, it is still unclear which thresholds are most suitable to remove contaminants of any sort. These spurious OTUs inflate the ‘species’ list with increasing sample numbers of a study and sequencing depth. Thus, benchmarking und validation of an analysis pipeline is important to generate reliable and reproducible outcomes.

2. Aims and objectives

The aim of this work is the characterization of the human gut microbiota and its effect on metabolic health. In this study we analyzed three large-scale population cohorts with a total of 4,131 participants. The primary focus was to find associations of the bacterial composition of the gut with T2D. In addition to the compositional diversity of individuals' gut microbiota, we identified robust diurnal oscillations in the faecal bacterial composition. We identified taxa with disrupted circadian rhythm and classified disease risk in individuals with T2D based on those taxa. We hypothesize that daytime-related oscillations of gut bacteria not only contribute to inter-individual variation, but also provide a link to metabolic health.

3. Methods

In the following, the used population-based human cohorts and data sets are described. The pipeline for the 16S rRNA gene sequencing is explained in detail, including bioinformatic and statistical methods.

3.1. Population-based cohorts

Using three large human cohorts ($KORA_{2013}$ and $KORA_{2018}$, TwinsUK, and FoCus cohorts), two smaller ones (enable, FS cohorts) and longitudinal data from two individuals (S1, S2), a combined data set of 5,366 subjects (n = 6,434 samples) was assembled to describe the composition of human gut microbiota in this study. Additionally, artificial mock communities were included in this work.

3.1.1. Cross-sectional $KORA_{2013}$ cohort

The longitudinal population-based cohort KORA (Cooperative Health Research in the Augsburg Region) started in 1999 and was designed to investigate cardio-metabolic health, with the focus on diabetes. Fecal samples of 2,076 individuals were collected in year 2013 during the third follow-up period of the cohort ($KORA_{2013}$). For the collection, participants were asked to use a sampling kit which included stool collection tubes and storing the samples in their household refrigerator as short as possible. The tubes have a spoon attached to the lid and 1-2 spoons of feces shall be collected. Each tube contains 5 ml stool stabilizer (Invitek DNA Stool Stabilizer, No. 1038111100). Participants were asked to either bring the samples personally to the study center or send it via mail. All samples received were stored at -80°C at the study center in Augsburg. A comprehensive data set on social-demographical characteristics, risk factors profiles, dietary habits and medical history was ascertained amongst others. Detailed study design and methods have been published previously (Holle et al. 2005). 100 stool samples from year 2013 ($KORA_{2013}$) were excluded due to medication issues (e.g., antibiotics intake) and gut-related diseases (e.g., colorectal cancer). T2D and a prediabetic phenotype were defined for each participant of $KORA_{2013}$ by oral glucose tolerance test (OGTT) or physicians-confirmation and classified by WHO in 2013. Individuals diagnosed with Type 1 Diabetes were excluded. The investigations were carried out in accordance with the Declaration of Helsinki, including written informed consent of all participants. All study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich ($KORA_{2013}$ from year 2013 EC No. 06068 and $KORA_{2018}$ from year 2018 EC No. 17040). The informed consent given by KORA study participants does not allow depositing of the data in public databases. However, data are available upon request by means of a project agreement (https://epi.helmholtz-muenchen.de/). Stool samples were paired-end sequenced on an Illumina HiSeq 2500 targeting the V3V4 region of the 16S rRNA genes performed by the ZIEL –

Core Facility Microbiome (Technical University of Munich, Freising) (Reitmeier et al. 2020). The Institute for Clinical Microbiology (IKMB) from Kiel (Germany), received DNA sample aliquots from the $KORA_{2013}$ cohort and sequenced the V1V2 rRNA gene on an Illumina MiSeq machine in paired-end mode (Relling et al. 2018).

3.1.2. Prospective $KORA_{2018}$ cohort

Consecutive samples were collected after 5 years in $KORA_{2018}$ of a subset of 699 individuals from the cross-sectional $KORA_{2013}$ cohort. Inclusion criteria for the second sampling was based on age, i.e., participants older than 75 years were excluded. T2D was diagnosed based on HbA1c value (%), a measurement which determines the long-term average blood sugar level. An incident T2D case was defined as HbA1c < 6.5% in 2013 and HbA1c ≥ 6.5% in 2018. An OGTT was not available in year 2018. Sample preparation and sequencing of the V3V4 region in paired-end mode on an Illumina MiSeq was performed by the ZIEL – Core Facility Microbiome, as described above.

A subset of 100 individuals (paired from $KORA_{2013}$ and $KORA_{2018}$) was selected for metagenomic shotgun sequencing. The dataset included incident T2D cases, persisting T2D cases and nonT2D controls, which were otherwise metabolically healthy. The shotgun metagenomic sequencing was performed by Prof. Paul O'Tool's group at the School of Microbiology and APC Microbiome Ireland, University College Cork, Ireland. The group received sample aliquots of isolated DNA. The analysis of the shotgun data was performed mainly by Dr. Tarini Gosh and Dr. Eduardo Almeida, both University College Cork.

3.1.3. *enable* cohort

The *enable* cohort, a cross-sectional study, was performed between February 2016 and February 2018 in two centers, Freising and Nuremberg, Germany. The aim of the study was to recruit individuals of four age groups. Groups included 44 children from 3-5 years, 94 adolescents and young adults aged 18-25 years, 205 adults aged 40 to 65 years (middle agers) and 160 older persons aged 75-85 years. A subset of 100 adults with an equal distribution of male and female were invited for another recruitment phase in 2018 (n = 357 samples of N = 100 subjects, from four sampling time points). Samples collection was conducted as described previously for the $KORA_{2013}$ cohort. Stool sampling and measurements to evaluate metabolic syndrome took place at the cross-sectional recruitment phase as well as in the interventional phase (three consecutive sampling time points after 4 and after 12 weeks). Stool samples were paired-end sequenced on an Illumina MiSeq machine targeting the V3V4 region of the 16S rRNA gene performed by the ZIEL – Core Facility Microbiome as described above.

3.1.4. Longitudinal sample collection of single subjects (S1 and S2)

In a small longitudinal study (N = 2 subjects) multiple samples were collected over three years. Consecutive fecal samples (n = 58 samples) of a 52-year-old healthy male subject (S1) were collected over three years (July 2018 – February 2020). Multiple timepoint sampling was conducted in four phases (July-August 2017 n = 22 samples, December 2017 n = 12 samples, February-March 2018 n = 15 samples and January 2020 n = 10 samples). Additionally, consecutive samples (n = 34 samples) from a 30-year-old healthy female individual (S2) were collected over two years (July 2018-February 2019). Sample collection and sequencing was performed as described above.

3.1.5. Human validation cohort (FS cohort)

For a small human control cohort (FS cohort), participants (N = 9) were enrolled from Freising, Germany. They were asked to provide fecal samples including information about sampling time point. Samples were stored at 4°C. Native samples were portioned (2gr in each tube) in 5-ml stool collection tubes and diluted each with 600µl DNA stabilizer (Invitek DNA Stool Stabilizer, No. 1038111100). After homogenization, these aliquots were stored at -80°C. Sequencing was conducted as described above.

3.1.6. TwinsUK cohort

In the TwinsUK cohort, N = 1,399 individuals, including N = 46 incident T2D cases as well as N = 94 T2D cases that were classified using a combination of self-reported questionnaires as well as longitudinal glucose measurements were included. T2D was diagnosed based on blood glucose values and assignment of cases and controls was performed by Tim Spector's group from King's College, UK , who also conducted sampling and sequencing. The collection of fecal samples, DNA extraction, amplification of the V4 hypervariable region of the 16S rRNA gene (primers 515F and 806R), purification and pooling were performed as previously described (Goodrich, Waters et al. 2014). The pooled amplicons were sequenced using the Illumina MiSeq platform with 2×250bp paired-end sequencing. The raw sequencing data were accessed via the European Nucleotide Archive (ENA: PRJEB13747).

3.1.7. FoCus cohort

The cohort originated from the Food Chain Plus (FoCus) project, running from 2011 to 2014. Sample collection and preparation were performed by the IKMB in Kiel, Germany as described previously (Relling, Akcay et al. 2018). DNA aliquots of 1,529 fecal samples were provided by the IKMB. The V3V4 region of the 16S rRNA gene was sequenced by the ZIEL – Core Facility Microbiome on an

Illumina HiSeq, otherwise as described above. Samples with low read counts (<5,000 reads) were excluded from data analysis as well as samples with missing information in BMI, HOMA index values and sampling time. In total 1,488 subjects were considered for the analysis. T2D was classified based on HOMA index which determines the insulin resistance by considering the fasting insulin and the fasting glucose (T2D: HOMA > 5.0 Stern et al. 2005).

3.1.8. Studies from public repositories for validation

Data from two publicly available human cohorts were included in this work. Amplicon sequencing data were accessed by downloading FASTQ sequencing files from ENA archive. First, a longitudinal population-based cohort published by Flores et al. (2014) (N = 85 subjects, n = 255 samples) (EMBL accession number ERP005150-ERP005153). Second, a study published by Halfvarson et al. (2017), focusing on shifts in the human fecal microbiota over time in patients with IBD vs. controls, of 1 - 10 fecal samples for each individual (N = 137 subjects, n = 683 samples) (EBI accession number ERP020401) were included.

3.2. Artificial microbial communities

3.2.1. Mock communities

Two spiked mock communities were included in the analysis. Mock communities have a defined bacterial composition. The TUM-Mock was in-house prepared and consist of 13 strains (**Table 1**). The DNA of this community was diluted to 12 ng/ul and technical PCR replicates were taken for sequencing.

Table 1 Composition of TUM-mock

Genera	Expected rel. abundance (%)
Actinomyces	6.3
Alistipes	3.5
Bacillus	14.0
Bacteroides	8.5
Cellulosimicrobium	5.1
***Clostridium* XVIII**	14.2
Enterococcus	13.3
Enterorhabdus	4.3
Flavonifractor	3.0
Parabacteroides	6.8
Pseudomonas	4.1
Staphylococcus	11.4
Acetatifactor	5.4

The second mock community was bought from Zymo (Zymo-Mock), compromising eight bacterial strains **(Table 2)**. Of this pool, technical PCR triplicates were included for 16S rRNA gene sequencing.

Table 2 Composition of synthetically spiked Zymo-Mock.

Genera	Expected rel. abundance (%)
Bacillus	17.4
Enterococcus	9.9
Escherichia/Shigella	10.1
Lactobacillus	18.4
Listeria	14.1
Pseudomonas	4.2
Salmonella	10.4
Staphylococcus	15.5

Besides these two in-house used mock communities, we analyzed published datasets which included other mock communities (see **Table 3**). From those studies, raw sequencing data were downloaded from public repositories. There are differences in targeted V-region of the 16S rRNA gene as well as differences in number of bacterial strains being part of these mock communities (ranging from 10 to 58 bacterial strains). These mock communities were included to increase the general number of communities from different sources, enabling observation of any possible bias in our in-house pipelines or the targeted V-region of the 16S rRNA genes. The 16S rRNA gene amplicon datasets generated concerning mock communities are available in the European Nucleotide Archive (www.ebi.ac.uk/ena) under study accession number PRJEB34431.

3.2.2. Gnotobiotic mice

Gnotobiotic mice were colonized with minimal consortia of known bacteria. We created four gnotobiotic communities varying in number of used strains (4 to 12) (**Table 4**). For each community, 9 to 12 mice were inoculated. Extracted DNA from feces were paired-end sequenced by targeting the V3V4 region of the 16S rRNA genes. Sample preparation and sequencing was performed by the ZIEL – Core Facility Microbiome on an Illumina MiSeq as described above. The 16S rRNA gene amplicon datasets generated for this part are available in the European Nucleotide Archive (www.ebi.ac.uk/ena) under study accession number PRJEB34431 (data from the ZIEL Core Facility Microbiome.

Table 3 Mock communities used in the present study

Name	Seq. facility[b]	Gene region	Replicates	No. species	No. raw reads	No. sequences after processing	Total no. taxa[d] (no filtering)	1st spurious taxon[d] (% rel. abundance)	Reference
Mock-1	See ref.	V4	1	27	153,841	140,397	432	0.007	(Alanis-Lobato et al. 2017)
Mock-2	See ref.	V4	1	58	593,868	578,569	761	0.439	(Schirmer, Ijaz et al. 2015)
Mock-3	See ref.	V4	1	21	1,012,097	453,215	1,081	0.031	(Kozich, Westcott et al. 2013)
Mock-4	See ref.	V4	1	14	169,516	159,352	417	0.020	(Tourlousse et al. 2017)
Mock-5	See ref.	V4	1	21	613,091	108,414	802	0.439	(Kozich, Westcott et al. 2013)
Mock-6	See ref.	V4	1	21	602,819	231,685	732	0.160	(Kozich, Westcott et al. 2013)
Mock-7	See ref.	V4	1	20	306,773	42,746	95	0.008	(Bokulich et al. 2013)
Mock-TUM	1	V34	7	13	25,640 ± ,516	19,882 ± 8,163	77 ± 15	0.130 ± 0.138	This study
Zymo-Mock (cat. #D6300)[a]	1	V34	7	8[c]	67,465 ± 31,752	52,079 ± 24,759	177 ± 33	0.059 ± 0.026	This study

In case of replicates, data are shown as mean ± sd

The sequences and taxonomies of all species included in the respective Mocks are provided in ***Table 1*** and ***Table 2***

[a] D6300 corresponds to an evenly distributed mixture of the microbes

[b] For in-house generated data in this study: ZIEL Core Facility Microbiome, TU Munich, Freising, Germany

[c] The Mock community also includes two yeast species not considered here (i.e., 10 species in total).

[d] All values refer to Operational Taxonomic Units (OTUs) clustered at 97% sequence identity, for which values refer to Amplicon Sequence Variants (ASVs).

Table 4 Gnotobiotic mouse communities used in the present study

	Gene region	Replicates	No. species	No. raw reads	No. sequences after processing	Total no. OTUs (no filtering)	1st spurious OUT (% rel. abundance)	Reference
GNOTO1	V3V4	6	7	28,706 ± 4,904	25,869 ± 4,494	66 ± 4	0.101 ± 0.014	This study
GNOTO2	V3V4	9	12	30,261 ± 11,434	21,148 ± 8,313	172 ± 24	0.009 ± 0.004	This study
GNOTO3	V3V4	6	6	30,444 ± 42,325	27,632 ± 3,563	85 ± 10	0.116 ± 0.016	This study
GNOTO4	V3V4	7	4	25,217 ± 6,514	47,505 ± 6,106	68 ± 10	0.249 ± 0.041	This study

In case of replicates, data are shown as mean ± sd

3.3. Sample processing for 16S rRNA gene sequencing

Sample processing is divided into four main steps: DNA isolation, library construction by PCR, amplicon cleaning and dilution, and sequencing (**Figure 2**).

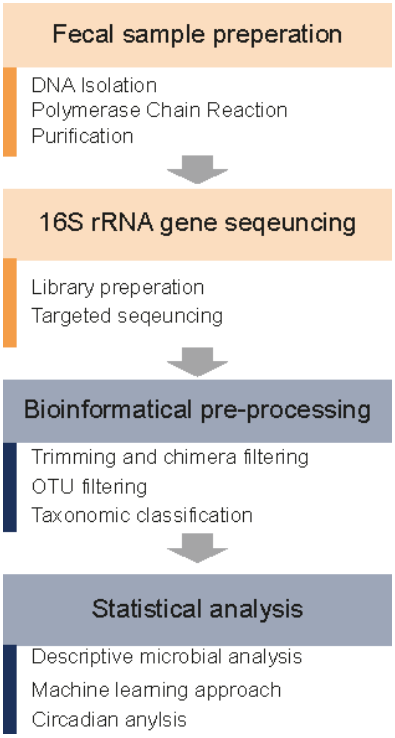


Figure 2 Overview of steps for 16S rRNA gene sequencing.

3.3.1. DNA isolation methods and protocols

The protocol used for the isolation of DNA from fecal samples is a modified version of Godon (1997). In total, 24 samples can be processed in parallel. Fecal samples (approx. 2 g in 5 ml stool stabilizer) are thawed for approximately 2 hours at room temperature (at 20°C – 22°C). Samples are vortexed until they are fully homogenized and 600 µl are transferred into a 2-ml screw cap tube containing 0.1 mm silica beads. To denature proteins, 250 µl 4 M guanidinium thiocyanate are added. As ionic surfactant, separating cellular components, 500 µl 5% N-lauroylsarcosine sodium salt is added. Samples are incubated for 60 min at 70 °C while shaking at 700 rpm. Remaining intact microbial cells are lysed by using a FastPrep-24 fitted with a CoolPrep adapter (filled with dry ice; program: 5 CY: 40 s; 6.5 m/s for 3 rounds). To remove phenolic and other fecal contaminants, 15 mg polyvinylpyrrolidone are added. Samples are vortexed and centrifuged for 3 min at 15,000×g and 4 °C. The supernatant is transferred to a new 2-ml tube. RNase (final concentration of 10 mg/ml) is added and incubated 20 min at 37 °C while shaking at 700 rpm. The DNA is purified using a silica-membrane based approach

following the manufacturer's instructions of the kit used (NucleoSpin gDNA Clean-up Kit, REF 740230.250 Machery-Nagel). After DNA purification, nucleic acid concentrations are measured using a NanoDrop 2000 (ThermoFisher scientific).

Besides the in-house protocol, the following kits were included for comparison (**Supplementary table 1**):

- PSP Spin Stool DNA Plus Kit
- OneStep™ PCR Inhibitor Removal Kit
- DNA Clean & Concentrator™ -5 (DNA C&C)
- NucleoSpin® gDNA Clean-up gDNA Clean & Concentrator™ -10 (gDNA C&C)
- QIAamp DNA Stool Mini Kit

The least kit was also used by an automatized DNA isolation pipeline provided by QIAGEN (QIAcube system No. 90012992). All kits were implemented as described in the manufacturer instruction.

3.3.2. Library construction and PCR

The amplification was performed in a 2-step PCR approach. (**Table 5**). The selection of primer added in the 1st step Mastermix depends on the targeted region of the 16S rRNA gene (**Table 6**). Besides, two different polymerase were used in this work: Phusion HotStart polymerase (Thermo Freezer; Cat No. F-549L) and Q5 polymerase (NEB; Cat. No. M0491S). Finally, 3 µl of each DNA sample aliquot (with 12 ng/µl) is mixed with 27µl of the Mastermix to perform the PCR (**Table 7**).

Table 5 Mastermix for 1st-step PCR

Reagents	Volume µl / sample
Phusion® HF Buffer (without Dye)	6
dNTPs (20µmol,)	0.6
341F-ovh Primer 20µM	0.1875
785r-ovh Primer 20µM	0.1875
Phusion® High-Fidelity DNA Polymerase Hotstart	0.15
DMSO 100% *	2.25
Water (for molecular biology, DEPC-treated and filter-sterilized)	17.625

Table 6 Primer sequences used to amplify different V-regions for short amplicon 16S rRNA gene sequencing.

V-region	Forward primer	Reverse primer	Forward sequence	Reverse sequence	Annealing Temp. (°C)	Reference
V12	27F	338R	AGAGTTTGATYMTGGCTCAG	GCTGCCTCCCGTAGGAGT	57	Salter et al. (2014)
V13	27F	534R	AGAGTTTGATYMTGGCTCAG	ATTACCGCGGCTGCTGG	57	Walker et al. (2015)
V34	341F	785R	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	55	Klindworth, Pruesse et al. (2013)
V4	515F	806R	GTGCCAGCMGCCGCGGTAA	GGACTACHVGGGTWTCTAAT	53	Caporaso et al. (2011)
V46	515F	944R	GTGCCAGCMGCCGCGGTAA	GAATTAAACCACATGCTC	53	Fuks, Elgart et al. (2018)
V68	939F	1378R	GAATTGACGGGGGCCCGCACAAG	CGGTGTGTACAAGGCCCGGGAACG	58	Lebuhn et al. (2014)
V79	1115F	1492R	CAACGAGCGCAACCCT	TACGGYTACCTTGTTACGACTT	51	Edwards et al. (2007)

Table 7 PCR cycling conditions for 1-step PCR. Rows in grey are performed for 15 cycles.

Steps	in °C	Time	Cycles
Initial Denaturation	98	30 sec	1
Denaturation	98	5 sec	15
Annealing	55	10 sec	
Extension	72	10 sec	
Final Extension	72	2 min	1
Hold	10	∞	

The Mastermix for the 2nd-step PCR includes forward and reverse index primer for sample barcoding with a double index (Kozich, Westcott et al. 2013). From the first PCR, 2 µl are mixed with 45.5 µl of the Mastermix, and 2.5 µl of each reverse index primer are added to conduct the PCR (**Table 9**). The 2nd-step PCR was conducted twice, and the final products of both reactions are pooled.

Table 8 Mastermix for 2-step PCR

Reagents	Volume in µl / Sample
Phusion® HF Buffer (without Dye)	10
dNTPs (20µmol, Bioline BIO-39043)	1
e.g. 341-ovh-HTS- SC501 Primer 20µM	0.313
Phusion® High-Fidelity DNA Polymerase Hotstart *	0.2
DMSO 100% *	1.5
Water (for molecular biology, DEPC-treated and filter-sterilized)	32.487

Table 9 PCR cycling conditions for 2-step PCR. Rows in grey are performed for 10 cycles.

Steps	Temperature	Time	Cycles
Initial Denaturation	98	30 sec	1
Denaturation	98	5 sec	10
Annealing	55	10 sec	
Extension	72	10 sec	
Final Extension	72	2 min	1
Hold	10	∞	

3.3.3. PCR purification and 16S rRNA gene sequencing

PCR purification was performed with AGENCOURT AMPure XP Beads (Beckman Coulter according to the manufacturer's instructions; **Supplementary file 1**) and was fully automatized by using a Beckman Coulter Biomek 4000 robot. DNA is measured by using a fluorimetry (Qubit measurement according to the manufacturer's instructions). Amplicons were pre-diluted to a concentration of 2 nM and then diluted to a final concentration of 0.5 nM. The molarity of each sample was calculated based on measured Qubit concentrations and library size for the targeted rRNA region of the 16S gene (e.g., V3V4 = 572bp) as follows:

$$concentration\ in\ nM = \frac{(concentration\ in\ ng/\mu l)\ x\ 10^6}{(average\ library\ size\ in\ bp * 660\ g/mol)}$$

The average insert/amplicon size for V3V4 is 572bp. For sequencing the amplicon DNA was denatured first and diluted to 20 pM. A pool of all samples (about 300 to 350 per run) was created. A fresh 0.2 nM NaOH solution and a 0.2 nM Tris HCl solution was prepared and 40 µl this solution was added to 40 µl of the 0.5 nM amplicon pool. After vortexing and centrifuging for 1 min (280×g), 880 µl cooled HT1-Buffer was added. The resulting library was diluted to a final concentration of a 10 pM. To increase complexity, necessary for amplicon sequencing, PhiX DNA was added to the final library (Illumina PhiX Control v3, Cat. No. FC-110-3001) of which 600 µl were transferred either to the Illumina MiSeq cartridge v3 600 cycles or to an Illumina HiSeq cartridge Rapid Cluster Kit v2.

3.4. Preprocessing of 16S rRNA gene sequencing data

The Illumina machine generates FASTQ files. For each sample, one forward-read file (R1) and one reverse-read file (R2) is generated. Raw sequencing data were preprocessed using the IMNGS pipeline (Lagkouvardos, Joseph et al. 2016) with the following settings: trim score 5, expected error rate of 1. In IMNGS, chimera are removed using UCHIME (Edgar, Haas et al. 2011) and the reads of de-multiplexed samples are merged and clustered by 97% similarity into operational taxonomic units (OTUs) using UPARSE v8.1.1861_i86 (Edgar 2013). Spurious OTUs were filtered with different approaches including singleton removal or using a relative abundance cutoff of < 0.25%.

3.4.1. RHEA - Open-source R-based scripts for downstream analysis

The downstream analysis pipeline was implemented in the programming language R **(Figure 3)**. R is an open-source language, which runs on all conventionally used operating systems (Windows, Unix and MacOS). All scripts are fully independent from each other and are structured in a uniform style. The pipeline is available on the GitHub repository (https://github.com/Lagkouvardos/Rhea) (Lagkouvardos, Fischer et al. 2017). Notable contributors in the development of the pipeline are Neeraj Kumar who was involved in the implementation of the scripts and Dr Ilias Lagouvardos who reviewed the manuscript.

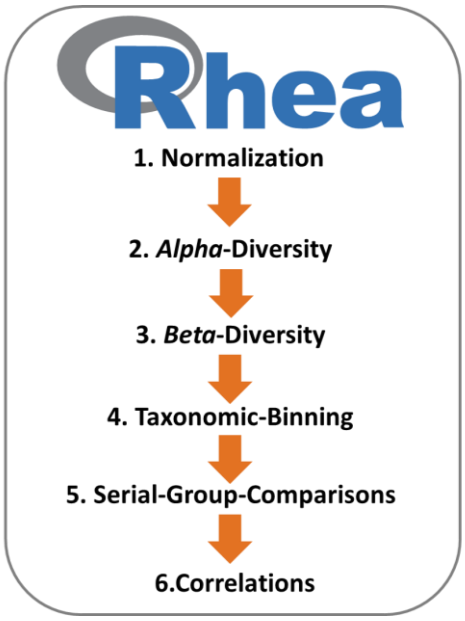


Figure 3 Rhea - R-based pipeline for the analysis of microbial 16S rRNA gene sequencing data.

1.4.1.1. Normalization of reads

Absolute read counts were normalized either by the minimum sum of reads observed within one sample or by a fixed normalization value of 10,000 reads.

3.4.1.1. *Alpha*-diversity

Alpha-diversity, the variation within one sample, was measured by either richness (number of observed OTUs) or Shannon effective counts (exponential function of Shannon index/bacterial diversity index). For the calculation of *alpha*-diversity a denoising threshold of 0.5 read counts per OTU is set, to remove low abundant OTUs which are probably not part of the community.

3.4.1.2. Taxonomic classification

Taxonomies of OTUs were classified according to RDP classifier (Wang, Garrity et al. 2007) and verified using other databases (SILVA (Quast, Pruesse et al. 2013), Greengenes (Yokono, Satoh et al. 2018), BLAST (Altschul et al. 1990)). The cumulative relative abundance was calculated of all taxonomic levels (kingdom, phylum, class, order, family, and genus) for each sample by summing up the relative abundance values from different OTUs assigned to the same taxonomy.

3.4.1.3. *Beta*-diversity

Similarity between samples was estimated based on a distance matrix using unweighted UniFrac (considering the phylogeny of sequences but without considering the abundance) and generalized UniFrac (considering the phylogenetic tree generated). For graphical illustration, projections of each sample in a two-dimensional space were used, either multidimensional scaling (MDS) plot or non-linear MDS (nMDS) plot. In addition, similarity was shown using samples as branches in a dendrogram. Significance between groups was determined by a permutational multivariate analysis of variances (*adonis* function of the *vegan* R-package) and p-values are corrected for multiple testing following Benjamini-Hochberg (Yekutieli and Benjamini 2001).

3.4.1.4. Groupwise differential abundance analysis

Differences in relative abundance of taxa and/or OTUs were determined by a Kruskal-Wallis Rank Sum Test for multiple groups and a Wilcoxon Rank Sum test for pairwise comparison. Differences in prevalence of taxa and/or OTUs were determined by a non-linear Fisher Exact test. P-values were corrected for multiple testing according to Benjamini-Hochberg correction. Changes over time were determined by a non-parametrical Friedman test and adjusted for multiple testing.

To provide information about possible associations between OTUs/taxa and continuous variables of interest (e.g. blood values, age, BMI), a correlation and network analysis were performed. The analysis calculates pairwise correlations based on Pearson correlation coefficient including p-value as well as an adjusted p-value. A heatmap was generated to provide an overview of all correlations, color coded for negative and positive correlation, as well as for significance. A network of variables with a correlation above a certain cutoff was generated to illustrate multiple connections and to find possible cluster of strongly connected taxa/OTUs.

3.4.2. Advanced statistical analysis for population-based cohorts

Effect modifier and confounder were determined by a permutational multivariate analysis of variances (*adonis* function of the *vegan* R-package). The explained variation of co-variables was determined by calculating R^2 values and were considered as significant with a p-value ≤ 0.05. For the cumulative explained variation, all significant covariates were included in a multivariate model. Data were adjusted according to confounding factors (gender, age, BMI, physical activity, PPI, metformin, and vitamin D intake) as well as stratified according to phenotypical characteristics. De-novo clustering was based on Ward hierarchical clustering. The selected number of clusters was chosen according to the Calinski and Harabasz index and was performed with the R package *NbClust* v.3.0. The circular

sample trees were based on generalized UniFrac distances and generated by the online tool EvolView v2 (He et al. 2016). Taxonomic trees were generated by Graphlan (Segata et al. 2013).

A random forest model was used to classify binary outcome variables based on a combination of co-variables and microbial composition with a 5-fold cross validation by using *randomForest* from the R package *randomForest* v4.6-14. To receive a robust and generalizable classification model, the machine-learning algorithm was applied 100-times iteratively assigning randomly individuals to either the training (80%) or test set (20%). For the training set, a subset of equally distributed cases and controls was taken to train the model. The model was further validated on the 20% test set. Based on out-of-bag error rates and Gini index, the most important features were selected for each iteration using *rfcv* from R package *randomForest* v4.6-14. Features, which appeared in at least 50% of all 100 random forest models, were considered as classification feature for the final model. Performance of the model was validated by several measurements generated by the R package performance v. 0.4.5. For each iteration, a received operational curve (ROC curves) with the AUC values was generated, as well as a precision – recall curve with the corresponding AUCPR showing the rate of positive predicted values among all true positives.

For the illustration of the results, the mean values of the curves were used as well as all AUC. The F1 score and Matthews Correlation Coefficient are both providing information about the model's prediction performance of the binary classifier (David Martin and Powers 2011, Chicco and Jurman 2020). Additionally, a learning curve was implemented which shows the out-of-box error (OOB error) rate for different number of samples included in the training set, to provide information about over- or underfitting. We further generated a mixed effect random forest model (MERF) to evaluate the effect of possible confounding variables towards the classification. Feature selection and performance of the MERF model was implemented as in random forest model by the R-package *MixRF* v 1.0. To address the issue of compositional data in microbiome analysis we transformed the original data input (relative abundance) via i) centered log scaled and ii) paired log ratio of OTUs and repeated the random forest approach. A generalized linear model (GLM) for binomial distribution and binary outcome (logit) was generated using the previously selected features. Therefore, two approaches were followed. First, the model was tested in a nested 80% - training and 20% - test scenario as described in the previous section. To verify the importance of the selected features, a generalized linear model for control OTUs (randomly selected OTUs, equal number of OTUs as feature list) are implemented repetitively 100-times. After validation of the model's quality, another GLM was generated based on all individuals. Predictions and classifications of other cohort studies are all performed on the generated GLM. Therefore, it is necessary to determine the corresponding OTUs representing the features or the GLM. FASTA sequences from the GLM OTUs are aligned against all FASTA sequences from the other cohort using BLAST (Edgar 2010). Based on the highest sequence similarity the selected s-arOTUs are assigned

to the corresponding sequences of the prediction/classification dataset. Sequences with an identity of 97%, coverage of 80% and an E value < 10^{-5} are considered as a hit. If there was no matching sequence available using the above thresholds, the best match is taken instead. Aiming to uniquely assign sequences to one reference sequence. A sequence is only allocated multiple times if no unique match above > 80% was available. The molecular species of the selected OTUs were further verified via EzBioCloud (Yoon et al. 2017).

All models used in this work are described in the **Supplementary Table 3** including information about binary classifier, included features, referred figures and performance. Scripts are integrated in the 16S amplicon shiny app Namco (online available in alpha version; https://exbio.wzw.tum.de/namco/) developed by the group of Dr. Markus List, Technical University Munich Chair of Experimental Bioinformatics.

3.4.3. Illustration and analysis of microbiota datasets for circadian rhythms

A high-resolution time course was generated by merging samples taken between 5:00 am and 24:00 pm in two-hour intervals, and a larger 4-hour interval at night from 00:01 am to 4:59 am to compensate for the low sample size in some groups during the nighttime points. The merged data points are illustrated using GraphPad Prism v6.01 (GraphPad Software) with the sample size of every groups per time point indicated below the data point in the individual graphs. To demonstrate the overall phase relationship and periodicity of all OTUs together, heatmaps have been generated using the online tool *Heatmapper* (http://www.heatmapper.ca). The raw data of each OTU were merged in the above-indicated intervals, sorted by the peak phase based on cosine-wave regression analysis (described below) and scaled in each row according to the highest abundance of the OTU.

Statistical analyses were conducted with GraphPad Prism v6.01 (GraphPad Software) and the R script JTK_CYCLE v3.1.R using Rstudio v1.1.456 (Rstudio Inc.). To efficiently identify and characterize diurnal oscillations in large datasets, circadian variation was tested by fitting a cosine-wave equation:

$$y = baseline + amplitude * \cos(2 * \pi \left(x - \frac{phase\ shift}{24}\right))$$

or a double harmonic cosine-wave equation:

$$y = baseline + (amplitude\ A * \cos(2 * \pi \left(\frac{x - phase\ shift\ A}{24}\right))) + (amplitude\ B * \cos(4 * \pi \left(\frac{x - phase\ shift\ B}{24}\right)))$$

on *alpha*-diversity and relative abundance, with a fixed 24-h period. The goodness of fit was corrected for multiple comparisons and the significance was determined using an F-test. Results from the cosine- and harmonic cosine-wave regression were compared with a widely used rhythmicity detection algorithms JTK_CYCLE which employs a non-parametric algorithm detecting sinusoidal signals (Hughes

et al. 2009), whereby JTK presents the highest false negative rates (Hughes et al. 2010). Each p-value was Bonferroni-adjusted for multiple testing. A statistically significant difference was assumed when p-value ≤0.05. The high-frequency time sampling allowed the identification of diurnal-regulated microbiota with a high statistical power (Hughes, Hogenesch et al. 2010).

The relative abundance of each OTU was assessed for a 24-h rhythmicity using the cosine-wave regression. OTUs which showed diurnal fluctuation were further analyzed for differential 24-h time-of-day patterns using the Detection of Differential Rhythmicity (DODR) R packages (Thaben and Westermark 2016). Resulting DODR p-value were corrected for multiple comparisons and at the corrected p-value ≤ 0.05 significance level.

3.4.4. Analysis of associated functional pathways

The analysis of the shotgun sequencing data was mainly performed by Dr. Tarini Gosh and Dr. Eduardo Almeida, School of Microbiology and APC Microbiome Ireland, University College Cork, Ireland

The taxonomic and functional annotation of the shotgun fecal metagenomic datasets were performed using the *metaphlan2* (Segata, Waldron et al. 2012) and HUMAnN2 (Franzosa et al. 2018) pipelines. Gene families detected using the HUMAnN2 approach were mapped to the KEGG pathway orthologs (Kanehisa and Goto 2000) using internal mappings within HUMAnN2. *Psych* R package v1.8.12 was used to compute the correlation between the OTU markers with the clinical markers and the KEGG pathways (Spearman correlation filtered with Benjamini-Hochberg corrected FDR ≤ 0.1). For the identification of the top disease-predictive pathways, we used an iterative random forest approach *randomForest* from R package *randomForest* v4.6-14, where we performed 100 iterations, each time taking 50% of all cases(from both 2013 and 2018) and an equal number of controls and tested the same model on the remaining 50%, again with an equal number of controls. Mean AUC and mean feature importance scores were computed across iterations using standard R functions. Results were further compared with the results of two published studies: a cohort of Qin et al. (2012) as well as a cohort of Beli et al. (2019). Pathways that were only detected in either cases, controls or had more than two-fold increase of representation in case or controls were identified as enriched.

4. Results

The first part of the result section focuses primarily on the validation of the 16S rRNA gene sequencing pipeline. Different methods and protocols for fecal sample preparation were compared and validated to set up a standard protocol for the preparation of stool samples for targeted amplicon sequencing. Preprocessing steps of raw sequencing data were critically evaluated and extensively tested to assure high quality data.

In the second part, population-based cohort studies were conducted and analyzed. The three large population-based cohort studies are described and compared against each other, as well as longitudinal data sets evaluating the dynamic and stability of the gut microbiota in and between cohorts. Finally, the daytime-related changes in the bacterial composition of the gut was investigated. The aim of the analysis was to evaluate the impact of metabolic health, especially T2D, on the bacterial composition (either generally or during different daytimes) in order to obtain prognostic and diagnostic tools derived from the microbiota.

4.1. Comparison of DNA isolations, polymerase chain reactions, and 16S rRNA gene sequencings

4.1.1. The influence of DNA extraction methods

Several publications already showed that the sample preparation is one of the strongest influencing factors for the 16S gene rRNA amplicon sequencing pipeline (Berry, Ben Mahfoudh et al. 2011, Claassen, du Toit et al. 2013, Wesolowska-Andersen, Bahl et al. 2014, Wagner Mackenzie, Waite et al. 2015). According to literature (see Introduction), the DNA isolation method seems to have a strong impact on the results. We tested seven different protocols in ten different settings, including one fully automated approach (see Methods, **Supplementary table 1**). Stool samples from eight human subjects (N = 8 subjects, n = 24 samples (technical triplicates) from the FS cohort were used to evaluate the intra – and inter individual effects for the different DNA isolations. First, the quality of the PCR product was evaluated by gel electrophoresis, yield (**Figure 4**) and purity was measured.

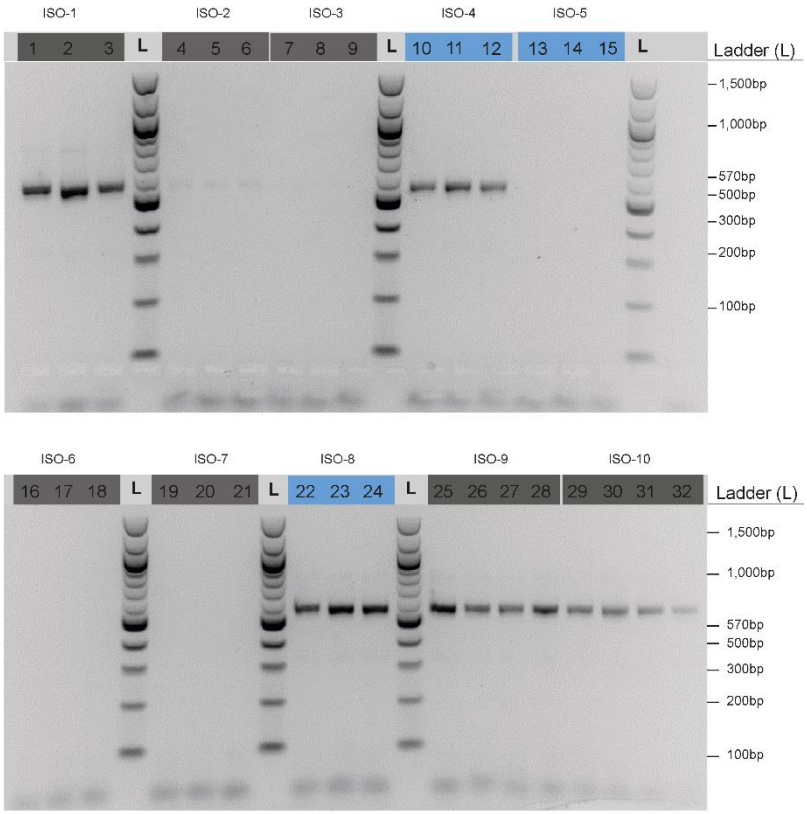


Figure 4 Gel electrophoresis of 2nd PCR products showing the targeted amplicon length for 16S rRNA gene sequencing. Obtained from DNA isolated with different methods.

Top panel: Isolation methods 1 to 5 (ISO-1 – ISO-5). Bottom panel: Isolation methods 6 to 10 (ISO-6 – ISO-10) (compare to **Supplementary table 1**). Shown are PCR products after the second PCR. Missing products indicate insufficient amounts of DNA. L; ladder (3 µl): 100-bp ThermoDNA indicating the aspired size of 570 bp for the PCR product. 10 µl sample and 1 µl dye on a 1.5% agarose gel (150V for 30min).

DNA isolation protocols producing insufficient amounts of DNA (**Figure 4**) were excluded in the following. Thus, fecal samples of the FS cohort (N = 8 subjects, n = 24 samples) were isolated using ISO-4, ISO-5, and ISO-8. The DNA was sequenced on an Illumina MiSeq (paired-end) and analyzed using IMNGS and Rhea (see Methods) with standard settings. To determine the similarity of microbial composition within triplicates, the mean generalized UniFrac distance within one subjects was calculated (**Table *5***). The smaller the mean distance within triplicates, the higher the similarity between samples. Smallest intra-individual variation was observed when using ISO-4, while ISO-5 showed the highest intra-individual distance. Thus, ISO-4, a modified version of the protocol published by Godon (1997), had the highest similarity within each triplicate of DNA isolation conducted. Richness and Shannon effective counts were less influenced and showed good agreement with the majority other the other protocols tested. Considering usability, time and costs for consumables, ISO-4 was chosen as the standard DNA isolation method.

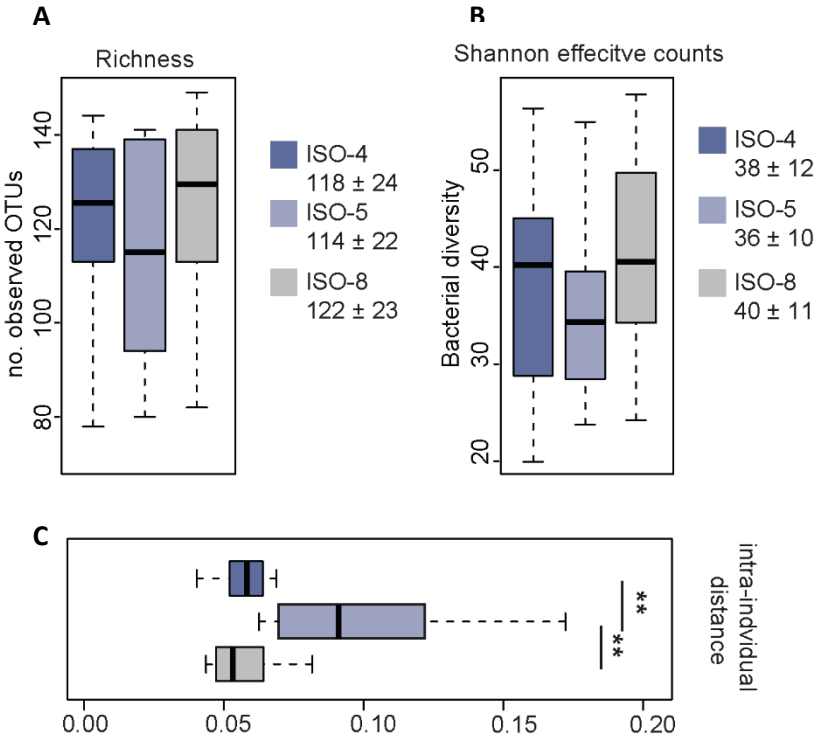


Figure 5 Compositional analysis of 16S rRNA gene sequencing data based on selected DNA isolation methods for the same samples.
The different DNA isolation methods are indicated by different colors as shown in each panel. (**A**) *Alpha*-diversity for each sample (mean over duplicates) for three methods compared. (**B**) Boxplots are Shannon effective number of species (**C**) Comparison of intra-individual generalized UniFrac distances; significant differences between the methods are shown.

4.1.2. The impact of PCR cycles and polymerases on sequence quality and results

For the validation of the PCR settings eight different combinations of cycles were tested for the 2-step PCR, as well as two different polymerases. The aim was to minimize the number of observed artefacts and to increase the number of detected molecular species (i.e., OTUs). Therefore, samples from the FS cohort (N = 8 subjects) and samples from the NGS-Mock were isolated according to ISO-4. DNA was amplified with 5 to 20 cycles in the first PCR and 5 to 20 cycles in the second PCR in different combinations. PCR products were sequenced. Results were compared according to relative abundance of OTUs between samples. Cycle numbers resulting in similar relative abundance values were selected and further tested by using two different polymerases (Phusion HotStar and Q5). Using the mock community, the influence of PCR cycles on the number of true positives and false positives was validated. A cutoff of 0.25% (see below) was found to be a threshold, which separates spurious OTUs (e.g., contaminations, crosstalk between barcodes, etc.) from true OTUs. Thus, this filtering cutoff of 0.25% was applied for the PCR-test conditions. The results showed that the number of observed OTUs differed between cycles and polymerase (**Table 6**).

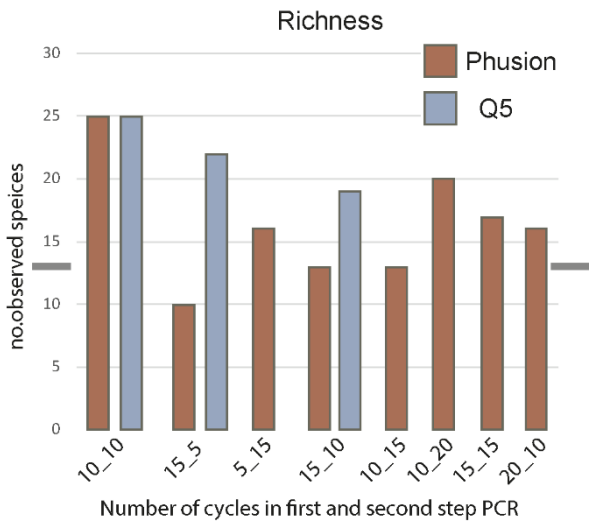


Figure 6 Influence of PCR reagents and number of cycles.

Number of observed species for NGS-Mock. Two different polymerases are used for the analysis (brown = Phusion high fidelity, blue = Q5); varying cycle numbers for the first and second PCR are indicated. The number of bacterial species (13) present in the NGS-Mock is marked with grey bars.

Compared to the original number of 13 species in the mock community, best results were generated when using 25 cycles in total and the Phusion polymerase (**Table 6**).

To observe the number of false positives, an OTU table without filtering was generated. The highest fraction of the first false OTU was recorded. The fraction of the false positive with the highest amount varied between cycles (**Table 10**). However, for some combinations cycling or polymerases, few species were missed (grey area), while the number of false OTUs varied. The lowest number of false positives was observed with 15 and 10 cycles for the first and second step, respectively, and using Q5. Nevertheless, the relative abundance of the most prominent false positive was much lower with the same number of cycles but using the Phusion polymerase.

Table 10 Appearance of artefacts with different polymerase and different number of cycles for 2-step PCR.

Cycles	1-Cycle	2-Cycle	Polymerase	Rel. abundance of most prominent artefact	No. of artefacts	No. of observed species
25	15	10	Phusion	0.013%	14	13
25	15	10	Q5	0.025%	9	13
25	10	15	Phusion	0.015%	17	13
20	15	5	Phusion	0.013%	13	13
20	15	10	Q5	0.020%	12	12
20	10	10	Phusion	0.042%	18	10
20	10	10	Q5	0.077%	34	10
20	5	15	Phusion	0.022%	18	10

4.1.2. Sample size, sequencing depth and number of reads

Next, to evaluate the influence of sample numbers included in OTU table generation, samples from known subjects (N = 10 subjects from $KORA_{2013}$) were integrated and combined with samples from different studies. The number of included samples ranged from 1,649 to 3,133 samples within one OTU table. With an increase in the total number of included samples, an increase in number of OTUs was observed for each of the ten subject samples. Thus, for the mentioned sample number range, the amount of observed OTUs approximately doubled within one individual resulting in a strong association with richness observed (**Figure 7**). The number of samples included in a study is influencing the results. This effect needs to be addressed while comparing results from different studies.

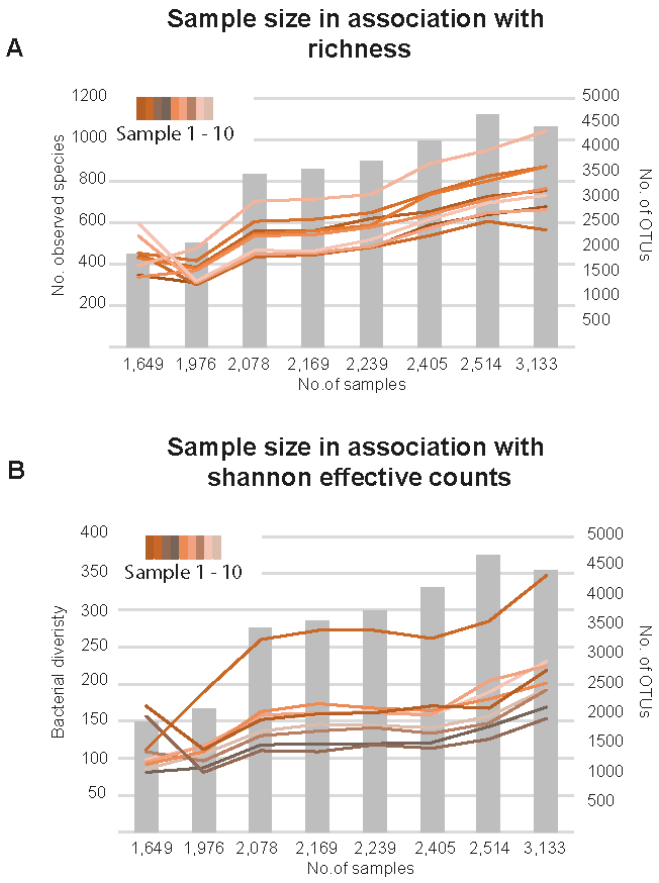


Figure 7 The influence of sample size, generated OTUs and *alpha*-diversity.
Grey bars indicate the number of samples included to generate OUT tables (x-axis to the right). Colored lines indicate the richness (i.e., number of observed 'species' as OTUs, **A**) or Shannon effective counts (i.e., bacterial diversity, **B**) for each sample of the 10 subjects. The color of the ten samples is indicated.

4.2. Appearance and origin of spurious OTU

The aim was to find a relative abundance filtering cutoff for 16S rRNA gene sequencing data from stool samples to remove sequences which are not matching to any of the reference species. These falsely detected sequences are referred to as spurious OTUs. At first, the origin of spurious OTUs was investigated. Next, remaining amounts of spurious OTUs, were compared for different filtering methods. The following results shall be published in manuscript currently under peer-review (Reitmeier et al. 2020).

