Genetic mapping of host loci determining gut microbiota in hybrid mice

Dissertation in fulfilment of the requirements
for the degree of Doctor in Natural Sciences
of the Faculty of Mathematics and Natural Sciences
at Kiel University

submitted by

Shauni Doms

Kiel, 2021

First examiner: Prof. Dr. John F. Baines

Second examiner: Prof. Dr. Thomas Roeder

Date of oral examination: December 15, 2021

"*The role of the infinitely small in nature is infinitely great.*"

Louis Pasteur

Summary

All animals and plants are colonized by microorganisms, whereby different host species contain different microbial populations. These microbial communities form long-term relationships with their hosts. Understanding the genomic basis underlying these relationships provides insight into the possible coevolution between hosts and their microbiota. This thesis aims to contribute to a deeper understanding of the forces shaping the microbiome.

In **Chapter 1**, we identify host genomic regions influencing bacterial traits in mice. We implement a unique panel of hybrids of two house mouse subspecies, the Eastern *Mus musculus musculus*, and the Western *Mus musculus domesticus*. This panel enables us to simultaneously investigate within- and between-species host genetic effects with high resolution due to many generations of natural recombination. We genotyped 320 second generation hybrid intercrossed mice and preformed 16S rRNA profiling at the DNA and RNA level, which represent the standing and the active communities, respectively. We identify a high number of mucosa-associated bacterial taxa with significant heritability estimates, particularly for 16S rRNA transcript-based traits. Interestingly, heritability estimates also positively correlate with cospeciation rate estimates. By using a genome-wide association study (GWAS) with bacterial abundances as traits, we were able to identify 443 loci contributing to variation in 123 taxa and identify promising candidate genes and pathways. Moreover, we show significant overlap with previous gut microbiome QTL studies performed in reconstituted lab mice. Taken together, these results indicate a unique genetic architecture for cospeciating taxa, a clear enrichment for several classes of human disease, and identify important functional categories including innate immunity and G-protein-coupled receptors, whose role in host-microbe interactions diverge as new species form.

In **Chapter 2**, we embark on functionally characterizing the association between ASV35 (*Bacteroides acidifaciens/uniformis*) and a candidate gene, Sirtuin 5 (*Sirt5*). *Sirt5* belongs to a family of NAD+-dependent deacylases and is involved in numerous metabolic pathways, with its most characterized role in urea cycle activation. *B. acidifaciens* is a species known to oscillate depending on feeding time and it belongs to the phylum Bacteroidetes, which is linked with host metabolism. We characterized the gene's role in association with bacteria in a circadian context in two model organisms, *Mus musculus* (house mouse) and *Drosophila* species (fruit fly). By combining the results of both model organisms, we provide evidence for a conserved role of sirtuins in regulation of bacterial abundance, possibly through metabolic (nitrogen) homeostasis, whereby SIRT5 acts as a metabolic sensor in a circadian NAD+-dependent manner.

Chapter 3 explores the possibility of using shotgun metagenome sequencing on cecum tissue samples by comparing the efficiency of three commercially available microbial DNA enrichment kits (LOOXSTER, NEBNext, and Molzym). The relative lack of bias, combined with the possibility to determine the functional capabilities as well as the taxonomic composition make shotgun metagenome sequencing an effective tool for characterizing the microbiome. As the microbial community associated with the cecum tissue has proven to be more stable and more heritable, cecum tissue samples would be ideal for quantitative studies, such as microbiome GWAS. However, the high amount of host DNA in tissue samples makes it unfeasible to achieve enough coverage in a cost effective manner. Thus, the host DNA must be removed prior to shotgun sequencing. Overall, we found that the enrichment kits did not sufficiently remove host DNA in order to reach adequate sequencing depth. Moreover, the kits introduce a bias on the microbial community composition, making them not suitable for use in quantitative studies.

In sum, these results suggest a strong impact of host-genetics on murine gut microbiome variation. We showed that host genes can influence the bacterial abundances through possible metabolic (*i.e.*, nitrogen) homeostasis and that this interaction is conserved across species. However, the exact functional mechanism remains unresolved. These results may lead to further in-depth investigation of host loci associated with specific bacterial taxa in order to determine underlying functional pathways.