4.2.1. Origin of spurious OTUs

To determine the origin of spurious OTU, >100,000 datasets were combined from 16S rRNA gene sequencing data available in IMNGS. Considering a total number of 108,881 samples (human n = 46,153 samples; soil n = 29,864 samples; freshwater n = 13,977 samples; mouse n = 10,409 samples; marine n = 8,478 samples), a broad variety of different molecular species were observed appearing in these datasets. This variety suggested that contaminants are occurring within sequencing runs and are dependent on studies/samples sequenced in parallel. There seemed to be no artificial taxa commonly found in amplicon data. The majority of spurious OTU are from phyla Firmicutes, followed by Bacteroides and Proteobacteria. The taxa found are mainly derived from samples of human and mice. However, some spurious OTUs derive from soil or water samples and include mainly bacteria from the *Pseudomonacae* (pyhlum Proteobacteria).

4.2.2. Filtering cutoff for 16S rRNA gene sequencing data of the human gut microbiota

Nine mock-communities (Mock) (**Table 3**) and four gnotobiotic communities (Gnoto) (**Table 4**) were analyzed to determine the highest relative abundance value of a spurious OTU. As well as the number of truly detected species for each community.

The common approach to filter OTU tables is to remove singletons. Singletons are OTUs where only one single read over all samples is assigned. The results of this filtering approach are compared with results from an unfiltered OTU table. It showed that the number of spurious OTUs without filtering was 508 ± 355 OTU (min. 52; max. 1,081) per Mock (10 to 58 species present) and 105 ± 50 OTU (min, 55; max. 215) per Gnoto (4 to 12 species present) (

Figure *8* A). There was no significant difference in the proportion of spurious OTU comparing the two methods (for Mock 50.8 vs. 64.3%, P-value = 0.23; for Gnoto 57.5% vs. 65.7%, P-value = 0.70). It is important to mention that even without any filtering applied, not all species present in each community were detected. On average only 94.9% of the community members are detected. After applying singleton removal, the proportion of positive hits dropped slightly to 92.3% (

Figure *8* A). Nevertheless, the number of spurious OTUs was still high after singleton removal and about 87% of all OTUs were contaminants. However, the cumulative relative abundance for such contaminants is generally low. For instance, spurious OTUs found in the gut community of gnotobiotic mice have a cumulative relative abundance of below 1%, suggesting that spurious OTUs only marginally influence the total community but inflate richness (

Figure *8* B).

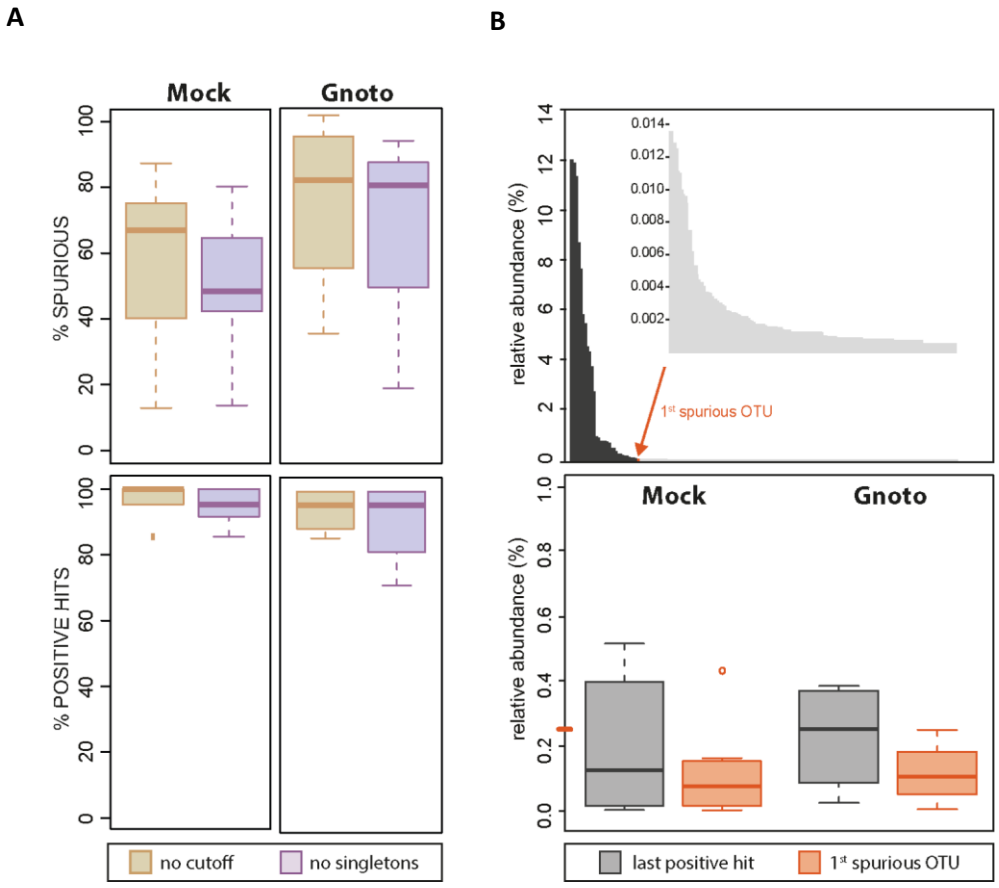


Figure 8 Determination of filtering thresholds using artificial communities of known composition in-vitro (Mock; N = 9 mock communities; n = 21 samples) and in mice (Gnoto; N = 4 different communities; n = 28 samples).

(A) Comparison of various standard filtering cutoffs regarding the percentage of spurious OTUs (i.e., those molecular species not matching sequences of the known species contained in the artificial communities). Lower boxplot shows the corresponding percentages of positive hits retained by the different filtering strategies; positive hits being defined as the reference sequences found in the respective amplicon datasets. **(B)** Example of the relative abundance distribution of total OTUs detected without filtering in the gut of a gnotobiotic mouse. The arrow indicates the position of the first spurious OTUs, all following OTUs being considered as having a high risk of being spurious (light grey bars).

The relative abundance of the spurious OTU with the highest abundance an occurrence below 0.25%. Based on this result, a relative abundance-based filtering cutoff of 0.25% was suggested to remove most OTUs, which are not part of the community (

Figure *9*). Only one outlier, which appeared in an abundance above the recommended threshold, was observed and, thus was not removed.

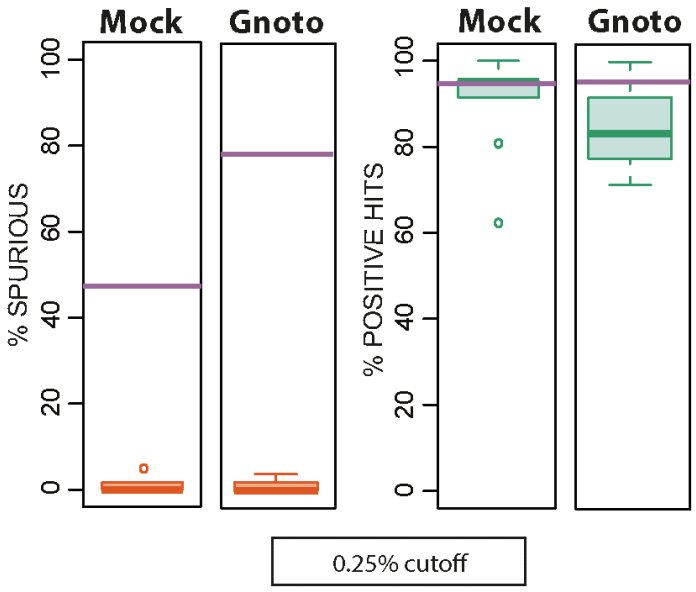


Figure 9 Determination of a 0.25% representative filtering threshold.
Lower boxplot shows the corresponding percentages of positive hits retained by the 0.25% filtering strategies; positive hits being defined as the reference sequences found in the respective amplicon datasets. The purple lines on the y-axis indicate the consensus threshold of singleton removal.

After applying the 0.25% cutoff, the number of spurious OTUs dropped significantly to 4.0% for Mock (compared to 58% for singleton removal; P-value < 0.001) and to 1.0% for Gnoto (compared to 62% for singleton removal; P-value < 0.001). By removing spurious OTUs, apparently only few low abundant but true members of the community are filtered out. However, most of the molecular species are retained without causing any significant differences in the final results compared to singleton removal (87.2% vs. 93.7% for Mock and 82.4 vs. 88.7% for Gnoto; P-value > 0.50). For Mock and Gnoto, a notable drop in the percentage of positive hits (8% - 25%) was observed. One true member of the microbial ecosystem was not detected due to a low relative abundance of this taxa (

Figure *9*). This effect is due to the low number of microbial members (n = 4-12) per community, which does not play a role in natural stool samples.

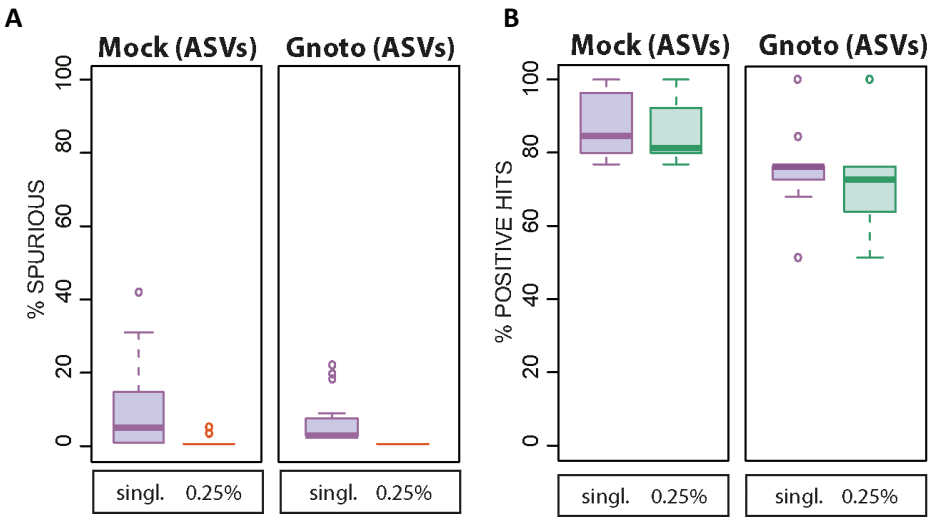


Figure 10 Contaminants in ASV data sets.

(A) Percentage of spurious OTUs and positive hits for the determined filtering cutoff of 0.25% in mock and mouse communities. (**B)** Percentage of spurious OTUs and positive hits in the same reference communities using different filtering thresholds and the DADA2 pipeline for analysis based on Amplicon Sequence Variants (ASVs).

Additionally to the OTU approach, the filtering cutoff of 0.25% was tested for ASVs (Callahan, Sankaran et al. 2016). In total, 42 ± 25 ASVs (min. 16; max. 98) for Mock (10 to 58 target species in theory) and 14 ± 8 ASVs (min. 4; max. 25) for Gnoto (4 to 12 target species in theory) were found. In the ASV approach a marked decrease in spurious taxa was observed compared to OTU clustering, with an average of 8.6% ± 11.8 and 4.4% ± 6.4 spurious sequences after singletons removal for Mock and Gnoto, respectively. However, the DADA2 pipeline used to create the ASVs, cannot be used without any filtering, the default approach removes singletons. Comparing these data with the OTU approach showed comparable results. On average, the spurious ASV with the highest relative amount had a relative abundance of 0.10% ± 0.32 (removing singletons only). Applying the cutoff of 0.25% relative abundance completely removed spurious sequences (except for three samples with outliers), again causing a slight drop in positive hits (

Figure *10*).

4.2.3. Filtering methods influence the results of within and between samples comparisons

To estimate the influence of different filtering strategies on data not generated at our facility, two published studies were analyzed (Flores, Caporaso et al. (2014) and Halfvarson, Brislawn et al. (2017)). The introduced filtering cutoff of 0.25% was compared with filtering based on singleton removal. The results showed significant differences in richness (p-value < 0.001) and Shannon diversity

(p-value < 0.001). Thus, data preprocessing influences the outcome and conclusions drawn out of it. Both studies included here focused on the intra – and inter-individual variation of the human gut microbiota. Therefore, it was possible to compare *alpha*-diversity over time within one individual. One would expect that the variation in richness stays the same, independent of applied filtering method. The inter-quartile range (IQR) of richness calculated for all available time points within one individual showed major differences between filtering methods. Applying singleton removal resulted in a wider spread of richness and the microbiota composition seemed to be more dynamic compared to the 0.25%-cutoff variation, where richness seemed to be comparable between samples within one individual.

Besides the unequal distribution of intra-individual richness, a significantly smaller heterogeneity was observed after applying an abundance-based filtering cutoff (Flores, Caporaso et al. (2014): IQR = 28.0 ± 17.8 vs. 70.6 ± 34.1, p-value < 0.001; Halfvarson, Brislawn et al. (2017): IQR = 17.0 ± 3.2 vs. 49.0 ± 10.4, p-value = 2.5e-13). A similar trend was observed by comparing the results for Shannon effective numbers.

In the study from Flores, Caporaso et al. (2014) and Halfvarson, Brislawn et al. (2017) distances were calculated using unweighted UniFrac distances. While there was a large intra-individual variation after singleton removal, a much greater similarity was observed in 0.25% filtered data, with an average of 0.31 compared to 0.60. This result could be due to the fact that low abundant OTUs were not excluded after singleton removal. Thus, these low abundant OTUs were considered to be important species in the unweighted approach. Applying the generalized UniFrac distance on de-novo picked OTUs showed a similar intra-individual distance for both filtering methods. The average distance for the 0.25% filtering was 0.31 and for the singleton removal 0.23 and, thus reversed compared to the results obtained from the unweighted approach.

In summary, the first part of the results introduced a validated pipeline for sample preparation, processing and sequencing. The use of bioinformatic methods, in particular for filtering 16S rRNA gene sequencing data, were intensively tested. The results obtained from this section were used for future analysis of population cohorts.

4.3. The analysis of the fecal microbiota of population-based cohorts

4.3.1. Microbial composition of the human gut in KORA$_{2013}$

Stool samples of the KORA$_{2013}$ cohort were sequenced using amplicons of the V3V4 region of the 16S rRNA genes. The fecal bacterial ecosystem was globally dominated by the two phyla, Firmicutes and Bacteroides, making up to 91% of the composition (**Figure 11 A**). In total, the bacterial community

compromised 13 phyla with 101 genera. The most diverse phylum was Firmicutes, which could be separated into 11 families and 50 genera. The cumulative relative abundance is 56%. Second to Firmicutes was Bacteroides, which contains four families and 104 genera with a cumulative relative abundance of 35% (**Figure 11 A**). Eighteen genera were present in ≥90% of all individuals. Their mean cumulative relative abundance was 18.67%, which showed that most of the high abundant OTUs was prevalent in nearly all individuals (**Figure 11 B**). The 18 most prevalent OTUs belong to following genera: *Bacteroides, Blautia, Agathobacter, Faecalibacterium, Fusicatenibacter, Clostridium, Agathobaculum, Anaerostipes, Anaerobutyricum,* and *Eubacterium*. Only some of them were found to be part of the 'core microbiome' introduced by other studies, suggesting that there is no global 'core microbiome' shared by all humans (Turnbaugh, Hamady et al. 2009, Huse, Ye et al. 2012). As already shown in previous studies, the gut microbiota composition could be influenced by a variety of factors, such as geographical region, lifestyle, host genetics, and host health, all of which change the composition of the intestinal bacterial ecosystem and increase the inter-individual variation (Qin, Li et al. 2012, Le Chatelier, Nielsen et al. 2013, Falony, Joossens et al. 2016, Zhernakova, Kurilshikov et al. 2016).

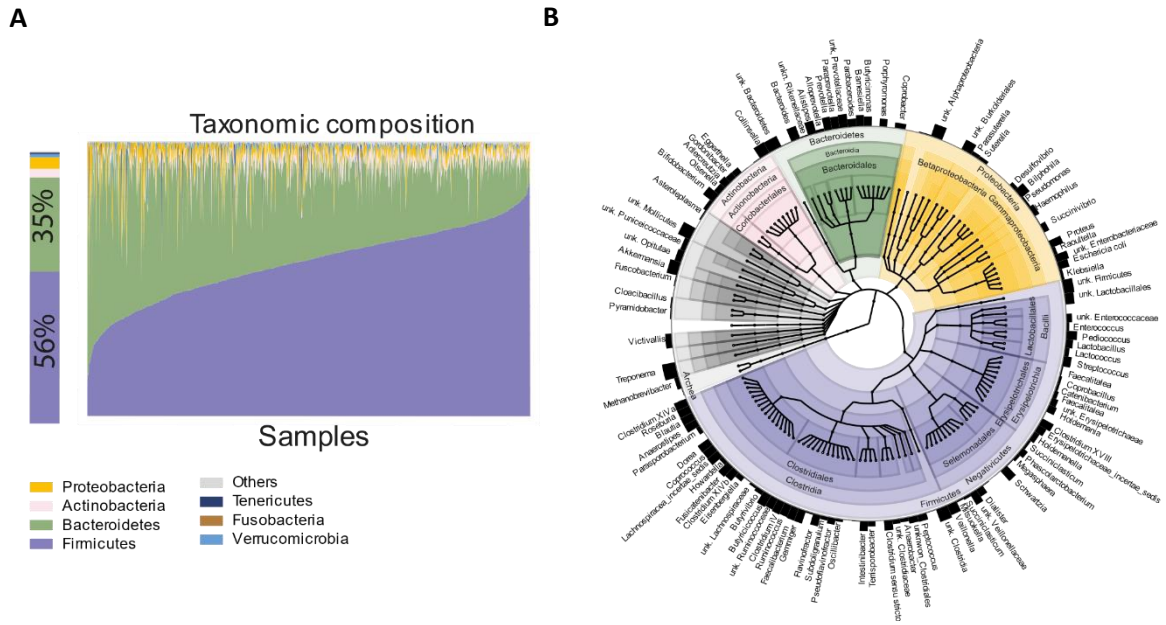


Figure 11 Taxonomic composition of the KORA$_{2013}$ cohort.
(A) Relative abundances of phyla across the whole cohort. Samples are ordered according to increasing relative abundances of Firmicutes. **(B)** Taxonomic tree of the gut microbiota in 1,976 KORA$_{2013}$ subjects. Colors indicate phyla. Taxonomic ranks are from kingdom (center) to genera indicated by the individual branches. Black bars indicate the prevalence of each genus, the name of which are shown if found in > 10% of individuals.

4.3.2. Targeting different variable regions of the 16S rRNA gene

All samples were examined targeting two different variable regions of the 16S rRNA genes, V1V2 and V3V4. Data for V1V2 had been sequenced with higher sequencing depth compared to V3V4, which resulted in an increased number of total OTUs found per sample (**Table 11**).

Table 11 Sequencing depth of the 16S rRNA gene sequencing for different targeted regions and different cohort studies.

	16S rRNA gene sequencing 2013 (V3V4)	16S rRNA gene sequencing 2013 (V1V2)	16S rRNA gene sequencing 2018
Samples wo resequencing	2,186	2,135	900
Average no. reads	87,655	149,209	29,880
Max. reads	211,648	380,092	60,141
Min reads	16	72	7,363
Sum all reads	187,144,068	3,185,262,516	26,892,897
Samples re-sequenced	207	not re-sequenced	not re-sequenced
Average no. reads	145,328		
Max reads	581,380		
Min reads	36,164		
Sum all reads	300,083,040		
Samples incl. resequencing**	2,137	2,135	900
Average no. reads	99,386	149,209	29,880
Max reads	581,380	380,092	60,141
Min reads	36,164	72	7,363
Sum all reads	212,289,588	3,185,262,516	26,892,897
Mean sequence quality score (phred score)	Q_{R1} = 35 Q_{R2} = 36	Q_{R1} = 32 Q_{R2} = 33	Q_{R1} = 36 Q_{R2} = 38
Mean error rate	Q_{R1} = 0.00032 Q_{R2} = 0.00025	Q_{R1} = 0.00063 Q_{R2} = 0.00032	Q_{R1} = 0.00025 Q_{R2} = 0.00016
Number OTUs	2,089	2,383	1,695

**** removing samples with low read counts (< 4,700)**

Comparing the taxonomy targeting the V1V2 and the V3V4 region showed differences in the overall bacterial composition. However, similar to V3V4, the two dominant phyla are Firmicutes (mean rel. abundance 54%) and Bacteroides (mean rel. abundance 31%) also for V1V2. In contrast, Proteobacteria seemed to be more abundant when targeting V1V2.

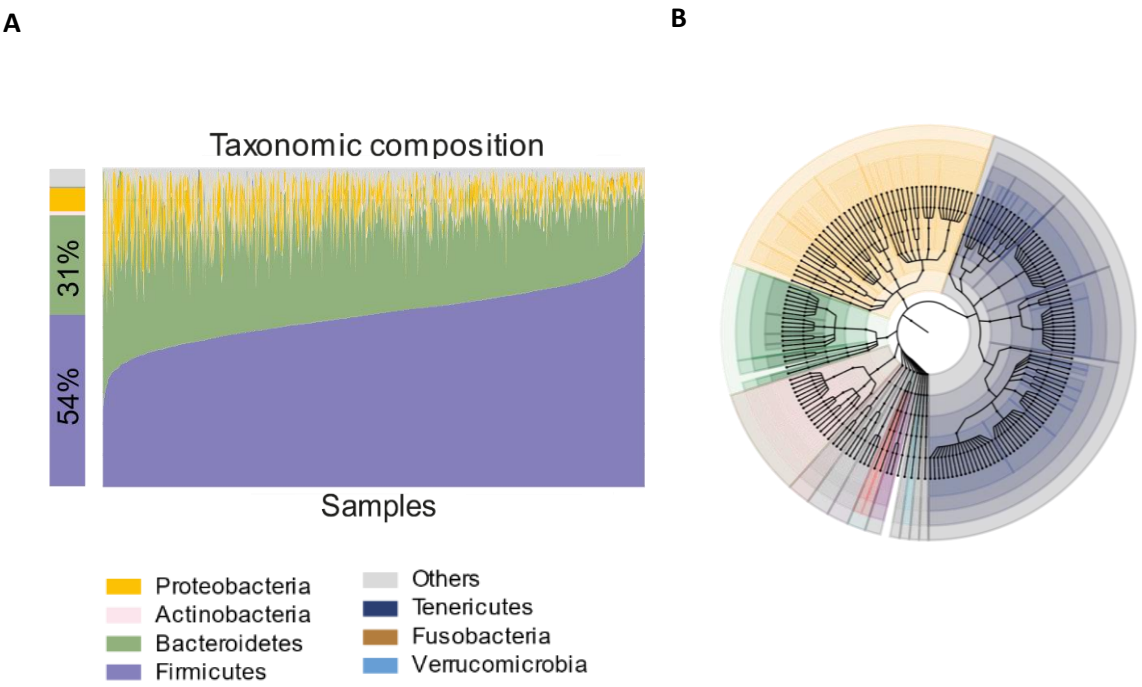


Figure 12 Taxonomic composition of the KORA$_{2013}$ cohort targeting the V1V2 region.
(A) Relative abundances of phyla across the whole cohort. Samples are ordered according to increasing relative abundances of Firmicutes. **(B)** Taxonomic tree of the gut microbiota in 1,976 KORA subjects. Colours indicate phyla. Taxonomic ranks are from kingdom (centre) to genera indicated by the individual branches. Black bars indicate the prevalence of each genus, the name of which are shown if found in > 10% of individuals.

In total, for both target V-regions of the 16S gene the same 90 genera were detected. Out of these, 27 genera were significantly different in their relative abundance between V1V2 and V3V4 (**Figure 13**, adj. p-value ≤ 0.05). Overall, 11 taxa, e.g. *Bifidobacteria***,** were absent in V1V2, but 48 genera were only found when targeting this V-region. Overall, highest differences were observed for the two phyla Actinobacteria and Proteobacteria.

Figure 13 Taxonomic differences between V1V2 and V3V4 V-region of the 16S rRNA gene.

Taxonomic differences targeting the V1V2 or the V3V4 region of the 16S rRNA gene. Taxonomies are grouped according to taxonomic level (phyla level on the top, genus level on the bottom). Same order applies for the presence (black) and absence (white) plot. Colors are groups of taxonomies form the same phyla level. Grouped by taxonomic level, the differences between V3V4 (first row) and V1V2 (second row) in presence/absence (black and white). Taxonomies were ordered according to phyla level and were color-coded respectively.

4.3.3. Intra-individual diversity and composition as reference marker for health

Independent of the targeted V-region, a large inter-individual variation was observed. This variation resulted in a broad distribution of the main phyla Firmicutes and Bacteroides, ranging from 10.62% to 94.73% for Firmicutes (F) and 0.04% to 84.09% for Bacteroidetes (B), respectively. One descriptive marker of the gut microbiota is the F/B ratio, which had provoked a lot of controversy in the past. However, previous studies suggested a decrease of the F/B ratio in patients with obesity (Backhed, Ding et al. 2004, Ley et al. 2005). In $KORA_{2013}$, the average F/B ratio was 1.6 for V3V4 (average F/B ratio V1V2 = 1.7), but with a large variation between individuals **(Figure 14 A**). Further, a large variation was observed for richness and the Shannon effective number of species. With an average richness of 381 ± 77, most individuals fall within the range of 338 and 437 assigned species, yet not following a normal distribution (Shapiro test < 0.05). Based on Shannon effective number, the diversity within an individual compromised on average 188 ± 37 species and the distribution was skewed to the right, and thus not normal distributed among the cohort **(Figure 14 B)**. Again, the comparison of the two targeted V-regions for two samples of one individual showed differences in the *alpha*-diversity without a correlation to richness or Shannon effective number of species (**Figure 14 C).** The diverse *alpha*-diversities observed for either V1V2 or V3V4 within one individual made it difficult to conclude anything about host health based on richness. The latter was shown to be an early and common markers for metabolic health (Stern, Williams et al. 2005, Gillies, Lambert et al. 2008, Wu, Ding et al. 2014, Skyler, Bakris et al. 2017).

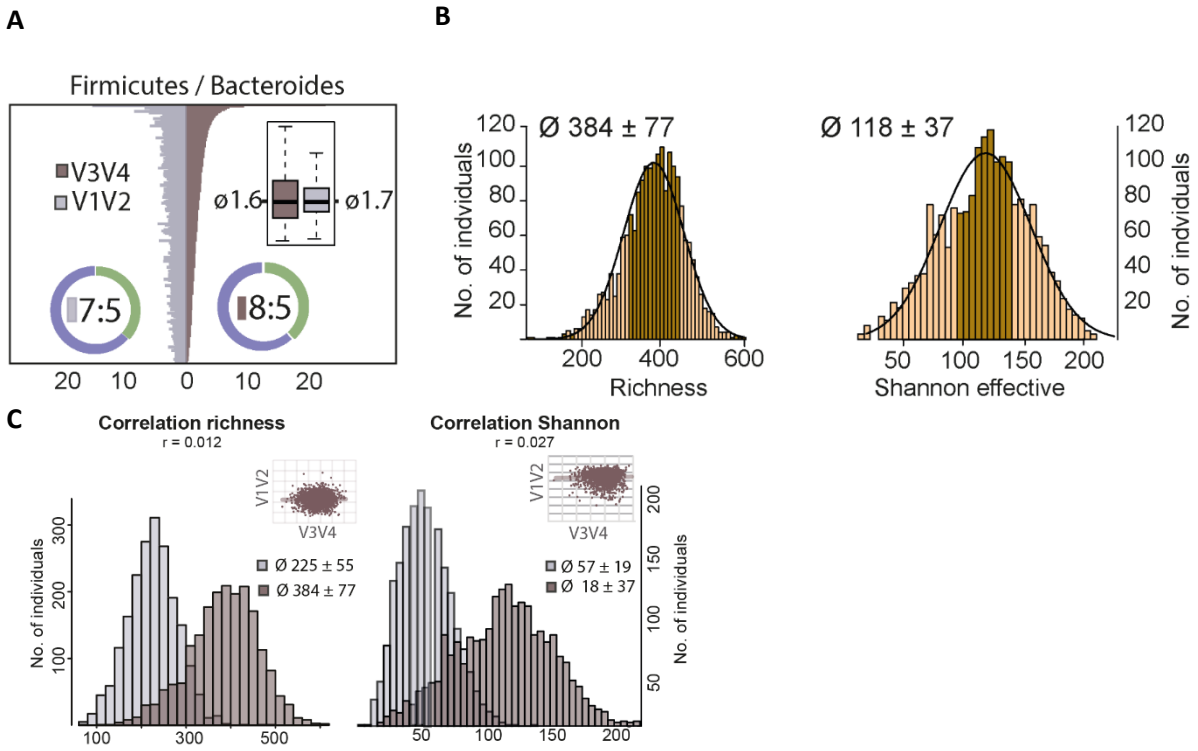


Figure 14 Compositional differences between targeted V-regions.
(A) The ratio of Firmicutes (purple) and Bacteroides (green) is shown for targeting the V1V2 (grey) and the V3V4 (brown) V-region. Barplots are sorted according to decrease in F/B ratio for V3V4. Even with a comparable mean F/B ratio of 1.6 and 1.7 the variation in V3V4 is larger. **(B)** *Alpha*-diversity of the fecal microbiota in $KORA_{2013}$. Richness (left; 384 ± 77) and Shannon effective counts (right; 118 ± 37), which were both not normally distributed across the whole cohort (Shapiro test < 0.05). **(C)** Differences in *alpha*-diversity between V1V2 (grey) and V3V4 (brown) without correlation for one sample between the two V-regions.

4.3.4. Clustering of individuals with similar microbial composition

Individuals of the $KORA_{2013}$ cohort were grouped according to similar gut microbiota composition using on an unsupervised clustering approach. The optimal number of clusters is defined by forming well separated clusters with maximal differences. The applied generalized UniFrac distances between samples was determined by Calinski Harabasz index (Calinski and Harabasz 1974). Based on this index, individuals were assigned to three distinct clusters (**Figure 15**). C2 was the largest cluster (N = 981 subjects) and is significantly dominated by the genus *Ruminococcus*. The second largest cluster was C1 (N = 744 subjects) with a higher relative abundance of *Bacteroides*. The smallest cluster C3 (N = 249 subjects) contained more *Prevotella*. Besides these three genera each cluster was also characterized by a unique higher relative abundance of other genera. Nevertheless, the above mentioned three genera where previously shown to be part of the three main genera in the concept of 'enterotypes' (Arumugam, Raes et al. 2011, Wu, Chen et al. 2011, Costea, Hildebrand et al. 2018). Despite, this terminology should be used carefully since three genera cannot describe the complexity of a complete microbial ecosystem. It is yet unclear, if on a global scale three ´enterotypes´ with the

mentioned three genera prevail. However, the microbial diversity of individuals within C2 was significantly higher compared to C1 and C3, while C3 was the one with the lowest bacterial diversity (**Figure 15 B**). As already mentioned in the previous section, the bacterial diversity of a sample was assumed to be a marker reflecting individuals health. Nevertheless, we did not observe any correlation between a reduced richness and an increased disease status (**Figure 15 A**).

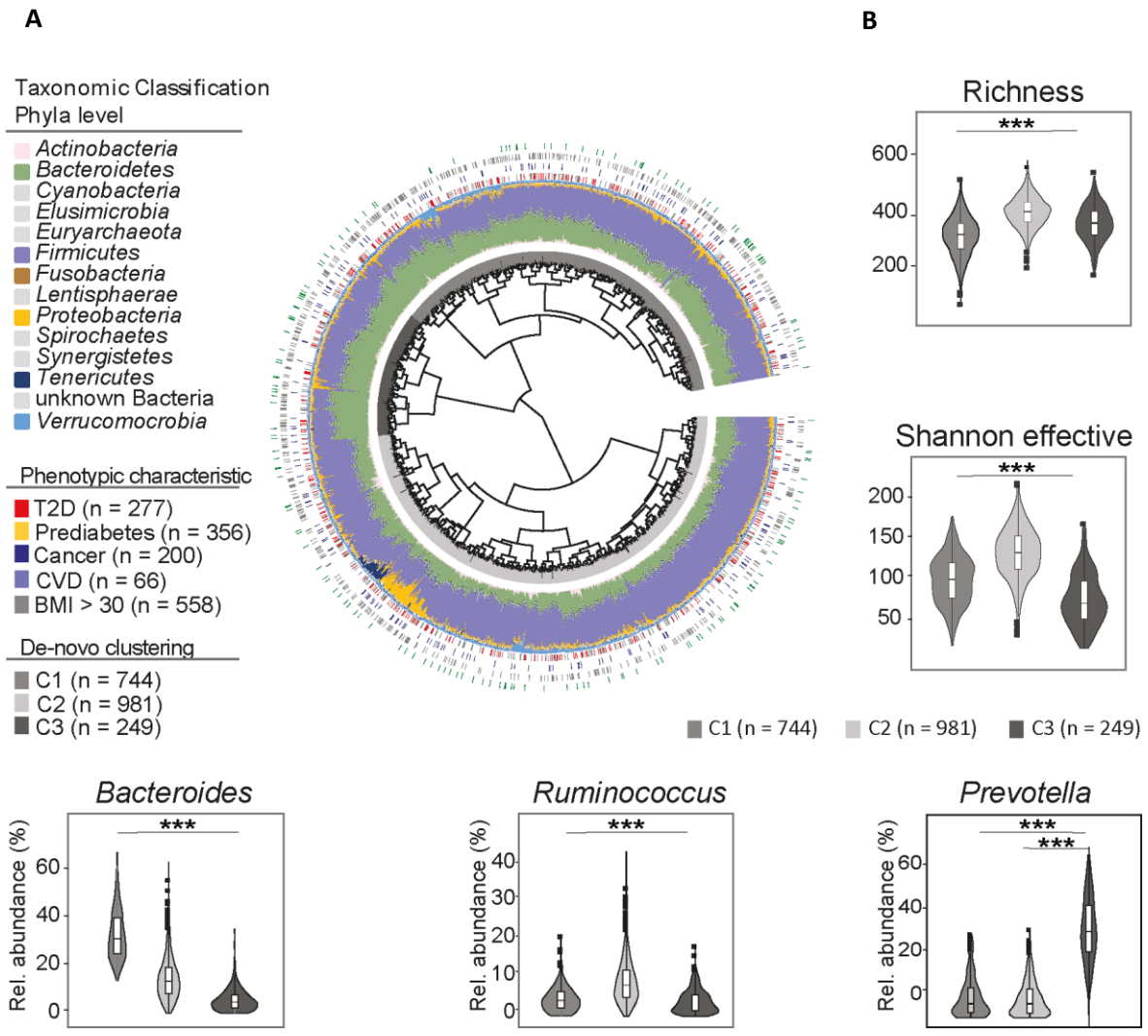


Figure 15 Homogeneous distribution of metabolic health among the cohort.
(A) Beta-diversity of the fecal microbiota in KORA$_{2013}$. The dendrogram shows similarities between microbiota profiles based on generalized UniFrac distances between 1,976 subjects represented by individual branches. Unsupervised hierarchical clustering identified three main clusters of individuals (grey-scale next to branches). Individual taxonomic composition at the phylum level is shown as stacked bar plots around the dendrogram and follows the color code as in panel A. Bars in the outer part of the figure indicate disease status: first ring, Diabetes status (red, T2D; orange, Prediabetes ; no color, nonT2D); second ring, cancer (blue); third ring, obesity (grey, BMI ≥ 30); fourth ring, cardiovascular diseases (green). **(B)** Differences in *alpha*-diversity between the de-novo clusters from panel A. **(C)** Differences in relative abundances of the three genera *Bacteroides*, *Ruminococcus* and *Prevotella* for the three microbiota clusters as in A. P-value < 1·10-5.

In order to understand if dietary habits possibly contribute to the differences in bacterial composition, its association with the three cluster was further analyzed. The general intake of macronutrients showed no significant differences between the clusters. Based on a 24-h food recall, the subjects were grouped in 'non-vegetarian' and 'possible-vegetarian', since C1 was previously characterized to be associated with intake of animal products (Hildebrandt, Hoffmann et al. 2009, Costea, Hildebrand et al. 2018). However, no variable clearly differentiated between these two groups. Thus, we classified those individuals that denoted to have eaten animal products into ´non-vegetarian´ (N = 1,496 individuals). Individuals having eaten no animal products in the last 24 h were assumed to be ´possible-vegetarian´, (N = 463 individuals). However, we could not detect a significant enrichment of individuals harboring C1 for the ´non-vegetarian´ group. Further, we could not find any link between diet and the cluster found. At this stage, the underlying reason for the stratification of individuals into three major clusters in $KORA_{2013}$ remains unexplained. Still, even with limited knowledge, an unsupervised clustering approach should be considered when studying the human gut microbiota composition, since it could be an important factor for an individual's response towards medical or dietary interventions.

4.3.5. The impact of environmental factors on the gut microbiota composition

Previous studies have shown that factors, such as lifestyle, health, physiology, medication and genetics are responsible for differences in the bacterial composition of the gut (Qin, Li et al. 2012, Le Chatelier, Nielsen et al. 2013, Falony, Joossens et al. 2016, Pedersen, Gudmundsdottir et al. 2016, Zhernakova, Kurilshikov et al. 2016, He, Wu et al. 2018). Thus, we wanted to understand which factors cause the large inter-individual variation observed in $KORA_{2013}$. Based on a multivariate permutational analysis of the cohort's similarity, we determined several variables that were significantly associated with the microbial diversity and calculated their contribution towards explained variation. In total, 9.1% of the bacterial composition could be explained by the mentioned factors, of which 'physiology' explained most of it (4.72%), followed by 'lifestyle' (2.30%), 'medication' (1%), and 'health' (1%). Thus, the gut microbiota seems to be more influenced by difference in lifestyle and physiology than on health (**Figure 16 A**). In addition, some of the variation is explained by environmental factors which were not directly connected to the human gut e.g., geographical region. Results of a larger Chinese population-based study already showed that there is a strong influence due to the place of residence, explaining around 9% of gut bacteria variation (He, Wu et al. 2018). In contrast, the $KORA_{2013}$ cohort was regionally restricted to a single city and its outskirts in Southern Germany, Augsburg. Still, a significant association to 'geographical region´ was observed. Overall, 0.9% of the microbial composition was explained by the circumstance if an individual lives within or outside the city (**Figure 16 B**). This resulted in a small

shift in phyla distribution with a decreased abundance of Firmicutes and a slight decrease in Bacteroides in rural areas.

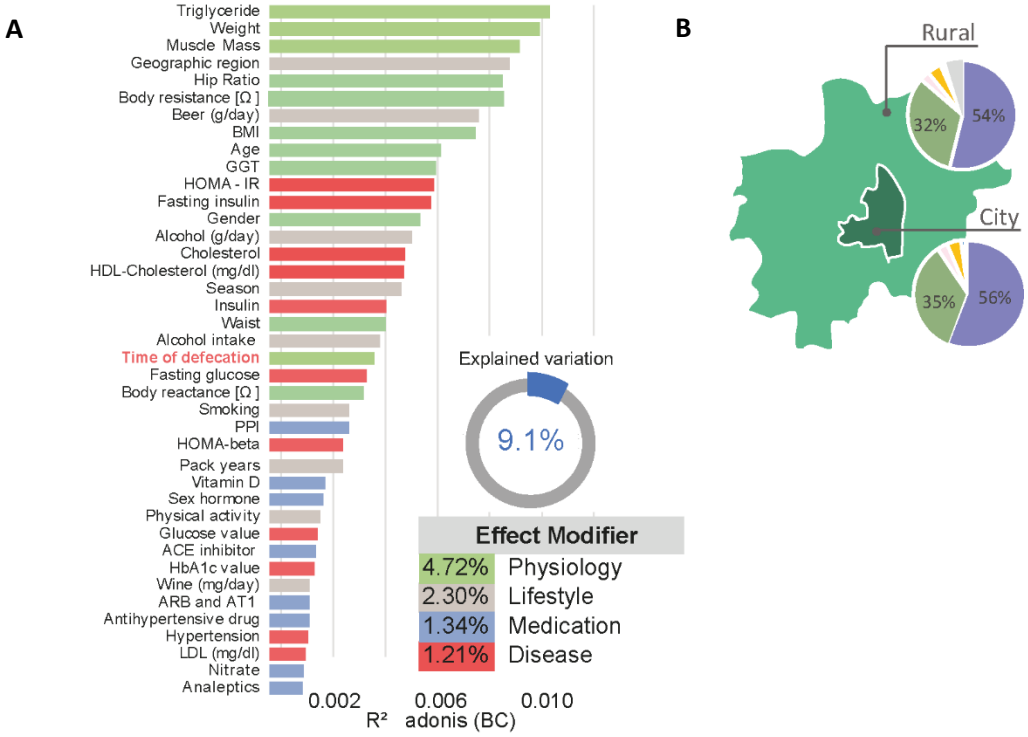


Figure 16 Influence of environmental factors in KORA$_{2013}$
(A) Explained variations in fecal microbiota composition by covariates. All variables shown had a significant influence (P-value ≤ 0.05) displayed as proportions of explained variations based on R^2 calculated by multivariate analysis of Bray-Curtis dissimilarity. **(B)** Geographical map of the city of Augsburg and its rural area. Subjects are grouped according to their place of residence. The pie charts show taxonomic distributions at the phylum level in individuals living outside (rural) or in the city.

Yet, one factor turned out to be important, but has been overlooked so far. The time of defecation significantly contributes to observed differences in the overall microbiota composition. Based on this finding, we started to study the diurnal rhythmicity of the gut microbiota in 1,943 individuals to better understand, how the observed microbial composition is influenced by sampling time (see below **Chapter 4.4**).

4.3.6. The FoCus cohort

In addition to the KORA cohort, we also analyzed the FoCus cohort (Food Chain Plus) from the Northern part of Germany. The recruitment for FoCus started in 2013 and included the collection of biological material and acquisition of phenotypical characteristics from 1,529 individuals. From the stool samples, 16S rRNA gene sequencing was conducted, targeting the V3V4 region. Isolated DNA

from FoCus was provided by Andre Franke's group from the Institute of Clinical Molecular Biology at Christian-Albrechts-University of Kiel. For 1,488 samples, sufficient high quality reads were obtained (i.e., excluding samples with low quality reads and low overall number of reads). Of note, results from the cohort are published by Relling, Akcay et al. (2018).

4.3.6.1. Taxonomic description of the fecal microbiota from the FoCus cohort

Taxonomic analysis of bacteria contained in the stool samples showed a similar distribution of the two main phyla Firmicutes (51.28%) and Bacteroides (36.85%) as already seen in the $KORA_{2013}$ cohort, which corroborated a large inter-individual variation between all samples.

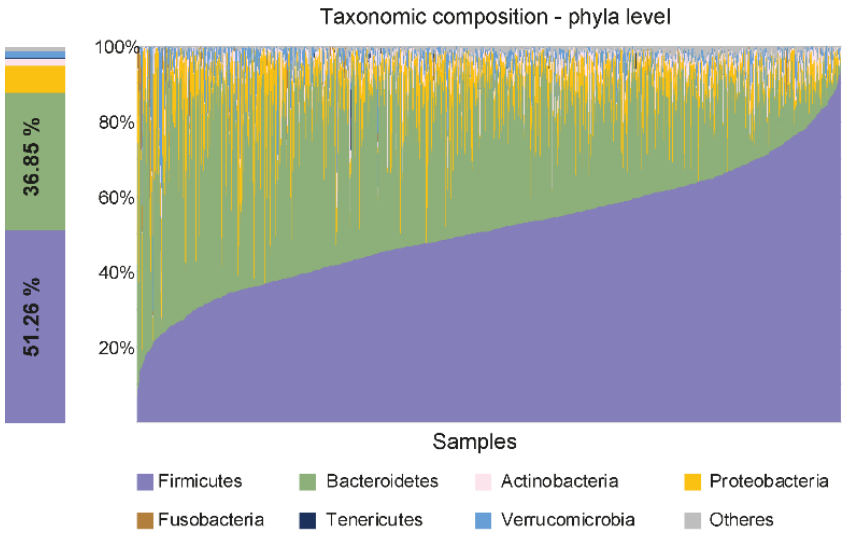


Figure 17 Taxonomic composition of the FoCus cohort based on phyla level.
Relative abundance values are order according to an increase in Firmicutes. The stacked barplot on the left shows the mean relative abundance.

The number of observed OTUs and the diversity within one sample were comparable to the results obtained from $KORA_{2013}$. For instance, the number of observed OTUs and diversity was not normally distributed (Shapiro test p-value < 0.0001). However, the overall diversity was higher in FoCus compared to $KORA_{2013}$, suggesting that there are differences between Northern and Southern Germany.

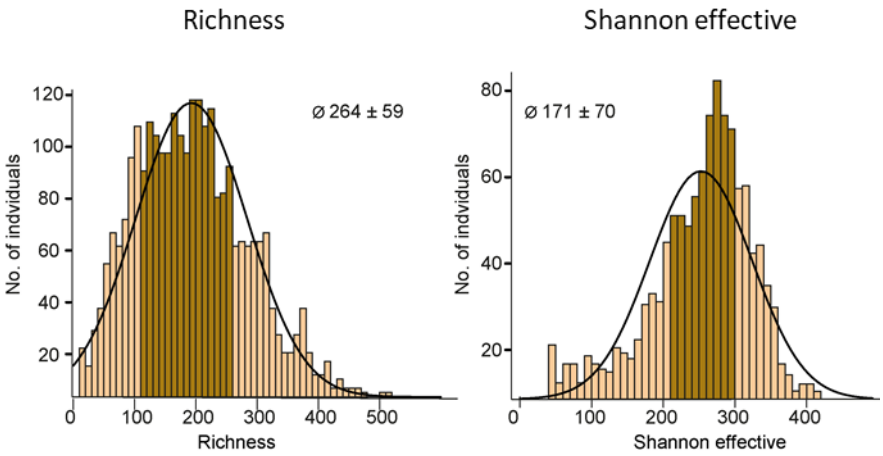


Figure 18 *Alpha*-diversity in FoCus cohort.
Left histogram shows the distribution of richness. Light colors bars are the 25%-quartile and the 75%-quartile. X-axis shows the number of observed OTUs and the y-axis refers to the number of observed individuals.

A phylogenetic tree, based on generalized UniFrac distances between individuals' microbiota, showed some differences in the taxonomic distribution compared to the $KORA_{2013}$ cohort. Performing an unsupervised clustering also generated three distinct clusters as seen in the $KORA_{2013}$ cohort. Each of these cluster was dominated by either *Prevotella*, *Bacteroidetes* or *Ruminococcus* and thus highlights again the presence of microbiota driven cluster formation. While in $KORA_{2013}$ the richness was highest in C2 (associated with *Prevotella*), in FoCus the richness was highest in C3 (associated with *Rumincoccus*). A similar trend was observed for Shannon effective numbers. Next, differences in the number of assigned individuals within one cluster were found between $KORA_{2013}$ and FoCus. As for $KORA_{2013}$, cluster C2 associated with *Prevotella* was the largest with N = 786 individuals, while the other two cluster contained more similar numbers (C1, N = 390 subjects; C2, N = 300 subjects).

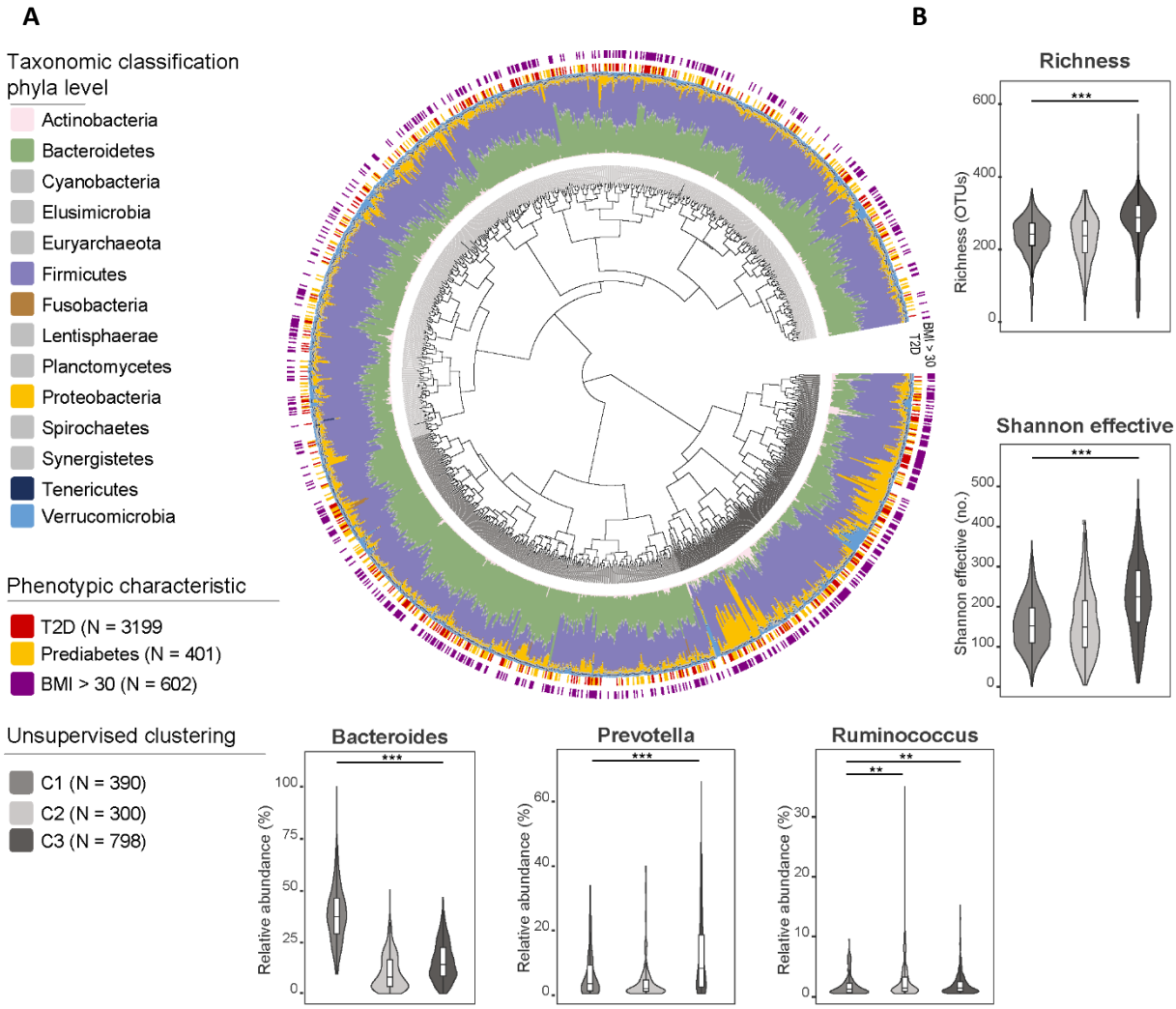


Figure 19 Homogeneous distribution of metabolic health among the FoCus cohort.
(A) Beta-diversity of the fecal microbiota in FoCus. The dendrogram shows similarities between microbiota profiles based on generalized UniFrac distances between 1,488 subjects represented by individual branches. Unsupervised hierarchical clustering identified three main clusters of individuals (grey-scale next to branches). Individual taxonomic composition at the phylum level is shown as stacked bar plots around the dendrogram and follows the color code as in panel A. Bars in the outer part of the figure indicate disease status: first ring, Diabetes status (red, T2D; grey, Prediabetes; no color, nonT2D); second ring, obesity (purple, BMI ≥ 30). **(B)** Differences in alpha-diversity between the de-novo clusters from panel A. Differences in relative abundances of the three genera Bacteroides, Rumincoccus and Prevotella for the three microbiota clusters as in A. P-value < 1·10-4.