Zusammenfassung

Alle Tiere und Pflanzen werden von Mikroorganismen besiedelt, wobei verschiedene Wirtsarten jewils unterschiedliche Mikrobenpopulationen beherbergen. Diese Gemeinschaften aus Mirkoorganismen stehen in einer langfristigen Beziehung zu ihren Wirten. Die Betrachtung dieser Interaktionen auf Ebene des Genoms gibt Aufschlüsse über die mögliche Koevolution zwischen Mikrobiom und Wirtsorganismus. Ziel dieser Arbeit ist es, zu einem tieferen Verständnis der Einflüsse auf das Mikrobiom beizutragen.

Kapitel 1 behandelt die Indentifizierung von Regionen im Wirtgenom, welche die Zusammensetzung des Mikrobioms bei Mäusen beeinflussen. Wir verwendeten eine spezifische Auswahl von Hybriden zweier Hausmaus-Unterarten, der im östlichen Europa verbreiteten *Mus musculus musculus* und zum anderen der im Westen dominaten *Mus musculus domesticus*, welche es uns aufgrund generationsübergreifender natürlicher Rekombination ermöglichte genetische Effekte innerhalb und zwischen den Wirtsarten zu beobachten. Wir genotypisierten 320 gekreuzte Hybridmäuse der zweiten Generation und erstellten ein 16S rRNA-Profil auf DNA- und RNA-Ebene der residenten und aktiven mikrobiellen Gemeinschaften. Wir konnten eine große Anzahl von Schleimhaut-assoziierten Bakterientaxa mit signifikanten Heritabilitätsschätzungen identifizieren, insbesondere für 16S rRNA-Transkript-basierte Merkmale. Erwähnenswert ist, dass die geschätzte Heritabilität positiv mit der geschätzten Kospeziationsrate korrelierte. Durch eine genomweiten Assoziationsstudie (GWAS) mit bakteriellen Abundanzen als Merkmal konnten wir 443 Loci identifizieren, die zu Variation in 123 Taxa beitragen, und vielversprechende Kandidatengene und Signalwege bestimmen. Darüber hinaus ergaben sich erhebliche Überschneidungen mit früheren QTL-Studien zum Darmmikrobiom, die an rekonstituierten Labormäusen durchgeführt wurden. Im Endeffekt deuten diese Ergebnisse auf eine einzigartige Genarchitektur für kospeziierende Taxa hin, auf eine deutliche Häufung mehrerer menschlicher Erkrankungen und auf wichtige funktionelle Bereiche wie die angeborene Immunität und G-Protein-gekoppelte Rezeptoren, deren Funktionen bei der Interaktion zwischen Wirt und Mikrobe mit der Entstehung neuer Arten diversifiziert wurden. Darmmikrobiom, die an rekonstituierten Labormäusen durchgeführt wurden. Insgesamt weisen diese Ergebnisse auf eine einzigartige genetische Architektur für kospeziierende Taxa hin, auf eine deutliche Anreicherung für mehrere Klassen menschlicher Krankheiten und auf wichtige funktionelle Kategorien wie die angeborene Immunität und G-Protein-gekoppelte Rezeptoren, deren Rolle bei den Interaktionen zwischen Wirt und Mikrobe mit der Entstehung neuer Arten divergiert.

In **Kapitel 2** befassen wir uns mit der funktionellen Charakterisierung der Verbindung zwischen ASV35 (*Bacteroides acidifaciens/uniformis*) und einem Kandidaten-Gen, Sirtuin 5 (*Sirt5*). *Sirt5* gehört zu einer Familie von NAD+-abhängigen Deacylasen und ist an zahlreichen Stoffwechselwegen beteiligt, wobei seine charakteristischste Rolle die Aktivierung des Harnstoffzyklus ist. *B. acidifaciens* ist eine Spezies, von der bekannt ist, dass sie in Abhängigkeit von der Nahrungsaufnahme oszilliert, und sie gehört zum Stamm der Bacteroidetes, der mit dem Wirtsmetabolismus verknüpft zu sein scheint. Wir charakterisierten die Rolle des Gens in Assoziation mit Bakterien in einem zirkadianen Kontext in zwei Modellorganismen, *Mus musculus* (Hausmaus) und *Drosophila* (Fruchtfliege). Durch die Kombination der Ergebnisse beider Modellorganismen belegen wir eine konservierte Funktion der Sirtuine bei der Regulierung der Bakterienzahl, möglicherweise durch metabolische (Stickstoff-) Homöostase, wobei SIRT5 als metabolischer Signalgeber in einer zirkadianen, NAD+-abhängigen Weise wirkt.

In **Kapitel 3** wird die Verwendung der Shotgun-Metagenom-Sequenzierung an Zökum-Gewebeproben durch den Vergleich der Effizienz von drei kommerziell erhältlichen Kits zur Anreicherung mikrobieller DNA (LOOXSTER, NEBNext und Molzym) untersucht. Die relative Fehlerfreiheit in Verbindung mit der Möglichkeit, sowohl die funktionellen Fähigkeiten als auch die taxonomische Zusammensetzung zu bestimmen, machen die Shotgun-Metagenom-Sequenzierung zu einem effektiven Instrument zur Charakterisierung des Mikrobioms. Da sich die mikrobielle Gemeinschaft, die mit dem Zökumgewebe assoziiert ist, als deutlich stabiler und heritabler erwiesen hat, wären Zökumgewebeproben ideal für quantitative Studien, wie z. B. Mikrobiom-GWAS. Aufgrund des hohen Anteils an Wirts-DNA in den Gewebeproben ist es jedoch nicht möglich, eine ausreichende Abgrenzung auf kosteneffiziente Weise zu erreichen. Daher musste die Wirts-DNA vor der Shotgun-Sequenzierung entfernt werden. Insgesamt stellten wir fest, dass die Anreicherungskits die Wirts-DNA nicht ausreichend reduzieren, um eine ausreichende Sequenzierungsgüte zu erreichen. Außerdem stellten die Kits die Komposition der mikrobiellen Gemeinschaft vezerrt dar, so dass sie sich nicht für quantitative Studien eignen.

Zusammenfassend deuten diese Ergebnisse auf einen starken Einfluss der Wirtsgenetik auf die Variabilität des Darmmikrobioms der Maus hin. Wir konnten zeigen, dass Wirtsgene die Abundanzen von Bakterien durch eine mögliche Stoffwechselhomöostase (z. B. Stickstoff) beeinflussen können und dass diese Interaktion artenübergreifend konserviert ist. Der genaue Funktionsmechanismus ist jedoch noch nicht geklärt. Diese Resultate könnten zu einer genaueren Untersuchung von Wirtsloci in Verbindung mit bestimmten Bakterientaxa führen, um die darunter liegenden Funktionsmechanismen zu klären.

Table of contents

Declaration

Hereby I declare that,

i. apart from my supervisor's guidance, the content and design of this thesis is completely my own work. Contributions of other authors are listed in the following section;

ii. this thesis has not been submitted either partially or completely as part of a doctoral degree to another examining institution. No materials are published or submitted for publication other than indicated in this thesis;

iii. this thesis was prepared in compliance with the "Rules of Good Scientific Practice" of the German Research Foundation (DFG).

Authors' contributions

Chapter 1: John Baines and Leslie Turner designed the study. Shauni Doms and Leslie Turner coordinated the mouse breeding. Shauni Doms, Hanna Fokt and Cecilia Chung performed the dissections and sample collection. Shauni Doms performed nucleic acid extractions and reverse transcriptase of RNA. Katja Cloppenborg-Schmidt prepared the 16S rRNA gene amplicon library for sequencing. All statistical analyses were performed by Shauni Doms, with technical guidance from Leslie Turner and Malte Rühlemann. Shauni Doms wrote the chapter with editing from Leslie Turner and John Baines.

Chapter 2: Shauni Doms designed the study with feedback from John Baines. Shauni Doms performed qPCR on the mouse samples. Christoph Kaleta used a metabolic model to predict the dependency of SIRT5 targets among *B. uniformis*-dependent genes. Shauni Doms collected mouse samples in a circadian fashion. Abdulgawaad Saboukh performed all experiments in *Drosophila* under the guidance of Thomas Roeder. Katja Cloppenborg-Schmidt prepared both 16S rRNA gene amplicon libraries for sequencing. Shauni Doms executed all analyses. Shauni Doms wrote the chapter with editing from John Baines.

Chapter 3: Shauni Doms designed the study, collected samples, performed the analyses and wrote the chapter with editing from John Baines.

Kiel, October 2021,

Shauni Doms

Prof. Dr. John F. Baines

General introduction

Animals evolved in a bacterial world, with bacterial cells predating animals by ~ 3 billion years (Fig. 1) (Knoll, 2003; McFall-Ngai et al., 2013). Therefore, it is not surprising that every animal harbors a complex community of microorganisms and that they are likely shaped by interactions with them throughout their evolution. (Margulis, 1991; Zilber-Rosenberg and Rosenberg, 2008; McFall-Ngai et al., 2013). However, these exact interactions remain largely unknown.