4.3.6.2. The impact of factors influencing the bacterial composition of the gut

In the FoCus cohort, 26 out of 56 co-variables (e.g., lifestyle, physiology, medication, and disease) showed a significant association with the observed microbial composition, explaining 5.38% of the differences in bacterial composition (**Figure 20**). Of those, 'disease' with the variables 'hypertension' and 'medication' explained 2%. 'Lifestyle' and 'physiology' contributed equally to the total microbial variation. The most important variables was found to be associated with 'lifestyle' were 'BMI' and 'type of recruitment' (i.e., individuals were either selected by the clinic or based on resident registration). As already observed in $KORA_{2013}$, the time of sampling significantly contributed to the

total bacterial variation of the human gut microbiota in FoCus. This observation corroborated the hypothesis that considering sampling time is an important factor for the analysis of the gut microbiota diversity.

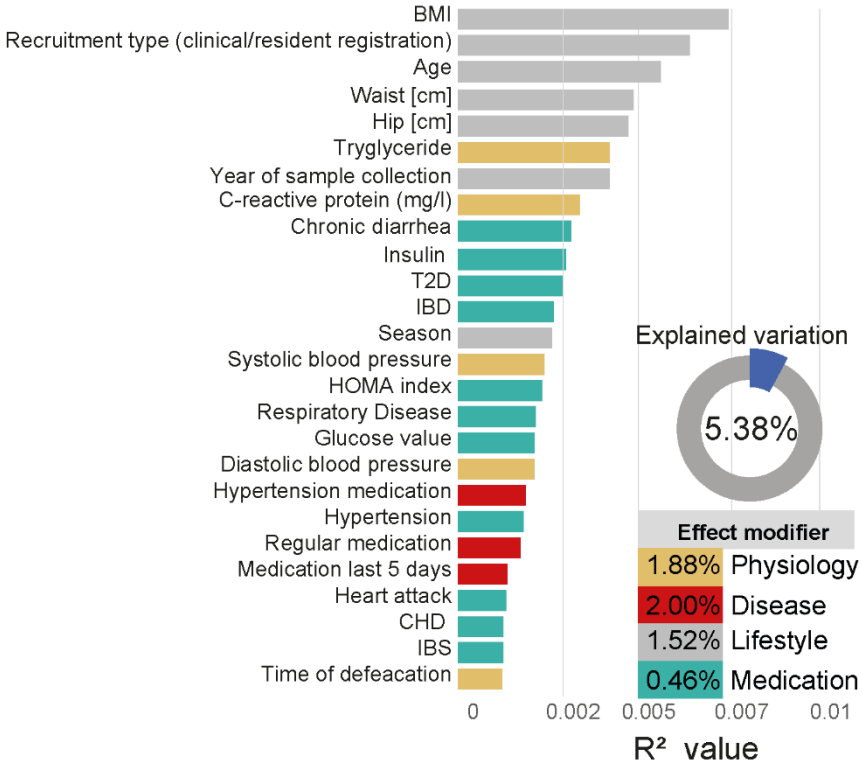


Figure 20 Influence of environmental factors in the FoCus cohort.
Explained variations in fecal microbiota composition by covariates. All shown variables have a significant influence (p-value ≤ 0.05) displayed as proportions of explained variations based on R^2 calculated by multivariate analysis of Bray-Curtis dissimilarity.

Of note, the abundance of Tenericutes was reduced compared to $KORA_{2013}$ (**Figure 19 A**). In $KORA_{2013}$, a cluster of individuals with higher abundances of Tenericutes was identified (**Figure 15 A**), but not in the FoCus cohort. Overall, both population-based cohort studies showed similar results in the taxonomic distribution and a large inter-individual diversity with minor differences in the relative abundance of the two phyla Tenericutes and Verrucomicrobia. Most co-variables significantly associated with differences in the gut microbiota composition overlapped between the two cohort studies. The reduced global impact of effect modifiers in FoCus compared to $KORA_{2013}$ can be explained by a lower number of co-variables available assessed for FoCus. Nevertheless, the hypothesis that sampling time (e.g., defecation or circadian rhythm) influenced the gut microbiota was strengthened.

4.3.7. The bacterial composition and influencing factors in the *enable* cohort

The *enable* cohort is a smaller cohort with only healthy participants recruited in Freising, Gemany. The analysis of 499 individuals' stool samples showed a large variation in each personal gut microbiota (**Figure 21**). The mean cumulative relative abundance of the two major phyla Firmicutes and Bacteroidetes contributed approximately 90% to the overall composition. The abundance of Firmicutes ranged from 27.41% to 84.39% over the whole cohort. For Bacteroidetes a minimum relative abundance of 7.58% and a maximum of 69.80% was found (**Figure 21**).

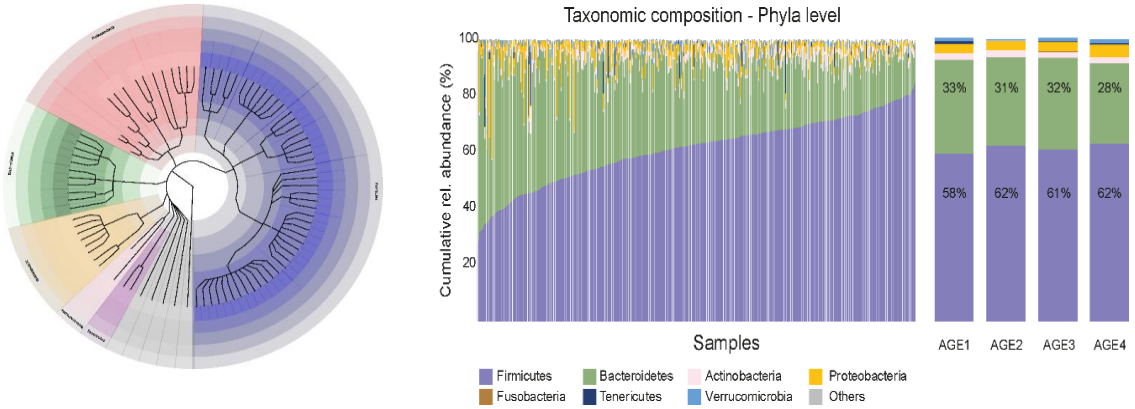


Figure 21 Taxonomic distribution on phyla level.
Taxonomic tree of the gut microbiota in 499 *enable* subjects. Colours indicate different phyla. Taxonomic ranks are from kingdom (centre) to genera indicated by the individual branches. Black bars indicate the prevalence of each genus, the name of which are shown if found in > 10% of individuals. Right diagram shows the cumulative relative abundance (y-axis) for the whole cohort plotted per sample (x-axis) ordered according to increase in Firmicutes. The four stacked barplots show the mean relative abundance of phyla grouped by age.

The large variation between samples was also reflected in richness, where on average 385 ± 85 OTUs were observed per individual. The bacterial diversity was 99 ± 31. Stratifying the cohort according to age showed no significant differences in phyla distribution between the four age-bins formed, even though, AGE1 had less Firmicutes and more Bacteroides compared to the other groups (**Figure 21 B**). Even though we did not observe obvious differences in taxonomic composition between age groups, age was among the most significant variables explaining 0.6% inter-individual variation (**Figure 22**). Nevertheless, 16% the observed taxonomic diversity of the *enable* cohort could be explained by co-variables (**Figure 22**). Most of these variables were related to blood-derived characteristics, e.g., blood cell count, hemoglobin values or disease-marker of the blood (e.g., ferritin index, uric acid) or hormones detected in blood (e.g., triiodothyronine, thyroxin). In total, 'physiology' explained nearly 11% of the variation, while 3.89% are explained by 'disease' and 3.85% by variables such as 'diet', 'occupation or sleeping pattern'. Since the cohort recruited healthy individuals only (i.e., no apparent illness), 'medication' was no significant co-variable. Of note, despite persons with metabolic health

syndromes and other severe diseases had been excluded from the cohort, some significant variables related to disease biomarkers, such as thyroxin (thyroid), uric acid (kidney), and insulin (Type 2 Diabetes) were found to contribute the observed taxonomic variation. Comparing the $KORA_{2013}$ effect modifiers with the *enable* effect modifiers showed some overlaps. Noteworthy, 'time of defecation' appeared again as a significant co-variable. Additionally, the dichotomous variable 'wake-up type' (i.e., if someone gets up with or without an alarm clock) showed some influence; probably because it is also associated with the circadian rhythm.

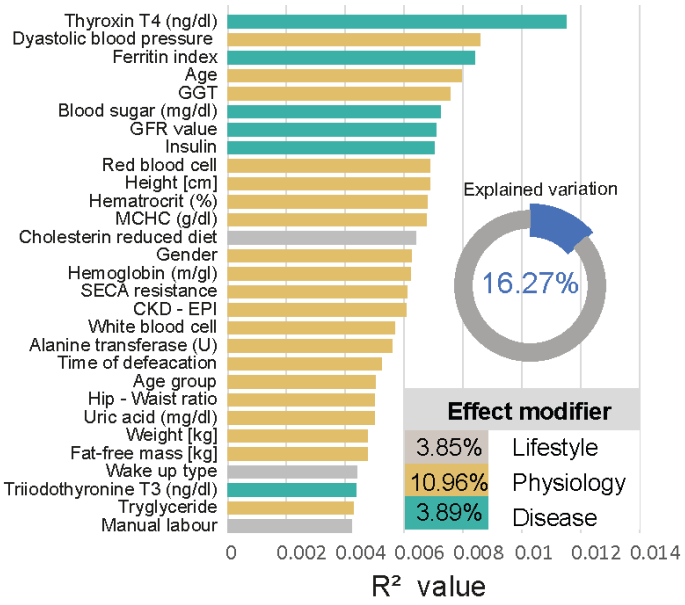


Figure 22 Explained variations in fecal microbiota composition by co-variates.
All variables shown had a significant influence (p-value ≤ 0.05) displayed as proportions of explained variations based on R^2 calculated by multivariate analysis of Bray-Curtis dissimilarity

As seen in the previously described cohorts, unsupervised clustering of generalized UniFrac distances for microbial diversity showed three distinct clusters. Each cluster is dominated by either *Bacteroides* (C1), *Ruminococcus* (C2) or *Prevotella* (C3) (**Figure 23 A**). Tenericutes (dark blue), Proteobacteria (yellow), and Verrucomicrobia (light blue) had higher abundances in C2, which could not be detected in the other two clusters and was only present in a few individuals there (**Figure 23**). In *enable*, richness and bacterial diversity were lower in C1 compared to the other two clusters (C1 mean richness = 319 ± 61, C1 mean Shannon effective number = 76 ± 21). However, in *enable*, a significant imbalance in the distribution of age groups among the cluster was observed. Age was higher in C2 (mean age = 54 ± 24), which was also the cluster with the highest bacterial diversity observed (mean Shannon effective number = 115 ± 26). C1 is dominated by younger participants, there is a higher proportions of individuals in AGE3 and AGE4 in C2 and C3. Thus, age seems to have an impact on microbial composition for healthy individuals (**Table 12**).

Table 12 Microbial associated clusters and their main descriptor found in a healthy population-based cohort.

	C1	C2	C3	P-value
No. of individuals	127	294	72	
Male (%)	49% (N = 63)	47% (N = 139)	63% (N = 46)	0.040
BMI	26.53 ± 4.68	25.71 ± 4.20	26.61 ± 4.89	0.035
Obesity				0.094
BMI < 30	78% (N =98)	83% (N = 53)	74% (N = 245)	
BMI > 30	22% (N = 29)	17% (N = 19)	26% (N = 49)	
Age	49 ± 18	56 ± 20	57 ± 21	0.0057
Age groups				0.0012
AGE1	12% (N = 15)	6% (N = 18)	15% (N = 11)	
AGE2	22% (N = 28)	18% (N = 52)	16% (N = 12)	
AGE3	48% (N = 61)	39% (N = 114)	36% (N = 26)	
AGE4	18% (N = 23)	37% (N = 110)	33% (N = 23)	
Richness	319 ± 61	416 ± 66	417 ± 62	0.008
Shannon effective	76 ± 21	115 ± 26	92 ± 29	0.006

Alpha-diversity between age groups was significantly different. Similar to results from previously published studies, which observed an age related shift in diversity (Yatsunenko et al. 2012, Buford 2017, Stewart et al. 2018). However, a reduction in richness and Shannon diversity observed in children or elderly (both are extreme age groups), could not be confirmed in the *enable* cohort. Nevertheless, the *enable* cohort showed a continuous increase in richness correlating with age, also for the elderly. Some genera showed significant differences in their relative abundance with increasing age, but only with few overlaps to other studies. Most of these taxa e.g., *Fecalibacterium prausnitzii, Roseburia* or *Dorea*, had been linked to metabolic health but were not found to be age associated.

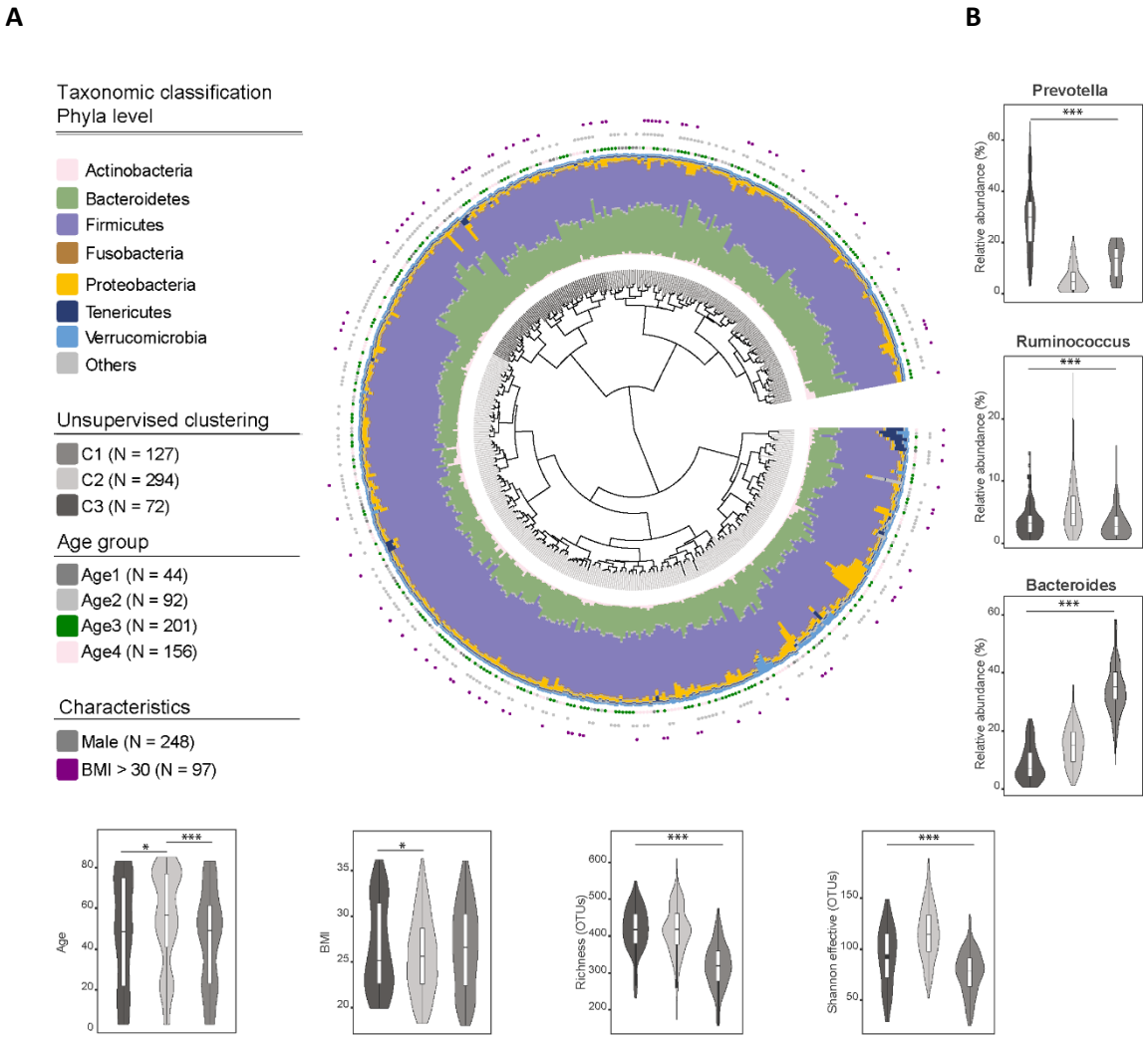


Figure 23 Sample tree of the cross-sectional cohort grouped according to age and BMI.
(A) *Beta*-diversity of the faecal microbiota in the cohort. The dendrogram shows similarities between microbiota profiles based on generalized UniFrac distances between all subjects represented by individual branches. Unsupervised hierarchical clustering identified three main clusters of individuals (grey-scale next to branches). Individual taxonomic composition at the phylum level is shown as stacked bar plots around the dendrogram and follows the colour code as in panel A. Bars in the outer part of the figure indicate age group, distribution of gender and BMI. **(B)** Differences in relative abundances of the three genera *Bacteroides*, *Ruminococcus* and *Prevotella* for the three microbiota clusters as in A. P-value < 1·10-5. **(C)** Differences in age and BMI as well as in alpha-diversity between the de-novo clusters from panel A.

4.3.8. Longitudinal studies

Several studies already showed that there is strong inter-individual variation of the gut microbiota composition (He, Wu et al. 2018). Only very few studies have looked at changed of the observed bacterial diversity overtime within one individual. For instance, Halfvarson, Brislawn et al. (2017) have shown that disease was an impact on intra-individual variation while Flores, Caporaso et al. (2014) also showed that medication and environmental changes affect the composition. Since some individuals within $KORA_{2013}$ have been resampled, prospective data from $KORA_{2018}$ was used for multi-timepoint comparison. Additionally, four sampling timepoint of a subset from the *enable* cohort as well as a longitudinal dataset from two individuals over three years were analyzed.

4.3.8.1. Prospective $KORA_{2018}$ cohort

Part of the $KORA_{2013}$ cohort was invited again after 5 years. Inclusion criteria for the prospective data collection was based on age (49-86 years) and in total 900 subjects were re-analyzed. Thus, the prospective data should reflect the development of the human gut microbiota over a 5-year period and its association with age and health. Finally, 699 subjects provided samples in year 2013 and year 2018, resulting in 1,398 samples examined here (**Figure 24**).

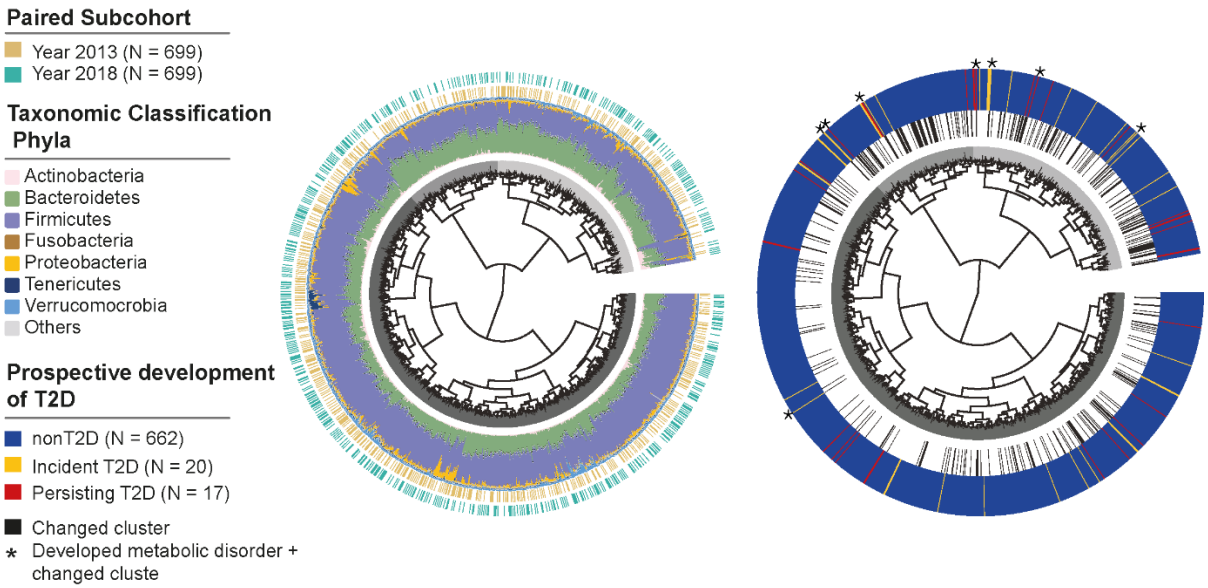


Figure 24 Phylogenetic tree of the KORA paired subcohort.
Beta-diversity of the faecal microbiota in paired subcohort. The tree shows similarities between microbiota profiles based on generalized UniFrac distances between 1,398 subjects represented by individual branches. Unsupervised hierarchical clustering identified three main clusters of individuals (grey-scale next to branches). Individual taxonomic composition at the phylum level is shown as stacked bar plots around the dendrogram and follows the colour code as in panel. There is sampling-year associated clustering (inner brown bars = year 2013, outer cyan bars = year 2018). Right tree shows the distribution of the development of Type 2 Diabetes within 5 years. Individuals which changed the cluster between 2013 and 2018 are marked with black bars. The outer circle shows nonT2D cases (blue), incident T2D cases (orange) and persisting T2D cases (red). An asterix indicates those individuals which changed diabetes status and microbial cluster.

To identify possible cohort bias (i.e., systematic differences between both sampling time points), we looked at the *alpha*-diversity and microbial distribution of both time points, but no differences in richness and Shannon effective count of species were detected. The phylogenetic tree showed no clustering concerning the collection year. This verifies that results obtained were not influenced by any cohort bias. As before, the phylogenetic tree built on the observed bacterial diversity could be divided in three distinct clusters, each with a higher relative abundance in either *Prevotella*, *Rumincoccus* or *Bacteroidetes*.

In total, about half of the individuals changed their previous cluster (change-yes = 56.1% vs. change-no = 43.9%). The change in bacterial composition was not due to an increased age (p-value = 0.98) or BMI (p-value = 0.22). A tendency of increased weight is associated with a change in cluster. In total, 36 individuals became obese over the 5 years but only half of them changed their cluster. The correlation of obesity and cluster change was not significant (p-value = 0.09). In order to determine similarities between individuals that changed cluster, a linear mixed effect model was implemented. The baseline relative abundance values of OTUs from $KORA_{2013}$ were considered as 'random effect' and the dichotomous variable of 'changing cluster' was considered as fixed effect. The analysis determined two genera, *Gemmiger* (P-value = 0.027) and *Roseburia* (P-value = 0.014), having significant differences in relative abundances between individuals which changed their cluster compared to individuals not

changing their cluster. Both genera were depleted in the dynamic group and with an even lower relative abundances already in $KORA_{2013}$ compared to $KORA_{2018}$. Considering BMI as an additional random-effect value showed similar results. Thus, an increased number of *Gemmiger* and *Roseburia* could hint towards a more stable gut ecosystem. These two genera are butyrate producers with anti-inflammatory properties and are depleted in individuals with metabolic disorders (e.g. obesity, Type 2 Diabetes; Qin, Li et al. 2012, Karlsson, Tremaroli et al. 2013). However, in $KORA_{2013}$ and $KORA_{2018}$ 'obese' versus 'non-obese' showed no differences concerning cluster change in association with these two genera.

4.3.8.2. Stability of the gut microbiota composition in a prospective subset of the *enable* cohort

Participants (N = 100) from the *enable* cohort were asked to provide three additional stool samples, spaced about 12 weeks apart (**Figure 25**). The composition of the microbiota in the *enable* cohort showed minor changes over time within one individual. Thus, the observed diversity of the gut microbiota composition was analyzed for each subject individually. The number of observed species per individual varied over time without sharing a similar trend. It was not possible to generate clusters of subjects showing similar changes in richness (**Supplementary figure 1**, **Figure 26**). Overall, for most samples individual's richness was lower at baseline suggesting an increase in richness with an increase in age. Nevertheless, the correlation between richness and age was not significant. The variation between individuals was high, shown by a high degree of branching in the cluster tree, i.e., most individuals were presented by one branch only. Branch length was generally long, reflecting intra-individual variation (**Supplementary figure 1**).

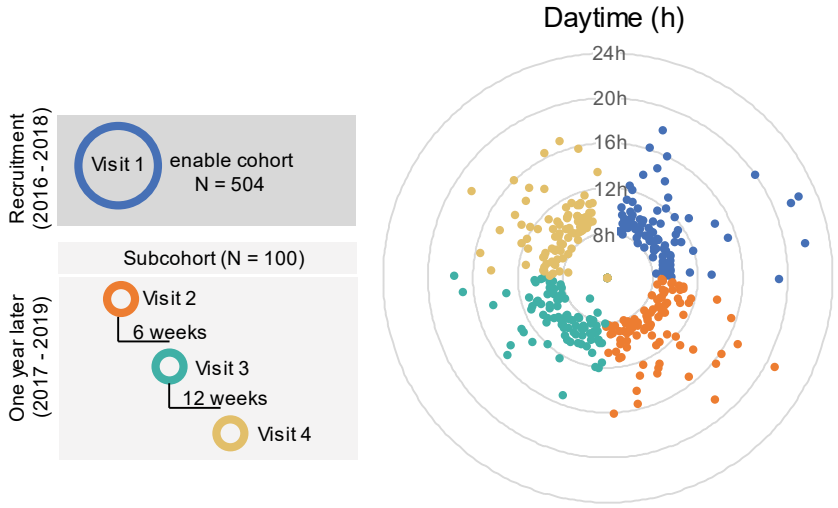


Figure 25 Flowchart for sampling period and daytime of the *enable* subcohort (N = 100 subjects).
Colors are indicating the sampling phase (blue = first visit, baseline; orange = second visit, green = third visit, yellow = fourth visit). The position of the colored dots is referring to sampling time in hours in a 24-h day. For each individual, there are up to four dots (one circle representing one sample).

The change of the gut microbiota composition in consecutive samples of one subject in *enable* were analyzed using generalized UniFrac distances (**Figure 26 A**). The mean distance observed was 0.29 ± 0.07, which is in concordance with intra-individual distances observed in other studies (Flores, Caporaso et al. 2014, Halfvarson, Brislawn et al. 2017). Some individuals displayed a greater within-sample distance of up to 0.66, suggesting that the composition of the gut microbiota was less stable (**Figure 26 B**). A negative correlation between caloric intake and increased distance (p-value = 0.05, Pearson r = -0.21) was observed. However, no correlation could be determined when analyzing macronutrients and no correlation with BMI and the *alpha*-diversity values was found.

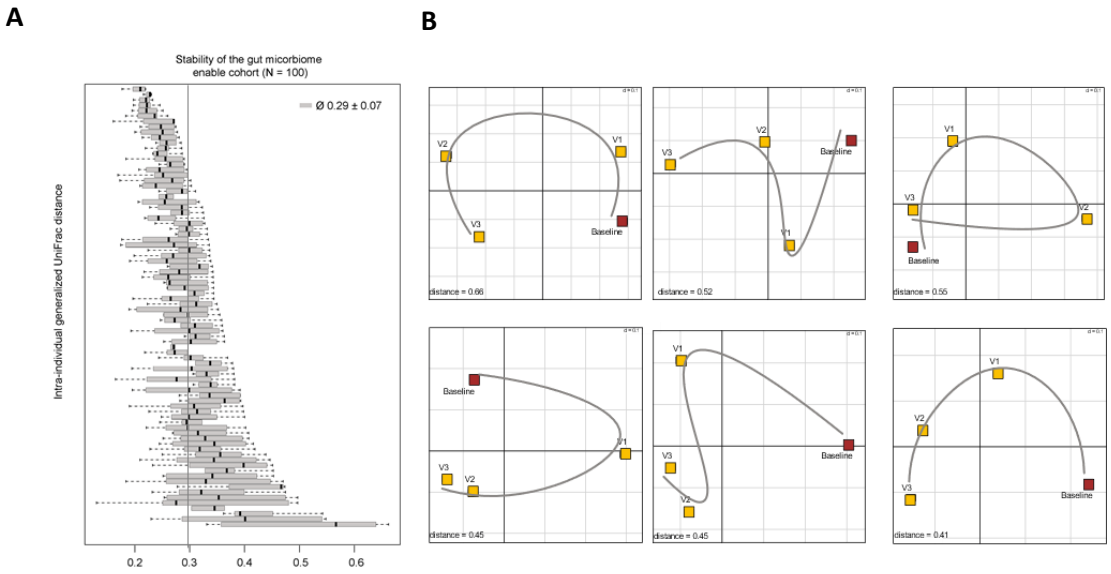


Figure 26 Stability of the gut microbiota composition of the *enable* subcohort.
Boxplot shows the intra-individual generalized UniFrac distance. Subjects are sorted according to maximum distance values. Grey line indicates the mean distance observed in the data set. Examples of six individuals with a high intra-individual distance. Dynamic of the samples microbial composition is indicated with a line (grey) starting at baseline sample from the recruitment phase (red) along the visits (Visit1 (V1), Visit2 (V2) and Visit3 (V3); yellow). Mean generalized UniFrac distances are shown on the left corner in each panel.

4.3.8.3. Longitudinal sampling over three years for two individuals

A time-series analysis of two individuals (S1 and S2) followed over several years was performed in order to better understand the dynamic of the human gut microbiota within a single person (**Figure 27**). There were no similar changes in the relative abundance values on phyla level seen in S1 and S2 (**Figure 28**, **Figure 29**, **Figure 30**). While the relative abundance of bacteria of S1 seemed to follow no clear pattern and varied without any correlation to weight, Bristol score, richness or sampling time (**Figure 28**), a progressive loss of Bacteroidetes in S1 was observed (**Figure 29**). This loss was also visible on genus level, while an increase of other, unknown genera.

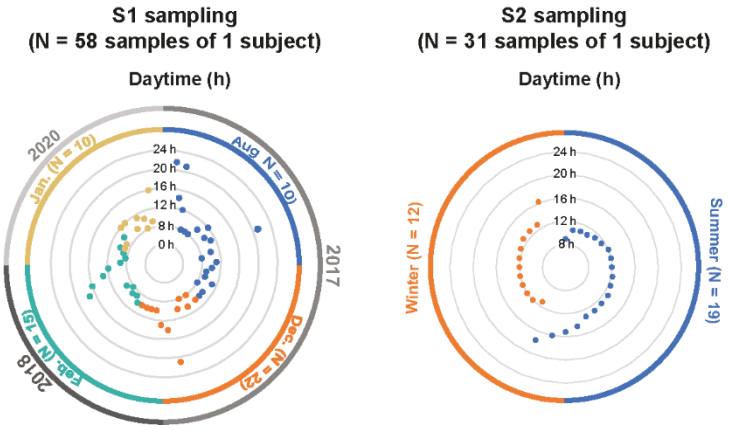


Figure 27 Flowchart of time series longitudinal samples of two Individuals (S1 and S2).

Colors are indicating the sampling periods. For S1, four sampling periods were conducted (blue = first period, baseline; orange = second period; green = third period; yellow = fourth period). For S2, only two sampling periods were conducted (blue = first period, baseline; orange = second period). The position of each dot refers to sampling time in hours for a 24-h daytime.

The *alpha*-diversity of S1 correlated with weekdays, which could be related to stress, sleep and diet or a combination thereof. The number of observed microbial species was reduced during the week, while it increased on the weekend (Pearson correlation = 0.35, p-value = 0.012). The effect was even stronger for Shannon effective numbers, ranging from the lowest diversity of 52.29 observed on Mondays to the highest Shannon effective number seen on Thursdays (Shannon effective counts = 127.13). Adjusting the used regression model for stool weight, under the assumption that it influences bacterial composition and diversity, still showed a significant difference between weekday and diversity (p-value < 0.001). There was no difference in *alpha*-diversity after stratifying the data according to sampling periods (T1, August 2017; T2 December 2018; T3 February 2019 and T4 February 2020), but a significant difference in the overall bacterial composition between the sampling periods was observed (p-value = 0.001). The bacterial composition was not influenced by medication (i.e., some probiotics taken) nor a medical intervention (i.e., colonoscopy). Comparing four sampling time points of S1, the mean distance was comparable between all four sampling phases (generalized UniFrac distance: T1 = 0.24 ± 0.06; T2 = 0.27 ± 0.07; T3 = 0.23 ± 0.06; and T4 = 0.24 ± 0.04, respectively), as well as between different sampling periods (generalized UniFrac distance: 0.27 ± 0.07).

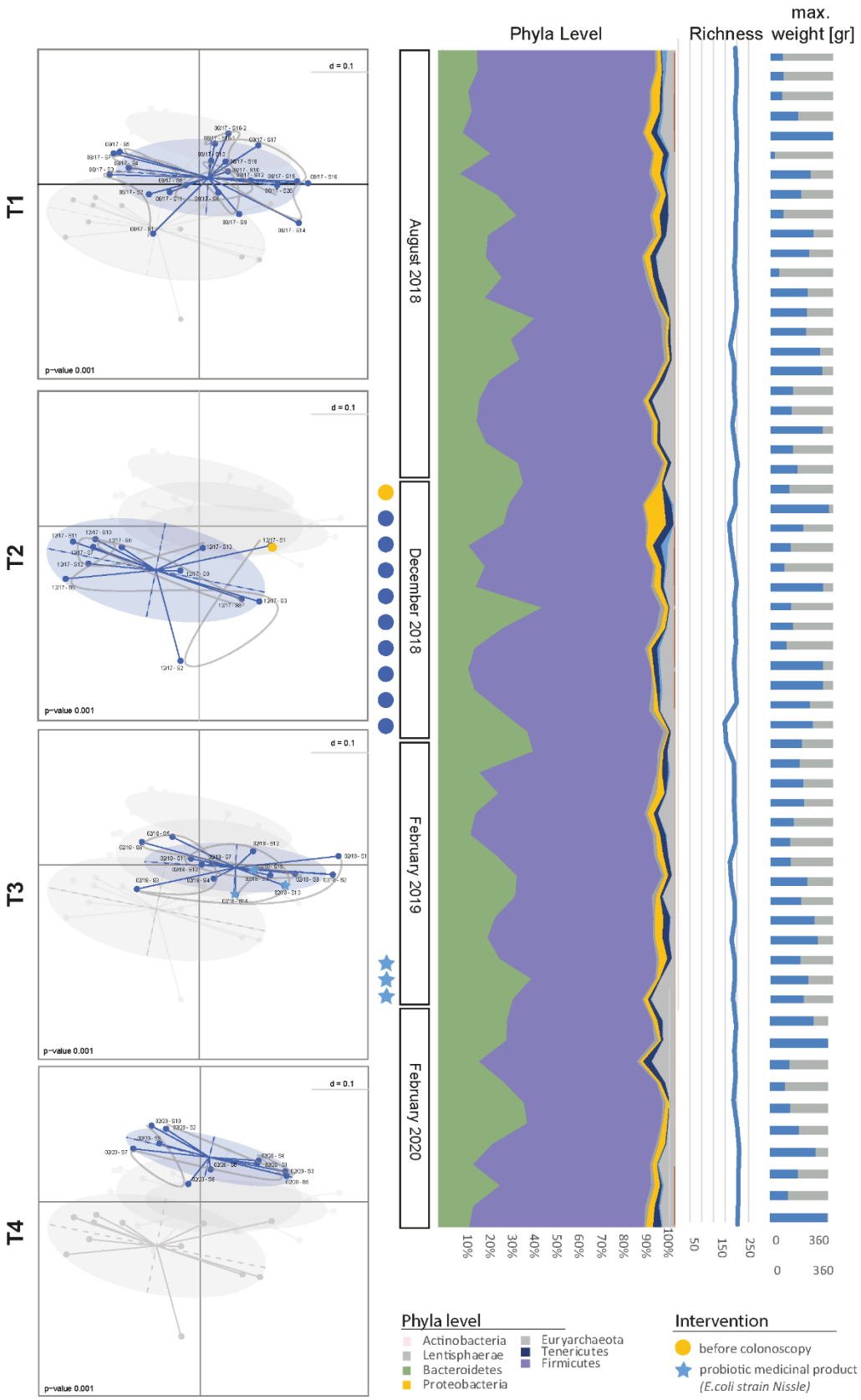

T1
T2
T3
T4
d = 0.1
p-value 0.001
August 2018
December 2018
February 2019
February 2020
Phyla Level
Richness
max. weight [gr]
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
50
150
250
0
360
0
360
Phyla level
Actinobacteria
Lentisphaerae
Bacteroidetes
Proteobacteria
Euryarchaeota
Tenericutes
Firmicutes
Intervention
before colonoscopy
probiotic medicinal product
(E.coli strain Nissle)

Figure 28 Dynamic of the gut microbiota composition in S1 over time.
(A) Beta-diversity between sampling periods. The three nMDS plots are showing the four different samples phases (T1, August 2017; T2 December 2018; T3 February 2019 and T4 February 2020). Each dot represents one collected sample, labelled according to time. The intra-individual dynamic of S1 is shown by a connected line between time points. In the second sampling phase, a sample before colonoscopy is marked by a yellow circle. In the third sampling phase, the intake of a probiotic product is indicated by a blue star. **(B)** Compositional changes over time within S1 illustrated on phyla level. Different sampling phases are indicated on the left. Interventions (colonoscopy) or probiotic intake are highlighted by coloured dots or stars as in (A). Next to the phyla distribution, the fluctuation of richness is shown (blue line) and the stool weight [g] is shown by a barplot, whereby a fully blue bar refers to the max. weight of 360 g and the other bars are coloured proportional to this maximum weight.

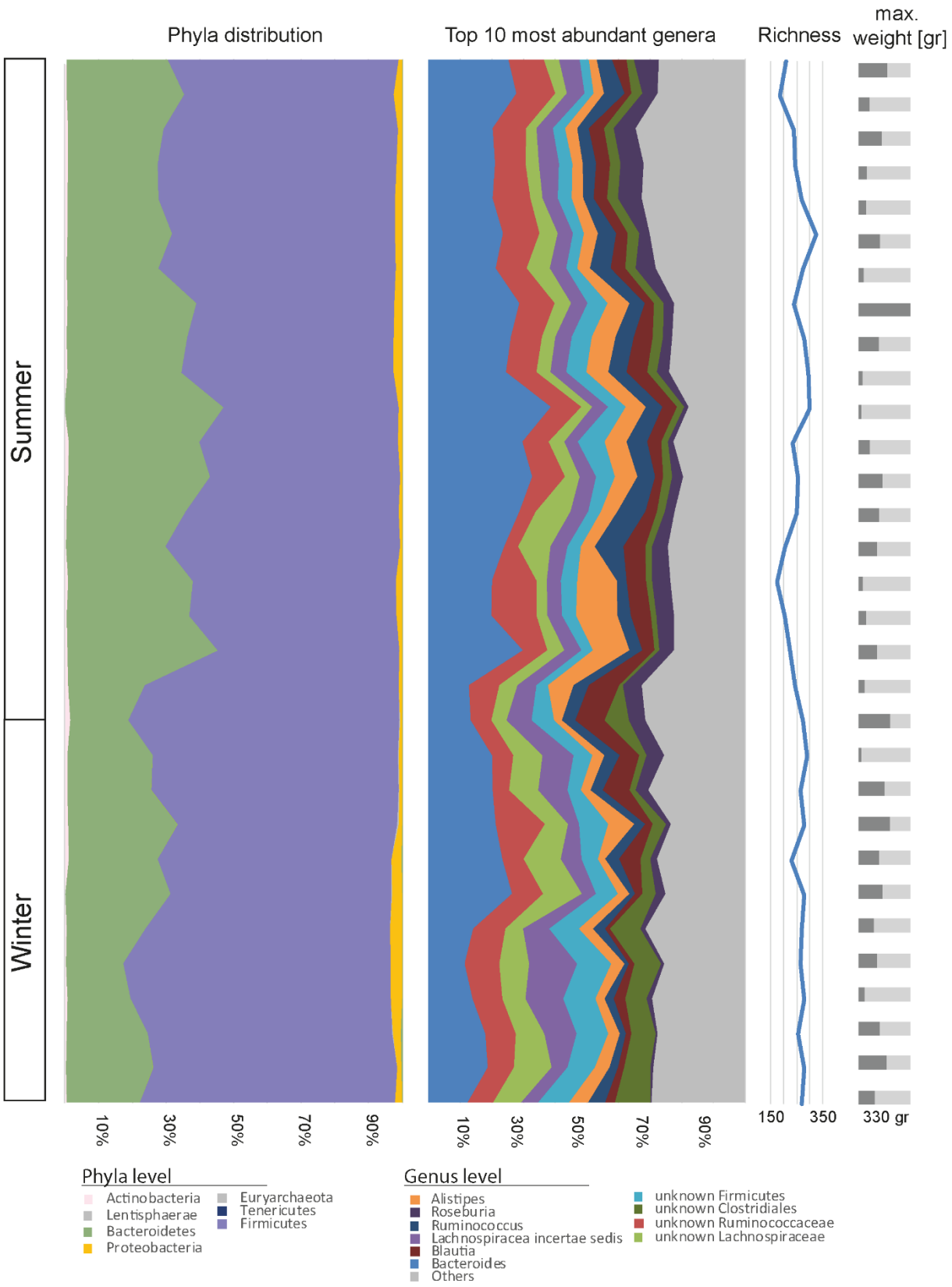


Figure 29 Dynamic of the gut microbiota composition in S2 over time.
Compositional changes over time within one individual illustrated by phyla (left) and genus (right) level. Different sampling phases are indicated by the labelled month on the left side. Next to the genus distribution the fluctuation of richness is shown by a line and the stool weight [gr] is shown by a barplot whereby a filled bar refers to the max. weight of 330 g and the others are filled proportional to the maximum weight.

The comparison of differences in relative abundance values on taxonomic level showed that between the four sampling phases two groups of taxa followed the same trend over the three years (**Figure 30**). On family level, *Ruminococcaceae* and *Clostridiales* were reduced at T1 and T3 but increased in T2 and T4. Members of both bacterial families are butyrate producers and are assumed to have a protective effect towards metabolic health, cardiovascular disease and cancer (Qin, Li et al. 2012, Canfora, Jocken et al. 2015). A similar trend was seen on genus level in *Alistipes* and *Butyricicoccus*, both promoting gut health (Thingholm, Ruhlemann et al. 2019). A different pattern was observed in the four genera *Roseburia*, *Blautia*, *Feacalibacterium* and *Lachnospiracea incertae sedis*, showing an increased abundance at T1 and T3 but a decreased in T2 and T4, respectively. Low abundances of these genera have been linked to increased BMI and metabolic dysbiosis (Lippert et al. 2017). These bacteria produce short-chain fatty acids, especially butyrate, affecting colonic motility, immune functions and anti-inflammatory properties (Reichardt et al. 2014, Ananthakrishnan, Luo et al. 2017). Other studies indicated a correlation with depression and the increased relative abundance of *Roseburia*, *Blautia*, *Feacalibacterium* and *Lachnospiracea incertae sedis*, suggesting that their abundances could be affected by stress (Franzosa et al. 2019). Some bacteria appeared in a very low prevalence or were even absent at certain time points. For instance, *Phascolarctobacterium* and *Acidaminococcaceae* are absent at T4. Both genera are associated with the production of the SCFA propionate from succinate (Reichardt, Duncan et al. 2014). Succinate, however, is related to Crohn's disease and IBD (Connors et al. 2018). Further, succinate is produced by *Veillonellaceae*, which was only present at T4. Its presence fits to the absence of *Phascolarctobacterium* and *Acidaminococcaceae*, since *Veillonellaceae* are negatively correlated with the other two phyla (Fernandez-Veledo and Vendrell 2019). Single sample analysis of the ten most abundance genera in S1 showed a strong increase of a member within the *Clostridiales*, e.g., during the winter sampling phase in 2017. The growth of this bacteria could be related to the colonoscopy, which took place a few days earlier. The data further revealed that there was an association of the relative abundance of this *Clostridiales* member with stool weight. The weight after colonoscopy was reduced, which possibly influenced bacterial composition. The increased abundance of the *Clostridiales* member could be due to decreased abundance of otherwise dominating genera due to the intervention. Thus, the decrease of "normal" gut inhabitants could result in an increased abundance of the low abundant genera *Clostridiales*. But since the abundances are relative, the number of *Clostridiales* would be overestimated even if no outgrowth occurs at all, but the otherwise dominating members only decrease. Further, each a member of the family *Ruminococcacea* and *Prevotella* also showed strong dynamic changes over time. A possible explanation for the fluctuation for these strains could be diet, since both were shown to be influenced by this (Arumugam, Raes et al. 2011, Wu, Chen et al. 2011, Koren, Knights et al. 2013).

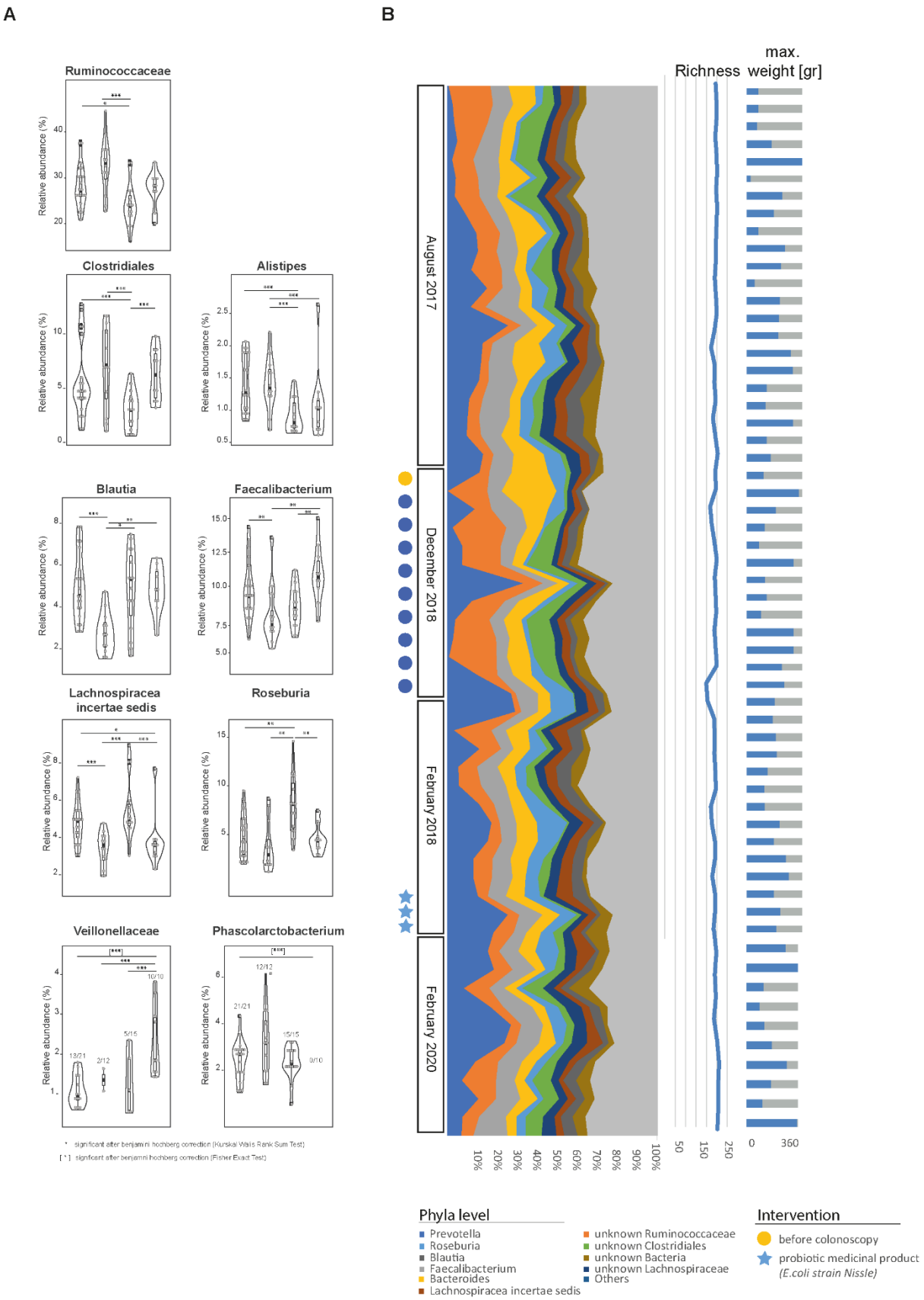

A
B
Ruminococcaceae
Clostridiales
Alistipes
Blautia
Faecalibacterium
Lachnospiracea incertae sedis
Roseburia
Veillonellaceae
Phascolarctobacterium
Relative abundance (%)
* significant after benjamini hochberg correction (Kurskal Wallis Rank Sum Test)
[*] significant after benjamini hochberg correction (Fisher Exact Test)
August 2017
December 2018
February 2018
February 2020
Richness
max. weight [gr]
0
360
50
150
250
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Phyla level
Prevotella
Roseburia
Blautia
Faecalibacterium
Bacteroides
Lachnospiracea incertae sedis
unknown Ruminococcaceae
unknown Clostridiales
unknown Bacteria
unknown Lachnospiraceae
Others
Intervention
before colonoscopy
probiotic medicinal product
(E.coli strain Nissle)

Figure 30 Intra-individual variation on genus level for S1.
(A) Significant differences in the relative abundance (adj. p-value, Kruskal Wallis Rank Sum test) or prevalence (adj. p-value, Fisher Exact Test) on taxonomic level between four different sampling phases. Groups are chronologically arranged. **(B)** Compositional changes over time within one individual illustrated by ten most abundant genera. Different sampling phases are indicated by the labelled month on the left. Interventions or medications are highlighted by colored dots or stars. Next to the phyla distribution, the fluctuation of richness is shown by a blue line and the stool weight [g] is shown by a barplot whereby a filled bar refers to the max. weight of 360 g and the others are filled proportional to the maximum weight.