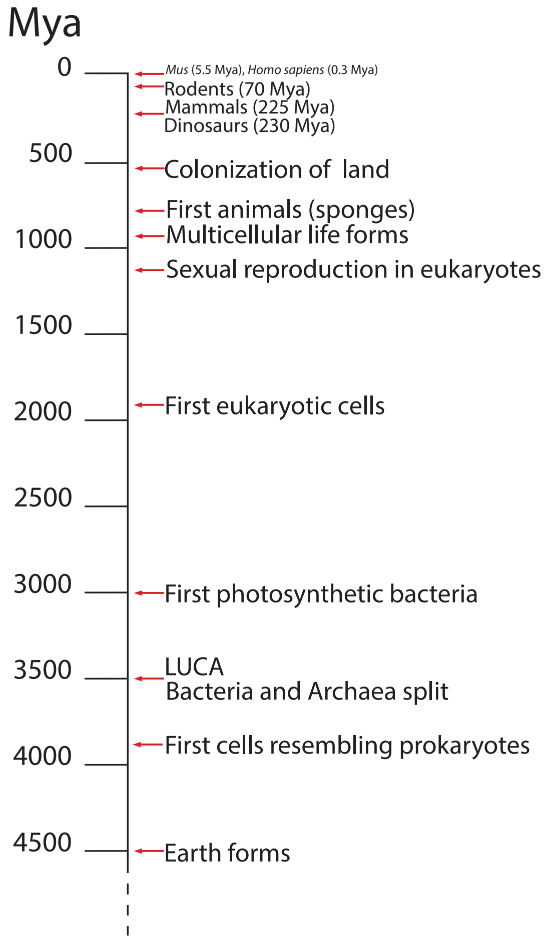


Figure 1: Timeline of the evolutionary history of life. The timeline illustrates the major events in the evolution of life starting with the formation of the earth until the emergence of mice (5.5 Mya) and modern humans (0.3 Mya). Figure adapted from Doms et al., 2018.

In this thesis, I first investigate the genetic basis underlying the host-microbiota relationships to provide insight into possible coevolution between hosts and their microbiota. Then, I perform fine-scale characterization of a candidate gene associated with bacterial abundance in two model organisms. Finally, I explore the possibility of using shotgun metagenome sequencing on mouse tissue samples to provide an alternative to feces samples, as the tissue-associated bacterial community is more stable and heritable.

1. On the origin of microbiome research

The word 'microbiome' was first mentioned by Whipps and colleagues in 1988 and who defined it as "a characteristic microbial community occupying a reasonably well-defined habitat which has distinct physio-chemical properties and not only refers to the microorganisms involved, but also encompasses their theaters of activity" (Fig. 2) (Whipps et al., 1988). Other definitions have surfaced and popularized the term, such as by Nobel laureate Joshua Lederberg in 2001, who describes microbiomes within an ecological context as "a community of commensal, symbiotic, and pathogenic microorganisms within a body space or other environment" (Lederberg and McCray, 2001), or by Marchesi and Ravel, who defined it as "the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions" (Marchesi and Ravel, 2015). However, the original definition by Whipps et al. is still the most comprehensive and captures the complexity of the microbiome and the diverse facets of its ecology and evolutionary biology (Fig. 2; Berg et al., 2020).