Some factors, for example weekday or season, only had a short-term impact on the bacterial diversity. Other factors, like aging or disease might change the bacterial composition profoundly. However, a number of factors are still unknown, and it is not yet clear what influences the taxonomic composition of the gut microbiota and to which degree the overall bacterial composition can be explained by environmental factors and phenotypical characteristics. The integration of S1 and S2 in the $KORA_{2013}$ cohort showed, that even though the longitudinal samples were clustering together, the distances observed within one sample were in some cases higher than they were between some samples. Suggesting that one individual could appear as diverse as two different subjects. Inversely suggesting that the microbiota composition of two different subjects could be more related to each other than one individual. In a data set of 4,181 individuals, 34 pairs of subjects were present with a generalized UniFrac distance ≤ 0.27 (i.e., the threshold was determined from the calculated mean distance for the samples of S1), which indicates a relatedness we normally could attribute to a single individual. Nevertheless, only one sample pair of S1 was falling below the similarity distance of samples over different time points for on individual. Thus, there is no “reference distance”, which can be used evaluating the similarity between and within individuals. This finding again highlights the constraints we face in comparing between different studies.

4.3.9. Integrating datasets: $KORA_{2013}$, FoCus and *enable* cohort

The individual analysis of each cohort provided information about the bacterial composition, the diversity, similarities and dissimilarities within one cohort. All cohorts showed a large variation between individuals, which could currently not be linked to host health or physiological differences. To better understand the differences between the cohorts, a data set containing all three cohort studies was assembled. This will allow to determine the specifics and commonalities between the three cohorts. The generated data set included 4,469 samples and 9,640 OTUs. The latter number was reduced to 167 OTUs after applying thresholds of 0.1% minimum relative abundance and a 10% prevalence. This great reduction in OTU number for such moderate thresholds shows the sparsity of 16S rRNA gene sequencing data. Unsupervised de-novo clustering of the combined data determined three distinct cluster, comparable to the cluster shown in the individual cohort analysis before (**Figure**

31, **Figure 15**, **Figure 19**, **Figure 23**). We observed an increased accumulation of individuals from the FoCus cohort in C3. Thus, samples of this cohort seemed to be closer related to each other than to samples from the other two cohorts. The regional differences between Northern and Southern Germany could be an explanation of the observed differences. In contrast, samples from $KORA_{2013}$ and *enable* were spread out evenly over the whole tree (**Figure 31**). Analysis of *alpha*-diversity provided no further insights.

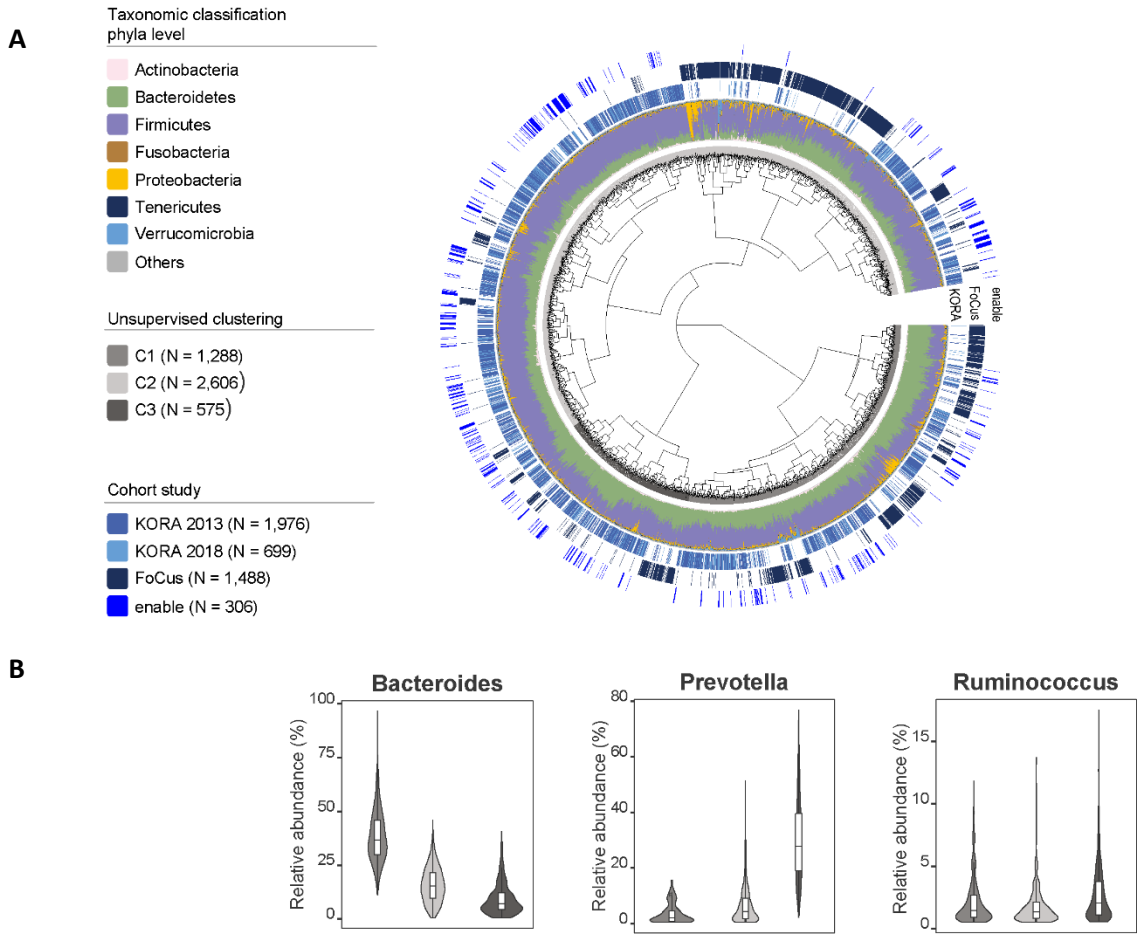


Figure 31 Sample tree of the integrative dataset ($KORA_{2013}$, enable and FoCus cohorts).
(A) Beta-diversity of the fecal microbiota in the cohort. The dendrogram shows similarities between microbiota profiles based on generalized UniFrac distances between all subjects represented by individual branches. Unsupervised hierarchical clustering identified three main clusters of individuals (grey-scale next to branches). Individual taxonomic composition at the phylum level is shown as stacked bar plots around the dendrogram. Bars in the outer part of the figure indicate cohort affiliation. The first label corresponds to the $KORA_{2013}$ cohort from year 2013 and year 2018, the second label marks individuals of the FoCus cohort and the outer label corresponds to the enable cohort. **(B)** Differences in relative abundances of the three genera *Bacteroides*, *Ruminococcus* and *Prevotella* for the three microbiota clusters as in A. p-value < 1·10-5.

4.4. Diurnal rhythmicity in fecal microbiota composition

The analysis of three large cohort studies as well as longitudinal time series analysis showed that defecation time, i.e., when the stool samples were collected, significantly explains some of the bacterial compositional. Based on this finding and a previously described circadian rhythm of the gut microbiota (Liang et al. 2015, Thaiss et al. 2016), we considered diurnal rhythmicity as important factor in our studies.

4.4.1. Description of the diurnal rhythmicity in the KORA$_{2013}$ cohort

Individuals of the KORA$_{2013}$ cohort were binned according to their denoted sampling time point and analyzed regarding possible fluctuations of microbial diversity throughout a 24h-day. Bins were predefined as intervals between 0 h, 3 h, 5 h, 7 h, 9 h, 11 h, 13 h, 15 h, 17 h, 19 h, 21 h, and 23 h. The distribution of subjects per sampling timepoint was sufficient to generate reliable results.

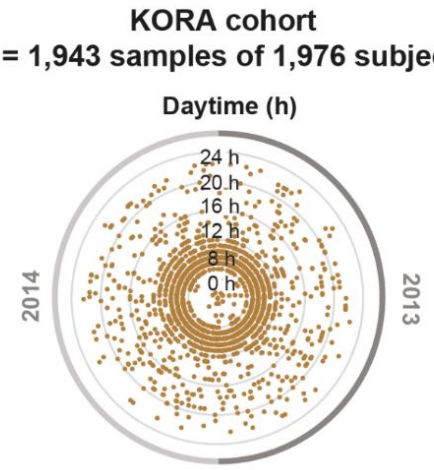


Figure 32 Flowchart of the distribution of sampling time in the KORA$_{2013}$ cohort (N = 1,943 samples).
Dots shows the sampling time point for the whole KORA cohort ordered according date of sampling. However, only 1,943 individuals of 1,976 participants provided information about the sampling time.

Sample collection time was available for 1,943 individuals from the KORA$_{2013}$ cohort. The circadian analysis examines possible circadian rhythmicity of numeric values, e.g., relative abundance, over a 24-h day by fitting the data to a harmonic cosines curve, a non-linear transformations curve used in the modelling of biological rhythms (Marler et al. 2006) (**Figure 32**). The relative abundance values for the OTUs in each bin were averaged and are shown by mean and standard error mean (SEM).

The *alpha*-diversity showed a fluctuation over the day, reaching its peak around noon and flatten out during night. Diurnal rhythmicity was also observed in the two main phyla Bacteroides and Firmicutes, oscillating in antiphase **(Figure 33 A**). Thus, the relative abundance of Bacteroides was lower during the day and showed a 6% higher mean relative abundance at night while Firmicutes

appeared vice versa. To observe bias due to unequal distribution of individuals per time point, randomly selected subsets of an equal number of subjects per time were generated 10-times and also analyzed by JTK_CYCLE. However, no bias was observed **(Figure 33 B).**

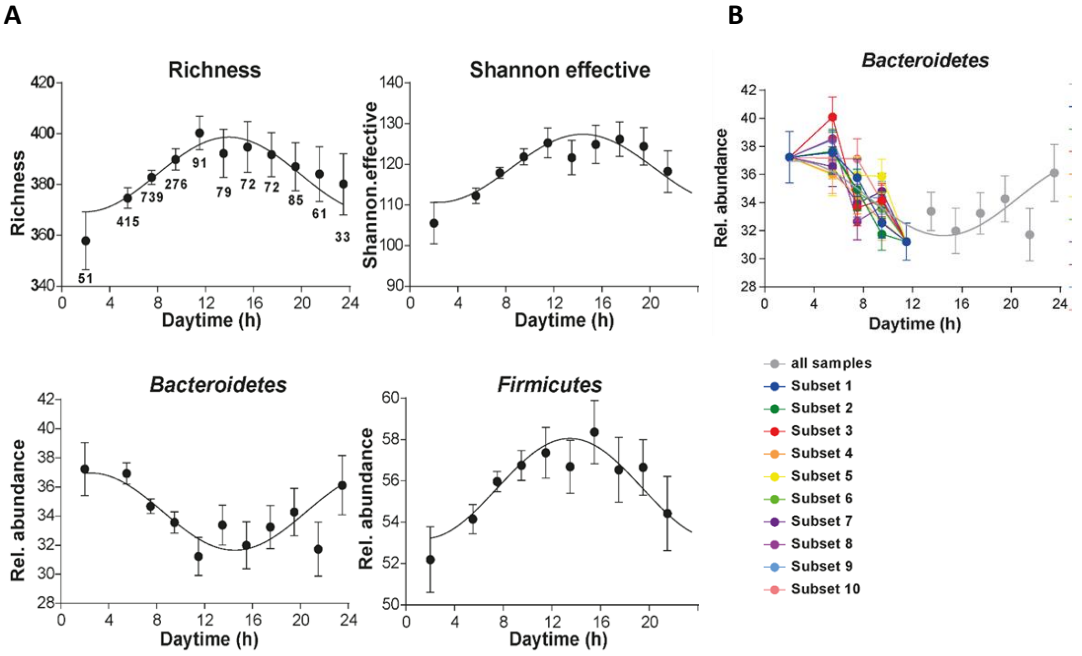


Figure 33 Diurnal profiles of *alpha*-diversity and of relative abundances of phyla.
(A) *Alpha*-diversity and the relative abundance of Bacteroidetes and Firmicutes in 1,943 subjects. Significant rhythms are illustrated with fitted cosine-wave curves (cosine-wave regression, p-value ≤ 0.05). **(B)** Circadian profiling of the relative abundance of Bacteroidetes based on 10 different randomly selected subsets with equal sample size (N = 25 subjects) between 5 and 11am. (relative abundance: 0.1%, prevalence: 10%).

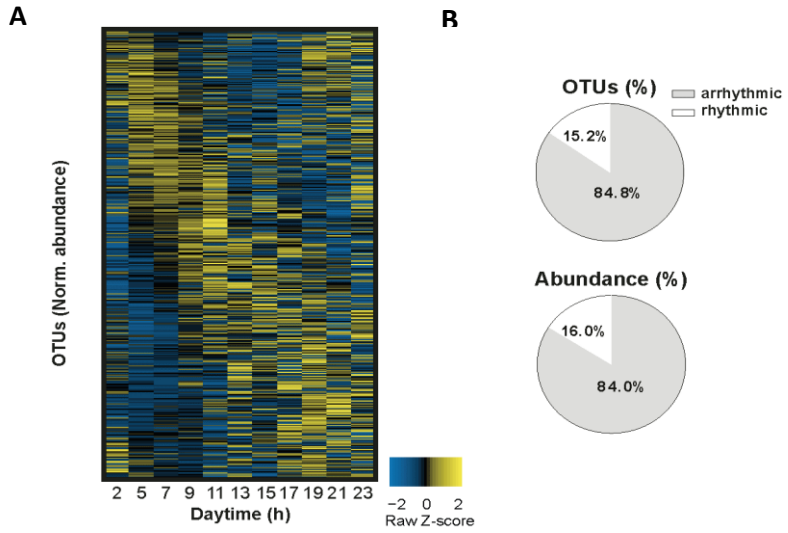


Figure 34 Phase relationship and periodicity of OTUs.
(A) Heatmap depicting the overall phase relationship and periodicity of OTUs (n = 422; mean relative abundance > 0.1%; prevalence > 10% individuals) ordered by their cosine-wave peak phase according to time of day and normalized to the peak of each OTU. **(B)** Amount of rhythmic (white) and arrhythmic (grey) OTUs and their relative abundance in percent (compare to panel B).

In order to identify rhythmicity of individual bacterial taxa, a subset of 422 prevalent (> 10%) and abundant (> 0.1%) OTUs was analyzed in more detail. OTUs were ordered according to their highest relative abundance from early morning to late night. The data were used to create a heatmap, with relative abundance values on the x-axis and the time on the y-axis. The depicted OTUs generated a pattern, forming groups of OTUs which were high abundant in the morning and low in the evening and *vice versa* (**Figure 34 A**). It turned out that 15.2% of these OTUs were rhythmic (rOTU) according to cosine-wave regression analysis, covering a cumulative abundance of 16.0% (**Figure 34 B**).

As shown before, individuals fell into three distinct clusters based on microbial similarities (**Figure 15**). Each cluster is dominated by a certain taxon, e.g., C1 contains higher amounts of *Bacteroides* and C3 more *Prevotella*. Both cluster show robust oscillations in *alpha*-diversity, while subjects in cluster C2 with enhanced amounts of *Ruminococcus* lack oscillations (**Figure 35 A,B).** The percentage of rhythmic OTUs was comparable between all three clusters, a minority of rhythmic OTUs was shared. The absence of rhythmicity in *alpha*-diversity in C2 was not due to low abundance of rOTUs (**Figure 35 C**), which could indicate that many OTUs oscillate with opposing phases in C2. This could explain the absence of the rhythm in *alpha*-diversity when using all OTUs together. OTUs, which were part of C3, tend to be more related, while the other OTUs are more spread out over the whole tree. Maximum relative abundance of all cluster-associated rOTUs was 2.5% while most of the remaining rOTUs showed a relative abundance below 1%.

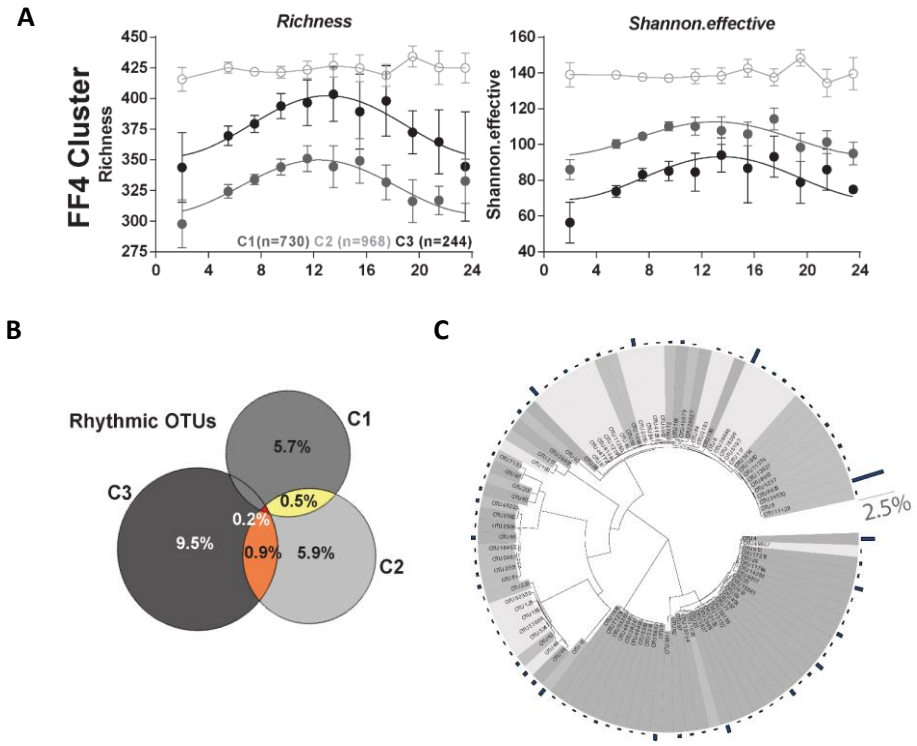


Figure 35 Diurnal rhythmicity in microbiota associated cluster.
(A) Diurnal profiles of *alpha*-diversity for the three determined cluster (**Figure 15 B**). Significant rhythms (cosine-wave regression, p-value ≤ 0.05) are illustrated with fitted cosine-wave curves; data points connected by straight lines indicate no significant cosine fit curves (p-value > 0.05) and, thus, no rhythmicity. **(B)** Amount of rhythmic OTUs found within one cluster. Cluster C1 and C2 share 0.5% of rhythmic OTUs (yellow) and C2 shares 0.9% with C3 (orange) as well as a 0.2% overlap between all three clusters. **(C)** Sequence similarity of all rhythmic OTUs grouped by cluster. Black bars around the tree illustrate the mean relative abundance of these OTUs (max. 2.5%).

Rhythmicity of the fecal microbiota based on targeting the V3V4 V-region of the 16S rRNA gene was next compared to data obtained by targeting the V1V2 V-region of the 16S rRNA gene (**Figure 36**). For V1V2, similar diurnal rhythms were detected in the phyla Firmicutes and on the amount of rhythmic OTUs (cosine-wave regression: 15.2%, harmonic cosine-wave regression: 16.9%), with a less pronounced OTU rhythmicity observed when analyzed by JTK_CYCLE (7.4%). Any oscillations in a*lpha*-diversity and Bacteroidetes were undetectable with either method to find rhythms for V1V2 data (compare **Figure 33** with **Figure 36**).

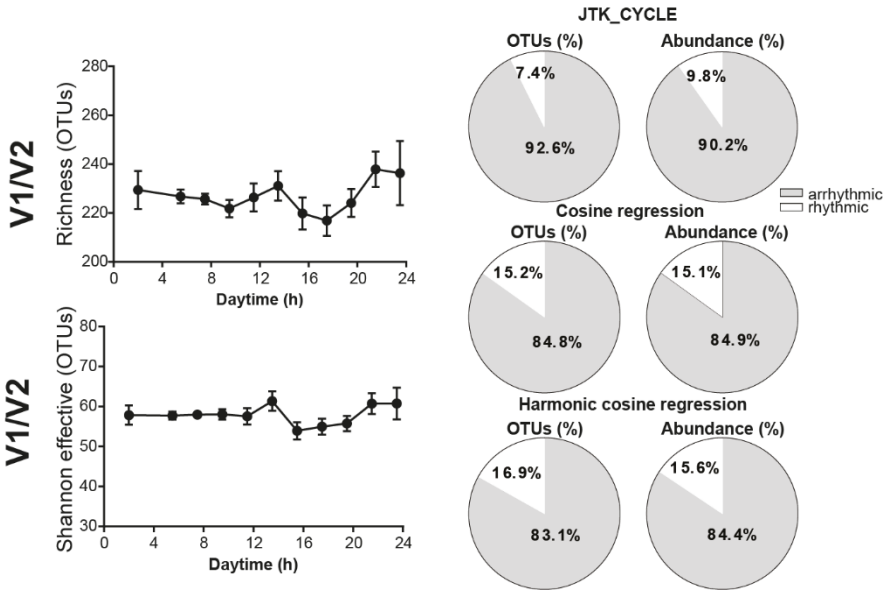


Figure 36 Diurnal rhythmicity in targeting another variable region of the 16S rRNA gene.
Diurnal profiles of *alpha*-diversity in 1,943 subjects by targeting the V1V2 V-region of the 16S rRNA gene. Significant rhythms are illustrated with fitted cosine-wave curves (cosine-wave regression, p-value ≤ 0.05).

Taken together these findings supported our hypothesis that 10 – 20% of the composition of fecal microbiota changes during the day and different times of the day are dominated by different microbial taxa. Thus, sampling time should be considered as a microbiota -associated factor when analyzing the gut microbiota.

4.4.1. Circadian rhythm in the prospective dataset from the KORA$_{2018}$ cohort

The prospective data from the KORA$_{2018}$ cohort in year 2018 can be used to show the reproducibility of circadian rhythms from KORA$_{2013}$. The analysis was performed on 703 individuals from the prospective KORA$_{2018}$ cohort. It could be shown that the bacterial composition follows circadian rhythmicity even after 5 years. A significant oscillation in *alpha*-diversity was determined with a similar curve as seen in the data from KORA$_{2013}$. Richness and Shannon effective number increase during the day and reach their peak around noon, as before (**Figure 37 A**). Finally, also the two main phyla Bacteroides and Firmicutes have a congruent rhythmicity as in KORA$_{2013}$.

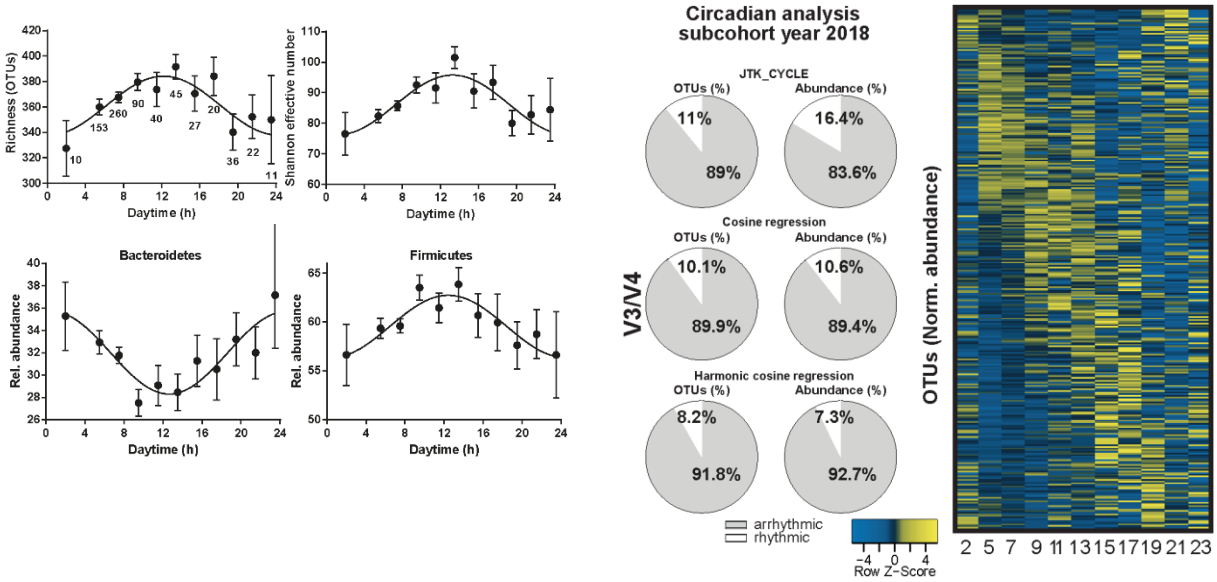


Figure 37 Diurnal rhythm and microbiota profiling of KORA$_{2018}$.
Circadian analysis of rhythmic OTUs in percent based on JTK_CYCLE, cosine-wave regression and harmonic cosine-wave regression of the subcohort from year 2018. On the right, the heatmap depicts the overall phase relationship and periodicity of OTUs. The map was generated by using the raw data of 425 OTUs passing thresholds (relative abundance: 0.1%, prevalence: 10%) and ordered by phase and normalized to the peak of each OTU.

In total, 10.1% of the OTUs were rhythmic according to cosine-wave regression analysis. Compared to the data from KORA$_{2013}$ with 15.2% rOTUs, less rhythmic OTUs are found. However, this can be attributed to the lower sample number included in the analysis. rOTUs have a cumulative abundance of 10.6%, compared to 16.0% in year 2013. The generated heatmap (see **Figure 34**) provides information about the periodicity of the OTUs, which is visible by the generated pattern as before (**Figure 37 B**).

4.4.2. Circadian rhythm in the FoCus cohort

To validate rhythmicity of the human gut microbiota in an independent population-based cohort study, a similar analysis was performed for the FoCus cohort (N = 1,492). Similar to KORA$_{2013}$ the co-variable 'time of defecation' contributed significantly to the explained variation of the gut microbiota (**Figure 20**). Rhythmicity for FoCus cohort was validated on multiple levels, such as *alpha*-diversity, phyla, and OTUs (**Figure 38**).

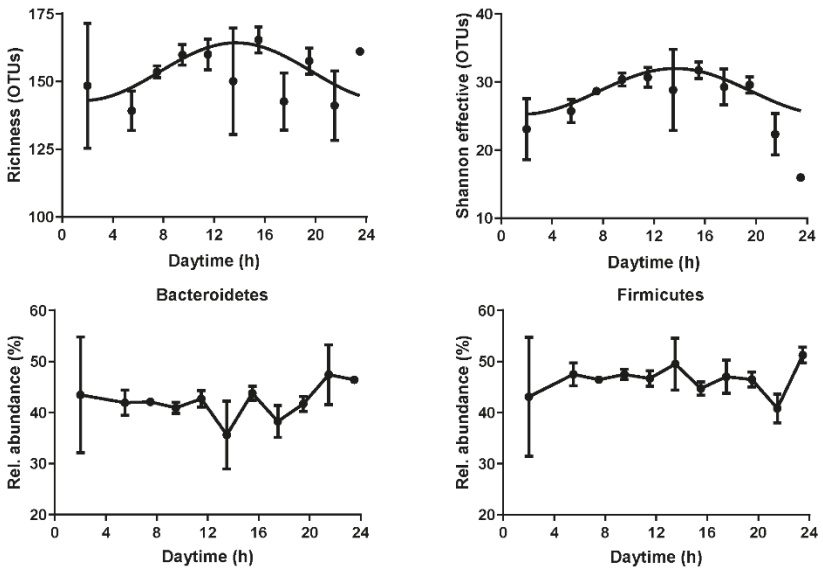


Figure 38 Diurnal profiles of *alpha*-diversity and of relative abundances of the phyla.
Alpha-Diversity and the relative abundance of Bacteroidetes and Firmicutes in 1,492 subjects. Significant rhythms are illustrated with fitted cosine-wave curves (cosine-wave regression, p-value ≤ 0.05).

Circadian rhythmicity in richness and Shannon effective counts could be confirmed for the FoCus cohort. Nevertheless, rhythmicity in Bacteroides and Firmicutes was not found. The reason for the absence of rhythmicity could be due to host health, insufficient distribution of sampling time point or other co-variables, which are still unknown (**Figure 38**).

4.4.3. Longitudinal studies

Multiple samples of one individual are reflecting the composition changes within one individual associated with sampling time and ruling out many co-variables or confounding factors. Thus, prospective data from the $KORA_{2018}$ cohort (two sampling time points) as well as the longitudinal subset of the *enable* cohort (four time points) were used here. In addition, the times series from the longitudinal dataset of S1 with 58 samples were analyzed.

4.4.3.1. Circadian rhythm in the *enable* cohort

The distribution of sampling time points in *enable* was sufficient to generate diurnal. Taking advantage of the additional sampling of further three consecutive samples from 80 individuals, the circadian analysis was performed on the prospective subset only (**Figure 25**).

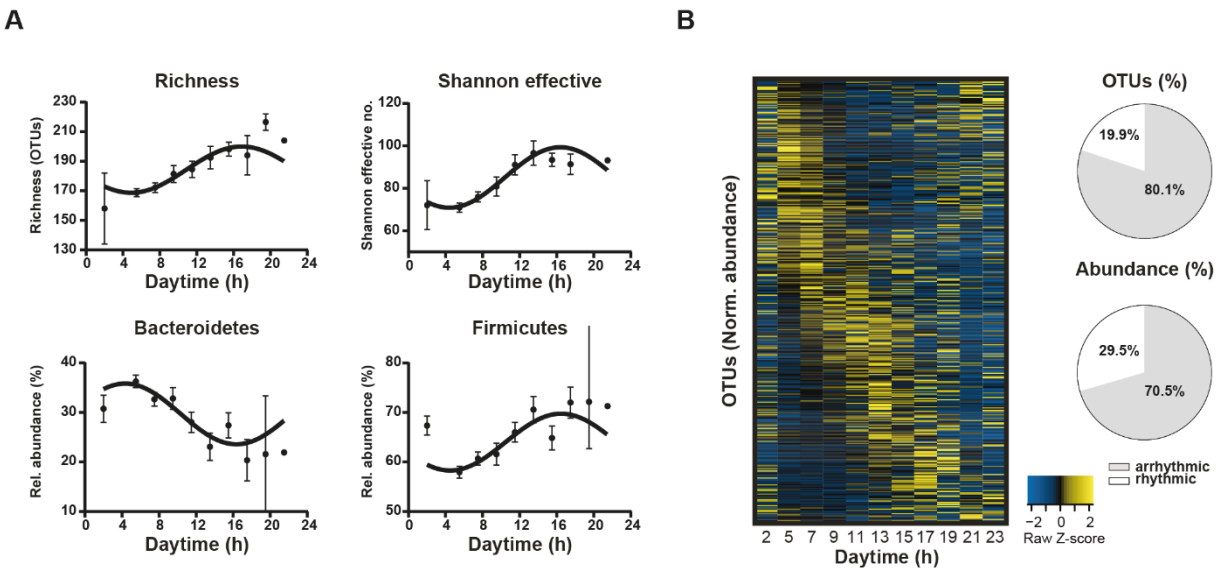


Figure 39 Diurnal rhythm and microbiota profiling of longitudinal samples from the *enable* subcohort (N = 80).

(A) *Alpha*-Diversity and the relative abundance of Bacteroidetes and Firmicutes in N = 80 subjects and N = 320 samples. Significant rhythms are illustrated with fitted cosine-wave curves (cosine-wave regression, P-value ≤ 0.05). **(B)** Circadian analysis of rhythmic OTUs in % based on JTK_CYCLE, cosine-wave regression and harmonic cosine-wave regression of the *enable* cohort. Heatmap depicting the overall phase relationship and periodicity of OTUs is generated by the raw data of OTUs (relative abundance: 0.1%, prevalence: 10%) ordered by phase and normalized to the peak of each OTU.

Significant rhythmicity was observed in *alpha*-diversity with a different peak in richness and Shannon effective counts compared to the $KORA_{2013}$ cohort. There was a small shift in number of observed species towards later time points resulting in an increased richness during the morning and noon and an increased richness in the afternoon and early evening (**Figure 39 A**). On phyla level, the same shift was observed with Bacteroides reaching its peak in the morning and showing the lowest relative abundance values in the afternoon. The relative abundance values of Firmicutes showed a similar cosine curve with similar peaking hours as observed for richness and Shannon effective counts (**Figure 39 A**). In total, 19.91% of the OTUs were rhythmic according to cosine-wave regression analysis, covering a cumulative relative abundance of 29.5%. Compared to KORA, the number of rhythmic OTUs ($KORA_{2013}$ = 15.2%, $KORA_{2018}$ = 10.1%) as well as their contribution to overall relative abundance ($KORA_{2013}$ = 16.0%, $KORA_{2018}$ = 10.6%) was much higher in the *enable* cohort - even though the number of analyzed samples was smaller. Increased values for rhythmicity might be explained by the repeated sampling of the same individuals. It turned out that for the analysis of circadian rhythmicity, the number of samples is even more important than the number of subjects. Thus, also these data confirm the presence of rhythmicity in the human gut microbiota.

4.4.3.2. Circadian rhythm within one subject – a longitudinal sampling over three years

In the description of the dynamic and stability of fecal microbiota composition, fluctuations in the relative abundances on phyla level and *alpha*-diversity were shown. For the analysis of time dependent changes in the fecal bacterial composition, a variety of sampling time points, covering approximately the complete 24-h day, is necessary. Since individual S2 had too many missing time points, only individual S1 was considered for the analysis (**Figure 27**).

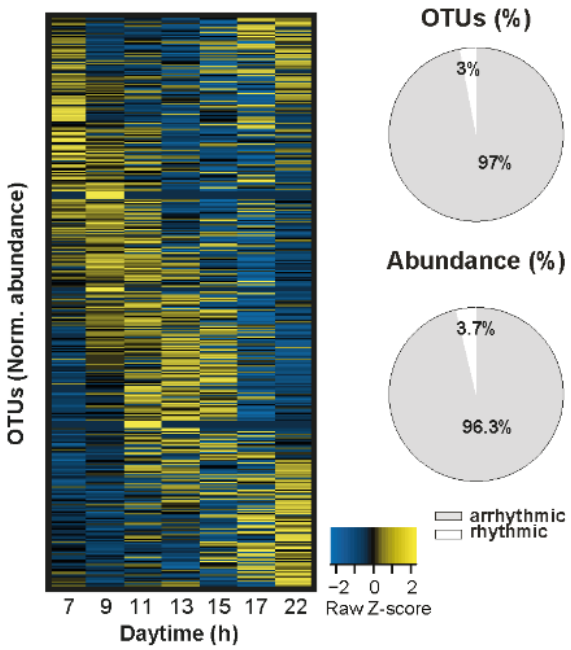


Figure 40 Heatmap depicting the overall phase relationship and periodicity of OTUs in S1.
OTUs are (n = 384 OTUs from n = 58 longitudinal samples; mean relative abundance > 0.1%; prevalence > 10% individuals) ordered by their cosine-wave peak phase according to time of day and normalized to the peak of each OTU. Amount of rhythmic (white) and arrhythmic (grey) OTUs and their relative abundance in percent.

In total, 3% of all 384 OTUs are rhythmic, covering 3.7% of the whole composition. The generated heatmap (see **Figure 37**) shows the OTU pattern oscillating over a 24-h day with differences in the peak relative abundance (**Figure 40**).

Taken together, all results supporting the hypothesis for the presence of a circadian rhythm in the gut microbiota. Already cross-sectional data are appropriate to analyze the circadian rhythm, but longitudinal data seem to provide even clearer signals. Based on these findings, the circadian rhythm was included in the comparative analysis between nonT2D and T2D patients.

4.5. Association of T2D and the human gut microbiota

To better understand the possible association of the gut microbiota and T2D, cross-sectional fecal samples from 1,976 individuals of the $KORA_{2013}$ cohort were analyzed. The aim of this analysis was to identify a bacterial risk signature in T2D using machine learning. In addition to this classification the predictability of the obtained bacterial risk signature was evaluated by including consecutive samples from a $KORA_{2018}$ subset of 699 individuals. Finally, additional cohorts (FoCus, TwinsUK, longitudinal data from S1) were integrated to validate the risk signature. In total, the analysis was performed on N = 4,435 subjects (n = 5,874 samples incl. different time points).

4.5.1. Metabolic health status causes a shift in the bacterial gut composition

Participants of the $KORA_{2013}$ cohort were stratified according to their health status, focusing on four major diseases: obesity (N = 558 subjects), T2D (T2D, N = 277 subjects) together with prediabetes (N = 356 subjects), cancer (N = 200 subjects), and myocardial infarction (N = 66 subjects). The distribution of these diseases among a tree based on taxonomy of the OTUs detected, showed no distinct clustering (**Figure 11**). Next, the previously determined de-novo clusters (C1 – C3) were examined separately. Still, no significant differences in the distribution of any disease (i.e., T2D, cancer, CVD and obesity) was found (**Table 13**).

Table 13 Distribution of major disease among microbial associated cluster.

	C1 (N = 744)	C2 (N = 981)	C3 (N = 249)	P-value
T2D	15.72% (N= 117)	12.74% (N = 125)	14.05% (N = 35)	0.04
Prediabetes	20.96% (N = 156)	16.20% (N = 159)	16.46% (N = 41)	0.25
Cancer	9.81% (N = 73)	10.60% (N = 104)	9.23% (N= 23)	0.78
CVD	3.22% (N = 24)	3.26% (N = 32)	3.61% (N = 9)	0.68
BMI ≥ 30	32.79% (N = 244)	23.24% (N = 228)	34.53% (N= 86)	0.06

Finally, the cohort was stratified either according to the diabetes status of each participant or BMI. Each group of the diabetes status is briefly described by its main differences including age, medication, blood measurements and BMI in **Table 14**. This time, both stratifications showed significant differences. Decreased richness and decreased Shannon effective numbers were observed (**Figure 41**).

Table 14 Description of the KORA$_{2013}$ cohort stratified according to diabetes status.

	nonT2D (N = 1270)	Prediabetes (N = 356)	T2D (N = 277)	P-value
Male (%)	566 (44.56%)	202 (56.74%)	164 (59.21%)	1.65E-07
Age	57 ± 11	65 ± 11	69 ± 9	< 2.2e-16
Waist (cm)	93 ± 13.09	103.06 ± 12.50	107.34 ± 13.13	1.16E-09
Hip (cm)	105.21 ± 8.89	109.15 ± 9.21	110.72 ± 10.63	< 2.2e-16
Bodyfat (%)	31.89 ± 7.04	34.51 ± 6.59	35.37 ± 7-03	3.98E-01
BMI	26.62 ± 4.46	29.75 ± 4.68	31.09 ± 5.38	9.92E-03
Triglyceride (mg/dl)	109.77 ± 66.51	141.21 ± 70.81	154.89 ± 84.77	1.58E-06
HbA1c (%)	5.32 ± 0.32	5.60 ± 0.33	6.49 ± 1.03	< 2.2e-16
HOMA-IR	2.09 ± 1.24	3.85 ± 2.26	5.48 ± 3.72	< 2.2e-16
Glucose (mg/dl)	94.25 ± 7.32	106.40 ± 9.72	136.79 ± 34.30	< 2.2e-16
Glucose (mg/dl) after OGTT	97.20 ± 20.09	148.13 ± 27.30	222.27 ± 65.88	< 2.2e-16
Insulin (pmol/l)	53.33 ± 29.54	87.43 ± 49.65	107.80 ± 93.70	< 2.2e-16
Insulin (pmol/l) after OGTT	283.55 ± 245.85	653 ± 496.19	794.53 ± 491.21	< 2.2e-16
HDL-Cholsterin (mg/dl)	69.48 ± 18.65	61.34 ± 18.11	57.84 ± 16.39	9.76E-06
LDL-Cholesterin (mg/dl)	135.49 ± 34.96	139.03 ± 36.69	128.32 ± 37.32	3.32E-01
Vitamin D (ng/ml)	25.87 ± 11.98	23.65 ± 11.89	22.62 ± 12.37	7.79E-01
Metformin	0	0	142 (48.73%)	< 2.2e-16
PPI	99 (7.80%)	56 (15.73%)	48 (17.32%)	3.92E-07

To further analyze the data unsupervised clustering was performed for each diabetes status. The number of clusters (varied from 3 to 4 clusters) differed according to health status. To avoid bias due to an unequal sample sizes 228 T2D age-matched healthy individuals were selected (without any metabolic health condition or any other major disease).

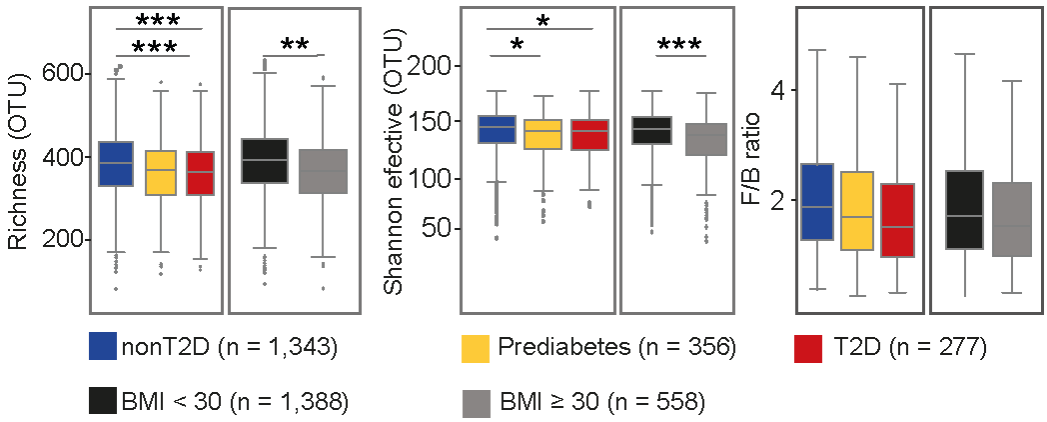


Figure 41 Differences in *alpha*-diversity and Firmicutes to Bacteroides ratios in T2D and obese individuals. Stratifying the whole cohort according to diabetes status (red, T2D; yellow prediabetes; blue nonT2D) and obesity (black, BMI < 30; grey, BMI ≥ 30) shows significant differences in richness and Shannon effective number of species. The decreasing diversity is significantly associated with a decrease in health status. The Firmicutes to Bacteroides (F/B) ratio was not affected by metabolic health.

The fecal microbiota in individuals with prediabetes and nonT2D was divided into two distinct clusters. Compared to the three detected clusters seen in the $KORA_{2013}$ cohort (**Figure 42**). The reduced number of microbiota driven clusters (C1-C2) could possibly be explained by a less diverse microbial composition found in a subset of selected healthy individuals. However, for T2D cases, four different clusters were found, suggesting an association of a more diverse and metabolic disorder. Two of these clusters were assigned to the *Bacteroides*- cluster, which was also observed in nonT2D subjects. For T2D and nonT2D one cluster was clearly dominated by *Prevotella*, while *Ruminococcus* was not found to be overrepresented in one cluster. For prediabetes the first cluster was dominated by *Bacteroides* and *Prevotella*, while the second cluster seemed to be *Ruminococcus*-associated (**Figure 42 A)**. The latter cluster also showed an increased bacterial diversity and richness. In nonT2D, richness was not significantly different between the two clusters, but the overall diversity was lower in the *Prevotella* cluster. Vice versa, the *Prevotella* cluster showed the highest diversity and richness for T2D individuals, while the *Bacteroides* cluster had less bacteria (**Figure 42 B**). Finally, the results gained above were compared to the clusters C1, C2 and C3, which were found when including all subjects of KORA. While in the overall cohort cluster *Bacteroides* seems to be the least diverse cluster, this was not found for the corresponding cluster of prediabetic and nonT2D individuals. Overall, we concluded that there are clusters, which can be formed according to host health corroborating previous studies (Arumugam, Raes et al. 2011, Costea, Hildebrand et al. 2018). However, these clusters were driven by certain genera, varying according to the metabolic health status. This again highlights the complexity of the microbiota composition of the gut and its possible dependency on individual host specific characteristics (**Figure 42 C**).

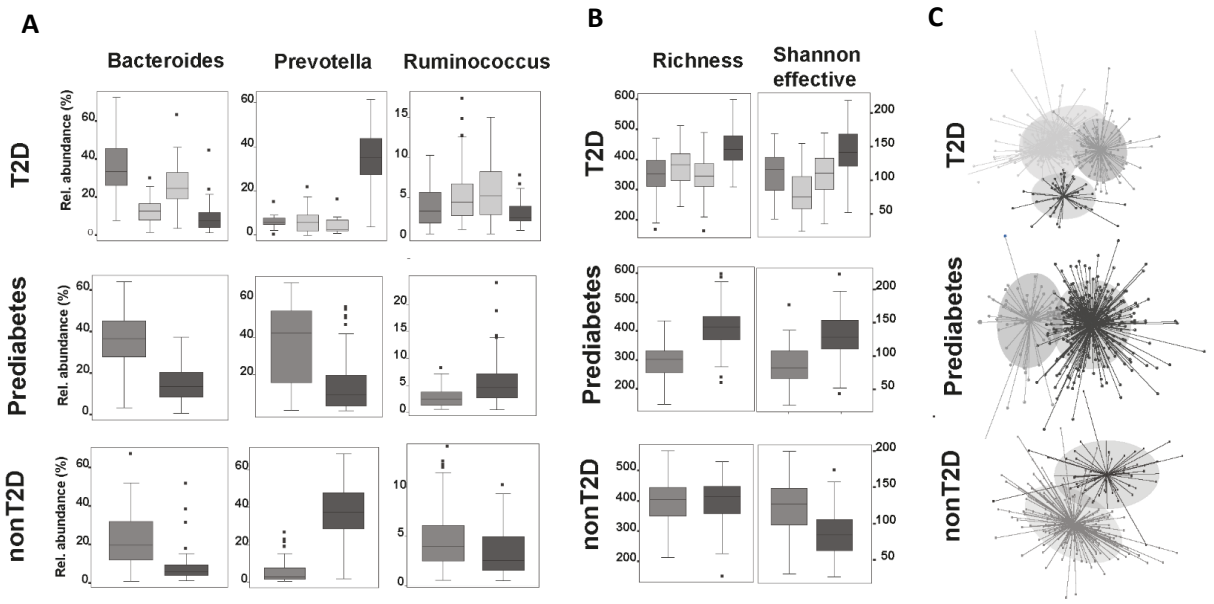


Figure 42 Unsupervised clustering for persons stratified by diabetes in the KORA$_{2013}$ cohort.

(A) The cohort is stratified into T2D (T2D-C1 n = 42 subjects, T2D-C2 n = 93 subjects, T2D-C3 n = 65 subjects, T2D-C4 n = 83 subjects), prediabetes (Pre-C1 n = 280 subjects, Pre-C2 n = 73 subjects) and a matched nonT2D control group (nonT2D-C1 n = 178 subjects, nonT2D – C2 n = 50 subjects). An unsupervised clustering was performed on the microbial profiles for each group. Number of bacterial associated clusters was determined based on the Calinski-Harabasz index. The three boxplots are showing the relative abundance values of the 'enterotype' associated genera for each cluster in different diabetes groups. **(B)** *Alpha*-diversity of the de-novo cluster found in the groups, shown as richness and Shannon effective number. **(C)** Non-parametric multidimensional scaling plots illustrating the *beta*-diversity of the de-novo cluster for each group. According to multivariate permutational analysis, a significant difference between the clusters for all groups were found (P-value T2D = 0.001; p-value prediabetes = 0.001; p-value nonT2D = 0.001).

4.5.2. Taxonomic differences in T2D subjects

Differences between the OTU relative abundance of T2D and nonT2D were determined. Next confounding factors were adjusted by conducting a multivariate permutation analysis on the taxonomic profile of T2D cases and a T2D-matched subcohort. A list of potential confounding factor was assembled based on literature (Belkaid and Hand 2014, Thingholm, Ruhlemann et al. 2019). Disease related variables, e.g. glucose or HbA1c values, were excluded. In total, six variables were found be confounding factors (**Supplementary table 2**) – metformin intake, physical activity, vitamin D intake, smoking, age, and gender. All statistical analysis were adjusted for these factors.

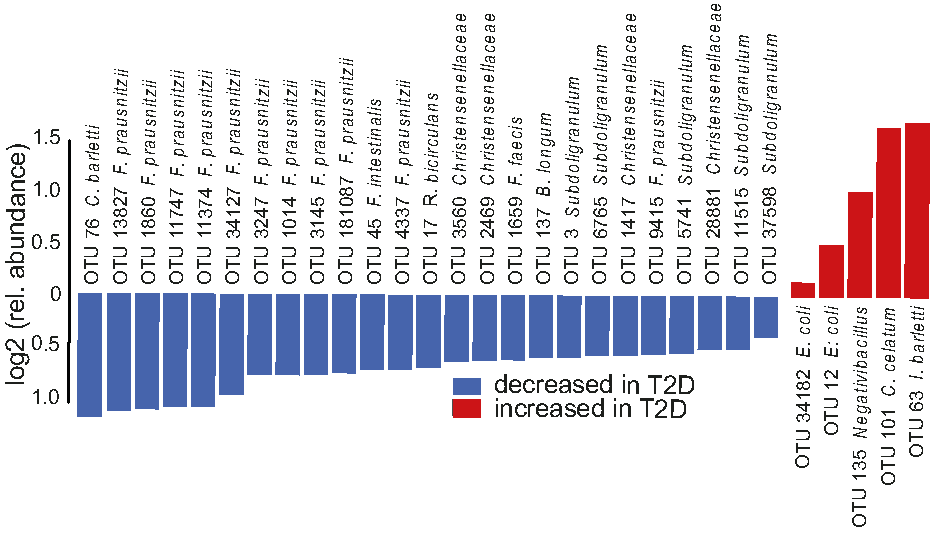


Figure 43 Thirty OTUs, which had significantly different relative abundances between T2D and nonT2D subjects.

A GLM model, following a binomial distribution, was applied to the relative abundance of all OTUs (N = 2,089) to determine significant differences between T2D and nonT2D. The status nonT2D was considered for subjects without any conspicuous diabetes related blood values, neglecting any comorbidities. In this analysis, 30 OTUs were identified to vary in their relative abundance (**Figure 43),** of which 5 were increased and 25 were decreased in T2D. Decreased OTUs mainly belonged to *Feacalibacterium prausnitzii* (11), *Subdoligranulum* (5), and *Christensenellaceae* (4). The remaining five OTUs were identified as *Faecalibacillus faecis*, *Bifdobacterium longum*, *Ruminococcus bicirculans*, *Faecalibacillus intestinalis*, and *Intestinibacter bartlettii*. The latter was also found to be increased in T2D, making up the most abundant two taxa in the analysis. Besides two OTUs seemingly *E. coli*, there was also an increased relative abundance in *Negativibacillus* sp. and *Clostridium celatum*, which both significantly differed between T2D and nonT2D.