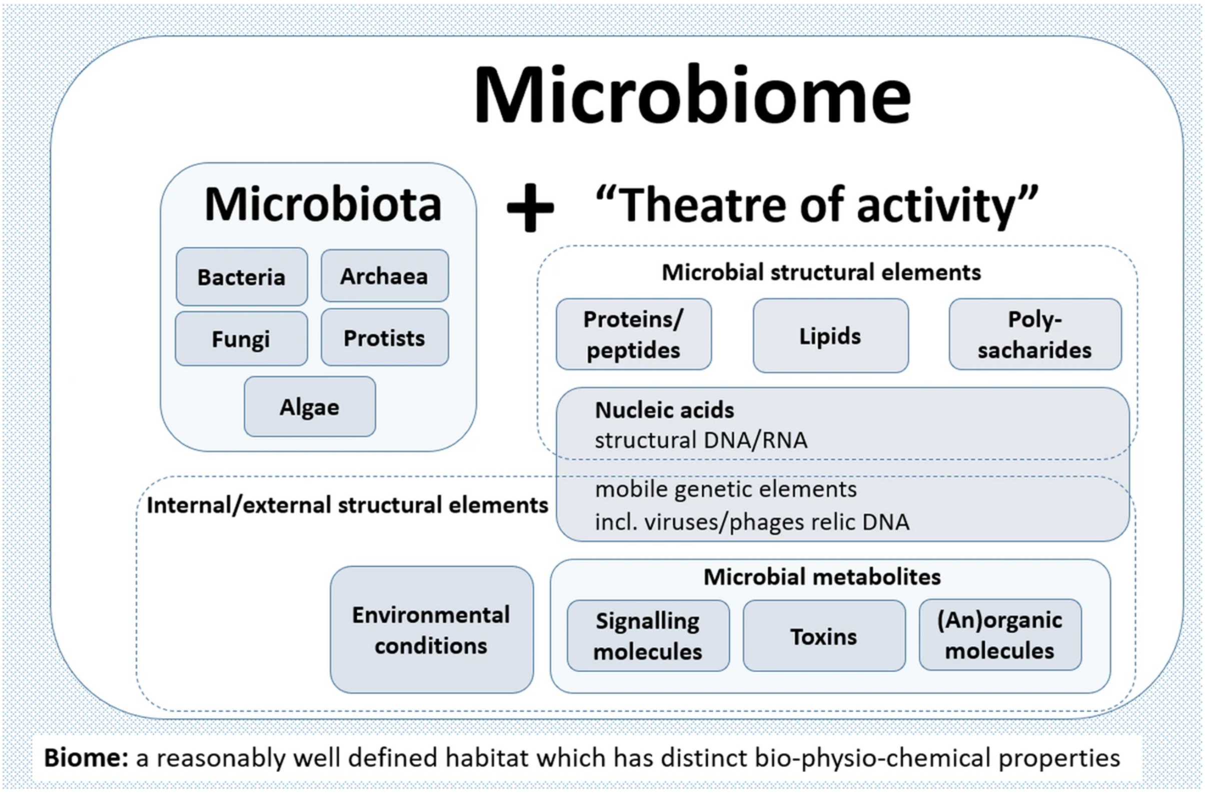


Figure 2: A diagram illustrating the makeup of the word microbiome, which includes both the microbiota (community of bacteria) and their "theatre of activity" (structural components, metabolites/signal molecules, and ambient circumstances). Figure from "Microbiome definition re-visited: old concepts and new challenges" by Berg G. et al., 2020, Microbiome (8), https://doi.org/10.1186/s40168-020-00875-0. Article is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

The term 'microbiota' can be traced back to as early as the 1800s (Lockyer, 1869) and has been in common use in the last 50 years. We define 'microbiota' here as the collection of microorganisms in a given environment. While the terminologies are relatively novel to our scientific language, the fundamental principles and significance of microbiome research date back to the early days of microbial ecology and to Sergei Winogradsky in the 1800s (Dworkin, 2012). While piecing together the microbial involvement in the nitrogen fixation cycle, Winogradsky realized the interconnectedness of microorganisms, where microbes occupy niches created by their neighbors activities and use the products of one metabolic pathways as substrates for another. He advocated for the need to study microbes in their natural environment (Winogradsky, 1949), essentially founding microbial ecology as an avenue of research.

Microbiome research has skyrocketed during the last decade. These advancements have been largely driven by the substantial cost reduction of high-throughput sequencing and the expansion of computational power, which has resulted in a mass of data that can be effectively handled on commonly available equipment. Our understanding of animal and environmental microbiomes has grown tremendously as a result of this data, and new discoveries are being produced on a daily basis.

2. Cospeciation and coevolution of the metaorganism

Essentially all animals and plants are colonized by microorganisms, with different host species having different microbial populations. These microbial communities form long-term relationships with their hosts and can have beneficial effect on the host's fitness, for example by providing nutrients (Rowland et al., 2018) or protection against pathogen invasion and colonization (Pickard et al., 2017). The bacteria, on the other hand, rely on the host to provide nutrients and maintain a stable environment (Rowland et al., 2018). These close interactions have often been used as evidence for coevolution between the host and their gut bacteria. However, coevolution occurs when two or more species reciprocally affect each other's evolution through the process of natural selection (Futuyma, 1983; Thompson, 1994; Page, 2003). A typical example of coevolution is seen in a predator-prey relationship, where prey develops adaptations to avoid predators and predators acquire additional adaptation in return. Cospeciation (or codiversification) is a form of coevolution in which one species' speciation dictates the speciation of another (Page, 2006). This is commonly studied in host-parasite relationships, when two hosts of the same species speciate, the parasite is no longer able to switch between those two hosts and thus, prevents the parasite populations from interacting and mating, and will ultimately lead to speciation within the parasite. In 1913, Heinrich Fahrenholz argued that when cospeciation occurs, the phylogenies of the host and parasite will eventually become congruent, or mirror each other (Robert C. King, 2007). Host switching, extinction, independent speciation, and other ecological processes, on the other hand, can change host-parasite phylogenies, making cospeciation more difficult to identify (Fig. 3). These co-phylogenetic patterns are also seen at the level of hosts and their entire community of microbes (microbiota) (Moeller et al., 2016; Brooks et al., 2016; Groussin et al., 2017; Gaulke et al., 2018; Youngblut et al., 2019) and is discussed under the term 'phylosymbiosis', where the similarity of the intestinal microbiota composition is positively associated with phylogenetic relatedness between hosts (Lim and Bordenstein, 2020). This can be the result of vertical inheritance, where ancestral linked microorganisms diversify in sync with the host throughout their evolutionary lineages, i.e. they codiversify. This is potentially, but not necessarily, due to their functioning as reciprocal selective pressures on one another, i.e. coevolving. However, other possible explanations exist to interpret congruent phylogenies. One is "ecological fitting", which was already proposed in 1985 by Daniel Janzen to oppose Fahrenholz's rule (Janzen, 1985; Agosta and Klemens, 2008). Janzen argues here that more closely related parasites will share similar traits that pertain to surviving on a particular host. This leans closely to Moran & Sloan's alternative explanation for phylosymbiosis, where more closely related host species simply are colonized by similar sets of microbial species from the environment ("ecological filtering") and also have greater likelihood to exchange microbes (Moran and Sloan, 2015). However, Mazel et al. showed that internal compartments of hosts, such as the gut, often display stronger phylosymbiosis than expected from a purely ecological filtering process, suggesting that other mechanisms, such as cospeciation and coevolution, are also involved (Mazel et al., 2018).

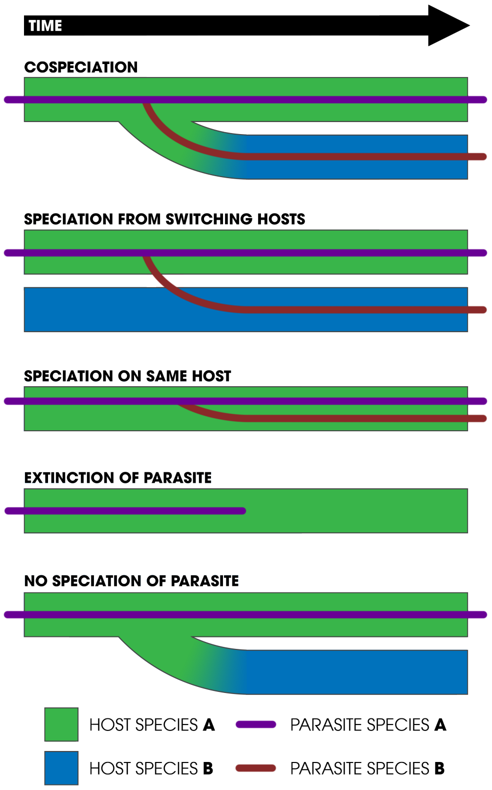


Figure 3: Cospeciation and host-parasite associations. From top to bottom: cospeciation: host and parasite speciate simultaneously; host switching: parasite switches hosts and evolves in reproductive isolation resulting in speciation; independent speciation: parasite speciates on same host due to reasons unrelated to host; extinction: parasite goes extinct on host; missing the boat: host speciates, but parasite does not end up reproductively isolated. Figure (and caption) copyright to Andrew Z. Colvin, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons.