4.5.3. Disrupted bacterial oscillations in obesity and T2D subjects

To analyze the association of circadian rhythm and metabolic health, the data were stratified according to the diabetes status. Additionally, the influence of obesity in subjects without T2D (BMI ≥ 30; N = 401 subjects) as well as a combination of T2D cases regardless of BMI (N = 269 subjects) was analyzed. Robust daily oscillations in *alpha*-diversity, phyla and molecular species were observed in subjects without T2D (nonT2D, N = 1,255 subjects), but was disrupted in T2D (**Figure 44 A**). When analyzing the two main phyla, Bacteroidetes and Firmicutes, the abundance in T2D and in prediabetes was distributed evenly during a day period, while it was following a significant cosine curve in nonT2D. In T2D, we observed an even stronger loss of rhythmicity compared to the intermediate phenotype, suggesting that the process of losing rhythmicity already starts in the early stages of T2D (**Figure 44 B**).

Heatmaps were generated for the stratified data sets to investigate the oscillation of OTUs in different phenotypes (**Figure 44 C**). OTUs were grouped according to their relative abundance peaks in nonT2D. The pattern seen in the heatmap for nonT2D (**Figure 45 B**) could not be replicated when using data from individuals with metabolic disorder. This suggests that in T2D, OTUs reaching their peak at different times, which results in a disruption of the distribution, as seen in non-diseased controls. While the pattern was still appearing to some extent in prediabetes, it was absent in T2D, indicating that degradation of metabolic health results in a different distribution of relative abundances during the day/night cycle. Overall, the number of rhythmic OTUs was reduced in T2D compared to nonT2D and prediabetes. While in nonT2D 9.0% of the OTUs were only rhythmic in this subgroup, approximately 1% was shared between nonT2D and prediabetes. The proportion of rhythmic OTUs dropped down to only 5% in prediabetes. Only 1.9% of all OTUs were rhythmic in T2D without any OTUs shared with prediabetes and/or nonT2D (**Figure 44 B).**

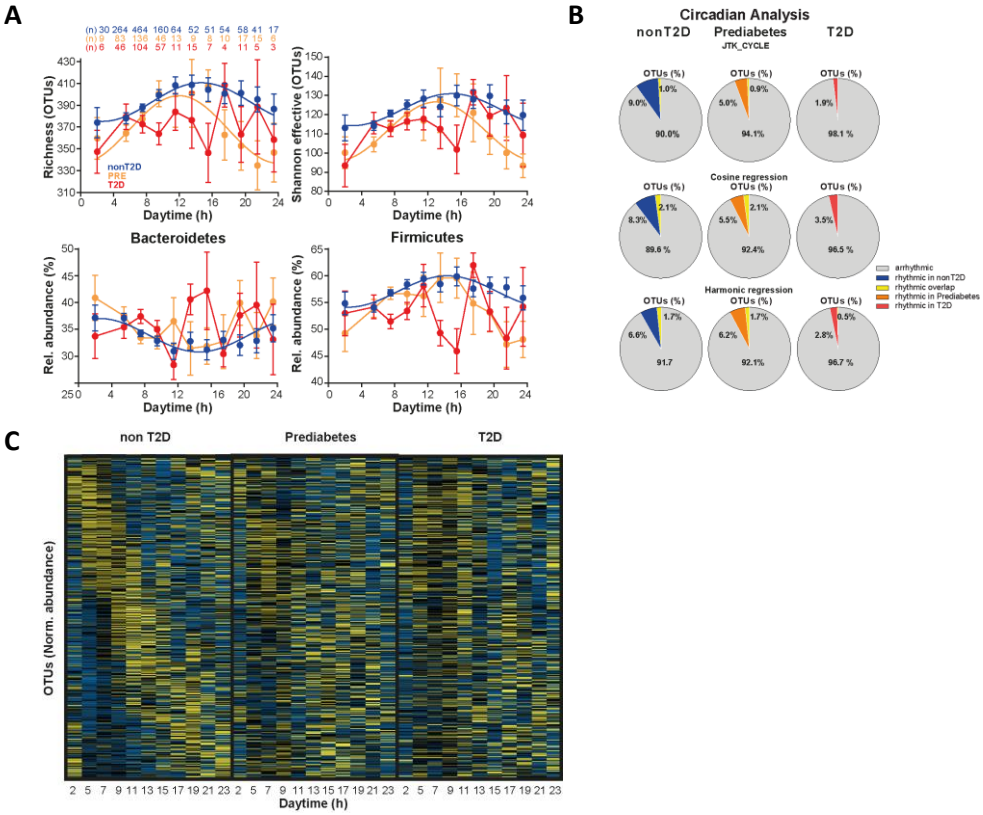


Figure 44 Diurnal rhythm in the human gut micorbiota in subjects with T2D, prediabetes or nonT2D.
(A) Diurnal profile of *alpha*-diversity and of relative abundances of the phyla Bacteroidetes and Firmicutes of subjects with diabetes (T2D, red; N = 269 subjects), with prediabetes (Pre, orange; N = 352 subjects) or without diabetes (non T2D, blue; N = 1,254 subjects). Significant rhythms (cosine-wave regression, p-value ≤ 0.05) are illustrated with fitted cosine-wave curves; data points connected by straight lines indicate no significant cosine fit curves (p-value > 0.05) and thus no rhythmicity. **(B)** Circadian analysis of rhythmic (coloured) and arrhythmic (grey) clustered OTUs and their relative abundance in percent based on JTK_CYCLE, cosine-wave regression and harmonic cosine-wave regression in nonT2D (left, blue; N = 1,255 subjects), prediabetes (middle, orange; N = 352 subjects) and T2D (right, red; N = 269 subjects). An overlap in rhythmic OTUs between each of these three metabolic states is given in yellow. **(C)** Heatmap of the normalized daytime-dependent relative abundance of OTUs based on 422 OTUs (see panel A). Data are normalized to the peak of each OTU and ordered by the peak phase according of subjects without diabetes (left, nonT2D) with prediabetes (middle) or with diabetes (right, T2D).

The comparison of subjects without metabolic health issues to subjects with metabolic disease (either T2D or obesity) identified OTUs, which are showing lack rhythmicity. Thus, rhythmicity in *alpha*-diversity and phylum proportions (Bacteroidetes and Firmicutes) were absent in subjects with either T2D (N = 401 subjects) or in individuals with a BMI ≥ 30 (N = 545 subjects) (**Figure 45 A**). OTUs with disrupted rhythmicity in T2D were largely (> 60%) not shared with arrhythmic OTUs in obese individuals, indicating a BMI-independent loss of rhythmicity in T2D. This observation is one of the key results confirming the presence of circadian rhythm in the human gut and its association with metabolic health.

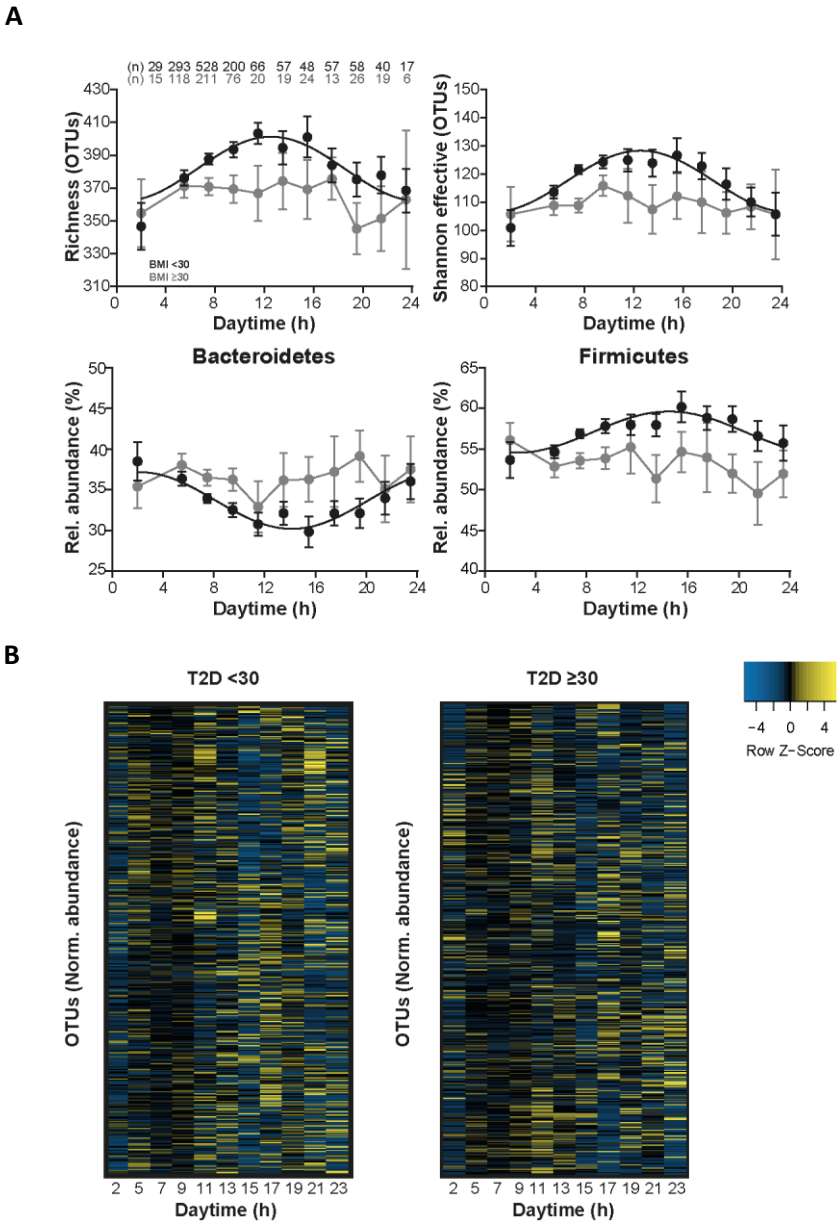


Figure 45 Diurnal rhythm in the human gut microbiota in subjects with T2D and with either BMI ≥ 30 or BMI < 30.

(A) Diurnal profile of alpha-diversity and of relative abundances of the phyla Bacteroidetes and Firmicutes of subjects with diabetes (T2D), either shown for a BMI < 30 (black, N = 1,393 subjects) or a BMI ≥ 30 (grey, N = 545 subjects). Significant rhythms (cosine-wave regression, p-value ≤ 0.05) are illustrated with fitted cosine-wave curves; data points connected by straight lines indicate no significant cosine fit curves (p-value > 0.05) and thus no rhythmicity. **(B)** Heatmap of the normalized daytime-dependent relative abundance of OTUs based on 422 OTUs (see panel A). Data are normalized to the peak of each OTU and ordered by the peak phase according of subjects with diabetes (T2D) and a BMI < 30 (left N = 124 subjects) or a BMI ≥ 30 (right N = 145 subjects).

A subset of OTUs, associated with metabolic health, showed disrupted rhythmicity (in T2D, prediabetes and obesity) (**Figure 46)**. Out of this OTU subset, we identified 87 OTUs, which had lost rhythmicity in T2D. Most of these OTUs belonged to the genera *Akkermansia*, *Bacteroides*, *Bifidobacterium*, *Blautia*, *Clostridium*, *Coprococcus*, *Dorea*, *Prevotella*, *Roseburia* and *Ruminococcus* (**Figure 46**). These findings

were in line with a previously published study by Thaiss et al., who analyzed the circadian rhythm in two individuals over time (Thaiss, Zeevi et al. 2014).

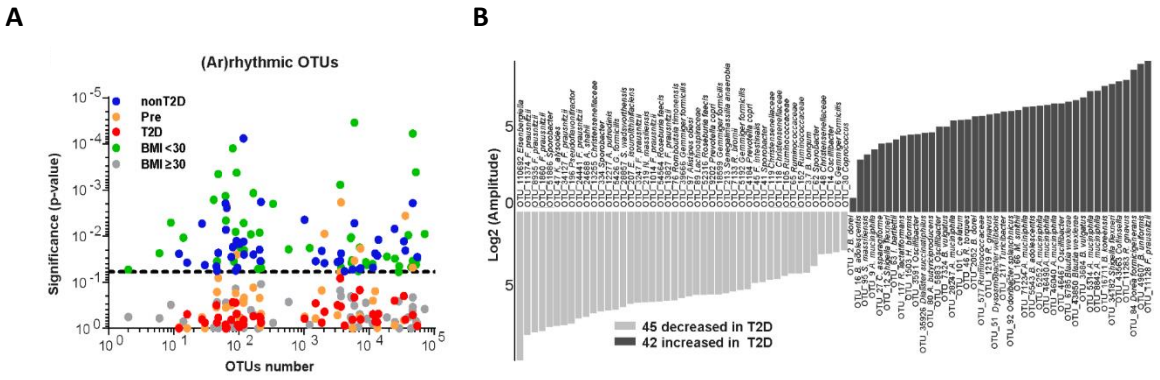


Figure 46 Diurnal rhythm associated with metabolic health.
(A) OTUs show diurnal rhythms in control groups (nonT2D , dark green; N = 1,254 subjects; BMI < 30, light green; N = 546 subjects), but are arrhythmic in disease stages like prediabetes (prediabetes, orange; N = 352 subjects), diabetes (T2D,red; N = 269 subjects), or obesity (BMI ≥ 30, grey; N = 1,396 subjects). Significance of rhythmicity (y-axis) is indicated by p-value below the dashed line (p-value ≤ 0.05; cosine-wave regression). **(B)** amplitude of OTUs rhythmic in healthy controls but arrhythmic in T2D. Of 422 OTUs 87 OTUs oscillated in controls only. These OTUs are ordered according to their fluctuation amplitude in healthy controls: light grey, decreased; dark grey, increased relative abundance in T2D.

In conclusion, these results clearly indicated a time dependent shift in the relative abundance of microbial species over a 24-hour day. Metabolic health is associated with the diurnal oscillation of certain bacteria and results in a disruption of the oscillation pattern in T2D and obese. It was possible to determine specific bacteria, which are losing their rhythmicity in individuals with metabolic issues.

4.5.4. Influence of diet and eating behavior on daytime dependent bacterial oscillation

Diet, the number of consumed meals, and time of eating could influence the gastro-intestinal system and, consequently, the time of defecation, which may have an influence of microbial composition detected in stool samples. Based on a 24-hour food recall, it was possible to draw conclusions of individual's food intake as well as number of consumed meals, and which meals were normally consumed. Additionally, we included data from a Food Frequency Questionnaire (FFQ) to estimate the overall intake of macronutrients such as fiber, saturated and unsaturated fat, and proteins.

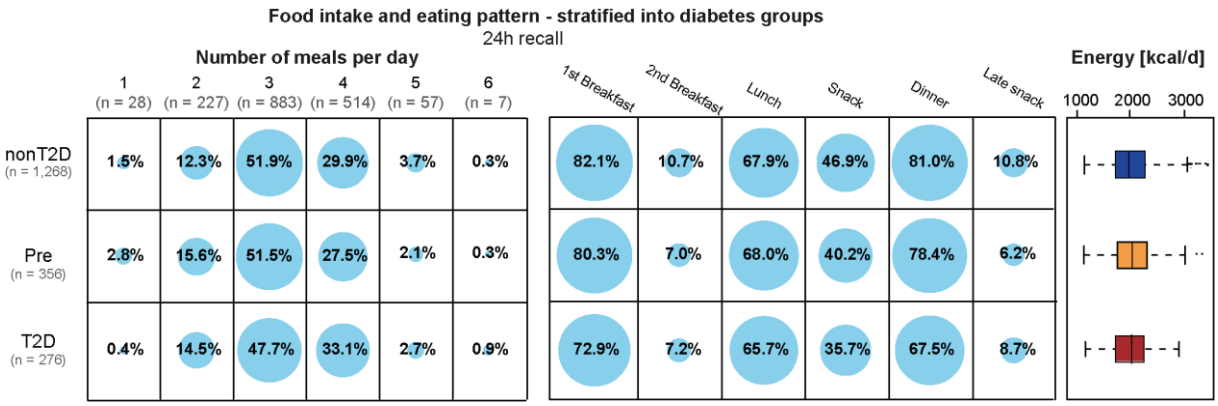


Figure 47 Food intake and eating pattern.
Food behavior and intake between nonT2D, prediabetes and T2D. Left diagram shows individuals self-reported number of meals per day ranging from 1 to 6, number of subjects assigned to the mean-number are shown below in grey. They are grouped according to diabetes status number of individuals per group are below the status in grey. Size of the blue circles are referring to the proportion of subjects within each diabetes group. Panel in the middle shows the type of meal. The size of the blue circles indicates the number of individuals within each diabetes group reported to have had this type of meal. The right boxplot shows the calorie intake in kcal per day for each group (color code as in panel A). In total 7.1% of nonT2D, 10.1% of prediabetes and 20.2% of T2D provided no information. For nonT2D 0.17% reported to have no meal at all and 0.45% for T2D.

Under the assumption that there are difference in eating habits between nonT2D, prediabetes and T2D, the distribution and correlations between these groups was analyzed. Starting with the number of meals consumed on a regular day showed that there were no differences in the proportion of consumed meals between the groups. The majority in all groups consumed three meals per day (**Figure 47**). There was also no observed difference in the type of meal (e.g. breakfast, lunch etc.). One would assume that individuals suffering from a metabolic disorder may tend to eat more and to eat later during the day, including snacks. Nevertheless, there was no such pattern observed in the $KORA_{2013}$ cohort. Of note, in the analysis of eating habits and health the variation in underreporting needs to be considered. There were only 7.1% missing information in nonT2D, but the number increased to 10.1% for prediabetes and to 20.2% for T2D. This finding shows that the self-reported questionnaires likely result in underreporting, especially for the groups obese individuals (King et al. 2016). However, previous studies did not find that mealtime influences the hosts' circadian rhythm, because of the presence of unregular food habits without clear patterns (Collado et al. 2018). In our study, the overall energy intake did not differ between the groups (**Figure 47**). The number of consumed meals influences the overall energy intake. An additional stratification concerning physical activity showed that there was a negative correlation between number of consumed meals and physical activity. Thus, the more meals were consumed the less sport was done (**Supplementary figure 2**). The global intake of macronutrients, such as fiber, saturated and unsaturated fat, and proteins was estimated based on a FFQ. We found no correlation between food intake, global nutrients as well as macronutrients' intake with diabetes blood markers (fasting blood glucose, HbA1c, HOMA index), suggesting that the observed arrhythmicity in fecal microbiota composition was independent of dietary habits (**Figure 48**).

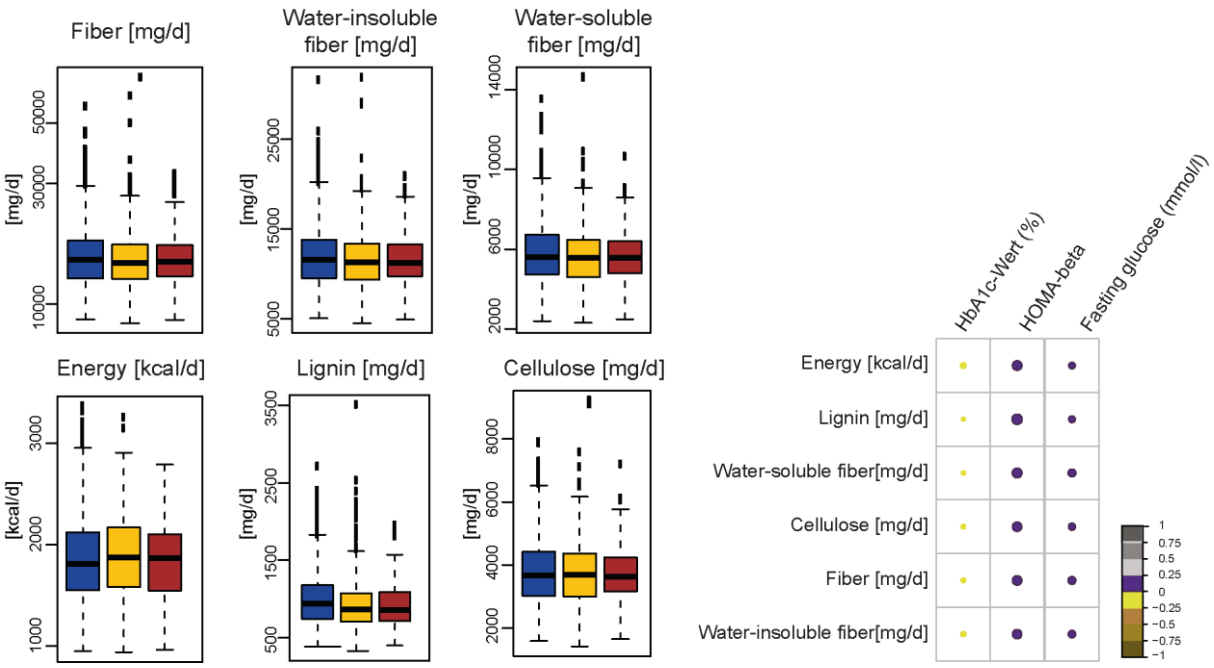


Figure 48 Differences and correlation of general macro-nutrients and diabetes risk markers.
Left boxplots illustrated the general intake of macro nutrients stratified according to the diabetes status. There is no significant difference between the groups. The correlation heatmap shows that these macronutrients do not correlate with diabetes markers.

4.5.5. Classification and prediction of T2D based on microbial signatures

With the aim to identify bacteria, which were important in the onset and progression of T2D, we started to focus on the identification of risk signatures for T2D. Therefore, the $KORA_{2013}$ cohort was split in two subsets to avoid possible misinterpretations due to overfitting of the data. The following results were all conducted on a subset of N = 1,340 individuals. Subjects which were re-collected in the second sampling phase were excluded. The excluded subjects were further used in the prospective analysis to test model's predictability for T2D. To increase the number of incident T2D cases in year 2018 subjects, which were excluded due to antibiotic intake or gastrointestinal related disease, were included (**Table 15**).

Table 15 Overview of subjects included in the analysis.

	Year	N sequenced	N after exclusion	No. OTUs analyzed
Cross-sectional analysis	2013	2,137	1,976	422
Classification model	2013	2,137	1,340	425
Cross-sectional analysis	2018	900	699	318
Prediction model	2013	2,137	699*	2,275**
	2018	900	699	
			* different exclusion criteria	
			** no OTU filtering for BLAST search	

4.5.6. Arrhythmic microbial signature to classify T2D

The preassigned 87 arrhythmic OTUs (**Figure 46**) were further analyzed to determine a T2D associated risk signature based on disrupted rhythmicity in T2D compared to nonT2D. We determined 13 out of 87 OTUs, which showed differences in their harmonic fitting. Comparing these 13 OTUs (s-arOTUs) with the OTUs that were showing significant differences in their relative abundance (**Figure 43**), showed an overlap of 100% **(Figure 49 B).**

With the aim to identify a microbial signature differentiating between T2D and nonT2D, we further trained a generalized linear model (GLM) following a binomial distribution of a logistic regression. In the GLM model, the relative abundance values of the 13 s-arOTUs were considered as dependent variables, as well as BMI.

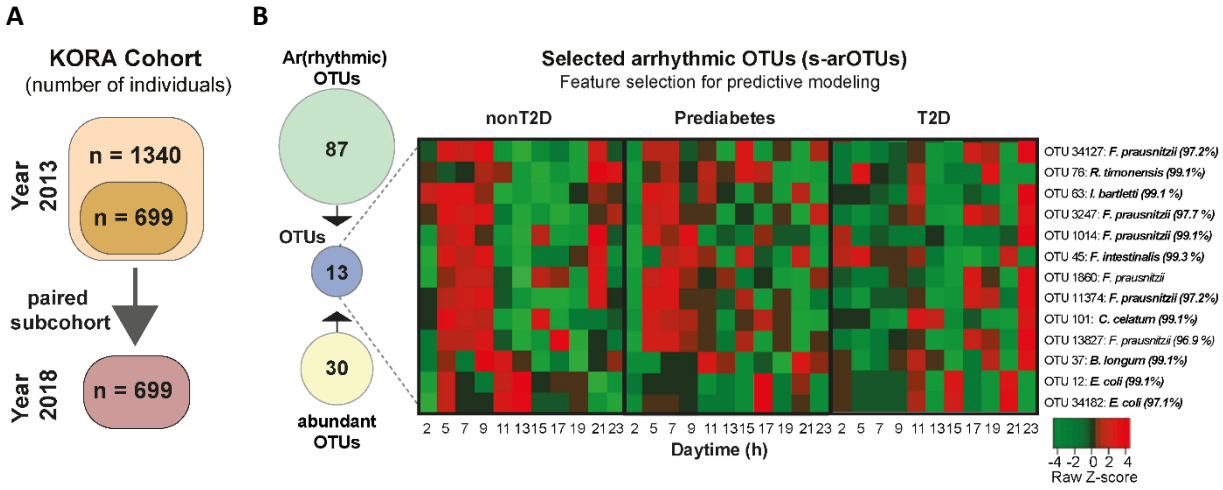


Figure 49 Prediction and classification of T2D based on selected arrhythmic risk signature.
(A) Number of KORA subjects of the prospective sub-cohort with samples from both year 2013 and 2018 (N = 699 subjects). **(B)** Left panel, Among the 87 arOTUs (green circle), which oscillated in controls (**Figure 46**) but are arrhythmic in subjects with T2D or BMI ≥ 30, 13 OTUs (blue circle) showed differential rhythmicity based on DODR analysis and overlapped with the previously defined 30 OTUs (yellow circle) with a significantly different relative abundance (panel A). These 13 selected arrhythmic OTUs (s-arOTUs) were used for further generalized linear model. Right panel, heatmap representing the relative abundance of the s-arOTUs according to the time of day.

The results were collated in a ROC curve reflecting the model's performance in terms of sensitivity and specificity (**Figure 50**). To highlight the importance of the s-arOTUs, we generated a control set of OTUs, which were not becoming arrhythmic in T2D. Results showed a significant better performance of s-arOTUs with a mean AUC of 0.79 for s-arOTUs compared to mean AUC of 0.59 (p-value = 2.03^{-8}), highlighting the validity of the arrhythmic OTUs.

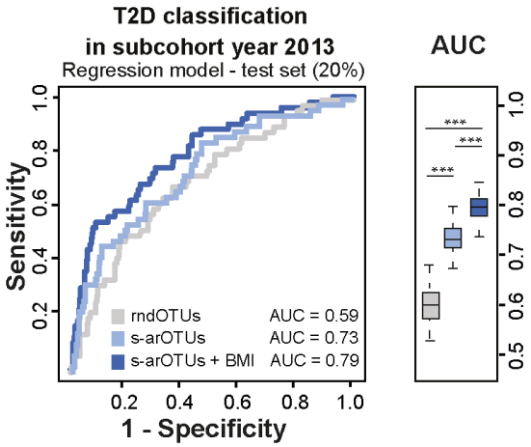


Figure 50 ROC curve of generalized linear model for T2D classification based on s-arOTUs.
Curves of receiver operating characteristics (ROC) for classification of T2D in an independent test set. The generalized linear model is based on 13 s-arOTUs +/- BMI (blue curves) as well as on 100 randomly selected sets of 13 OTUs (grey curve). The distribution of AUCs are shown by boxplots and are significantly different between the types of models.

4.5.7. Implementation of a machine learning approach for T2D classification

We further implemented a random forest model to distinguish between T2D and nonT2D. To increase validity, reproducibility and reliability, we performed the machine learning approach multiple times over a nested 20% test and 80% training set including a 5-fold cross validation. Input of the random forest model were 425 OTUs which were indentified in ≥ 10% of the population with at least 0.1% relative abundance. Of note, the 425 OTUs were determined based on 1,340 subjects, which differed from the 422 OTUs (N = 1,967 subjects), which were considered for circadian analysis in the previous section (**Table 15**). After each iteration, a subset of the most important OTUs was selected based on the lowest error rates generated in the 5-fold cross-validation. For the final model, OTUs, which were picked at least 50-times, were selected.

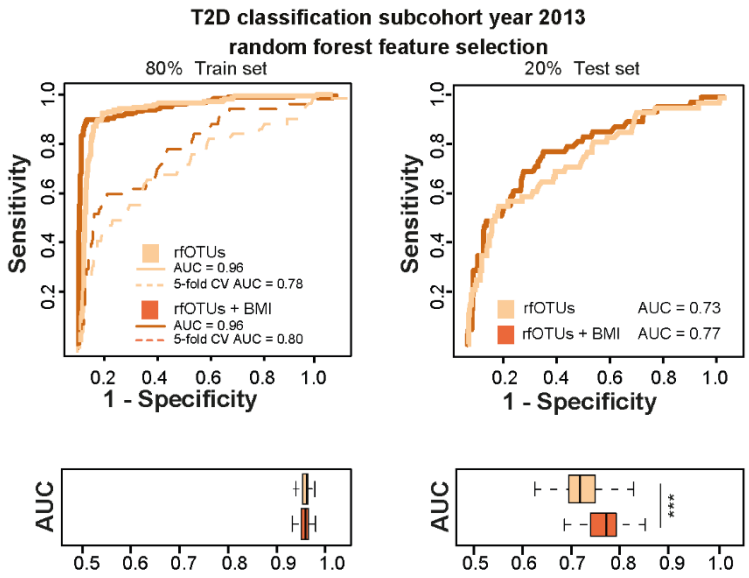


Figure 51 Random forest model for T2D classification.
(Curves of receiver operating characteristics (ROC) for a random forest model using a training set (train set) of 80% of the data (dashed lines in the left panel) as well as using a test set with the remaining 20% of the data (ROC curves in the right panel). The mean AUC over 100 random data splits is shown. The boxplots below the curve panels show the distribution of AUCs across all generated models for the corresponding training and test sets, respectively.

With an AUC of 0.73, the model was able to distinguish between nonT2D and T2D. BMI as an additional variable increased the AUC to 0.77. To verify that the classification was driven by the selected OTUs and not by BMI alone, we generated a GLM out of random selected OTUs. The random model without BMI worked nearly as well as a random classifier (**Figure 52 A**), while BMI increased the performance from AUC of 0.59 to 0.73 **(Supplementary Table 3**).

Nevertheless, the model performed significantly worse than the rfOTUs and/or s-arOTUs, which highlights the validity of the selected rfOTUs and s-arOTUs. Additionally, we trained a random forest model with the outcome 'obesity'. The performance of the random forest model was poor and even on the 80% training set, the model only reached an AUC of 0.84 (5-fold CV AUC = 0.67). Applying the selected features of the random forest model on the 20% test set showed a poor performance of AUC = 0.63 differentiating between individuals with a BMI ≥ 30 and BMI < 30 (**Figure 52 B**). Further, the obesity random forest feature list was used to classify T2D and was compared with a random set of OTUs for the same number of features. Both models were lacking behind in their classification of T2D and showed similar poor results, highlighting that a random OTU list performs only as good as an obesity-trained list (**Figure 52 B**). Considering BMI not only as an additional feature in the classification of T2D but also as a random effect in the random forest model, we implemented a mixed effect random forest model (MERF) including BMI as random effect. As in the random forest model, we split the data in training (80%) and test (20%) set, performed a 5-fold cross validation and repeated the procedure 100 times (**Figure 53**).

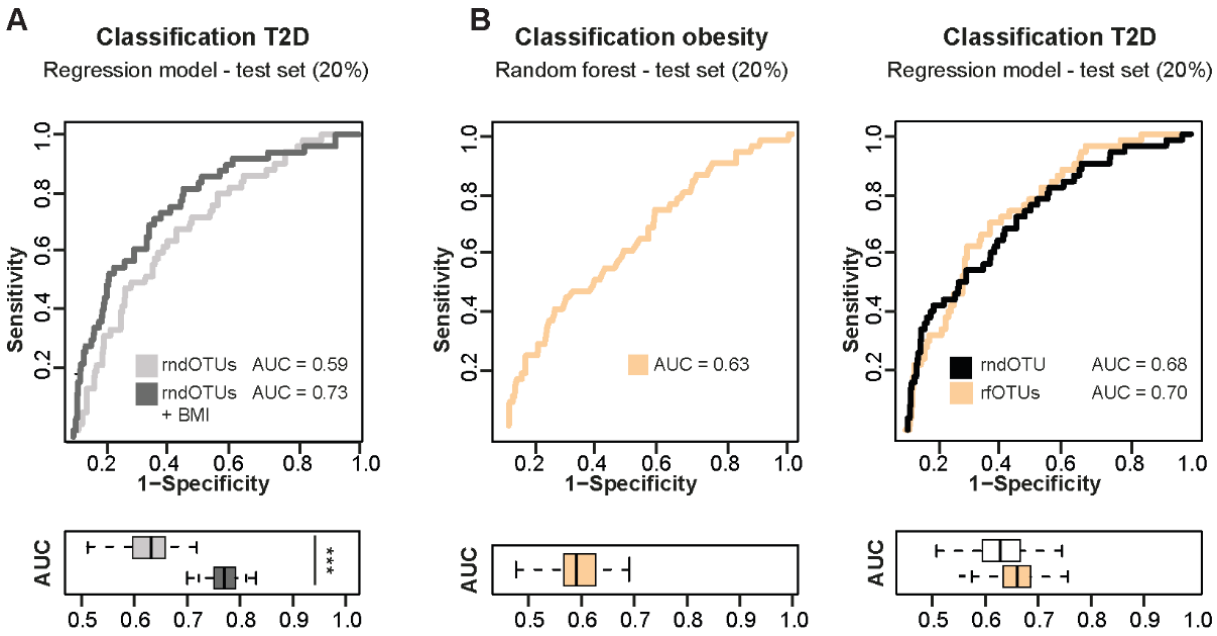


Figure 52 Random forest model for classification of obesity.
(A) ROC curve for classification of T2D in an independent test set. Generalized linear model is based on 13 randomly selected control OTUs +/- BMI (rndOTUs, grey curves). The iterative calculated AUCs are shown by boxplots and are significantly different between the models. (**B)** ROC curve of random forest trained model for classification of individuals with BMI ≥30 and BMI <30 with mean AUC for the training set and 5-fold cross validation. Left ROC curve represents the results out of random forest model for an independent validation set. Iterative AUC values are shown in the boxplot below the curve.

The model (mixedRF) performed significantly worse than the s-arOTUs and was only able to classify T2D with an AUC of 0.69, which is even worse than BMI only with a random set of OTUs (**Figure 52 A**, **Supplementary Table 3**). Interestingly, the mixedRF selected 63 OTUs (out of 425 OTUs) in total, among which all s-arOTUs were present.

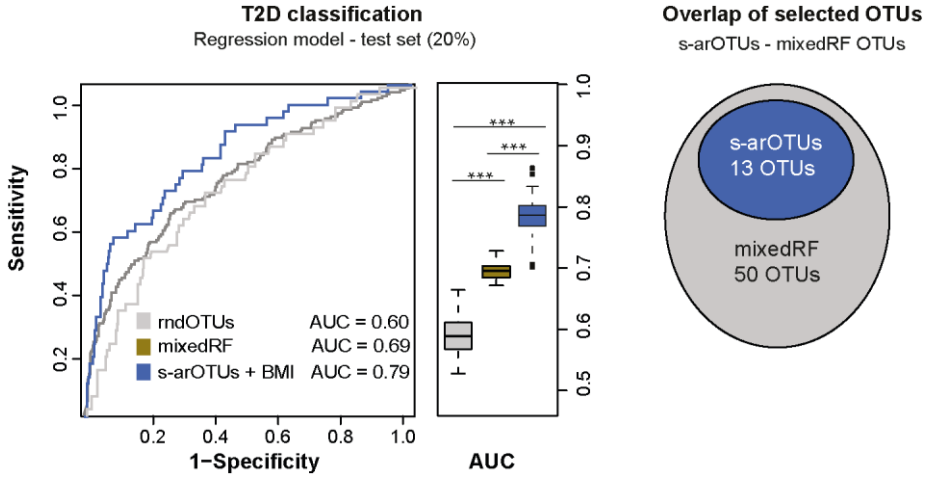


Figure 53 Mixed effect random forest model for T2D classification.
The mixed effect random forest model, taking BMI as random effect, selects 63 OTUs as important features. The mixedRF-OTUs encompass the 13 s-arOTUs.

Same number of OTUs (63 OTUs) were selected by a random forest model which was only trained on OTUs (rfOTUs) being important in the differentiation of diabetes from nonT2D. This, model, when including BMI, reduced the from 63 OTUs to 14 OTUs (rfOTUs+BMI). The inclusion of the BMI did not only increase the predictability of the model, but it also promoted a proper feature selection, resulting in OTUs that were all determined as significantly different in their relative abundance (**Figure 43**).

4.5.8. Validation of the robustness of the random forest model

The strength of the machine learning algorithm is its bootstrapping approach generating multiple decision trees, which enables the handling of unbalanced and large data sets.

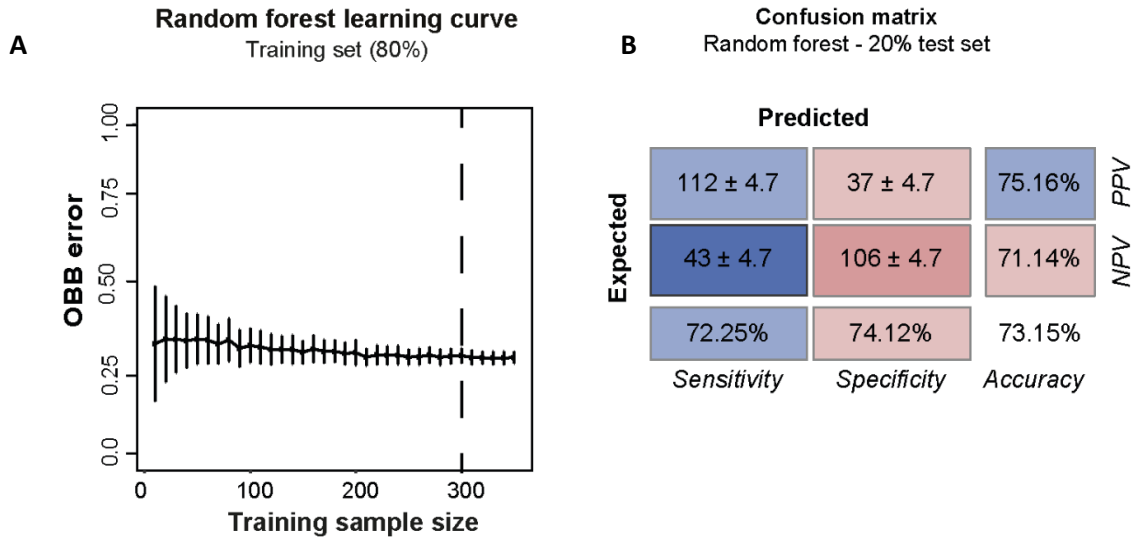


Figure 54 Performance of the random forest model for selected number of samples.
(A) Learning curve showing how sample size, ranging from 20 to 450, influences model robustness. Note that data sets used for training are always balanced for equal numbers of T2D and nonT2D patients. Models were trained with n = 300 samples (dashed line), where the out of bag (OOB) error is stable. **(B)** Confusion matrix showing the mean number of patients and standard deviations across 100 test splits assigned to T2D and nonT2D correctly (blue) or incorrectly (red).

We generated an iterative nested 5-fold cross-validation approach, repetitively selecting a random train (20%) and test subset (remaining 80%), for which the training set consists of an equal distribution of cases and controls (as in the complete data set). The training was performed on an equal number of T2D and nonT2D subjects (N = 300) and cross validated in a 5-fold approach (**Figure 54 A**). The selection of an appropriate number of samples, used for the training set, was important to minimize the error rate. In a repetitive setting, the number of samples included for training was increased to generate a learning curve. The curve, which reached a steady state, resulted in the smallest error rate with 200 samples (nonT2D N = 100 subjects; T2D N = 100 subjects), and decreased the variation of error with an increase in included samples (**Figure 54 A**). Here, feature selection was based on the

cross-selected features ranked by the Gini index. At the end of the iterative feature selection approach, only the features which were selected in at least 50% of all random models are considered as being important.

Besides the already shown AUC value for the random forest model implemented for T2D classification (rfOTUs + BMI), we further validated the performance of the model. Overall, the model was able to truly detect T2D cases with a sensitivity of 72.25% - on average 112 out of 150 T2D cases were identified by the model. With a specificity of 74.12%, the number of falsely classified T2D cases was high; approximately 106 nonT2D cases were classified as having T2D. The number of individuals assigned to a health group according to the model's prediction varied between the 100 randomly repeated models in approximately five individuals and showed an accuracy of 73.15%. (**Figure 54 B**)**.** Focusing on the sensitivity of the model, we generated a precision-recall curve, which showed the positive predicted value against the true positives. When reaching an average AUCPR of 0.77, a decent classification of T2D and nonT2D was achieved. The curve for the GLM of the s-arOTUs performed slightly better with an AUCPCR of 0.79, but with much fewer false positives then the random forest model. According to Matthews's correlation coefficients and F1 scores, both models performed best with a probability threshold between 0.2 and 0.3 resulting in a F1 score of 0.5 for both models. The correlation coefficient of the random forest model was slightly higher when reaching a score of 0.5 for a predictive value of 0.21, while the score of the GLM was 0.38 for a predictive value of 0.24.

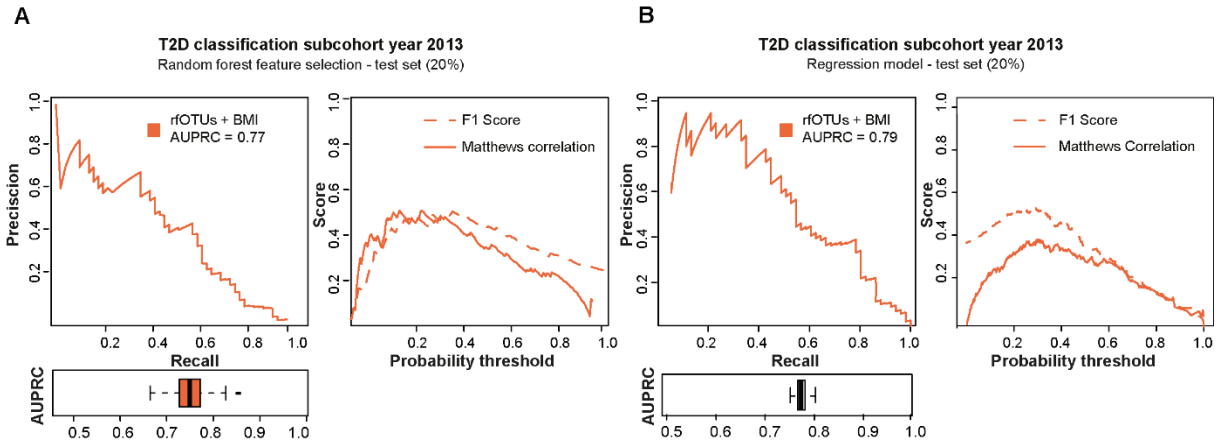


Figure 55 Validation of the performance of the random forest model and GLM.
(A) Left panel: Precision recall curve of the random forest model for T2D classification trained on 425 OTUs and BMI (compare to **Figure 51**). We repeatedly (100 times) split the data set randomly into training (80%) and test (20%) sets and computed the mean AUPRCs for each split. The distribution of the resulting AUPRCs across all generated models for the corresponding test sets is shown below. Right panel: mean F1 scores (orange dashed) and Matthews correlation coefficients (orange solid) computed on the test sets. (**B)** As panel A, but using a GLM of the 13 s-arOTUs including BMI.

To deal with compositionality of bacterial data, we repeated the training of the model with two log transformed input data sets. We performed the training on a centered scaled log transformed data set and on the paired log ratio of OTUs (N = 70,000). Following the same approach as for the original random forest model, the classifier was trained and tested with the transformed data input. Results showed that the AUC values for both data sets, with and without considering BMI as an additional classifier, performed worse in the identification of T2D cases compared to the relative abundance-based random forest model (**Figure 56**, **Supplementary Table 3**). This showed that, in this study, the compositionality did not affect the performance of the random forest model. Further, better results are achieved when including relative abundance values.

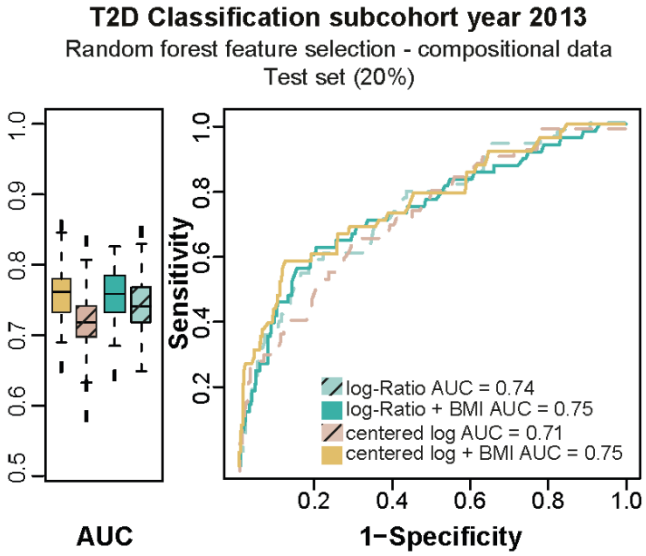


Figure 56 Implementation of random forest for log-transformed data adjusting for compositionality.
Random forest model accounting for possible compositionality bias by training on centered log ratio transformed data (brown) or log ratios (n = 70,000 pairs, green). The models were trained with (solid) and without (dashed) BMI as variable. All other settings, e.g. data splitting, is the same as in the random forest model trained on relative abundances (see **Figure 51**). The boxplots show the range of AUC values for all models computed by using 100 random splits.

4.6. Arrhythmic bacterial risk signature for T2D risk profiling

Comparing the two approaches for the selection of T2D differentiation OTUs showed that the selected 13 s-arOTUs were fully overlapping with the 14 rfOTUs+BMI. This signature included *Bifidobacterium longum* (OTU 37), *Clostridium celatum* (OTU 101), *Intestinibacter bartletti* (OTU 63), *Romboutsia ilealis* (OTU 76), and several taxa closely related to *Faecalibacterium prausnitzii* (OTU 34127, OTU 3247, OTU 1860, OTU 11374, OTU 1014) and *Escherichia coli* (OTU 12, OTU 34182). The above mentioned 13 OTUs did not only differ in their relative abundance between T2D and nonT2D (**Figure 57**), but they also showed a disruption in rhythmicity resulting in different peaking hours according to health status (**Figure 49**).

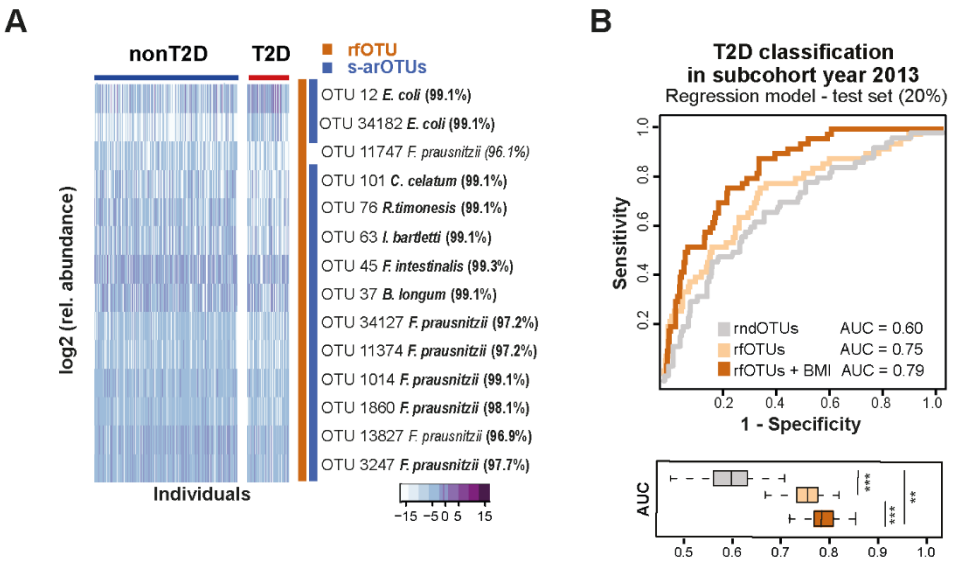


Figure 57 Selected OTU indicating a T2D bacterial risk signature.
(A) Heatmap showing the log-transformed relative abundance of the OTUs (y-axis) selected by the random forest model. Individuals (x-axis) are grouped according to their T2D status. Taxonomic classification of OTUs is shown by their species names and 16S rRNA gene sequence identities (%); bold letters indicate ≥97% identity (proxy for species level). To the right, overlap representation of the 14 rfOTUs (orange, according to C) and the 13 s-arOTUs (dark blue). **(B)** ROC curve for classification of T2D in an independent test set consisting of 20% of the data. The generalized linear model is based on 14 selected OTUs +/- BMI (rfOTUs, orange curves) as well as on the average performance of 100 random sets of 14 OTUs (rndOTUs, grey curve). The distribution of AUCs of all models is shown by boxplots and differed significantly between the three model types.

With the aim to generate a model which could be used as a prediction/classification models for other studies, we trained a GLM including rfOTUs+BMI differentiating between T2D and nonT2D. Random sets of 14 OTUs were selected 100 times and went into a GLM. The results show that a random set of OTUs performed significantly worse in the classification of T2D than the random forest selected OTUs (AUC rfOTUs+BMI = 0.78; mean AUC of random OTUs = 0.60). As expected, the GLM performed as good as the s-arOTU model. For further analysis, we selected s-arOTUs as bacterial risk signature (**Supplementary Table 3**).

4.6.1. Additional diabetes risk markers as classifier for T2D

The s-arOTUs GLM showed an increased AUC when including BMI as an additional classifier. In the commonly models used for prognostic and diagnostic of T2D, risk markers, such as bodyweight, lifestyle, and blood values, are important markers. Based on these factors, the classification of T2D is reliable and precise (Gillies, Lambert et al. 2008, American Diabetes 2017, Skyler, Bakris et al. 2017). The inclusion of microbial data and the observation of different compositions of the human gut microbiome associated with metabolic health or issues thereof, should not be seen as a replacement of the known diabetes risk classifiers. It should serve as an additional factor helping to understand the impact of metabolic health better. Generating different random forest models, based on the metabolic risk markers with and without OTUs, showed the added value of bacterial taxa in the classification of T2D (**Figure 58**). The introduced s-arOTU risk signature, based on 13 OTUs and BMI, was able to differentiate T2D with an AUC of 0.79. The model included BMI or Hip-Waist (HW) ratios. The performance of models using any miscellaneous risk markers (blood pressure, parental T2D, age, hip-waist ratio (HW), smoking, and physical activity) were improved, when OTUs were included. In a stepwise procedure, we added BMI and HW as additional classifiers in the random forest model. The inclusion of BMI instead of HW performed slightly better (AUC: 0.82 vs. AUC 0.83). Overall, when considering all miscellaneous risk makers, and BMI and HW, the inclusion of OTUs additionally improved the performance to an AUC of up to 0.87. The selected 9 OTUs of the random forest model were all overlapping with the 13 s-arOTUs. This verified the importance of at least 9 the 13 s-arOTUs in T2D classification and suggests the remaining 4 are also important.

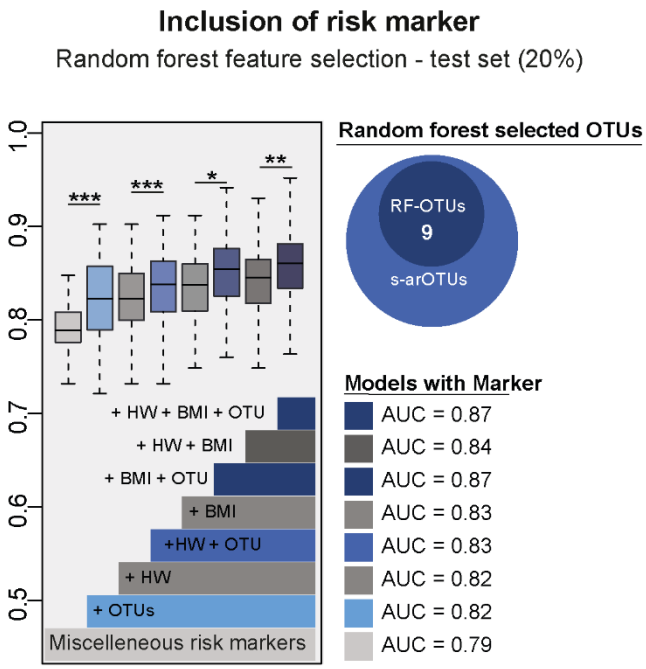


Figure 58 The influence of bacterial taxa in T2D classification based on commonly used risk marker.
A random forest model trained on miscellaneous diabetes risk markers – blood pressure, parental T2D, age, hip-waist ratio (HW), smoking, and physical activity – to classify T2D. Boxplots depict the increase of AUC values by including either BMI or HW, as well as microbial taxa. The AUC increased from 0.79 to 0.87 by including OTUs in the classification model. The random forest models with OTUs including all additional diabetes risk markers selected 9 OTUs as important features, all of them are overlapping with the 13 s-arOTUs.

To better understand the impact of each metabolic risk factor towards the classification of T2D, all variables were tested. It was shown that the variable 'age' was the most influencing one. For instance, one model was built on risk markers only, the other included of s-arOTUs (**Figure 59 A).** The corresponding mean AUC values increased with s-arOTUs included (**Figure 59 A**, table: first column = AUC of miscellaneous marker only, second column with dashes = AUC of miscellaneous risk marker including s-arOTUs). The AUC increased from 0.76 to. 0.79 for BMI and HW, and from 0.77 to 0.79 by adding 'gender' as an additional variable (**Figure 59 A**, left graph). After considering age, the inclusion of OTUs did not improve the models' performance substantially, suggesting that age is one of the most important factors in the classification of T2D. Thus, excluding age from the set of miscellaneous risk markers helped to evaluate the impact of s-arOTUs in the T2D classification better. Results of these models showed that the inclusion of the 13 s-arOTUs resulted in an increased AUC for all models excluding age (**Figure 59 A** right graph). The benefit of the 13 selected s-arOTUS became clear when including randomly selected OTUs instead. With random OTUs, any model performed equally or worse independent of the specific OTUs used (**Figure 59 B)**.

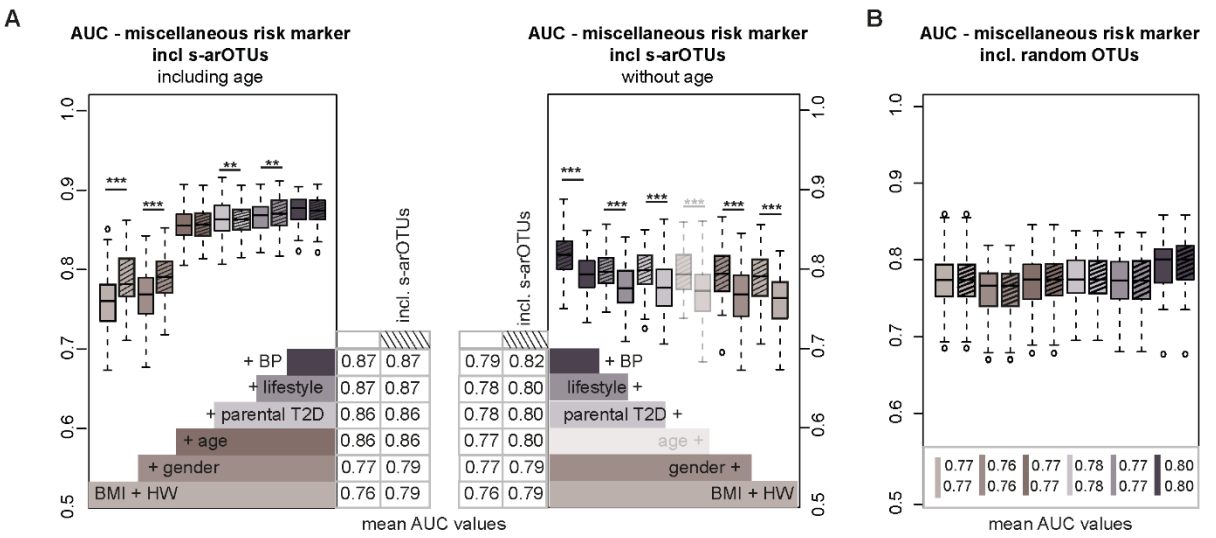


Figure 59 Combination of s-arOTUs and miscellaneous risk marker.
(A) Left Boxplot shows the AUC values of a GLM generated for the shown miscellaneous risk marker (BMI, Hip-Waist (HW), age, parental T2D, lifestyle (smoking, physical activity), blood pressure (BP)) without OTU (left filled boxes) and with s-arOTUs (right shaded boxes). Bars in the lower part of the plot shows the included risk marker. Stepwise inclusion of another risk marker. Boxes and bars are labelled in the same color. Tables shows the AUC values with and without the inclusion of s-arOTUs. The right boxplot as the left boxplot but without inclusion of the variable age. (**B)** Classification of T2D including 13 randomly selected OTUs (shaded boxes) and stepwise inclusion of miscellaneous risk marker as in A.

4.7. Impact of Metformin on microbiota changes

Metformin is a blood glucose level lowering drug prescribed for T2D especially if patients are obese. Metformin is recommended as the first-line treatment in T2D (American Diabetes 2017). It has been shown that intake of metformin alters the human gut microbiota, suggesting that compositional changes observed for T2D patients were mainly driven by the medication and not by the disease (Forslund, Hildebrand et al. 2015, Pryor, Norvaisas et al. 2019). To evaluate the influence of metformin in the $KORA_{2013}$ cohort, and, thus, in the selection of microbial species related to T2D, we analyzed this medication in more detail.

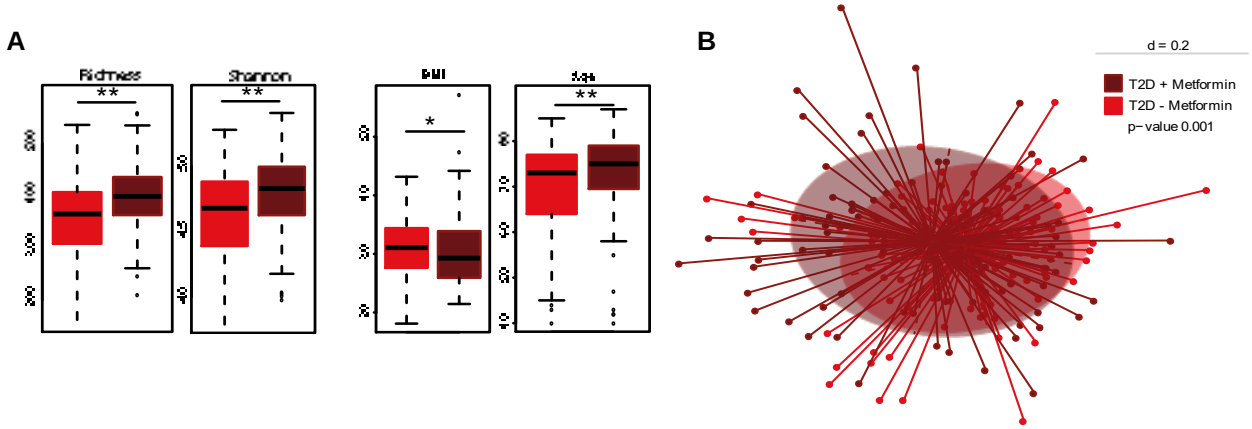


Figure 60 Influence of metformin in classification of T2D.
(A) Differences in alpha-diversity between T2D cases taking metformin (T2D+MET, darkred; N = 132 subjects) and T2D cases without metformin treatment (T2D-MET, red; N = 136 subjects). **(B)** Multivariate nonparametric permutational analysis showed significant differences in the composition of the gut microbiota between metformin treated T2D cases and not treated (p-value = 0.001). The nMDS plot shows a shift of the microbiota, but with several individual overlaps between the groups and no clear distinction.

The cohort was stratified according to T2D either taking or not taking metformin (T2D+MET; N = 132 subjects, T2D-MET; N = 136 subjects). The number of observed 'species' and bacterial diversity was significantly increased in T2D+MET, which was in accordance with previous studies (Forslund, Hildebrand et al. 2015, Pryor, Norvaisas et al. 2019). There was also a significant imbalance in BMI and age, which indicated that T2D+MET had an increased BMI but were younger compared to the T2D-MET group (**Figure 60 A**). This result seemed to be in contradiction with the known reduced richness and diversity in obesity and age (Yatsunenko, Rey et al. 2012, Le Chatelier, Nielsen et al. 2013, O'Toole and Jeffery 2015). However, comparing the composition of the gut microbiota between individuals with and without metformin showed a significant separation (p-value = 0.0001) into two groups (**Figure 60 B**).

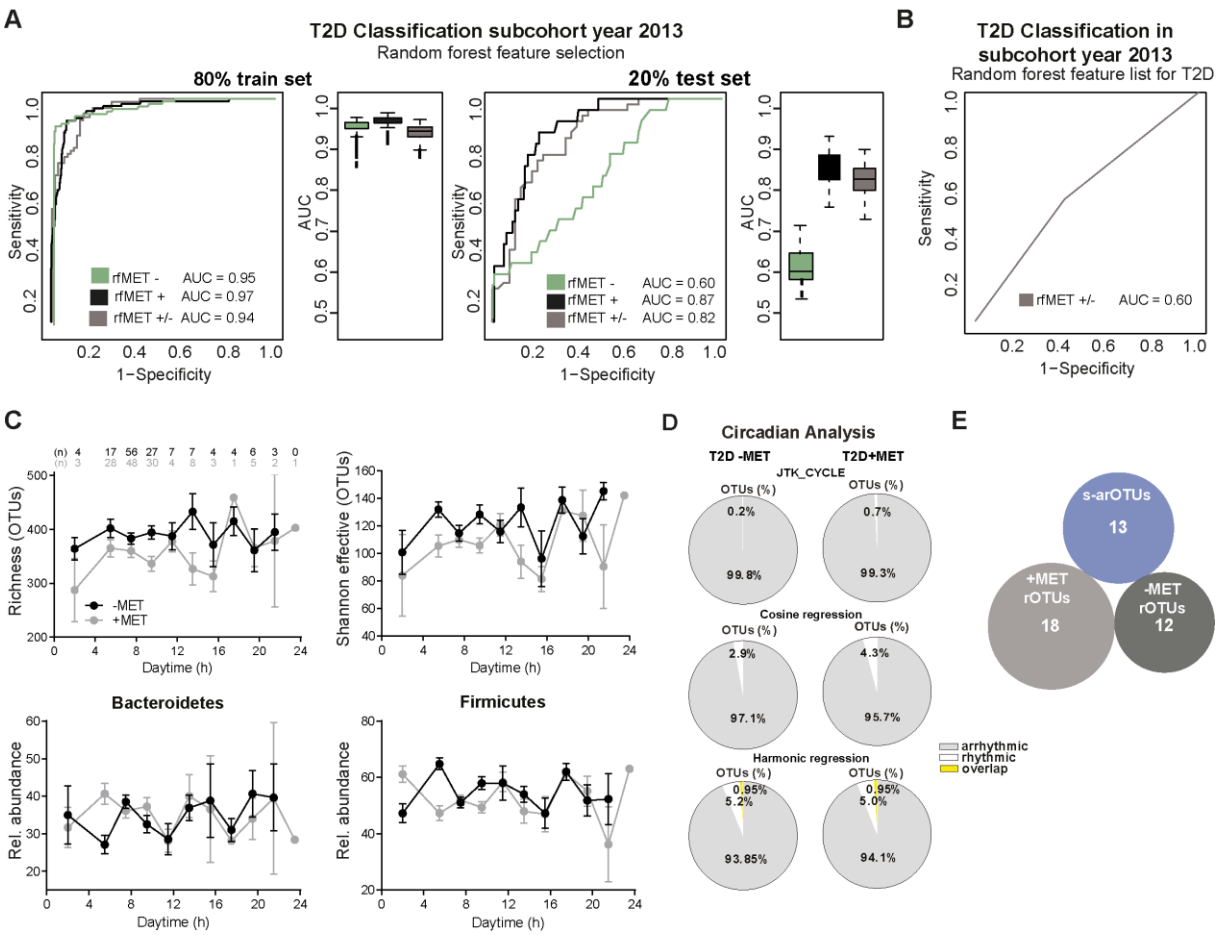


Figure 61 Association of metformin and the selected risk signature.
(A) ROC curves of random forest model to evaluate the effect of metformin intake in T2D. Individuals are stratified in three groups, T2D + MET compared to nonT2D compared to nonT2D (black), T2D – MET (green) and T2D +/– metformin (grey). In case of the last group, classification is based on the outcome 'metformin'. Iterative AUCs are represented in boxplots for training and independent test sets, as well as the mean AUC of all groups for training and independent test. **(B)** ROC curve based on generalized linear regression model for 13 selected arrhythmic OTUs (**Figure 55**, s-arOTU + BMI) for differentiation of metformin in T2D. **(C)** Diurnal profile of the relative abundance of alpha-diversity indicated by richness and Shannon effective number of species and the phyla Bacteroidetes and Firmicutes of T2D subjects who take metformin (+MET, grey; N = 132 subjects) or not (–MET, black; N = 136 subjects). Data point were connected by straight lines to illustrate that no significant rhythm based on a fitted cosine-wave curve was found (cosine-wave regression, p-value > 0.05). **(D)** Circadian analysis of rhythmic (white) and arrhythmic (grey) OTUs and their relative abundance in percent based on JTK_CYCLE, cosine-wave regression or harmonic cosine-wave regression in T2D – MET (left) and T2D + MET (right). **(E)** Venn diagram illustrating no overlap between the s-arOTUs and all identified rOTUs in either T2D – MET or T2D + MET based on cosine-wave regression analysis (P ≤ 0.05).

Previous studies found that metformin synchronizes peripheral circadian clocks (Barnea et al. 2012). Thus, three random forest models were trained to classify either T2D for the two subgroups (T2D+MET and T2D-MET against nonT2D individuals) or to differentiate between T2D cases taking metformin or not taking the medication. The performance of these models was comparable to a classification model published by Forslund et al. (2015). The generated random forest model was not able to distinguish between nonT2D and T2D for individuals not taking metformin (rfT2D+ AUC = 0.60) (**Figure 61 A**). However, the differentiation of nonT2D and T2D in individuals taking metformin (rfT2D+ AUC = 0.87)

was much better. A model trained on T2D cases only, but for the endpoint metformin intake, was able to classify individuals with T2D according to taking metformin or not taking metformin (rfMET AUC = 0.82). Finally, comparing the selected feature list for the metformin-based analysis with the 13 s-arOTUs, an overlap of 8 OTUs was found for selected OTUs present in of all three models. Further, one OTU was additionally picked up in the rfT2D+ model.

Comparing the beta estimator, the margin by which an s-arOTU contributed to the classification of T2D, showed no differences when stratifying the cohort according to metformin intake before. Next, the generated GLM for the classification of T2D was used to differentiate metformin intake in T2D cases. This GLM model was not able to distinguish between T2D cases, whether taking metformin or not. This verified that even though a random forest model trained on the outcome 'metformin' performed with an AUC of 0.82 in distinguishing +MET-T2D and +MET-T2D, the s-arOTUs are not biased for metformin (**Figure 61 B)**. We further analyzed the data obtained from T2D cases, and divided them into the two treatment groups, i.e., with and without metformin. Interestingly, there was a 1.4% increase for rhythmic OTUs in T2D+MET when comparing T2D with or without metformin (i.e., +MET versus –MET). Of those, 14 OTUs gained rhythmicity in T2D regardless of taking metformin or not. When analyzing the T2D+MET and T2D-MET groups separately, 9 OTUs of the 14 OTUs showed significant rhythms, but they were restricted to T2D+MET and not present in T2D-MET. These results indicate that MET treatment may induce rhythmicity of specific taxa. Nevertheless, none of the 14 taxa were among the 13 s-arOTUs used for classification or prediction of T2D in the validation and prospective cohorts before. Although metformin was found to synchronize peripheral circadian clocks (Barnea, Haviv et al. 2012), indicating that metformin may directly interfere with the circadian analysis, this substance did not change rhythmicity of the *alpha*-diversity and any phylum, nor the overall percentage of rOTUs in subjects with T2D was changed in our study (**Figure 61 C, D**). Importantly, none of the rOTUs identified in +/-MET-T2D overlapped with the 13 selected arrhythmic OTUs (s-arOTUs) used for the classification of T2D (**Figure 61 E**). Thus, we were able to replicate the results that the human gut microbiota is altered after metformin intake, but the before selected risk signature (s-arOTUs) was not driven by the medication.

4.8. Validation of T2D risk signature in other population-based cohorts

In the previous part it has been shown that the composition of the human gut microbiota follows a circadian rhythm. Further, a bacterial risk signature for the classification of T2D was determined. The selection of specific OTUs was conducted using two approaches. For the next part, it was examined if the risk signature can be applied to other cohorts and used to classify T2D.

4.8.1. T2D classification in an independent German cohort, FoCus

The methods used for $KORA_{2013}$ were applied to FoCus, which is an independent German cohort with the study center in Kiel (FoCus: nonT2D N = 1,070 subjects, T2D N = 293 subjects) (Relling, Akcay et al. 2018). The above obtained results for $KORA_{2013}$ could be replicated. For instance, loss of daily oscillations in richness and *alpha*-diversity was associated with a significant reduction of rhythmic OTUs in T2D (1.4% rOTUs) compared to nonT2D (8.5% rOTUs) (**Figure 62**).

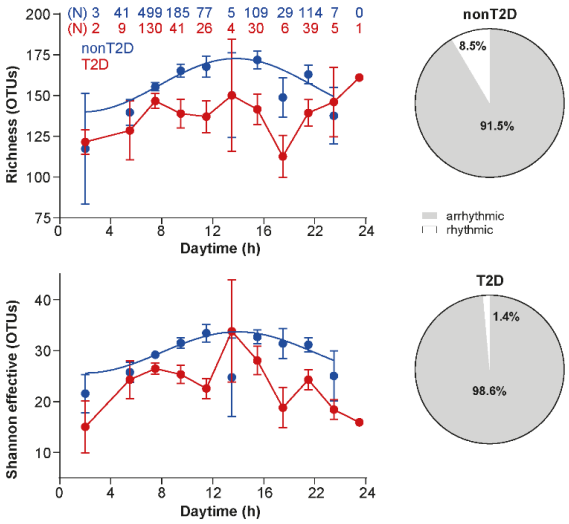


Figure 62 Circadian analysis of the FoCus cohort.
Diurnal profiles of *alpha*-diversity in 1,363 subjects. Significant rhythms are illustrated with fitted cosine-wave curves (cosine-wave regression, p-value ≤ 0.05, non-significance is shown by straight lines between data points. Right panel: Pie chart showing the amount of rhythmic (white) and arrhythmic (grey) OTUs and their relative abundance in percent.

Based on a BLAST search, we determined which OTUs from the FoCus cohort were reflecting the s-arOTUs of the $KORA_{2013}$ cohort. Their differences in relative abundances between nonT2D and T2D were examined (**Figure 63**, **Table 16**). A possible explanation for differences in relative abundance between OTUs of $KORA_{2013}$ and the assigned s-arOTUs of FoCus could be the limitation of the short amplicon sequenced and, subsequent, limited BLAST results. As seen in **Table 10**, some differences in the taxonomic classification of the assigned OTUs causing some discrepancy in species classification. In combination with BMI, the relative abundance values of FoCus were considered as input for the

imported $KORA_{2013}$ GLM, classifying T2D with an AUC of 0.76 and values for sensitivity and specificity of 75% and 69%, respectively.

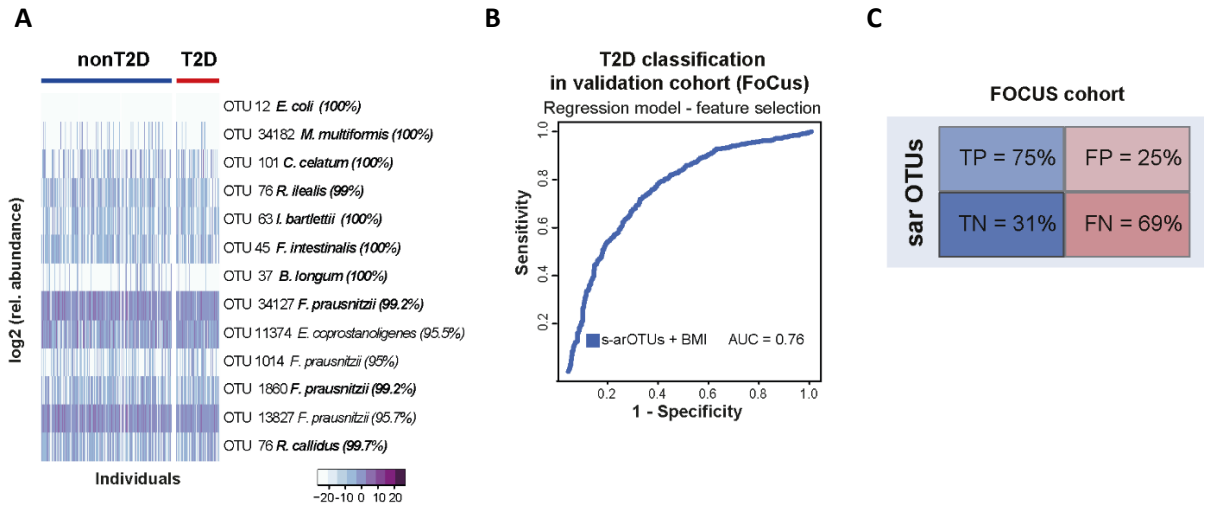


Figure 63 Performance of s-arOTUs in the classification of T2D in FoCus.
(A) Heatmap showing the assigned OTUs for the s-arOTUs based on BLAST search. Individuals are grouped according to diabetes status. Taxonomies are determined by EzBioCloud database search, best matches are shown beside each OTU with the designated taxonomic name and identity value in brackets. **(B)** ROC curve for s-arOTUs' GLM model trained on the KORA cohort in 2013 used to classify T2D in FoCus. **(C)** Confusion matrix showing the performance of the KORA model applied to FoCus.

The confusion matrix of the results obtained by the GLM also showed that 75% of the T2D cases were correctly predicted. However, the model lacked in performance to identify T2D cases correctly, with a large number of false negatives (**Figure 63 C**).

Table 16 Results of BLAST search for s-arOTUs in the FoCus cohort.

KORA s-arOTUs	Species name	Similarity	Diff/Total	Completeness (%)
OTU_101	*Clostridium celatum*	100	0/402	100
OTU_1014	*Faecalibacterium prausnitzii*	95.02	20/402	99.9
OTU_11374	*Eubacterium coprostanoligenes*	95.53	18/403	99
OTU_12	*Shigella flexneri*	100	0/427	100
OTU_13827	*Faecalibacterium prausnitzii*	95.77	17/402	99.9
OTU_1860	*Faecalibacterium prausnitzii*	99.25	3/402	99.9
OTU_3247	*Ruminococcus callidus*	99.75	1/402	100
OTU_34127	*Faecalibacterium prausnitzii*	99.25	3/402	99.9
OTU_34182	*Mesosutterella multiformis*	100	0/427	100
OTU_37	*Bifidobacterium longum* subsp. *longum*	100	0/407	100
OTU_45	*Faecalibacillus intestinalis*	100	0/426	97.4
OTU_63	*ntestinibacter bartlettii*	100	0/401	100
OTU_76	*Romboutsia ilealis*	99	4/401	100
OTU_101	*Clostridium celatum*	100	0/402	100

The reduced performance of the $KORA_{2013}$ s-arOTUs classifying T2D in FoCus could be due to regional differences, which were already shown to have an impact on overall bacterial composition of the gut (**Figure 31**) (He, Wu et al. 2018).

4.8.2. Prediction of incident T2D cases in $KORA_{2018}$

Taking advantage of the prospective sampling in $KORA_{2018}$, we aimed to predict incident T2D cases. Microbial data from $KORA_{2013}$ were used in the GLM s-arOTU model to differentiate between nonT2D and individuals under risk. Since samples were taken from 699 individuals again in 2018, it was possible to assign newly diagnosed T2D cases (N = 20 subjects). Using these, we were able to evaluate the predictive power of the model. To assign the s-arOTU from $KORA_{2013}$ to corresponding OTUs in $KORA_{2018}$, a BLAST search was performed, mapping the sequences of s-arOTUs all OTUs of the $KORA_{2018}$ subcohort. The top hits were further verified by EzBioCloud ensuring correct assignments (**Table 17**).

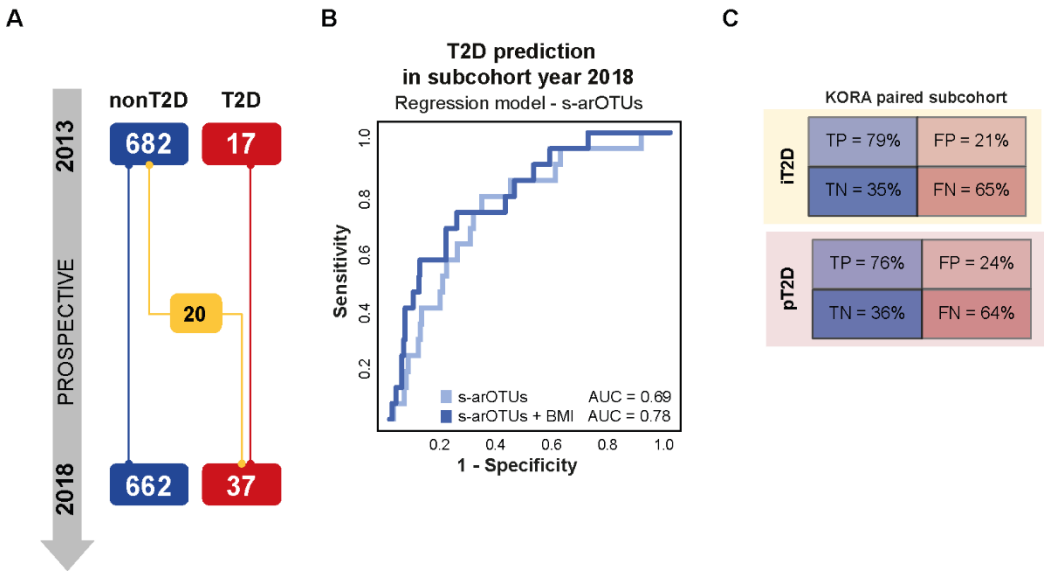


Figure 64 Arrhythmic microbial signature for prediction of T2D.
(A) Predictive analysis of T2D in the paired $KORA_{2018}$ subcohort (samples from 2013, top, and 2018, bottom). **(B)** The prediction of incident T2D cases (iT2D) is based on the 13 s-arOTUs +/– BMI in the regression model to the right. Baseline data is from nonT2D individuals from 2013. Endpoint is iT2D in 2018 (N = 20 subjects). **(C)** Confusion matrices for the classification and prediction of T2D. Upper matrices show the True Positive (sensitivity) and True Negative values (blue) as well as the False Positive and False Negative (specificity) values (red) for the prediction of incident T2D (iT2D). Lower matrices show the values for the classification of T2D (pT2D).

The model reached an AUC of 0.69 for s-arOTUs and 0.79 for s-arOTUs + BMI. Based on a microbial risk signature of 13 OTUs, the majority of persisting T2D cases (pT2D, N = 17 subjects) and incident T2D cases (iT2D, N = 20) were determined, resulting in a sensitivity of 79% and 76% for pT2D and iT2D, respectively. Specificity was high, but resulted in many false positives, which is a limitation (specificity pT2D = 35%; specificity iT2D = 35%). Thus, nonT2D individuals were wrongly assigned to be under risk

(**Figure 64 C**). The lack in specificity was already apparent in the classification of T2D in the $KORA_{2013}$ data and became even more obvious for the FoCus cohort.

Table 17 Results of BLAST search of s-arOTUs and matched $KORA_{2018}$ subcohort

KORA s-arOTUs	Species name	Similarity	Diff/Total	Completeness (%)
OTU_45	*Faecalibacillus intestinalis*	99.56	2/454	97.4
OTU_12	*Shigella flexneri*	99.34	3/455	100
OTU_76	*Romboutsia sedimentorum*	98.6	6/429	94
OTU_63	*Intestinibacter bartlettii*	99.3	3/429	100
OTU_37	*Bifidobacterium longum* subsp. *longum*	99.31	3/435	100
OTU_1014	*Faecalibacterium prausnitzii*	97.44	11/430	99.9
OTU_101	*Clostridium celatum*	99.07	4/430	100
OTU_3247	*Faecalibacterium prausnitzii*	96.74	14/430	99.9
OTU_1860	*Faecalibacterium prausnitzii*	99.07	4/430	99.9
OTU_13827	*Faecalibacterium prausnitzii*	96.05	17/430	99.9
OTU_34182	*Shigella flexneri*	99.34	3/455	100
OTU_11747	*Faecalibacterium prausnitzii*	96.28	16/430	99.9
OTU_11374	*Faecalibacterium prausnitzii*	96.28	16/430	99.9
OTU_34127	*Faecalibacterium prausnitzii*	96.28	16/430	99.9

4.8.3. Validation of arrhythmic risk signature in the TwinsUK cohort

TwinsUK is a cohort studies in Great Britain thus, deviations should be pronounced compared to the above mention FoCus cohort, since geography influences the results. For TwinsUK, the prediction was conducted on baseline OTU relative abundance values from 1,399 individuals with known T2D status (nonT2D, N = 1,259; pT2D, N = 94; iT2D, N = 46). As before for FoCus, s-arOTUs of $KORA_{2013}$ taken as query for a BLAST search in the TwinsUK-OTUs and top hits were verified by EzBioCloud. However, in TwinsUK, the V4 16S rRNA gene region was targeted for sequencing (~250bp amplicon length) which complicated the correct assignment of corresponding s-arOTUs (**Table 18**). Still, both cohorts' sequencing data overlapped in the V4 region. To avoid using misclassified OTUs we further checked the V4 sequences in EzBioCloud verifying taxonomic assignment. Next, the relative abundance of the selected OTUs between nonT2D and T2D for TwinsUK were analyzed (**Figure 65**).

Table 18 Results of BLAST search of s-arOTUs and TwinsUK OTUs

KORA s-arOTUs	TwinsUK	Name	Similarity	Diff/Total n	Completeness
OTU_76	OTU_63	*Romboutsia ilealis*	98.76	3/242	94
OTU_63	OTU_35	***Intestinibacter bartlettii***	100	0/242	100
OTU_45	OTU_24	*Faecalibacillus intestinalis*	100	0/242	97.4
OTU_37	OTU_27	***Bifidobacterium adolescents***	100	0/243	100
OTU_34182	OTU_4	*Escherichia coli*	100	0/243	98
OTU_34127	OTU_6054	***Faecalibacterium prausnitzii***	97.94	5/243	99.9
OTU_3247	OTU_2975	*Faecalibacterium prausnitzii*	96.3	9/243	99.9
OTU_1860	OTU_1555	***Faecalibacterium prausnitzii***	97.12	7/243	99.9
OTU_13827	OTU_7	*Faecalibacterium prausnitzii*	98.77	3/243	99.9
OTU_12	OTU_8045	***Escherichia coli***	96.69	8/242	98
OTU_11374	OTU_2639	*Faecalibacterium prausnitzii*	95.47	11/243	99.9
OTU_1014	OTU_2325	***Faecalibacterium prausnitzii***	95.06	12/243	99.9
OTU_101	OTU_59	*Clostridium celatum*	100	0/243	100

Based on the previously trained GLM of $KORA_{2013}$, the model was able to classify T2D (pT2D) with an AUC of 0.68 and predict T2D (iT2D) with an AUC of 0.69 using the risk signature of $KORA_{2013}$. Again, the number of false positives was high, which was reflected by a poor sensitivity of 30% for pT2D and 20% for iT2D. The true positive rate for iT2D was 56% and, thus slightly higher for pT2D (sensitivity = 63%).

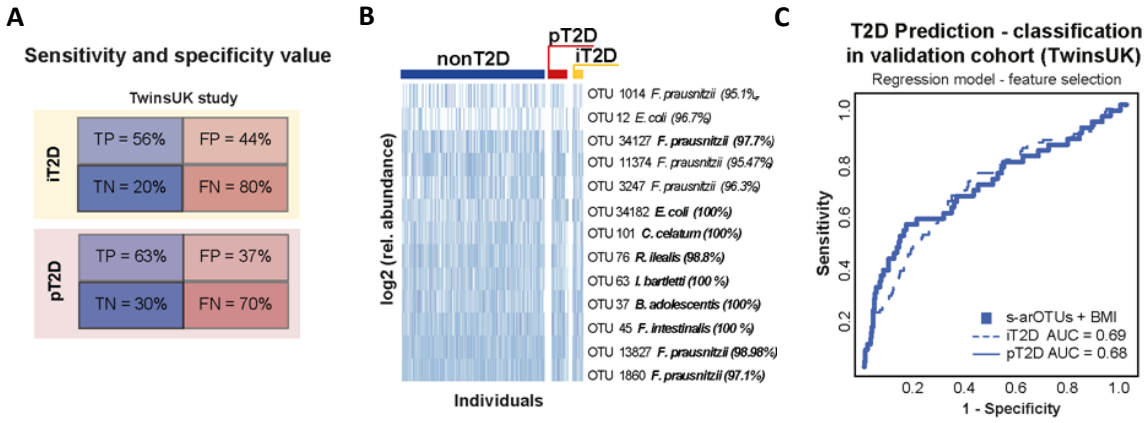


Figure 65 Validation of risk signature for predication and classification in the TwinsUK cohort.
(A) Confusion matrix for the classification and prediction of T2D. Upper left matrix shows the True Positive (sensitivity) and true negative values (blue) as well as the false positive and False Negative (specificity) values (red) for the prediction of incident T2D (iT2D) in the paired KORA subcohort. Lower matrix shows the values for the classification of T2D (pT2D). **(B)** Heatmap showing the log-transformed relative abundance of the TwinsUK OTUs (y-axis) identified as representatives of the a-sprouts. Individuals (x-axis) are grouped according to their T2D status. Taxonomic classification of OTUs is shown by their species names and 16S rRNA sequence similarity (%), bold indicates ≥97% similarity. **(C)** Classification (pT2D, straight blue line, AUC = 0.69) and predictive analysis of T2D (iT2D, dashed blue line, AUC = 0.68) in independent TwinsUK cohort. The prediction of incident T2D cases (iT2D) is based on the 13 s-arOTUs +/- BMI.

We showed that the $KORA_{2013}$ based bacterial risk signature and GLM model could be transferred to three cohorts ($KORA_{2013}$, FoCus and TwinsUK) to classify and predict T2D. Nevertheless, it was not possible to obtain similar results in all three cohorts.

4.9. Functional analysis of arrhythmic bacterial risk signature and its association with T2D

Shotgun metagenomic sequencing for the analysis of the bacterial composition on a strain level was conducted by Prof. Paul O'Tool's group from the School of Microbiology and APC Microbiome Ireland, University College Cork, Ireland. This allows to perhaps classify bacteria in beneficial and harmful, since each diabetes status might overrepresent certain taxa. This could allow to understand the underlying functional mechanism of the microbes in the gut and, thus, co-occurrence of certain bacteria might link health with causality.

A subset of individuals from the KORA cohort (N = 50 subjects) with data from both years, 2013 and 2018, were selected for metagenomic sequencing (n = 100 samples). The individuals (**Table 19**) had been classified as follows:

- No metabolic syndrome in year 2013 and 2018 (nonT2D)
- No metabolic syndrome in year 2013, but developed T2D in 2018 (incident T2D, iT2D)
- prediabetic in year 2013 without developing T2D in 2018 (nonT2D)
- prediabetic in year 2013 with developing T2D in 2018 (iT2D)
- T2D in year 2013 and 2018 (remaining T2D, rT2D)

After conducting metagenome sequencing and analyzing the data, 16 bacterial metabolic pathways were identified to be significantly different (p-value ≤ 0.1) in gene amounts between nonT2D and pT2D. Of those, 11 were shown to be increased in pT2D (**Supplementary table 4**). Only seven pathways were significantly different between iT2D and nonT2D (

Supplementary table *5*), whereas only three were increased in iT2D without any overlaps between pT2D and iT2D.

Assuming that different pathways are associated with a different stage of the diabetes. It has to be noted that the number of samples within one comparison group (pT2D or iT2D) was low which could be one possible explanation for the lack of overlaps. Significant differences found in one-to-one comparisons were not detected anymore after adjusting for multiple comparisons.

Table 19 Samples selected for Metagenome Shotgun Sequencing

	Year	no.	mean Age	Gender	BMI	Mean ± SD HbA1c
nonT2D	2013	8	57	50% male	50% obese	5.41 ± 0.31
	2018	8	62		50% obese	5.42 ± 0.30
iT2D	2013	20	59	60% male		6.06 ± 0.33
	2018	20	64			7.10 ± 0.86
pT2D	2013	14	62	47% male		7.64 ± 1.19
	2018	14	67			7.75 ± 1.19
Prediabetes	2013	8	59	50% male	50% obese	5.33 ± 0.35
	2018	8	64		50% obese	5.755 ± 0.36
TOTAL		100				

Next, a random forest based approach was applied to choose the most important pathways differentiating T2D from nonT2D. After multiple nested and cross-validated random forest analyses, a list of 30 pathways was considered for further analysis. Based on these 30 selected pathways, the model was able to distinguish between T2D and nonT2D with a mean AUC of 0.81 (**Figure 66 A, B**). To show the contribution of the 30 selected pathways towards T2D classification, a random set of 30 pathways, excluding the 30 previous selected ones, was generated. This random set of pathways was using in a random forest approach as before. However, this resulted in a significant reduction of the AUC to 0.60 ($P < 8 \cdot 10^{-10}$) (**Figure 66 B**). In order to address the compositionality of microbial data (Tsilimigras and Fodor 2016) , we repeated the above steps using a centered log-scale transformation of the pathway abundances for the 30 originally selected pathways above. The pattern remained invariant, with 27 of the 30 pathways retained (**Figure 66 B**). This indicates a strong association between these pathways and T2D and included pathways for the metabolism of amino acids (phenylalanine, cysteine, methionine, alanine, glutamine and aspartate), aromatic compounds (toluene, fluorobenzoate, chlorocyclohexane), and fatty acids (α-linoleic acid, riboflavin) (**Figure 66 C ,Table 20**). Some of the pathways were significantly different in relative abundance between nonT2D and T2D.

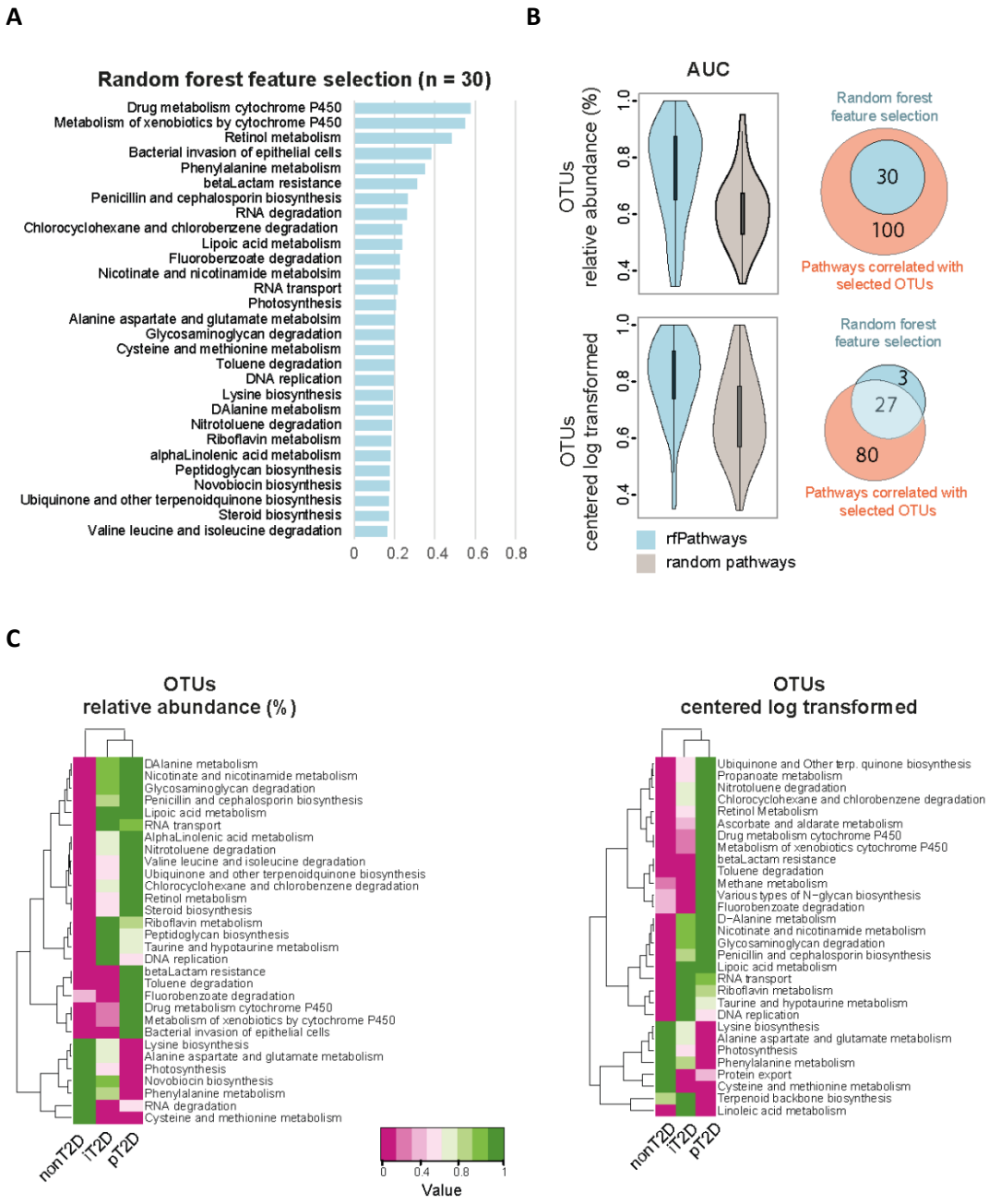


Figure 66 Random forest model for classification of T2D with metabolic pathways.
(A) Left panel, list of top 30 optimal marker pathways along with their feature importance score. The feature importance scores increased sharply for the top 30 pathways and decreased linearly for all pathways below the top 30. Consequently, the top 30 pathways were identified as the optimal set of functional markers here. (**B)** Left panel, comparison of AUC distribution of iterative random forest models obtained by including only the top 30 pathways and excluding the top 30 pathways. Models incorporating only the top 30 pathways could predict T2D individuals from controls with a mean AUC of 0.81. Excluding these pathways resulted in a significant decrease ($P < 3 \cdot 10^{-9}$) to mean AUCs around 0.6. In total, 130 KEGG pathways are significantly associated with at least one of the selected OTUs. Comparing the random forest selected KEGG pathways with the risk signature associated pathways, shows a 100% overlap. The right panel shows the results for log-scaled transformed data for compositionality analysis. (**C**) Left heatmap shows the associated of KEGG pathways and T2D. After stratifying the individuals into nonT2D, iT2D and T2D it is shown that some of the pathways are associated with nonT2D while other are more prevalent in T2D. Right heatmap shows the results for log-scaled transformed data for compositionality analysis.

To validate the determined bacterial taxa of the s-arOTUs and to correctly classify the OTUs on strain level, a correlation analysis of OTU relative abundance values against shotgun sequences was conducted (**Table 17**). Some previously assigned taxonomies were henceforth adapted (**Table 21**). Twenty-six of the originally chosen 30 pathways associated with T2D were found to correlate (p-value ≤ 0.1) with at least 2 of the 13 previously identified predictive s-arOTUs. It could be shown that T2D-associated pathways significantly correlate with the selected arrhythmic OTU list of 13 s-arOTUs, strengthens the importance of these s-arOTUs in T2D classification. This finding also connects taxa (i.e., OTUs) with functionality (i.e., genes from metagenomes).

Table 20 Functional Pathways in association with T2D.

	P-value	Disease	Corrected P-value	Mean in T2D	Mean in nonT2D
Retinol metabolism	0.003	T2D	0.094	0.0002	0.0001
Drug metabolism cytochrome P450	0.005	T2D	0.156	0.0002	0.0002
Metabolism of xenobiotics by cytochrome P450	0.005	T2D	0.156	0.0002	0.0002
Bacterial invasion of epithelial cells	0.014	T2D	0.373	0.0000	0.0000
Penicillin and cephalosporin biosynthesis	0.018	T2D	0.466	0.0001	0.0000
betaLactam resistance	0.020	T2D	0.511	0.0004	0.0003
Lipoic acid metabolism	0.020	T2D	0.511	0.0012	0.0009
Fluorobenzoate degradation	0.023	T2D	0.533	0.0001	0.0001
Lysine biosynthesis	0.025	nonT2D	0.550	0.0141	0.0147
DAlanine metabolism	0.046	T2D	0.960	0.0018	0.0016
Alanine aspartate and glutamate metabolism	0.047	nonT2D	0.960	0.0210	0.0219
Ubiquinone and other terpenoidquinone biosynthesis	0.049	T2D	0.960	0.0032	0.0026
Phenylalanine metabolism	0.049	nonT2D	1	0.0030	0.0034
Riboflavin metabolism	0.063	T2D	1	0.0041	0.0038
AlphaLinolenic acid metabolism	0.063	T2D	1	0.0002	0.0001
Photosynthesis	0.067	nonT2D	1	0.0101	0.0108
Valine leucine and isoleucine degradation	0.072	T2D	1	0.0035	0.0032
Steroid biosynthesis	0.076	T2D	1	0.0001	0.0000
Cysteine and methionine metabolism	0.082	nonT2D	1	0.0175	0.0180
Glycosaminoglycan degradation	0.084	T2D	1	0.0014	0.0011
Nitrotoluene degradation	0.087	T2D	1	0.0009	0.0007
Nicotinate and nicotinamide metabolism	0.119	T2D	1	0.0083	0.0080
Novobiocin biosynthesis	0.119	nonT2D	1	0.0030	0.0031
RNA transport	0.122	T2D	1	0.0014	0.0013
Toluene degradation	0.163	T2D	1	0.0027	0.0024
Chlorocyclohexane and chlorobenzene degradation	0.175	T2D	1	0.0001	0.0001
RNA degradation	0.202	nonT2D	1	0.0068	0.0069
Peptidoglycan biosynthesis	0.212	T2D	1	0.0122	0.0119
DNA replication	0.327	T2D	1	0.0087	0.0087
Taurine and hypotaurine metabolism	0.551	T2D	1	0.0031	0.0030

The 30 pathways found to correlate with T2D classification clustered into two groups. The first group is associated with *F. prausnitzii* (G1) and the other shows stronger correlation with *E. coli* (G2). Next, the 13 s-arOTUs were correlated with clinical markers. It turned out that the G1 OTUs are negatively correlated with HbA1c, HOMA and glucose values. In contrast, the G2-related OTUs are positively correlated with diabetes markers and increased presence of *E.coli* (**Figure 67**).

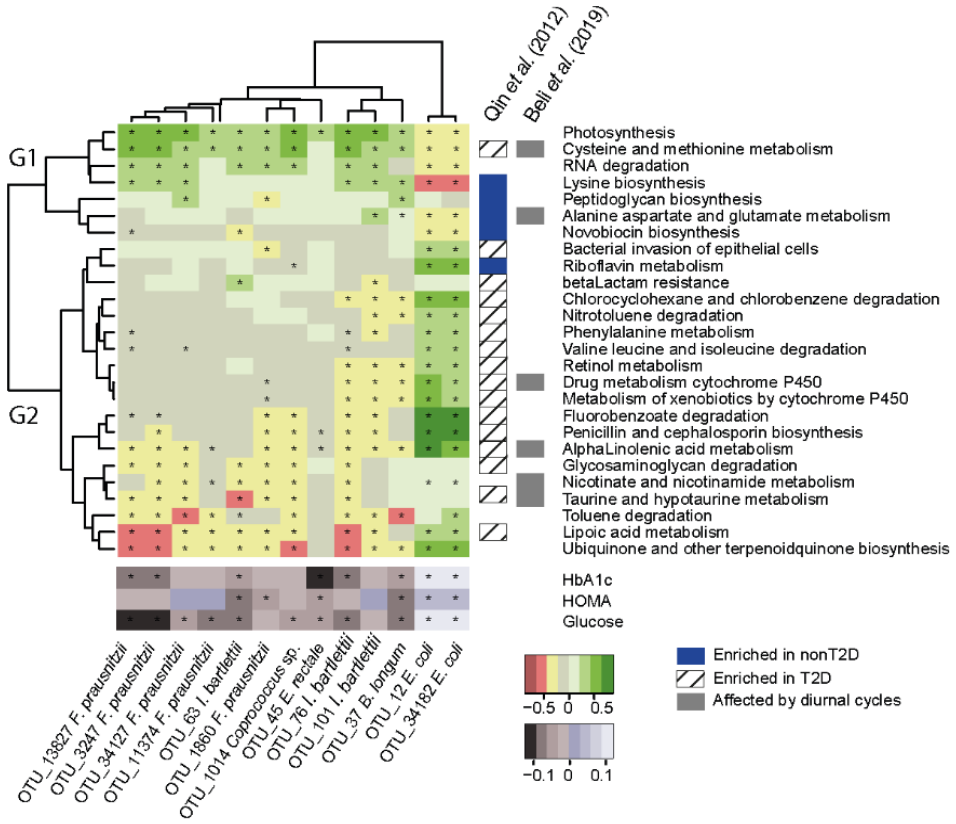


Figure 67 Metabolic pathways and arrhythmic risk signature.
The heatmap shows the Spearman correlations for 26 disease-predictive microbial pathways within the 13 OTUs becoming arrhythmic in T2D (determined from 100 shotgun sequenced samples). Only pathways that were significantly correlated with at least one of the 13 arrhythmic OTUs are shown. The heatmap on the bottom left shows the association pattern for the 13 OTUs with clinical markers that characterize T2D (determined from the entire cohort). For both heatmaps, significant associations were corrected using the Benjamini-Hochberg procedure. Corrected p-values ≤ 0.1 are indicated with * in each field. On the right, a one-dimensional heat-plot shows relative representations of each of the previous pathways observed in the cohort of Qin et al. (2012), either enriched in T2D (dashed) or enriched in nonT2D (blue). Furthermore, pathways predicted to be influenced by diurnal cycles in Beli et al. (2019) are indicated in grey if affected.

Strong correlations were observed between *E. coli* and xenobiotic metabolism pathways, the latter of which were also negatively associated with short-chain fatty acid biosynthesis as well as metabolism of co-factors and vitamins. The taxon-group of *E. coli* was negatively associated with the taxon-group frome by *F. prausnitzii*, the latter of which mostly occurred in combination with *C. barletti (***Figure 67***)*.

In contrast, *B. longum* was not associated with the presence of other taxa, but negatively correlated with xenobiotic biodegradation pathways.

Table 21 Taxonomic classification of s-arOTUs by metagenomic shotgun sequencing data

s-arOTUs	Metagenomic classification	Species name	Pearson correlation
OTU_76	OTU_102	*Clostridium bartlettii*	0.482
OTU_101	OTU_105	*Clostridium bartlettii*	0.434
OTU_11747	OTU_1149	*Faecalibacterium prausnitzii*	0.730
OTU_34182	OTU_13	*Escherichia coli*	0.958
OTU_12	OTU_13672	*Escherichia coli*	0.905
OTU_3247	OTU_27295	*Faecalibacterium prausnitzii*	0.721
OTU_37	OTU_37	*Bifidobacterium longum*	0.952
OTU_11374	OTU_4027	*Faecalibacterium prausnitzii*	0.601
OTU_34127	OTU_5	*Faecalibacterium prausnitzii*	0.810
OTU_45	OTU_51	*Eubacterium rectale*	0.365
OTU_1860	OTU_60140	*Faecalibacterium prausnitzii*	0.621
OTU_1014	OTU_66374	*Coprococcus* sp ART55 1	0.434
OTU_63	OTU_75	*Clostridium bartlettii*	0.772
OTU_13827	OTU_9108	*Faecalibacterium prausnitzii*	0.614

We then determined whether the functional associations identified could be validated in datasets from previous studies. First, we compared differentially abundant genes from the study by Qin et al (Qin, Li et al. 2012), with the relative representation of the corresponding pathways in the T2D and nonT2D individuals of the KORA cohort (identified as either enriched in T2D or nonT2D). Notably, 19 of the 26 pathways could be validated (**Figure 67**). Second, we checked if any of these identified pathways were previously observed showing arrhythmic behaviors (or diurnal cycles) in a recent study published by Beli et al. (2019). Notably, 7 of the 26 pathways identified in the study, including xenobiotic metabolism, cysteine and methionine metabolism, α-linoleic metabolism, and taurine and hypertaurine metabolism, were also reported to undergo diurnal rhythmicity (**Figure 67**). These results show that the metagenomic analysis is able to link functions of the arrhythmic risk signature (s-arOTU) with metabolic features identified in individuals with T2D (**Figure 68**).

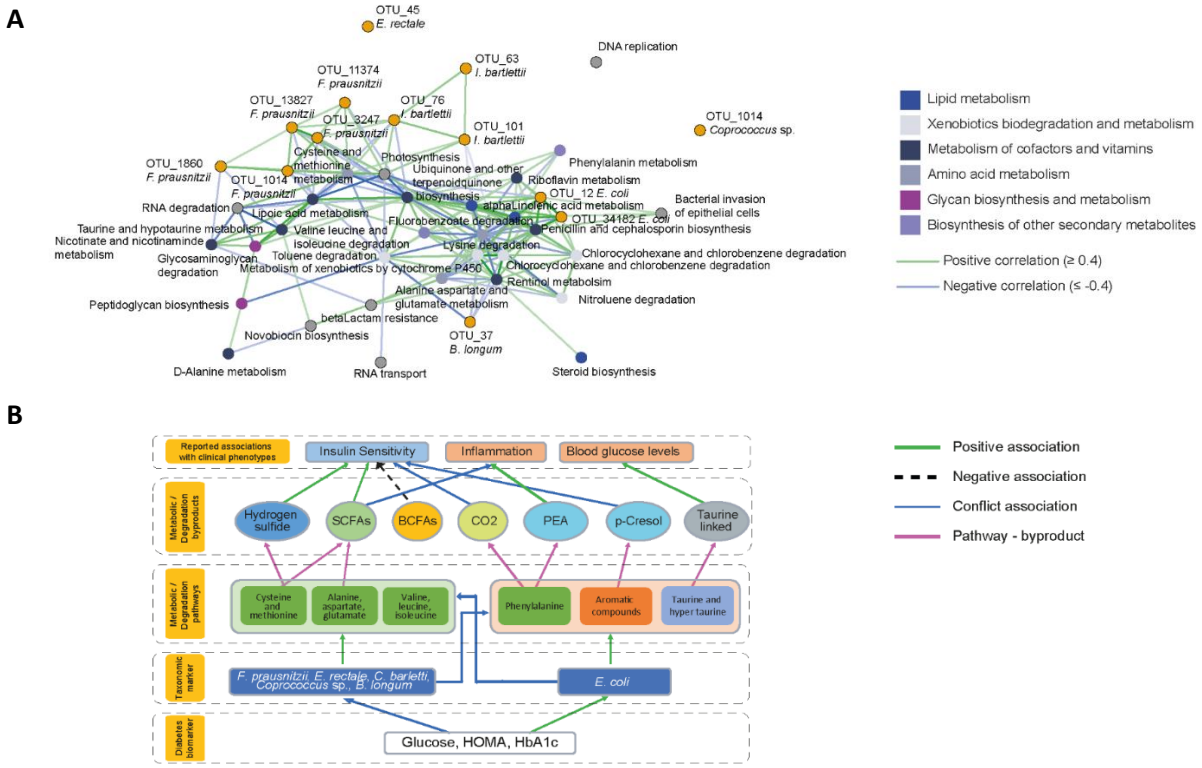


Figure 68 Schematic overview of the associations between metabolic pathways, diabetes marker and arrhythmic risk signature.

Correlation network of HbA1c associated pathways in relation to selected T2D markers (s-arOTUs; orange circles). Green and blue lines indicate a positive and negative association, respectively. The thickness of lines is proportional to correlation coefficients. The color of circles refers to functional classes as indicated.

Schematic flow integrating T2D signature, rhythmicity and microbiome function in the metagenomes. The flow correlates the major disease-associated metabolic pathways identified in the current study, their associations with the disease/health-associated taxonomic markers, the metabolic byproducts originating from these pathways, and the previously known associations of these byproducts with the various clinical phenotypes. The green arrows are positive association, dashed is negative, conflicting associations are shown as blue arrow (BCFA has been shown to be positively associated insulin sensitivity in some paper, while others show no association). Purple-arrows link pathways to their metabolic by-products.

5. Discussion

The main focus of this thesis was to determine a bacterial risk signature for the classification and prediction of T2D. As an initial step, we systematically evaluated DNA extraction, 16S rRNA gene sequencing methods and downstream bioinformatics and statistical analysis tools to comprehensively benchmark the protocols used in this work. In order to avoid the appearance of spurious bacterial taxa in 16S rRNA gene sequencing data an abundance-based filtering cutoff of 0.25% OTU relative abundance was introduced.

Our results showed that under physiological conditions, the gut microbiota follows a circadian rhythm, which is partly disrupted in T2D and obesity. The human gut microbiota is characterized by a large inter-individual variation, but it also shows a dynamic behavior within one individual. This inter-individual variation can be explained by a number of co-variables, including BMI, age, physical activity, vitamin D intake and medication. Our results showed that triglyceride, BMI and geographical region account for most of the variation among the population.

5.1. Benchmarking of DNA extraction and sequencing protocols

The analysis of complex microbial communities by high-throughput 16S rRNA gene amplicon sequencing has become very popular. Taking advantage of the rapid development of next generation sequencing it is possible to analyze the human gut microbiota of thousands of individuals in a short time frame and with relatively low costs. However, there is still a substantial gap between the claim of high quality data as well as complex study question which needed to be addressed and the limited expertise available in many users' laboratories (Hamady and Knight 2009, Human Microbiome Project 2012, Avershina and Rudi 2015, Tsilimigras and Fodor 2016, Gloor et al. 2017). As already shown by others before (Hiergeist, Reischl et al. 2016), we also confirmed that the data generated by targeted sequencing is strongly affected by sample preparation methods (e.g. DNA isolation, PCR cycles, etc.) as well as by applied bioinformatical and statistical methods (Salter, Cox et al. 2014, Schirmer, Ijaz et al. 2015, Clavel, Lagkouvardos et al. 2016). In particular, the method for DNA isolation was shown to have a strong influence on the sequencing outcome (Claassen, du Toit et al. 2013, Wagner Mackenzie, Waite et al. 2015). Testing for different isolation methods showed that, even with a high amount of bacterial DNA as starting material, some DNA extraction kits or certain methods provided insufficient yield of DNA for 16S rRNA gene sequencing, which could be due to the inability of inactivating inhibitors in the stool sample (Claassen, du Toit et al. 2013, Wesolowska-Andersen, Bahl et al. 2014).We also determined some discrepancies in taxonomy for certain OTUs between different methods, but also within replicates of the same method. This shows that the applied DNA extraction protocols and library

preparation steps influence the results. Therefore, it is important to provide as much information as possible about the applied method to achieve better reproducibility.

The results of the cohorts used for validation, where the samples were prepared from collaborating partners (FoCus and TwinsUK), using different DNA extraction protocols, showed variation in *alpha*-diversity and taxonomic composition, even though the same bioinformatic pipeline was used for the statistical analysis (Lagkouvardos, Joseph et al. 2016, Lagkouvardos, Fischer et al. 2017). In the FoCus cohort a different the DNA isolation protocol was applied (QIAamp DNA stool mini kit, automated on the QIAcube Qiagen). Keeping this in mind a reduction in richness (FoCus of 264 ± 59, $KORA_{2013}$ 384 ± 77) but an increase in Shannon effective number wcould be determined (FoCus of 171 ± 70, $KORA_{2013}$ 118 ± 37). The TwinsUK cohort targeted the V4 region of the 16S rRNA gene which could have had an influence on the taxonomic classification due to different mapping regions of the 16S rRNA gene.

Some of these discrepancies are mainly explained by known microbial influencing factors such as geographical region (He, Wu et al. 2018), health (Karlsson, Tremaroli et al. 2013, Khan et al. 2014, Pascal, Pozuelo et al. 2017, Thingholm, Ruhlemann et al. 2019) and lifestyle (O'Toole and Jeffery 2015, Falony, Joossens et al. 2016, Zhernakova, Kurilshikov et al. 2016), but variations could also be due to the differences in sample preparation protocols as well as methodological differences in targeted sequencing (Goodrich, Waters et al. 2014, Relling, Akcay et al. 2018).

To increase the comparability of different studies and to avoid misinterpretations of the data due to artefacts (Bokulich, Subramanian et al. 2013, Fuks, Elgart et al. 2018), it is further important to filter the sequencing outcome properly (Avershina and Rudi 2015). Filtering 16S rRNA gene sequencing data is not a novel concept for the preprocessing of raw sequencing data. Nevertheless, the commonly used filtering to remove singletons has disadvantages. This filtering is strongly dependent on the actual OTUs and read number (Schloss, Westcott et al. 2009, Caporaso, Kuczynski et al. 2010, Callahan, McMurdie et al. 2016). We showed that the number of generated OTUs dependents on the number of samples included in a study. The appearance of bacteria, which are not part of the microbial community, was still high after singleton removal. An incomplete filtering of spurious OTUs results in a higher number bacterial taxa which are not part of the community which could lead to wrong conclusions e.g. associations with health (Halfvarson, Brislawn et al. 2017). The proposed relative abundance-based filtering cutoff is currently limited to 16S rRNA gene sequencing OTU data from stool samples, targeting the V3V4 region, but future experience will broaden the applicability. In addition, bacterial members of the gut microbiota are determined using ASVs (Callahan, McMurdie et al. 2016) instead of OTUs. The clustering to build ASVs is based on sequences with ≥ 99% similarity and, thus, results in a higher number of ASVs compared to OTUs (≥ 97% similarity), which influences for example the *alpha*-diversity (Callahan, McMurdie et al. 2017). Of note, currently neither of the two short-

amplicon clustering methods should replace the other. Both methods provide results in a synergetic manner and are of high quality, and each method has advantages and disadvantages. The advantage of OTU picking is the biologically meaningful clustering of a sequence similarity of 97%, but this has the disadvantage that sequences are not identical and therefore no real representative sequence can be shown for an OTU. This is different with ASVs, where sequences are first modified by a denoising step and then identical sequences are combined. This results in a database of ASV sequences that can be compared between studies.

Next to preprocessing of sequencing data, the statistical analysis should be evaluated critically. Data tables of OTUs and ASVs should be normalized before statistical analysis are performed. From our experience, we recommend a normalization to a fixed cutoff of 10,000 reads per sample for larger studies. A normalization to a given value will allow the integration of samples with fewer number of reads (< 10,000), but still enough to be considered in the analysis (> 5,000 reads). Samples with a read count around the recommend value of 10,000 will not lose information due to normalization of e.g. the minimum read number assigned in the study (Lagkouvardos, Fischer et al. 2017). The aggregation of OTUs / ASVs with the same taxonomic classification varies according to the database used (e.g. SILVA (Quast, Pruesse et al. 2013), Greengenes (Yokono, Satoh et al. 2018), and RDP (Wang, Garrity et al. 2007)). Disadvantage of binning OTUs with the same taxonomic classifier (e.g, phylum, genus) is the high error rate, especially at lower taxonomic levels, such as genus or strain level (Hamady and Knight 2009, Clavel, Lagkouvardos et al. 2016) and the missing overlap in nomenclature in different database (Edgar 2018). The use of relative abundance instead of absolute counts could also lead to wrong interpretations. Data transformation (Tsilimigras and Fodor 2016) or the add-in of spike communities (Stämmler et al. 2016) may be one possible solution to deal with this problem.

For the analysis of microbial differences between groups, a multidimensional scaling plot is one of the most commonly used illustration tools. Nevertheless, the applicability of two-dimensional scaling methods is limited to data sets of about 300 samples at maximum, due to the linear increase in dimensions (which all need to be projected on the two dimension of a paper or screen surface) with an increase in sample size. Thus, above about 300 samples, a similarity tree is more appropriate representation, which should be used.

Taken together, a standardized and well documented and evaluated pipeline increases transparency, enables reproducibility, and generates the high-quality outcome needed.

5.2. The association of changes in the bacterial gut composition with health and co-variables

This work provided novel insights to the composition of the human gut microbiota implementing three large population-based cohort studies. Based on more than four thousand individuals, it was evident that the composition of the human gut microbiota is largely influenced by multiple factors which drive changes in the abundance of the presence of bacteria within the gut (including age, BMI, health, and medication). The human gut microbiota showed a high degree of variability between the different human cohorts, but also between individuals within one cohort, as well as within one individual. The strong inter-individual variation makes it difficult to draw conclusions based on the appearance and abundance of certain bacteria, especially without consider additional factors. Thus, it is important to identify variables influencing the stability and function of the gut bacterial ecosystem. One of these mechanism seemed to be the circadian clock. Previous studies showed that the composition is influenced by human genetics (Goodrich, Waters et al. 2014), environment (Falony, Joossens et al. 2016, Zhernakova, Kurilshikov et al. 2016, He, Wu et al. 2018), and host health (Turnbaugh, Ley et al. 2006, Qin, Li et al. 2012, Le Chatelier, Nielsen et al. 2013, Thingholm, Ruhlemann et al. 2019). We confirmed the contribution of above-mentioned co-variables to the overall taxonomic composition, nevertheless, a large amount of approximately 90% of the variation remained unexplained.

The major strength of the *enable* cohort was its large number of co-variables assessed, including several bio samples (e.g. blood) as well as information related to dietary habits. Taking all the available covariables in consideration the proportion of explained variation increased from 9.1% ($KORA_{2013}$ cohort) to 16.27% (*enable* cohort). Combined results of the three cohorts identified a set of variables, including age, BMI, insulin and glucose values, but also variables related to lifestyle and physiology of the body e.g. time of defecation, impact the microbiome's composition. In a study from an Israel cohort, 20% of bacterial diversity could be explained by factors related to diet, drugs and anthropometric measurements (Rothschild et al. 2018). This was also confirmed in the study from Falony et al. with a combined effect size of 16% highlighting again the association of covariates and microbiota composition (Falony, Joossens et al. 2016). It was shown that medication explained the largest variance with over 10%. A study from the United States showed that sociodemographic variables e.g. national origin and geographic region explained around 4% (Kaplan et al. 2020). The strong effect of geography was also seen in a Chinese cohort, where the host location was the strongest explanatory factor (approx. 9%) (He, Wu et al. 2018). Overall, there are some effect modifiers (e.g. BMI, alcohol intake, age, gender, blood glucose), which are overlapping in all population-based cohort studies. Overall, the results of our study as well as from previously described large population-

based cohort studies provided profound knowledge about influencing factors (Qin, Li et al. 2010, Goodrich, Waters et al. 2014) and diet associated pattern (Clarke, Murphy et al. 2012, David, Maurice et al. 2014, Kaczmarek, Musaad et al. 2017).

In addition to the taxonomic description, we addressed the dynamic and stability of the gut microbiota by including longitudinal data. For the individual S1, 58 samples were collected and analyzed to investigate changes over time. The bacterial composition of the gut seemed to react towards many environmental changes, showing continuous fluctuation over time. Nevertheless, it was not possible to determine what causes these fluctuations. For instance, the intake of probiotics did not result in significant changes compared to other samples, collected before and after this intervention. Differences on genus level could possibly be linked to stress and changes in dietary habits. Generally, the dynamics of the gut microbiota seemed to be specific for individuals and varied significantly between different individuals. For instance, the longitudinal analysis for two individuals showed that there was a high variability in bacterial compositional in subject S1, while the analysis of subject S2 showed less pronounced variability. This was also confirmed using a time series analysis of 93 subjects from the enable cohort, where four consecutive fecal samples were collected. Some of them showed large variation, while others were relatively stable in their microbial composition. Other studies had detected a comparable variation within an individual over time (Flores, Caporaso et al. 2014, Halfvarson, Brislawn et al. 2017, Zhou, Sailani et al. 2019). It was suggested that the intra-individual variation could dependent on health status e.g. a less stable microbial composition in individuals diagnosed with IBD (Halfvarson, Brislawn et al. 2017).

But even after numerous studies trying to describe the function, variability, and influencing factor of the gut microbiome and its influence towards host health, research into the gut microbiome is still in its infancy. A strong global variation in bacterial composition, a lack in commonly shared species including methodological settings as well as possible association with co-factors, increases the complexity gut microbiome analysis.

5.3. Functional pathways associated with circadian rhythmicity and metabolic health

Oscillations of the gut microbiome were shown to be absent in mice with a genetically dysfunctional circadian clock (Thaiss, Zeevi et al. 2014, Liang, Bushman et al. 2015), indicating that a functional circadian clock of the host is required to maintain rhythmicity of microbiota composition.

The rhythm observed in the two main phyla could be associated with food intake. The peaking hours of Bacteroidetes are in the early morning and late evening/night. Bacteroidetes is known to be associated with gene responsible to break down carbohydrates from dietary sources (Turnbaugh, Ley et al. 2006). The absorption process of carbohydrates takes around 3 - 6 hours which could be a possibly explanation for the reduced relative abundance of Bacteroidetes during the day. Intracellular transport of e.g., sugars is associated with the phyla Firmicutes and thus, could be a possible explanation for decreased abundance during the afternoon/evening whereas the absorption of carbohydrates is complete (Louis et al. 2007).

It was also shown that circadian signals from the microbiota affect diurnal rhythmicity of histone acetylation in intestinal epithelial cells controlling metabolic responses in the host (Kuang et al. 2019). These previous findings supported our hypothesis that circadian rhythms in microbiota-host interactions contribute to metabolic homeostasis of the host. Accordingly, loss of rhythmicity of bacterial taxa in T2D subjects, identified in this study, likely resulted in arrhythmicity of their metabolic products. Incorporating, shotgun metagenomic analysis identified 26 microbial pathways associated with ‘xenobiotic’, ‘branched-chain amino acids’, ‘fatty acids’, as well as ‘taurine metabolism’, forming a functional link between the diurnal oscillation of bacteria in the gut and metabolites. Branched-chain amino acids were previously documented to follow circadian rhythmicity in blood samples from humans kept under 40 hours constant routine, but rhythmicity is lost in subjects with T2D (Skene et al. 2018). However, 19 of 26 pathways (73%) identified in the metagenomic analysis were validated in an independent cohort of T2D individuals (Qin, Li et al. 2010), suggesting that the functionality of arrhythmic microbiota and thus sampling time points are relevant biomarkers. Notably, among pathways, several were linked to the metabolism of certain amino acids and degradation of aromatic compounds. Positive associations with health-associated taxonomic markers microbes and the amino-acid metabolism of alanine, aspartate, glutamate, and cysteine have been found before. The major (by-)products of the microbial fermentation of alanine, aspartate, glutamate and cysteine are short-chain fatty acid (SCFAs) and hydrogen sulfide (H_2S), respectively (Oliphant and Allen-Vercoe 2019). While SCFAs are known for their health benefits including amelioration of insulin resistance, H_2S has been suggested as a positive regulator of insulin sensitivity especially (Khan, Nieuwdorp et al. 2014). In contrast, a negative association with metabolic health was observed for ‘phenylalanine metabolism’,

which had been attributed to phenylethylamine and carbon dioxide in the past (Khan, Nieuwdorp et al. 2014, Oliphant and Allen-Vercoe 2019). Similar, p-cresol, derived from the degradation of aromatic compounds as toluene-related substances, was suggested to be a negative regulator of insulin sensitivity (Koppe et al. 2013). Interestingly, there were 9 OTUs that gained rhythmicity in T2D when patients were treated with metformin. Seven of those OTUs were assigned to the genus *Collinsella*. Thus, suggesting that species that recovered rhythmicity with metformin treatment are associated with metabolic pathways and genes.

Taken together, the signature of arrhythmic bacteria did not only contribute to the classification and prediction of T2D, but also suggests a functional link between circadian rhythmicity and the microbiome in metabolic diseases.

5.4. Evaluation of the predictability and transferability of the arrhythmic bacterial risk signature

In the present work, we were able to determine a bacterial signature, which not only acts as a classifier but also works as a predictor for T2D. Using the prospective data of KORA (sampling in 2013 and 2018), we evaluated the predictability of T2D with the previously defined risk signature. Indeed, T2D was correctly predicted with an accuracy of 78%. However, the application of this bacterial risk signature to another cohort (TwinsUK) was only partly successful. The power to predict T2D was lost for the TwinsUK cohort. Regional differences in individual microbiota profiles and sampling time possibly affect the identification of arrhythmic bacteria in T2D and the applicability of risk signatures. These data emphasizes the need to acquire the time points of stool sampling (i.e. time of defecation).

The classification of disease had a large effect on the outcome. For instance, the applicability of T2D risk signatures was aggravated by the different diabetes classifiers used in different studies. In the prospective part of the $KORA_{2018}$ cohort, which included re-sampling in year 2018, T2D was only classified based on HbA1c (Holle, Happich et al. 2005), while TwinsUK used fasting blood glucose (Goodrich, Waters et al. 2014), and FoCus the HOMA index (Relling, Akcay et al. 2018). Even though all these measurements are based on values measured in blood and represent accepted biomarkers (Stern, Williams et al. 2005), some of the border lining T2D cases might be misclassified. Despite the extensive efforts to define the role of the gut microbiome in metabolic diseases, especially obesity and T2D, limited reproducibility and specificity of disease-associated taxa, e.g. members of *Christensenellaceae*, *Collinsella* and *Escherichia coli* are also associated with Crohn's disease (Pascal, Pozuelo et al. 2017), across cohorts complicates the identification of microbial risk factors. Differences in the reported arrhythmic taxa between our study and the report from Thaiss et al. (2016) may again be explained by variations in microbiota composition of the different subjects at the individual level

(Falony, Joossens et al. 2016, Zhernakova, Kurilshikov et al. 2016) and regional level (Shin et al. 2016, He, Wu et al. 2018), but also methodological differences in the study design (e.g. DNA isolation, library preparation, primer selection) (Klindworth, Pruesse et al. 2013, Dominianni, Wu et al. 2014, Salter, Cox et al. 2014, Schirmer, Ijaz et al. 2015, Costea et al. 2017, Johnson et al. 2019). The lack of documentation of stool sampling time in prior studies in addition to the well-documented regional and individual differences in microbiota profiles may account for discrepancies between studies. We therefore suggest considering circadian oscillations to better understand the underlying mechanisms of disease-associated microbiome alterations and to validate risk profiles in prospective cohorts.

Differences could be also attributed to the different methods used in the 16S rRNA gene sequencing by differences in the targeted gene region for amplicon sequencing. Validation of the adaptability to independent cohort studies as well as to results obtained by published data was complicated by this variation in sequencing. KORA, FoCus, and enable used V3V4, TwinsUK relied on V4 only, and Thaiss et al. had chosen V1V2. Comparing the sequencing data generated by targeting the V1V2 region of the 16S rRNA gene with the data from targeting the V3V4 region of the 16S rRNA gene showed differences in *alpha*-diversity without correlation in richness or Shannon effective number of species within one individual. We further found no rhythms while targeting the V1V2 V-region in *alpha*-diversity and substantially less rhythmic OTUs, compared to the data obtained from the same individuals but sequenced on the V3V4 region from the same samples. Comparison of two regions targeted (V1V2 and V3V4) showed discrepancies in the results, even though DNA was extracted using the same protocol.

5.5. Longitudinal sample and integrative studies increase knowledge about the bacterial composition of the human gut

To characterize key differences between and within cohorts, we analyzed three large population-based studies (e.g., $KORA_{2013}$, $KORA_{2018}$, *enable*, and FoCus cohort). Using cross-sectional cohorts, we provided evidence that the sampling time point influences the composition of the gut microbiota and showed that metabolic health could be linked to a disrupted circadian rhythmicity in certain bacteria. We further identified functional pathways associated with the arrhythmic microbial species and were able to verify the results from previous studies (Thaiss, Zeevi et al. 2014, Thaiss, Levy et al. 2016, Skene, Skornyakov et al. 2018, Beli, Prabakaran et al. 2019). Nevertheless, cross-sectional data have its limitations especially in the analysis of rhythmicity. To verify the concept of circadian rhythmicity and its association with health, strong consideration should be given to longitudinal study designs include as many sampling time points as possible and/or necessary to provide sufficient

statistical power. Stool sampling from human subjects has its limitations since subjects are rarely able to provide multiple samples throughout the day and night.

Furthermore, metagenomic shotgun analysis provides better insight into the functional capacities of different bacterial groups. Using shotgun metagenomic and strain-level diversity analysis allows the characterization causal relationships between microbial and metabolic dysbiosis and its association with the disease. Further, rare or novel organisms in the community could be identified. The taxonomic classification obtained through shotgun sequencing is more reliable and specific, allowing an increased comparability of results across studies (Heintz-Buschart and Wilmes 2018, Zhang et al. 2019).

6. Concluding remarks

We identified daytime-related oscillations in bacterial abundance of specific intestinal community members and most importantly, diurnal rhythmicity was disrupted in patients with T2D. Nevertheless, 16S rRNA amplicon sequencing is limited in terms of taxonomic resolution and is not able to inform about specific genes and pathways involved in these oscillations.

Therefore, shotgun metagenomic sequencing from large-scale population-based cohorts needs to be performed to generate strain-specific and functional characterization to identify clock-related target genes as well as inter-species diversity of selected rhythmic bacterial species. This aims to identify daytime-dependent bacterial functions and diversity to understand causality of circadian disruption in individuals with T2D. Comparative and integrative analyses of stool samples will provide a powerful tool to assess the phylogenetic, strain and functional diversity of the human gut microbiome. Metabolite profiles from stool will help to further stratify functionally relevant mediators of daytime-specific microbiome-host interactions. However, the diversity and fluctuation of individual microbiomes still hinders the identification of universal risk signatures. In addition to the population-based analysis it is important to follow the individual trajectory of rhythmic bacterial species over time. In this respect longitudinal sampling of healthy and diseased individuals is indispensable.

Our findings clearly highlight the need to consider diurnal changes in the analysis of the gut microbiota composition. Including sampling time as parameter in the analysis of upcoming studies will contribute to the better understanding of the impact of gut associated circadian rhythmicity with health.

To increase applicability of identified risk signatures a regionally restricted core microbiome could be achieved e.g., for Southern Germany. A core microbiome could serve as reference microbiome estimating the microbial health of newly integrated individuals of this area. It would further help to define cluster, associated with certain co-variables and/or functional pathways, and to categorize samples under investigation. We attempt to expand the integrative dataset, by adding more samples with adequate preprocessing and addressing the same target region of the 16S rRNA genes. In addition, dietary recommendations (e.g., amount of fiber) and defined microbial related preventive treatments (e.g., probiotics) should be defined. Suggesting that individuals may differ in their response to certain food the association of bacteria, food and time should be considered as possibility to prevent metabolic diseases like T2D. Towards this end, food-related intervention studies, which are personalized according to each individual's' gut microbial composition would be very important.

7. Supplementary

7.1. Supplementary tables

Supplementary table 1 Overview of different DNA isolation methods.

	Nanodrop (mean ng/µl ± SD)	Qubit (mean ng/µl ± SD)	Position on agarose gel (Figure 4)	DNA protocol/kit
ISO-1	252.2 ± 105.9	85.8 ± 21.9	1 – 3	PSP Spin Stool DNA Plus Kit
ISO-2	595.0 ± 233.7	not available	4 – 6	OneStep™ PCR Inhibitor Removal Kit
ISO-3	45.9 ± 10.9	not available	7 – 9	DNA Clean & Concentrator™ -5 (DNA C&C)
ISO-4	153.6 ± 41.5	85.9 ± 3.9	10 – 12	GODON excl. isopropanol step (Godon 1997)
ISO-5	216.7 ± 40.3	104.2 ± 2.7	13 – 15	GODON incl. isopropanol step (Godon 1997)
ISO-6	61.8 ± 7.8	1.5 ± 0.4	16 – 18	NucleoSpin® gDNA Clean-up
ISO-7	52.8 ± 36.8	not available	19 – 21	gDNA Clean & Concentrator™ -10 (gDNA C&C)
ISO-8	10.9 ± 1.9	9.6 ± 2.7	22 – 24	QIAamp DNA Stool Mini Kit + QIAcube (200µl)
ISO-9	not available	18.2 ± 1.0	25 – 28	QIAamp DNA Stool Mini Kit (400µl)
IOS-10	not available	10.7 ± 2.3	29 – 32	QIAamp DNA Stool Mini Kit + QIAcube (200µl)

Supplementary table 2 Co-variables acting as confounder in T2D.
Permutational multivariate analysis of variances to determine confounding factors. Control group are individuals with BMI < 30 and without T2D or prediabetes.

Covariable	P-value
Gender	0.001
Age	0.001
Vitamin D intake	0.002
Physical activity	0.004
PPI	0.023
Metformin intake	0.001

Supplementary Table 3 Overview of generated machine learning models for classification and prediction.

Model	Binary Outcome	Included Features	Classification Model	Related Figure	AUC
s-arOTUs	nonT2D / T2D	13 arrhythmic OTUs	GLM	Figure 51	0.73
s-arOTUs+BMI	nonT2D / T2D	13 arrhythmic OTUs + BMI	GLM	Figure 51	0.79
rfOTUs	nonT2D / T2D	random forest selected OTUs	RF	Figure 52	0.73
rfOTUs + BMI	nonT2D / T2D	random forest selected OTUs + BMI	RF	Figure 52	0.77
rndOTUs	nonT2D / T2D	13 randomly selected OTUs	GLM	Figure 53 A	0.59
rndOTUs + BMI	nonT2D / T2D	13 randomly selected OTUs + BMI	GLM	Figure 53 A	0.73
model obesity	BMI < 30 / BMI > 30	random forest selected OTUs	RF	Figure 53 B	0.63
rfOTUs	nonT2D / T2D	model obesity random forest selected OTUs	RF	Figure 53 B	0.7
mixedRF	nonT2D / T2D	mixed effect random forest selected OTUs	MERF	Figure 54	0.69
centered log model	nonT2D / T2D	random forest selected cantered log ratio transformed OTUs	RF	Figure 57	0.71
centered log model + BMI	nonT2D / T2D	random forest selected cantered log ratio transformed OTUs + BMI	RF	Figure 57	0.75
log-Ratio model	nonT2D / T2D	random forest selected log ratio OTUs	RF	Figure 57	0.74
log-Ratio model + BMI	nonT2D / T2D	random forest selected log ratio OTUs + BMI	RF	Figure 57	0.75
rfOTUs	nonT2D / T2D	14 random forest selected OTUs	GLM	Figure 58	0.75
rfOTUs + BMI	nonT2D / T2D	14 random forest selected OTUs + BMI	GLM	Figure 58	0.79
model with miscellaneous risk marker	nonT2D / T2D	models with random forest selected features (OTUs + miscellaneous risk marker)	RF	Figure 59 + Figure 60	
rfMET -	nonT2D / T2D without metformin treatment	random forest selected OTUs	RF	Figure 62 A	0.6
rfMET +	nonT2D / T2D with metformin treatment	random forest selected OTUs	RF	Figure 62 A	0.87
rfMET - / -	T2D without metformin	random forest selected OTUs	RF	Figure 62 A	0.82

	treatment / T2D with metformin treatment				
rfMET - / -	T2D without metformin treatment / T2D with metformin treatment	13 arrhythmic OTUs + BMI	GLM	Figure 62 B	0.6
s-arOTUs+BMI	nonT2D / T2D	13 arrhythmic OTUs BLAST matched in FoCus cohort + BMI	GLM	Figure 64 B	0.76
s-arOTUs	nonT2D / Prediabetes	13 arrhythmic OTUs	GLM	Figure 65 B	0.69
s-arOTUs+BMI	nonT2D / Prediabetes	13 arrhythmic OTUs + BMI	GLM	Figure 65 B	0.78
s-arOTUs+BMI	nonT2D / T2D	13 arrhythmic OTUs BLASt matched in TwinsUK cohort + BMI	GLM	Figure 66 C	0.68
s-arOTUs+BMI	nonT2D / Prediabetes	13 arrhythmic OTUs BLASt matched in TwinsUK cohort + BMI	GLM	Figure 66 C	0.69
rfPathways	nonT2D / T2D	random forest selected KEGG pathways	RF	Figure 67	0.81
rndPathways	nonT2D / T2D	random forest without rfPathways feature list	RF	Figure 67	0.6

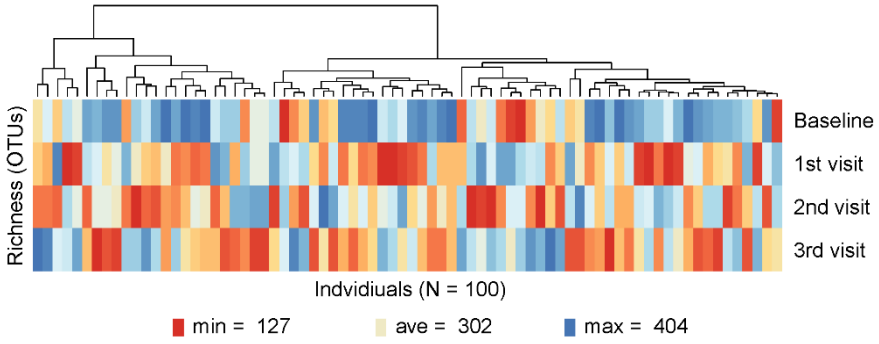
Supplementary table 4 MetayCyc pathways significantly different between persisting T2D cases and nonT2D.

MetaCyc Pathways	P-value	Disease	mean rel. abundance (%) pT2D	mean rel. abundance (%) nonT2D
Coenzyme A biosynthesis I	0.027	pT2D	0.004	0.003
Purine ribonucleosides degradation	0.034	pT2D	0.010	0.009
Superpathway of pyrimidine ribonucleosides salvage	0.040	pT2D	0.003	0.002
Superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis	0.040	pT2D	0.003	0.003
Pyruvate fermentation to acetate and lactate II	0.042	pT2D	0.004	0.004
Pantothenate and coenzyme A biosynthesis I	0.045	pT2D	0.006	0.005
Starch degradation V	0.071	pT2D	0.019	0.018
Superpathway of pyrimidine ribonucleotides de novo biosynthesis	0.074	pT2D	0.006	0.005
Galactose degradation I (Leloir pathway)	0.078	pT2D	0.008	0.007
D-galactose degradation V (Leloir pathway)	0.086	pT2D	0.008	0.007
Glycolysis IV (plant cytosol)	0.099	pT2D	0.013	0.011
Superpathway of L-lysine, L-threonine and L-methionine biosynthesis I	0.027	nonT2D	0.002	0.002
Superpathway of UDP-glucose-derived O-antigen building blocks biosynthesis	0.039	nonT2D	0.000	0.000
Superpathway of fatty acids biosynthesis (E. coli)	0.065	nonT2D	0.000	0.000
Aspartate superpathway	0.074	nonT2D	0.002	0.002
Sulfoglycolysis	0.084	nonT2D	0.000	0.000

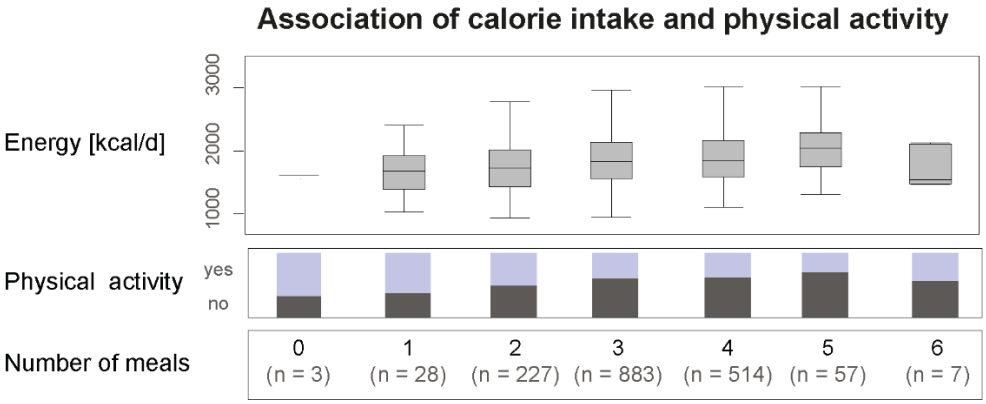
Supplementary table 5 MetayCyc pathways significantly different between incident T2D cases and nonT2D.

MetaCyc Pathways	P-value	Disease	mean rel. abundance (%) pT2D	mean rel. abundance (%) nonT2D
4-hydroxyphenylacetate degradation	0.032	nonT2D	4E-07	9E-05
Superpathway of menaquinol-8 biosynthesis II	0.037	nonT2D	3E-05	6E-05
Peptidoglycan maturation (meso-diaminopimelate containing)	0.038	nonT2D	3E-03	4E-03
Superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation	0.046	nonT2D	3E-03	3E-03
Guanosine nucleotides degradation II	0.043	iT2D	5E-04	3E-04
Superpathway of guanosine nucleotides degradation (plants)	0.052	iT2D	3E-04	2E-04
L-lysine biosynthesis II	0.063	iT2D	2E-03	2E-03

7.2. Supplementary figures



Supplementary figure 1 Changes in *alpha*-diversity within an Individual of the *enable* subcohort.
Heatmap shows richness within one individual for the four consecutive sampling time phases. Rows are ordered according to the dendrogram, which clusters individuals with similar richness values.



Supplementary figure 2 Association of calorie intake and physical activity.
Boxplots show the calorie intake [kcal/day] stratified according to average number of meals per day, indicated by the participants. Stacked barplot shows the distribution of individuals self-reported status of physical activity. Number of individuals per group is shown below the number of meals.

7.3. Supplementary file

Supplementary file 1 Instructions for PCR purification with AMPure XP Beads.

PCR purification is performed with AGENCOURT AMPure XP Beads (Beckman Coulter)

1. Remove the AMPure XP beads from -4 °C storage and leave to stand for at least 30 min to bring to room temperature.
2. Vortex the AMPure XP beads until they are well dispersed.
3. Transfer the content of each PCR tube into the wells of a 96 deep-well plate.
4. Add 1.8 µl AMPure XP beads per 1.0 µl of PCR product. Gently pipette the entire volume up and down 10 times to mix thoroughly.
5. Incubate at room temperature for 5 min.
6. Transfer the volume in the 96well-plate. Put the well-plate in the magnetic rack and leave to stand at room temperature for 2 min or until the liquid becomes clear in appearance.
7. Remove and discard supernatants.
8. Add 200 µl freshly prepared 70 % EtOH to each well without disturbing the beads.
9. Leave at room temperature for 30 sec and discard all supernatants. Take extra care not to disturb the beads.
10. Repeat steps 8 and 9 once more, for a total of two 70% EtOH washes.
11. Let the 96-well-plate stand at room temperature for 4-5 min to dry, then remove from the magnetic rack.
12. Re-suspend the dried bead pellet in each well with 31µl BE Elution (recommended volume of AMPure standard protocol). Gently pipette the entire volume up and down 10 times to mix thoroughly.
13. Incubate the 96-well-plate at room temperature for 2 min.
14. Place the 96-well-plate on the magnetic rack at room temperature for 2 min or until the liquid becomes clear in appearance.
15. Transfer 23-31 µl of the clear supernatant from each well to a PCR-tube for DNA measurement by fluorimetry (Qubit measurement according to the manufacturer's instructions).

8. Figure list

9. Table list

10. Supplementary figure and table list

10.1. Supplementary table list

10.2. Supplementary figure list

10.3. Supplementary file list

11. Abbreviations

AGE1	enable individuals between 3-5 years
AGE2	enable individuals between 18-25 years
AGE3	enable individuals between 40-46 years
AGE4	enable individuals between 75-85 years
arOTUs	Arrhythmic OTUs
ASV	Amplicon Sequence Variants
AUC	Area Under the Curve
AUCPR	Area Under the Precision-Recall Curve
BMI	Body Mass Index
bp	Base-pairs
C1	Ruminococcus cluster
C2	Bacteroides cluster
C3	Prevotella cluster
CD	combinatorial dual
CV	coefficient of observation
CVD	Cardiovascular Disease
ESV	Exact Sequence Variant
F/B	Firmicutes to Bacteroides ratio
FDR	False Discovery Rate
FFQ	Food Frequency Questionnaire
FoCus	Food Chain Plus
fold-CV	Cross Validation
GLM	Generalized Linear Model
Gnoto	Gnotobiotic mice
HMP	Human Microbiome Project
HW	Hip-Waist Ratio
IBD	Inflammatory Bowle Disease
IMNGS	Integrated Microbial Next Generation Sequencing
IQR	Inter Quartile Range
ISO-4	GODON modified excl. isopropanol step
ISO-5	GODON modified incl. isopropanol step
ISO-8	QIAamp DNA Stool Mini Kit + QIAcube (200µl)
iT2D	incident Type 2 Diabetes
JMF	Joint Microbiome Facility
KORA	Cooperative Health Research in the Augsburg Region
MDS	Multidimensional Scaling
MERF	Mixed Effect Random Forest Model
MetaHIT	Human Intestinal tract project
Mock	Synthetically spiked Mock community
nMDS	Non-linear Multidimensional Scaling
OOB	Out-of-Box error
OTU	Operational Taxonomic Unit
PCoA	Principle Coordinate Analysis
PCR	Polymerase Chain Reaction
PPI	Proton Pump Inhibitor
pT2D	Persisting Type 2 Diabetes

RDP	Ribosomal Database Project
rfMET	Selected random forest OTUs for Metformin intake
rfOTUs	Selected random forest bacterial risk signature
rfT2D+	Selected random forest OTUs for T2D + Metformin intake
rfT2D+	Selected random forest OTUs for T2D - Metformin intake
rndOTUs	Random OTUs
ROC	Receiver Operating Characteristic
rOTU	Rhythmic OTUs
S1	Healthy male subject
S2	Healthy female subject
s-arOTUs	Selected arrhythmic bacterial risk signature
Study-1	16S rRNA gene sequencing data from Flores et al (2014)
Study-2	16S rRNA gene sequencing data from Halfvarson et al. (2017)
T2D	Type 2 Diabetes
T2D+MET	T2D taking metformin
T2D-MET	T2D without metformin treatment
Taxa	Taxonomically classified bacteria
TUM-mock	In-house synthetically spiked community of 13 bacteria
UD	Unique dual
V1V2	Variable region 1 and 2 of the 16S rRNA gene
V3V4	Variable region 3 and 4 of the 16S rRNA gene
V4	Variable region 4 of the 16S rRNA gene
ZymoBIOMICs	commercial synthetically spiked community of 8 bacteria

12. References

Alanis-Lobato, G., M. A. Andrade-Navarro and M. H. Schaefer (2017). "HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks." Nucleic Acids Res **45**(D1): D408-D414.

Albert, K., A. Rani and D. A. Sela (2019). "Comparative Pangenomics of the Mammalian Gut Commensal Bifidobacterium longum." Microorganisms **8**(1).

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

American Diabetes, A. (2017). "Erratum. Pharmacologic Approaches to Glycemic Treatment. Sec. 8. In Standards of Medical Care in Diabetes-2017. Diabetes Care 2017;40(Suppl. 1);S64-S74." Diabetes Care **40**(7): 985.

Ananthakrishnan, A. N., C. Luo, V. Yajnik, H. Khalili, J. J. Garber, B. W. Stevens, T. Cleland and R. J. Xavier (2017). "Gut Microbiome Function Predicts Response to Anti-integrin Biologic Therapy in Inflammatory Bowel Diseases." Cell Host Microbe **21**(5): 603-610 e603.

Andrikopoulos, S., B. C. Fam, A. Holdsworth, S. Visinoni, Z. Ruan, M. Stathopoulos, A. W. Thorburn, C. N. Joannides, M. Cancilla, L. Balmer, J. Proietto and G. Morahan (2016). "Identification of ABCC8 as a contributory gene to impaired early-phase insulin secretion in NZO mice." J Endocrinol **228**(1): 61-73.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J. M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Dore, H. I. T. C. Meta, M. Antolin, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. Melo Minardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich and P. Bork (2011). "Enterotypes of the human gut microbiome." Nature **473**(7346): 174-180.

Avershina, E. and K. Rudi (2015). "Confusion about the species richness of human gut microbiota." Benef Microbes **6**(5): 657-659.

Backhed, F., H. Ding, T. Wang, L. V. Hooper, G. Y. Koh, A. Nagy, C. F. Semenkovich and J. I. Gordon (2004). "The gut microbiota as an environmental factor that regulates fat storage." Proc Natl Acad Sci U S A **101**(44): 15718-15723.

Barnea, M., L. Haviv, R. Gutman, N. Chapnik, Z. Madar and O. Froy (2012). "Metformin affects the circadian clock and metabolic rhythms in a tissue-specific manner." Biochim Biophys Acta **1822**(11): 1796-1806.

Bartram, A. K., M. D. Lynch, J. C. Stearns, G. Moreno-Hagelsieb and J. D. Neufeld (2011). "Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads." Appl Environ Microbiol **77**(11): 3846-3852.

Beli, E., S. Prabakaran, P. Krishnan, C. Evans-Molina and M. B. Grant (2019). "Loss of Diurnal Oscillatory Rhythms in Gut Microbiota Correlates with Changes in Circulating Metabolites in Type 2 Diabetic db/db Mice." Nutrients **11**(10).

Belkaid, Y. and T. W. Hand (2014). "Role of the microbiota in immunity and inflammation." Cell **157**(1): 121-141.

Berry, D., K. Ben Mahfoudh, M. Wagner and A. Loy (2011). "Barcoded primers used in multiplex amplicon pyrosequencing bias amplification." Appl Environ Microbiol **77**(21): 7846-7849.

Bleicher, A., T. Stark, T. Hofmann, B. Bogovic Matijasic, I. Rogelj, S. Scherer and K. Neuhaus (2010). "Potent antilisterial cell-free supernatants produced by complex red-smear cheese microbial consortia." J Dairy Sci **93**(10): 4497-4505.

Bloom, S. M., V. N. Bijanki, G. M. Nava, L. Sun, N. P. Malvin, D. L. Donermeyer, W. M. Dunne, Jr., P. M. Allen and T. S. Stappenbeck (2011). "Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease." Cell Host Microbe **9**(5): 390-403.

Bokulich, N. A., S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills and J. G. Caporaso (2013). "Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing." Nat Methods **10**(1): 57-59.

Buford, T. W. (2017). "(Dis)Trust your gut: the gut microbiome in age-related inflammation, health, and disease." Microbiome **5**(1): 80.

Calinski, T. and J. Harabasz (1974). "A dendrite method for cluster analysis." Communications in Statistics - Theory and Methods **3**(1): 1-27.

Callahan, B. J., P. J. McMurdie and S. P. Holmes (2017). "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis." ISME J **11**(12): 2639-2643.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson and S. P. Holmes (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nat Methods **13**(7): 581-583.

Callahan, B. J., K. Sankaran, J. A. Fukuyama, P. J. McMurdie and S. P. Holmes (2016). "Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses." F1000Res **5**: 1492.

Canfora, E. E., J. W. Jocken and E. E. Blaak (2015). "Short-chain fatty acids in control of body weight and insulin sensitivity." Nat Rev Endocrinol **11**(10): 577-591.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith and R. Knight (2012). "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms." ISME J **6**(8): 1621-1624.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer and R. Knight (2011). "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample." Proceedings of the National Academy of Sciences **108**(Supplement 1): 4516-4522.

Carroll, I. M., T. Ringel-Kulka, J. P. Siddle, T. R. Klaenhammer and Y. Ringel (2012). "Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage." PLoS One **7**(10): e46953.

Chen, J., K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman and H. Li (2012). "Associating microbiome composition with environmental covariates using generalized UniFrac distances." Bioinformatics **28**(16): 2106-2113.

Chicco, D. and G. Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC Genomics **21**(1): 6.

Claassen, S., E. du Toit, M. Kaba, C. Moodley, H. J. Zar and M. P. Nicol (2013). "A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples." J Microbiol Methods **94**(2): 103-110.

Claesson, M. J., I. B. Jeffery, S. Conde, S. E. Power, E. M. O'Connor, S. Cusack, H. M. Harris, M. Coakley, B. Lakshminarayanan, O. O'Sullivan, G. F. Fitzgerald, J. Deane, M. O'Connor, N. Harnedy, K. O'Connor, D. O'Mahony, D. van Sinderen, M. Wallace, L. Brennan, C. Stanton, J. R. Marchesi, A. P. Fitzgerald, F. Shanahan, C. Hill, R. P. Ross and P. W. O'Toole (2012). "Gut microbiota composition correlates with diet and health in the elderly." Nature **488**(7410): 178-184.

Clarke, S. F., E. F. Murphy, K. Nilaweera, P. R. Ross, F. Shanahan, P. W. O'Toole and P. D. Cotter (2012). "The gut microbiota and its relationship to diet and obesity: new insights." Gut Microbes **3**(3): 186-202.

Clavel, T., I. Lagkouvardos and A. Hiergeist (2016). "Microbiome sequencing: challenges and opportunities for molecular medicine." Expert Rev Mol Diagn **16**(7): 795-805.

Collado, M. C., P. A. Engen, C. Bandin, R. Cabrera-Rubio, R. M. Voigt, S. J. Green, A. Naqib, A. Keshavarzian, F. Scheer and M. Garaulet (2018). "Timing of food intake impacts daily rhythms of human salivary microbiota: a randomized, crossover study." FASEB J **32**(4): 2060-2072.

Connors, J., N. Dawe and J. Van Limbergen (2018). "The Role of Succinate in the Regulation of Intestinal Inflammation." Nutrients **11**(1).

Costea, P. I., F. Hildebrand, M. Arumugam, F. Backhed, M. J. Blaser, F. D. Bushman, W. M. de Vos, S. D. Ehrlich, C. M. Fraser, M. Hattori, C. Huttenhower, I. B. Jeffery, D. Knights, J. D. Lewis, R. E. Ley, H. Ochman, P. W. O'Toole, C. Quince, D. A. Relman, F. Shanahan, S. Sunagawa, J. Wang, G. M. Weinstock, G. D. Wu, G. Zeller, L. Zhao, J. Raes, R. Knight and P. Bork (2018). "Enterotypes in the landscape of gut microbial community composition." Nat Microbiol **3**(1): 8-16.

Costea, P. I., G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, F. Levenez, M. Tramontano, M. Driessen, R. Hercog, F. E. Jung, J. R. Kultima, M. R. Hayward, L. P. Coelho, E. Allen-Vercoe, L. Bertrand, M. Blaut, J. R. M. Brown, T. Carton, S. Cools-Portier, M. Daigneault, M. Derrien, A. Druesne, W. M. de Vos, B. B. Finlay, H. J. Flint, F. Guarner, M. Hattori, H. Heilig, R. A. Luna, J. van Hylckama Vlieg, J. Junick, I. Klymiuk, P. Langella, E. Le Chatelier, V. Mai, C. Manichanh, J. C. Martin, C. Mery, H. Morita, P. W. O'Toole, C. Orvain, K. R. Patil, J. Penders, S. Persson, N. Pons, M. Popova, A. Salonen, D. Saulnier, K. P. Scott, B. Singh, K. Slezak, P. Veiga, J. Versalovic, L. Zhao, E. G. Zoetendal, S. D. Ehrlich, J. Dore and P. Bork (2017). "Towards standards for human fecal sample processing in metagenomic studies." Nat Biotechnol **35**(11): 1069-1076.

Cuthbertson, L., G. B. Rogers, A. W. Walker, A. Oliver, T. Hafiz, L. R. Hoffman, M. P. Carroll, J. Parkhill, K. D. Bruce and C. J. van der Gast (2014). "Time between collection and storage significantly influences bacterial sequence composition in sputum samples from cystic fibrosis respiratory infections." J Clin Microbiol **52**(8): 3011-3016.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton and P. J. Turnbaugh (2014). "Diet rapidly and reproducibly alters the human gut microbiome." Nature **505**(7484): 559-563.

David Martin and W. Powers (2011). "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation." Journal of Machine Learning Technologies.

De La Cochetiere, M. F., T. Durand, V. Lalande, J. C. Petit, G. Potel and L. Beaugerie (2008). "Effect of antibiotic therapy on human fecal microbiota and the relation to the development of Clostridium difficile." Microb Ecol **56**(3): 395-402.

Dethlefsen, L., S. Huse, M. L. Sogin and D. A. Relman (2008). "The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing." PLoS Biol **6**(11): e280.

Dominianni, C., J. Wu, R. B. Hayes and J. Ahn (2014). "Comparison of methods for fecal microbiome biospecimen collection." BMC Microbiol **14**: 103.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson and D. A. Relman (2005). "Diversity of the human intestinal microbial flora." Science **308**(5728): 1635-1638.

Edgar, R. (2018). "Taxonomy annotation and guide tree errors in 16S rRNA databases." PeerJ **6**: e5030.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edgar, R. C. (2013). "UPARSE: highly accurate OTU sequences from microbial amplicon reads." Nat Methods **10**(10): 996-998.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince and R. Knight (2011). "UCHIME improves sensitivity and speed of chimera detection." Bioinformatics **27**(16): 2194-2200.

Edwards, J. E., S. A. Huws, E. J. Kim and A. H. Kingston-Smith (2007). "Characterization of the dynamics of initial bacterial colonization of nonconserved forage in the bovine rumen." FEMS Microbiology Ecology **62**(3): 323-335.

Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum and J. I. Gordon (2013). "The long-term stability of the human gut microbiota." Science **341**(6141): 1237439.

Falony, G., M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, M. J. Bonder, M. Valles-Colomer, D. Vandeputte, R. Y. Tito, S. Chaffron, L. Rymenans, C. Verspecht, L. De Sutter, G. Lima-Mendez, K. D'Hoe, K. Jonckheere, D. Homola, R. Garcia, E. F. Tigchelaar, L. Eeckhaudt, J. Fu, L. Henckaerts, A. Zhernakova, C. Wijmenga and J. Raes (2016). "Population-level analysis of gut microbiome variation." Science **352**(6285): 560-564.

Fernandez-Veledo, S. and J. Vendrell (2019). "Gut microbiota-derived succinate: Friend or foe in human metabolic diseases?" Rev Endocr Metab Disord **20**(4): 439-447.

Flores, G. E., J. G. Caporaso, J. B. Henley, J. R. Rideout, D. Domogala, J. Chase, J. W. Leff, Y. Vazquez-Baeza, A. Gonzalez, R. Knight, R. R. Dunn and N. Fierer (2014). "Temporal variability is a personalized feature of the human microbiome." Genome Biol **15**(12): 531.

Forslund, K., F. Hildebrand, T. Nielsen, G. Falony, E. Le Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H. Krogh Pedersen, M. Arumugam, K. Kristiansen, A. Y. Voigt, H. Vestergaard, R. Hercog, P. Igor Costea, J. R. Kultima, J. Li, T. Jorgensen, F. Levenez, J. Dore, H. I. T. c. Meta, H. B. Nielsen, S. Brunak, J. Raes, T. Hansen, J. Wang, S. D. Ehrlich, P. Bork and O. Pedersen (2015). "Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota." Nature **528**(7581): 262-266.

Franzosa, E. A., L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata and C. Huttenhower (2018). "Species-level functional profiling of metagenomes and metatranscriptomes." Nat Methods **15**(11): 962-968.

Franzosa, E. A., A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower and R. J. Xavier (2019). "Gut microbiome structure and metabolic activity in inflammatory bowel disease." Nat Microbiol **4**(2): 293-305.

Fuks, G., M. Elgart, A. Amir, A. Zeisel, P. J. Turnbaugh, Y. Soen and N. Shental (2018). "Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling." Microbiome **6**(1): 17.

Ghatak, S., Z. A. King, A. Sastry and B. O. Palsson (2019). "The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function." Nucleic Acids Res **47**(5): 2446-2454.

Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson (2006). "Metagenomic analysis of the human distal gut microbiome." Science **312**(5778): 1355-1359.

Gillies, C. L., P. C. Lambert, K. R. Abrams, A. J. Sutton, N. J. Cooper, R. T. Hsu, M. J. Davies and K. Khunti (2008). "Different strategies for screening and prevention of type 2 diabetes in adults: cost effectiveness analysis." BMJ **336**(7654): 1180-1185.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn and J. J. Egozcue (2017). "Microbiome Datasets Are Compositional: And This Is Not Optional." Front Microbiol **8**: 2224.

Godon, J.-J. Z., Emmanuelle; Dabert, Patrick; Habouzit, Frédêric; Moletta, René (1997). "Molecular Microbial Diversity of an Anaerobic Digestor as Determined by Small-Subunit rDNA Sequence Analysis." Appl Environ Microbiol **63**(7): 2802-2813.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight and R. E. Ley (2014). "Conducting a microbiome study." Cell **158**(2): 250-262.

Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark and R. E. Ley (2014). "Human genetics shape the gut microbiome." Cell **159**(4): 789-799.

Halfvarson, J., C. J. Brislawn, R. Lamendella, Y. Vazquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight and J. K. Jansson (2017). "Dynamics of the human gut microbiome in inflammatory bowel disease." Nat Microbiol **2**: 17004.

Hamady, M. and R. Knight (2009). "Microbial community profiling for human microbiome projects: Tools, techniques, and challenges." Genome Res **19**(7): 1141-1152.

He, Y., W. Wu, S. Wu, H. M. Zheng, P. Li, H. F. Sheng, M. X. Chen, Z. H. Chen, G. Y. Ji, Z. D. Zheng, P. Mujagond, X. J. Chen, Z. H. Rong, P. Chen, L. Y. Lyu, X. Wang, J. B. Xu, C. B. Wu, N. Yu, Y. J. Xu, J. Yin, J. Raes, W. J. Ma and H. W. Zhou (2018). "Linking gut microbiota, metabolic syndrome and economic status based on a population-level analysis." Microbiome **6**(1): 172.

He, Y., W. Wu, H. M. Zheng, P. Li, D. McDonald, H. F. Sheng, M. X. Chen, Z. H. Chen, G. Y. Ji, Z. D. Zheng, P. Mujagond, X. J. Chen, Z. H. Rong, P. Chen, L. Y. Lyu, X. Wang, C. B. Wu, N. Yu, Y. J. Xu, J. Yin, J. Raes, R.

Knight, W. J. Ma and H. W. Zhou (2018). "Regional variation limits applications of healthy gut microbiome reference ranges and disease models." Nat Med **24**(10): 1532-1535.

He, Z., H. Zhang, S. Gao, M. J. Lercher, W. H. Chen and S. Hu (2016). "Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees." Nucleic Acids Res **44**(W1): W236-241.

Heintz-Buschart, A. and P. Wilmes (2018). "Human Gut Microbiome: Function Matters." Trends Microbiol **26**(7): 563-574.

Hiergeist, A., U. Reischl, p. Priority Program Intestinal Microbiota Consortium/ quality assessment and A. Gessner (2016). "Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability." Int J Med Microbiol **306**(5): 334-342.

Hildebrandt, M. A., C. Hoffmann, S. A. Sherrill-Mix, S. A. Keilbaugh, M. Hamady, Y. Y. Chen, R. Knight, R. S. Ahima, F. Bushman and G. D. Wu (2009). "High-fat diet determines the composition of the murine gut microbiome independently of obesity." Gastroenterology **137**(5): 1716-1724 e1711-1712.

Holle, R., M. Happich, H. Lowel, H. E. Wichmann and M. K. S. Group (2005). "KORA--a research platform for population based health research." Gesundheitswesen **67 Suppl 1**: S19-25.

Hsiao, E. Y., S. W. McBride, S. Hsien, G. Sharon, E. R. Hyde, T. McCue, J. A. Codelli, J. Chow, S. E. Reisman, J. F. Petrosino, P. H. Patterson and S. K. Mazmanian (2013). "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders." Cell **155**(7): 1451-1463.

Huang, Z.-q., Y.-q. Liao, R.-z. Huang, J.-p. Chen and H.-l. Sun (2018). "Possible role of TCF7L2 in the pathogenesis of type 2 diabetes mellitus." Biotechnology & Biotechnological Equipment **32**(4): 830-834.

Hughes, M. E., L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda and J. B. Hogenesch (2009). "Harmonics of circadian gene transcription in mammals." PLoS Genet **5**(4): e1000442.

Hughes, M. E., J. B. Hogenesch and K. Kornacker (2010). "JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets." J Biol Rhythms **25**(5): 372-380.

Human Microbiome Project, C. (2012). "Structure, function and diversity of the healthy human microbiome." Nature **486**(7402): 207-214.

Huse, S. M., Y. Ye, Y. Zhou and A. A. Fodor (2012). "A core human microbiome as viewed through 16S rRNA sequence clusters." PLoS One **7**(6): e34242.

Ilett, E. E., M. Jorgensen, M. Noguera-Julian, G. Daugaard, D. D. Murray, M. Helleberg, R. Paredes, J. Lundgren, H. Sengelov and C. MacPherson (2019). "Gut microbiome comparability of fresh-frozen versus stabilized-frozen samples from hospitalized patients using 16S rRNA gene and shotgun metagenomic sequencing." Sci Rep **9**(1): 13351.

Johnson, J. S., D. J. Spakowicz, B. Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren and G. M. Weinstock (2019). "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis." Nat Commun **10**(1): 5029.

Kaczmarek, J. L., S. M. Musaad and H. D. Holscher (2017). "Time of day and eating behaviors are associated with the composition and function of the human gastrointestinal microbiota." Am J Clin Nutr **106**(5): 1220-1231.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kaplan, R. C., Z. Wang, M. Usyk, D. Sotres-Alvarez, M. L. Daviglus, N. Schneiderman, G. A. Talavera, M. D. Gellman, B. Thyagarajan, J. Y. Moon, Y. Vazquez-Baeza, D. McDonald, J. S. Williams-Nguyen, M. C. Wu, K. E. North, J. Shaffer, C. C. Sollecito, Q. Qi, C. R. Isasi, T. Wang, R. Knight and R. D. Burk (2020). "Author Correction: Gut microbiome composition in the Hispanic Community Health Study/Study of Latinos is shaped by geographic relocation, environmental factors, and obesity." Genome Biol **21**(1): 50.

Karlsson, F., V. Tremaroli, J. Nielsen and F. Backhed (2013). "Assessing the human gut microbiota in metabolic diseases." Diabetes **62**(10): 3341-3349.

Karlsson, F. H., I. Nookaew and J. Nielsen (2014). "Metagenomic data utilization and analysis (MEDUSA) and construction of a global gut microbial gene catalogue." PLoS Comput Biol **10**(7): e1003706.

Karlsson, F. H., V. Tremaroli, I. Nookaew, G. Bergstrom, C. J. Behre, B. Fagerberg, J. Nielsen and F. Backhed (2013). "Gut metagenome in European women with normal, impaired and diabetic glucose control." Nature **498**(7452): 99-103.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.

Khan, M. T., M. Nieuwdorp and F. Backhed (2014). "Microbial modulation of insulin sensitivity." Cell Metab **20**(5): 753-760.

King, B. M., A. N. Ivester, P. D. Burgess, K. M. Shappell, K. L. Coleman, V. M. Cespedes, H. S. Pruitt, G. K. Burden and E. S. Bour (2016). "Adults with Obesity Underreport High-calorie Foods in the Home." Health Behavior and Policy Review **3**(5): 439-443.

Klindworth, A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn and F. O. Glockner (2013). "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies." Nucleic Acids Res **41**(1): e1.

Koppe, L., N. J. Pillon, R. E. Vella, M. L. Croze, C. C. Pelletier, S. Chambert, Z. Massy, G. Glorieux, R. Vanholder, Y. Dugenet, H. A. Soula, D. Fouque and C. O. Soulage (2013). "p-Cresyl sulfate promotes insulin resistance associated with CKD." J Am Soc Nephrol **24**(1): 88-99.

Koren, O., D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower and R. E. Ley (2013). "A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets." PLoS Comput Biol **9**(1): e1002863.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform." Appl Environ Microbiol **79**(17): 5112-5120.

Kuang, Z., Y. Wang, Y. Li, C. Ye, K. A. Ruhn, C. L. Behrendt, E. N. Olson and L. V. Hooper (2019). "The intestinal microbiota programs diurnal rhythms in host metabolism through histone deacetylase 3." Science **365**(6460): 1428-1434.

Lagkouvardos, I., S. Fischer, N. Kumar and T. Clavel (2017). "Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons." PeerJ **5**: e2836.

Lagkouvardos, I., D. Joseph, M. Kapfhammer, S. Giritli, M. Horn, D. Haller and T. Clavel (2016). "IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies." Sci Rep **6**: 33721.

Langille, M. G., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko and C. Huttenhower (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nat Biotechnol **31**(9): 814-821.

Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jorgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clement, J. Dore, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J. D. Zucker, J. Raes, T. Hansen, H. I. T. c. Meta, P. Bork, J. Wang, S. D. Ehrlich and O. Pedersen (2013). "Richness of human gut microbiome correlates with metabolic markers." Nature **500**(7464): 541-546.

Lebuhn, M., A. Hanreich, M. Klocke, A. Schlüter, C. Bauer and C. M. Pérez (2014). "Towards molecular biomarkers for biogas production from lignocellulose-rich substrates." Anaerobe **29**: 10-21.

Lee, Y., E. D. Berglund, X. Yu, M. Y. Wang, M. R. Evans, P. E. Scherer, W. L. Holland, M. J. Charron, M. G. Roth and R. H. Unger (2014). "Hyperglycemia in rodent models of type 2 diabetes requires insulin-resistant alpha cells." Proc Natl Acad Sci U S A **111**(36): 13217-13222.

Ley, R. E., F. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight and J. I. Gordon (2005). "Obesity alters gut microbial ecology." Proc Natl Acad Sci U S A **102**(31): 11070-11075.

Li, Q., Y. Chang, K. Zhang, H. Chen, S. Tao and Z. Zhang (2020). "Implication of the gut microbiome composition of type 2 diabetic patients from northern China." Sci Rep **10**(1): 5450.

Liang, X., F. D. Bushman and G. A. FitzGerald (2015). "Rhythmicity of the intestinal microbiota is regulated by gender and the host circadian clock." Proc Natl Acad Sci U S A **112**(33): 10479-10484.

Lippert, K., L. Kedenko, L. Antonielli, I. Kedenko, C. Gemeier, M. Leitner, A. Kautzky-Willer, B. Paulweber and E. Hackl (2017). "Gut microbiota dysbiosis associated with glucose metabolism disorders and the metabolic syndrome in older adults." Benef Microbes **8**(4): 545-556.

Louis, P., K. P. Scott, S. H. Duncan and H. J. Flint (2007). "Understanding the effects of diet on bacterial metabolism in the large intestine." J Appl Microbiol **102**(5): 1197-1208.

MacConaill, L. E., R. T. Burns, A. Nag, H. A. Coleman, M. K. Slevin, K. Giorda, M. Light, K. Lai, M. Jarosz, M. S. McNeill, M. D. Ducar, M. Meyerson and A. R. Thorner (2018). "Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing." BMC Genomics **19**(1): 30.

Madeira, F., Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn and R. Lopez (2019). "The EMBL-EBI search and sequence analysis tools APIs in 2019." Nucleic Acids Res **47**(W1): W636-W641.

Maier, L., M. Pruteanu, M. Kuhn, G. Zeller, A. Telzerow, E. E. Anderson, A. R. Brochado, K. C. Fernandez, H. Dose, H. Mori, K. R. Patil, P. Bork and A. Typas (2018). "Extensive impact of non-antibiotic drugs on human gut bacteria." Nature **555**(7698): 623-628.

Marler, M. R., P. Gehrman, J. L. Martin and S. Ancoli-Israel (2006). "The sigmoidally transformed cosine curve: a mathematical model for circadian rhythms with symmetric non-sinusoidal shapes." Stat Med **25**(22): 3893-3904.

O'Toole, P. W. and I. B. Jeffery (2015). "Gut microbiota and aging." Science **350**(6265): 1214-1215.

Oliphant, K. and E. Allen-Vercoe (2019). "Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health." Microbiome **7**(1): 91.

Onaolapo, A. Y. and O. J. Onaolapo (2018). "Circadian dysrhythmia-linked diabetes mellitus: Examining melatonin's roles in prophylaxis and management." World J Diabetes **9**(7): 99-114.

Panda, S. (2019). "The arrival of circadian medicine." Nat Rev Endocrinol **15**(2): 67-69.

Park, S. C. and S. Won (2018). "Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community." Genomics Inform **16**(4): e24.

Pascal, V., M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, S. Vermeire, H. Sokol, F. Guarner and C. Manichanh (2017). "A microbial signature for Crohn's disease." Gut **66**(5): 813-822.

Pedersen, H. K., V. Gudmundsdottir, H. B. Nielsen, T. Hyotylainen, T. Nielsen, B. A. Jensen, K. Forslund, F. Hildebrand, E. Prifti, G. Falony, E. Le Chatelier, F. Levenez, J. Dore, I. Mattila, D. R. Plichta, P. Poho, L. I. Hellgren, M. Arumugam, S. Sunagawa, S. Vieira-Silva, T. Jorgensen, J. B. Holm, K. Trost, H. I. T. C. Meta, K. Kristiansen, S. Brix, J. Raes, J. Wang, T. Hansen, P. Bork, S. Brunak, M. Oresic, S. D. Ehrlich and O. Pedersen (2016). "Human gut microbes impact host serum metabolome and insulin sensitivity." Nature **535**(7612): 376-381.

Peppercorn, M. A. and P. Goldman (1972). "The role of intestinal bacteria in the metabolism of salicylazosulfapyridine." J Pharmacol Exp Ther **181**(3): 555-562.

Price, M. N., P. S. Dehal and A. P. Arkin (2009). "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix." Mol Biol Evol **26**(7): 1641-1650.

Pryor, R., P. Norvaisas, G. Marinos, L. Best, L. B. Thingholm, L. M. Quintaneiro, W. De Haes, D. Esser, S. Waschina, C. Lujan, R. L. Smith, T. A. Scott, D. Martinez-Martinez, O. Woodward, K. Bryson, M. Laudes, W. Lieb, R. H. Houtkooper, A. Franke, L. Temmerman, I. Bjedov, H. M. Cocheme, C. Kaleta and F. Cabreiro (2019). "Host-Microbe-Drug-Nutrient Screen Identifies Bacterial Effectors of Metformin Therapy." Cell **178**(6): 1299-1312 e1229.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Antolin, F. Artiguenave, H. Blottiere, N. Borruel, T. Bruls, F. Casellas, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, M. Forte, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, A. Jamet, C. Juste, G. Kaci, M. Kleerebezem, J. Knol, M. Kristensen, S. Layec, K. Le Roux, M. Leclerc, E. Maguin, R. Melo Minardi, R. Oozeer, M. Rescigno, N. Sanchez, S. Tims, T. Torrejon, E. Varela, W. de Vos, Y. Winogradsky, E. Zoetendal, P. Bork, S. D. Ehrlich and J. Wang (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H. I. T. C. Meta, P. Bork, S. D. Ehrlich and J. Wang (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J. M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen and J. Wang (2012). "A metagenome-wide association study of gut microbiota in type 2 diabetes." Nature **490**(7418): 55-60.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies and F. O. Glockner (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." Nucleic Acids Res **41**(Database issue): D590-596.

Raymond, F., A. A. Ouameur, M. Deraspe, N. Iqbal, H. Gingras, B. Dridi, P. Leprohon, P. L. Plante, R. Giroux, E. Berube, J. Frenette, D. K. Boudreau, J. L. Simard, I. Chabot, M. C. Domingo, S. Trottier, M. Boissinot, A. Huletsky, P. H. Roy, M. Ouellette, M. G. Bergeron and J. Corbeil (2016). "The initial state of the human gut microbiome determines its reshaping by antibiotics." ISME J **10**(3): 707-720.

Reichardt, N., S. H. Duncan, P. Young, A. Belenguer, C. McWilliam Leitch, K. P. Scott, H. J. Flint and P. Louis (2014). "Phylogenetic distribution of three pathways for propionate production within the human gut microbiota." ISME J **8**(6): 1323-1335.

Reitmeier, S., T. C. A. Hitch, N. Fikas, B. Hausmann, A. E. Ramer-Tait, K. Neuhaus, D. Berry, D. Haller, I. Lagkouvardos and T. Clavel (2020). "Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling." Microbiome.

Reitmeier, S., S. Kiessling, T. Clavel, M. List, E. L. Almeida, T. S. Ghosh, K. Neuhaus, H. Grallert, J. Linseisen, T. Skurk, B. Brandl, T. A. Breuninger, M. Troll, W. Rathmann, B. Linkohr, H. Hauner, M. Laudes, A. Franke, C. I. Le Roy, J. T. Bell, T. Spector, J. Baumbach, P. W. O'Toole, A. Peters and D. Haller (2020). "Arrhythmic Gut Microbiome Signatures Predict Risk of Type 2 Diabetes." Cell Host Microbe.

Relling, I., G. Akcay, D. Fangmann, C. Knappe, D. M. Schulte, K. Hartmann, N. Muller, K. Turk, A. Dempfle, A. Franke, S. Schreiber and M. Laudes (2018). "Role of wnt5a in Metabolic Inflammation in Humans." J Clin Endocrinol Metab **103**(11): 4253-4264.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H. W. Zhou, R. Knight and J. G. Caporaso (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences." PeerJ **2**: e545.

Rothschild, D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, D. Zeevi, P. I. Costea, A. Godneva, I. N. Kalka, N. Bar, S. Shilo, D. Lador, A. V. Vila, N. Zmora, M. Pevsner-Fischer, D. Israeli, N. Kosower, G. Malka, B. C. Wolf, T. Avnit-Sagi, M. Lotan-Pompan, A. Weinberger, Z. Halpern, S. Carmi, J. Fu, C. Wijmenga, A. Zhernakova, E. Elinav and E. Segal (2018). "Environment dominates over host genetics in shaping human gut microbiota." Nature **555**(7695): 210-215.

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman and A. W. Walker (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses." BMC Biology **12**(1): 87.

Salter, S. J., M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman and A. W. Walker (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses." BMC Biol **12**: 87.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan and C. Quince (2015). "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform." Nucleic Acids Res **43**(6): e37.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Appl Environ Microbiol **75**(23): 7537-7541.

Segata, N., D. Bornigen, X. C. Morgan and C. Huttenhower (2013). "PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes." Nat Commun **4**: 2304.

Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson and C. Huttenhower (2012). "Metagenomic microbial community profiling using unique clade-specific marker genes." Nat Methods **9**(8): 811-814.

Shaw, A. G., K. Sim, E. Powell, E. Cornwell, T. Cramer, Z. E. McClure, M. S. Li and J. S. Kroll (2016). "Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room?" Microbiome **4**(1): 40.

Shin, J. H., M. Sim, J. Y. Lee and D. M. Shin (2016). "Lifestyle and geographic insights into the distinct gut microbiota in elderly women from two different geographic locations." J Physiol Anthropol **35**(1): 31.

Skene, D. J., E. Skornyakov, N. R. Chowdhury, R. P. Gajula, B. Middleton, B. C. Satterfield, K. I. Porter, H. P. A. Van Dongen and S. Gaddameedhi (2018). "Separation of circadian- and behavior-driven metabolite rhythms in humans provides a window on peripheral oscillators and metabolism." Proc Natl Acad Sci U S A **115**(30): 7825-7830.

Skyler, J. S., G. L. Bakris, E. Bonifacio, T. Darsow, R. H. Eckel, L. Groop, P. H. Groop, Y. Handelsman, R. A. Insel, C. Mathieu, A. T. McElvaine, J. P. Palmer, A. Pugliese, D. A. Schatz, J. M. Sosenko, J. P. Wilding and R. E. Ratner (2017). "Differentiation of Diabetes by Pathophysiology, Natural History, and Prognosis." Diabetes **66**(2): 241-255.

Sonnenburg, J. L. and F. Backhed (2016). "Diet-microbiota interactions as moderators of human metabolism." Nature **535**(7610): 56-64.

Stämmler, F., J. Gläsner, A. Hiergeist, E. Holler, D. Weber, P. J. Oefner, A. Gessner and R. Spang (2016). "Adjusting microbiome profiles for differences in microbial load by spike-in bacteria." Microbiome **4**(1).

Stern, S. E., K. Williams, E. Ferrannini, R. A. DeFronzo, C. Bogardus and M. P. Stern (2005). "Identification of individuals with insulin resistance using routine clinical measurements." Diabetes **54**(2): 333-339.

Stewart, C. J., N. J. Ajami, J. L. O'Brien, D. S. Hutchinson, D. P. Smith, M. C. Wong, M. C. Ross, R. E. Lloyd, H. Doddapaneni, G. A. Metcalf, D. Muzny, R. A. Gibbs, T. Vatanen, C. Huttenhower, R. J. Xavier, M. Rewers, W. Hagopian, J. Toppari, A. G. Ziegler, J. X. She, B. Akolkar, A. Lernmark, H. Hyoty, K. Vehik, J. P. Krischer and J. F. Petrosino (2018). "Temporal development of the gut microbiome in early childhood from the TEDDY study." Nature **562**(7728): 583-588.

Sunagawa, S., D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, A. Stamatakis and P. Bork (2013). "Metagenomic species profiling using universal phylogenetic marker genes." Nat Methods **10**(12): 1196-1199.

Thaben, P. F. and P. O. Westermark (2016). "Differential rhythmicity: detecting altered rhythmicity in biological data." Bioinformatics **32**(18): 2800-2808.

Thaiss, C. A., M. Levy, T. Korem, L. Dohnalova, H. Shapiro, D. A. Jaitin, E. David, D. R. Winter, M. Gury-BenAri, E. Tatirovsky, T. Tuganbaev, S. Federici, N. Zmora, D. Zeevi, M. Dori-Bachash, M. Pevsner-Fischer, E. Kartvelishvily, A. Brandis, A. Harmelin, O. Shibolet, Z. Halpern, K. Honda, I. Amit, E. Segal and E. Elinav (2016). "Microbiota Diurnal Rhythmicity Programs Host Transcriptome Oscillations." Cell **167**(6): 1495-1510 e1412.

Thaiss, C. A., D. Zeevi, M. Levy, G. Zilberman-Schapira, J. Suez, A. C. Tengeler, L. Abramson, M. N. Katz, T. Korem, N. Zmora, Y. Kuperman, I. Biton, S. Gilad, A. Harmelin, H. Shapiro, Z. Halpern, E. Segal and E. Elinav (2014). "Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis." Cell **159**(3): 514-529.

Thingholm, L. B., M. C. Ruhlemann, M. Koch, B. Fuqua, G. Laucke, R. Boehm, C. Bang, E. A. Franzosa, M. Hubenthal, A. Rahnavard, F. Frost, J. Lloyd-Price, M. Schirmer, A. J. Lusis, C. D. Vulpe, M. M. Lerch, G.

Homuth, T. Kacprowski, C. O. Schmidt, U. Nothlings, T. H. Karlsen, W. Lieb, M. Laudes, A. Franke and C. Huttenhower (2019). "Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition." Cell Host Microbe **26**(2): 252-264 e210.

Toner, M. (2005). "Revenge of the Microbes: How Bacterial Resistance is Undermining the Antibiotic Miracle." Emerging Infectious Diseases **11**(10): 1650-1650.

Tourlousse, D. M., S. Yoshiike, A. Ohashi, S. Matsukura, N. Noda and Y. Sekiguchi (2017). "Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing." Nucleic Acids Res **45**(4): e23.

Tsilimigras, M. C. and A. A. Fodor (2016). "Compositional data analysis of the microbiome: fundamentals, tools, and challenges." Ann Epidemiol **26**(5): 330-335.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis and J. I. Gordon (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.

Valles-Colomer, M., G. Falony, Y. Darzi, E. F. Tigchelaar, J. Wang, R. Y. Tito, C. Schiweck, A. Kurilshikov, M. Joossens, C. Wijmenga, S. Claes, L. Van Oudenhove, A. Zhernakova, S. Vieira-Silva and J. Raes (2019). "The neuroactive potential of the human gut microbiota in quality of life and depression." Nat Microbiol **4**(4): 623-632.

Wagner Mackenzie, B., D. W. Waite and M. W. Taylor (2015). "Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences." Front Microbiol **6**: 130.

Waldschmitt, N., A. Metwaly, S. Fischer and D. Haller (2018). "Microbial Signatures as a Predictive Tool in IBD-Pearls and Pitfalls." Inflamm Bowel Dis **24**(6): 1123-1132.

Walker, A. W., J. C. Martin, P. Scott, J. Parkhill, H. J. Flint and K. P. Scott (2015). "16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice." Microbiome **3**: 26-26.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Appl Environ Microbiol **73**(16): 5261-5267.

Wesolowska-Andersen, A., M. I. Bahl, V. Carvalho, K. Kristiansen, T. Sicheritz-Ponten, R. Gupta and T. R. Licht (2014). "Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis." Microbiome **2**: 19.

Wilson, K. H. and R. B. Blitchington (1996). "Human colonic biota studied by ribosomal DNA sequence analysis." Appl Environ Microbiol **62**(7): 2273-2278.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-4579.

Wood, D. E., J. Lu and B. Langmead (2019). "Improved metagenomic analysis with Kraken 2." Genome Biol **20**(1): 257.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y. Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman and J. D. Lewis (2011). "Linking long-term dietary patterns with gut microbial enterotypes." Science **334**(6052): 105-108.

Wu, H., E. Esteve, V. Tremaroli, M. T. Khan, R. Caesar, L. Manneras-Holm, M. Stahlman, L. M. Olsson, M. Serino, M. Planas-Felix, G. Xifra, J. M. Mercader, D. Torrents, R. Burcelin, W. Ricart, R. Perkins, J. M. Fernandez-Real and F. Backhed (2017). "Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug." Nat Med **23**(7): 850-858.

Wu, Y., Y. Ding, Y. Tanaka and W. Zhang (2014). "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention." Int J Med Sci **11**(11): 1185-1200.

Wymore Brand, M., M. J. Wannemuehler, G. J. Phillips, A. Proctor, A. M. Overstreet, A. E. Jergens, R. P. Orcutt and J. G. Fox (2015). "The Altered Schaedler Flora: Continued Applications of a Defined Murine Microbial Community." ILAR J **56**(2): 169-178.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A.

Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight and J. I. Gordon (2012). "Human gut microbiome viewed across age and geography." Nature **486**(7402): 222-227.

Yekutieli, D. and Y. Benjamini (2001). "The control of the false discovery rate in multiple testing under dependency." The Annals of Statistics **29**(4): 1165-1188.

Yokono, M., S. Satoh and A. Tanaka (2018). "Comparative analyses of whole-genome protein sequences from multiple organisms." Sci Rep **8**(1): 6800.

Yoon, S. H., S. M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo and J. Chun (2017). "Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies." Int J Syst Evol Microbiol **67**(5): 1613-1617.

Zarrinpar, A., A. Chaix, S. Yooseph and S. Panda (2014). "Diet and feeding pattern affect the diurnal dynamics of the gut microbiome." Cell Metab **20**(6): 1006-1017.

Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalova, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav and E. Segal (2015). "Personalized Nutrition by Prediction of Glycemic Responses." Cell **163**(5): 1079-1094.

Zhang, X., L. Li, J. Butcher, A. Stintzi and D. Figeys (2019). "Advancing functional and translational microbiome research using meta-omics approaches." Microbiome **7**(1): 154.

Zhernakova, A., A. Kurilshikov, M. J. Bonder, E. F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A. V. Vila, G. Falony, S. Vieira-Silva, J. Wang, F. Imhann, E. Brandsma, S. A. Jankipersadsing, M. Joossens, M. C. Cenit, P. Deelen, M. A. Swertz, s. LifeLines cohort, R. K. Weersma, E. J. Feskens, M. G. Netea, D. Gevers, D. Jonkers, L. Franke, Y. S. Aulchenko, C. Huttenhower, J. Raes, M. H. Hofker, R. J. Xavier, C. Wijmenga and J. Fu (2016). "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity." Science **352**(6285): 565-569.

Zhou, W., M. R. Sailani, K. Contrepois, Y. Zhou, S. Ahadi, S. R. Leopold, M. J. Zhang, V. Rao, M. Avina, T. Mishra, J. Johnson, B. Lee-McMullen, S. Chen, A. A. Metwally, T. D. B. Tran, H. Nguyen, X. Zhou, B. Albright, B. Y. Hong, L. Petersen, E. Bautista, B. Hanson, L. Chen, D. Spakowicz, A. Bahmani, D. Salins, B. Leopold, M. Ashland, O. Dagan-Rosenfeld, S. Rego, P. Limcaoco, E. Colbert, C. Allister, D. Perelman, C. Craig, E. Wei, H. Chaib, D. Hornburg, J. Dunn, L. Liang, S. M. S. Rose, K. Kukurba, B. Piening, H. Rost, D. Tse, T. McLaughlin, E. Sodergren, G. M. Weinstock and M. Snyder (2019). "Longitudinal multi-omics of host-microbe dynamics in prediabetes." Nature **569**(7758): 663-671.

Zhou, Y., K. A. Mihindukulasuriya, H. Gao, P. S. La Rosa, K. M. Wylie, J. C. Martin, K. Kota, W. D. Shannon, M. Mitreva, E. Sodergren and G. M. Weinstock (2014). "Exploration of bacterial community classes in major human habitats." Genome Biol **15**(5): R66.

13. Publications and Presentations *

Peer-reviewed Manuscripts

Reitmeier, S., T. C. A. Hitch, N. Fikas, B. Hausmann, A. E. Ramer-Tait, K. Neuhaus, D. Berry, D. Haller, I. Lagkouvardos and T. Clavel (2020). "Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling." ISME J.

Published Manuscripts

Matchado, M. S, Lauber, M., Reitmeier, R., Kacprowski, T.,Baumbach, J., Haller, D., List, M. (2021). "Network analysis methods for studying microbial communities: A mini review." Computational and Structural Biotechnology Journal 19: 2687-2698.

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J.,List, M., Neuhaus K. (2021). "Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing." mSphere 6(1).

Breuninger, A.T., Wawro, N.,Breuninger, J., Reitmeier, S., Clavel, T., Sixt-Merker, J.,Pestoni, G.,Rohrmann, S., Rathmann, W., Peters, A., Grallert, H., Meisinger, C., Haller, D., Linseisen, J. (2020). "Associations between habitual diet, metabolic disease, and the gut microbiota using Latent Dirichlet Allocation", Microbiome

Reitmeier, S., Kießling, Neuhaus, K., Haller, D. (2020).” Comparing circadian rhythmicity in the human gut microbiome”. STAR Protocols.

Reitmeier, S., Kießling, S., Clavel, T, List, M., Almeida, E. L., Gosh, T. S., Neuhaus, K., Grallert, H., Linseisen, J., Skurk, T., Brandl, B., Breuninger, A.T., Troll, M., Rathmann, W., Linkohr, B., Hauner, H., Laudes, M., Franke, A., Le Roy, C. I., Bell, J. T., Spector, T., Baumbach, J., O’Toole, P. W., Peters, A., Haller, D. (2020).” Arrhythmic Gut Microbiome Signatures Predict Risk of Type 2 Diabetes.” Cell Host & Mircobe 28, 1-15

Troll, M., S. Brandmaier, S. Reitmeier, J. Adam, S. Sharma, A. Sommer, M. A. Bind, K. Neuhaus, T. Clavel, J. Adamski, D. Haller, A. Peters and H. Grallert (2020). "Investigation of Adiposity Measures and Operational Taxonomic unit (OTU) Data Transformation Procedures in Stool Samples from a German Cohort Study Using Machine Learning Algorithms." Microorganisms **8**(4).

Lagkouvardos, I., S. Reitmeier [Fischer], N. Kumar and T. Clavel (2017). "Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons." PeerJ **5**: e2836.

Ott, B., T. Skurk, L. Hastreiter, I. Lagkouvardos, S. Reitmeier [Fischer], J. Buttner, T. Kellerer, T. Clavel, M. Rychlik, D. Haller and H. Hauner (2017). "Effect of caloric restriction on gut permeability, inflammation markers, and fecal microbiota in obese women." Sci Rep **7**(1): 11955.

Ott, B., T. Skurk, L. Lagkouvardos, S. Reitmeier [Fischer], J. Buttner, M. Lichtenegger, T. Clavel, A. Lechner, M. Rychlik, D. Haller and H. Hauner (2018). "Short-Term Overfeeding with Dairy Cream Does Not Modify Gut Permeability, the Fecal Microbiota, or Glucose Metabolism in Young Healthy Men." J Nutr **148**(1): 77-85.

Bamberger, C., A. Rossmeier, K. Lechner, L. Wu, E. Waldmann, S. Reitmeier [Fischer], R. G. Stark, J. Altenhofer, K. Henze and K. G. Parhofer (2018). "A Walnut-Enriched Diet Affects Gut Microbiome in Healthy Caucasian Subjects: A Randomized, Controlled Trial." Nutrients **10**(2).

Mak'Anyengo, R., P. Duewell, C. Reichl, C. Horth, H. A. Lehr, S. Reitmeier [Fischer], T. Clavel, G. Denk, S. Hohenester, S. Kobold, S. Endres, M. Schnurr and C. Bauer (2018). "Nlrp3-dependent IL-1beta inhibits CD103+ dendritic cell differentiation in the gut." JCI Insight **3**(5).

Waldschmitt, N., A. Metwaly, S. Reitmeier [Fischer] and D. Haller (2018). "Microbial Signatures as a Predictive Tool in IBD-Pearls and Pitfalls." Inflamm Bowel Dis **24**(6): 1123-1132.

Published Abstracts

DGE 2020: ***Arrhythmic gut microbiota members as diagnostic risk factors for Type-2 Diabetes,***

Reitmeier, S., Kießling, S., Clavel, T., Peters, A., Haller, D. Proc. Germ. Nutr. Soc., Vol. 26 (2020)

DDW 2019: ***Gut Microbiota Profiling in a Prospective Population Cohort in Relation to Metabolic Health***, Reitmeier, S., Clavel, T., Thingholm, L., Troll, M., Sommer, A., Grallert, H., Mueller, C., Franke, A., Peters, A., Haller, D. Gastroenterology, Vol. 156, Issue 6, S-50

qPCR dPCR & NGS 2019**: *Handling of Spurious Molecular Species Dictates the Outcome of High-throughput 16S rRNA Gene Amplicon Profiling***, Reitmeier, Sandra.

Oral Presentations

Digestive Disease Week (DDW) (San Diego, USA)

Gut Microbiota Profiling in a Prospective Population Cohort in Relation to Metabolic Health. Reitmeier, S., Clavel, T., Thingholm, L., Troll, M., Sommer, A., Grallert, H., Mueller, C., Franke, A., Peters, A., Haller, D

9th Gene Quantification Event qPCR dPCR & NGS 2019 (Freising, Germany)

Handling of Spurious Molecular Species Dictates the Outcome of High-throughput 16S rRNA Gene Amplicon Profiling, Reitmeier, Sandra.

12th Seeon Conference, Microbiota, Probiota and Host (Seeon, Germany)

Arrhythmic microbiota improves diagnostic profiling of Type-2-Diabetes in a prospective cohort. Reitmeier, S., Kießling, S., Clavel, T., Neuhaus, K., Grallert, H., Peters, A., Haller, D.

Poster Presentations

2019

United European Gastroenterology (UEG) (Barcelona, Spain)

Arrhythmic microbiota signatures improve diagnostic profiling of Type-2 diabetes in a prospective population cohort**.** Reitmeier, S., Kießling, S., Clavel, T, List, M., Heddes, M., Almeida, E. L., Gosh, T. S., Neuhaus, K., Grallert, H, T., O'Toole, P. W., Peters, A., Haller, D.

Keystone Symposia Microbiome: Therapeutic Implications (T1) (Killarney, Ireland)

Arrhythmic microbiota signatures improve diagnostic profiling of Type-2 diabetes in a prospective population cohort**.** Reitmeier, S., Kießling, S., Clavel, T, List, M., Heddes, M., Almeida, E. L., Gosh, T. S., Neuhaus, K., Grallert, H, T., O'Toole, P. W., Peters, A., Haller, D.

2018

11th Seeon Conference, Microbiota, Probiota and Host (Seeon, Germany)

KORA – The intestinal microbiome of a prospective population-based cohort. Reitmeier [Fischer], S., Clavel, T., Lagkouvardos, I., Neuhaus, K., Peters, A., Haller, D.

2017

2nd ASM Conference on Rapid Applied Microbial Next Generation Sequencing and Bioinformatic Pipelines (New York, USA)

Handling of Spurious Molecular Species in High-throughput 16S rRNA Gene Amplicon Data,

* Sandra Fischer and Sandra Reitmeier are the same person, name changed to Sandra Reitmeier in 2018

14. Curriculum Vitae

Msc. Sandra Reitmeier

formerly Sandra Fischer

Academic Education

08/2015 – 08/2020	**Doktor der Naturwissenschaften (Dr. rer. nat.) (Technical University Munich)** Title: "Arrhythmic gut microbiome signature for Type 2 Diabetes risk profiling." Chair of Nutrition and Immunology (Prof. Dr. Dirk Haller), Technical University Munich.
09/2013 - 08/2015	**Master of Science Epidemiology (Ludwig-Maximilians-University Munich)** Title "Gender-specific differences in predictive factors of psychiatric disorders in the German Armed Forces: A retrospective case study of the military psychiatric ambulance" Medical academy (Dr. Roland Vogl), German Armed Forces.
09/ 2010 - 09/2013	**Bachelor of Science Bioinformatics (Technical-University Munich)** Title " Quantification of marker expression for mesoderm and endoderm segregation" Institute of Computational Biology (Prof. Dr. Fabian Theis), Helmholtz Zentrum Munich – German, Research Center for Environmental Health.
09/1997 - 06/2010	**High School** Maria-Ward Gymnasium, München Nymphenburg

Work experience

08/2015 – 08/2020	**Research assistant** Core Facility Microbiome, ZIEL - Institute for Food & Health, Technical University Munich
10/2013 – 08/2015	**Student assistant** The Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-University Munich
08/2011 – 09/2013	**Student assistant** Institute of Diabetes and Regeneration Research , Helmholtz Zentrum München German Research Center for Environmental Health

List of cited publications

Reitmeier, S., Kießling, S., Clavel, T, List, M., Almeida, E. L., Gosh, T. S., Neuhaus, K., Grallert, H., Linseisen, J., Skurk, T., Brandl, B., Breuninger, A.T., Troll, M., Rathmann, W., Linkohr, B., Hauner, H., Laudes, M., Franke, A., Le Roy, C. I., Bell, J. T., Spector, T., Baumbach, J., O'Toole, P. W., Peters, A., Haller, D. (2020)." Arrhythmic Gut Microbiome Signatures Predict Risk of Type 2 Diabetes." **Cell Host & Microbe** 28, 1-15

Reitmeier, S., T. C. A. Hitch, N. Fikas, B. Hausmann, A. E. Ramer-Tait, K. Neuhaus, D. Berry, D. Haller, I. Lagkouvardos and T. Clavel (2020). "Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling." Microbiome.(under revision)

Hiermit erkläre ich unter Eid, dass ich alleiniger, federführender Hauptautor der zwei oben genannten Publikationen und Studien bin, die in dieser Arbeit wörtlich zitiert wurden. Die betreffenden Passagen wurden ausschließlich von mir verfasst.

Acknowledgement

First, I would like to thank my supervisor Prof. Dirk Haller. Thank you for believing in me and my project. I had the opportunity to really learn a lot during my PhD, to travel around, to improve myself, to know what it means to fail but also to know what it means to have success. Thank you for being patient with me, pushing me even though I took a lot of energy for both sides. This work would not have been as it is now, without his support. But not only the support from Prof. Haller increased the quality of this work, one main part was done in cooperation with Dr. Silke Kiesling. Thank you for our 24/7 work, for hard working on the project. Thank you for trying to understand a statistical point of few and helping me to understand chronobiology. It was an intense but really nice time working with you. Next I would like to thank Prof. Klaus Neuhaus who spend a lot of time in proof reading and discussing several aspects with me. Same thank goes to a very good friend Dr. Amira Metwaly, reading again and again every page helped me a lot to avoid mistakes and to improve my writing skills.

Without my friends and colleagues Angela Sachsenhauser and Caroline Ziegler, I would not have been able to generate the data I needed for the analysis. Thank you for showing me every single step of the sample preparation – for performing hundreds of DNA isolations and PCRs. Without you I would have been able to do my work. Also many thanks are going to my colleagues and friends Franziska Giehren and Isabel Abellan-Schneyder. You made office-life the best environment to work.

This work was not only a job – this was also dominating my private life. And this is why my greatest thank goes to my husband, my family and my friends. Thank you to my parents and my parents in law who always believed in me and let me feel how proud they are. To my very best friend Dr. Sarah Perschbacher, who went through the same phase during this time. Thank you for understanding me and my problems and for always being straight and lovely.

Most important and last person to mention is my husband Philipp Reitmeier. He should be the one who deservers this title. I would have quit so many times if you would not have been on my side. You did not only support me, you knew that I would finish this project, you knew that this is going to be a great work and you never doubted that I would fail. Thank you for making me happy even if I did not believe in myself.

Thank you for having me in your life!