2.1. The hologenome concept of evolution

In 2008, Zilber-Rosenberg and Rosenberg introduced the hologenome theory of evolution as a holistic view on coevolution between the host and its associated microbiota (Zilber-Rosenberg and Rosenberg, 2008). They propose that the holobiont (the host and its associated microbiota)(Meyer-Abich, 1943; Margulis, 1991) with its hologenome (the sum of the genetic information of the host and its microbiota), operating in consortium, should be considered a unit of selection in evolution, and that relatively rapid variation in the diverse microbiota can play a key role in the adaptation and evolution of the holobiont (Zilber-Rosenberg and Rosenberg, 2008). For natural selection to act upon the holobiont, it needs to fulfill the key conditions required by Darwin's Theory of Evolution via Natural Selection: (1) phenotypic variation occurs among individuals, (2) the phenotype is (at least partially) heritable and (3) is associated with the probability of survival and/or fertility between individuals. Translated to the holobiont, this implies that if there is (1) inter-individual variation in microbial composition, which is heritable (2), and that this variation is associated with a change in the host's fitness (3), then natural selection can act upon the holobiont (Hurst, 2017).

2.1.a. Inter-individual variation and the effect on host fitness

Many studies have demonstrated that individuals of the same host species show variation in their microbiome (condition 1) and that this variation is associated with a change in host fitness (condition 3), for example in a recent study, Zeevi et al. showed that structural variation in the gut microbiome was associated with host health (Zeevi et al., 2019).

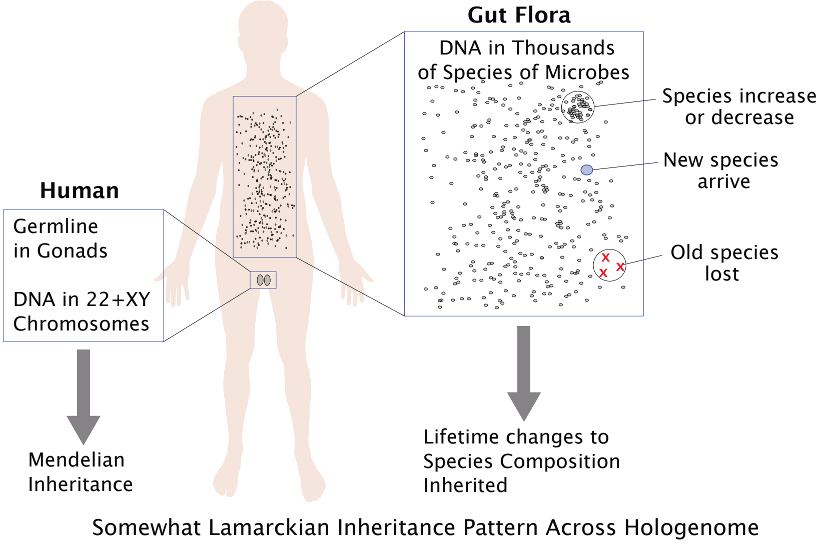

Gut Flora
DNA in Thousands of Species of Microbes
Species increase or decrease
New species arrive
Old species lost
Human
Germline in Gonads
DNA in 22+XY Chromosomes
Mendelian Inheritance
Lifetime changes to Species Composition Inherited
Somewhat Lamarckian Inheritance Pattern Across Hologenome

Figure 4: The hologenome is inherited in a Neo-Lamarckian manner. Microbe species in the intestinal flora can increase or decrease, be added to or removed. These alterations can have a hereditary impact, similar to Lamarckism rather than Mendelian genetics. Figure copyright to Ian Alexander, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons.

2.1.b. Transmission of the microbiome

The hologenome concept incorporates Lamarckian aspects within a Darwinian framework (Fig. 4). The nuclear genome is inherited mainly in a Mendelian process, while the microbiome is originally acquired from the environment, but may become inherited (Bordenstein and Theis, 2015). The gut microbiome can be inherited from parent to offspring in two ways. The first form is *direct* or *vertical* transmission of the microbe. Here, the microbe in the offspring is a direct descendant of the microbe in the parent. Transmission can also be *indirect* or *horizontal*, where the microbe is acquired from the environment instead of the parent. Vertical transmission can be brought about in several ways. Many plants and some animals are able to reproduce vegetatively, e.g. budding and fragmentation, causing the microbiota to be vertically transmitted (Adiyodi and Adiyodi, 1983; Hart, 2002; Vaughn, 2010). Endosymbionts, such as *Wolbachia* in numerous insects and *Buchnera* in aphids, are vertically transmitted via oocytes (Baumann et al., 1995; Veneti et al., 2004). Several animals perform coprophaghy, where the offspring consumes the mother's feces, in order to be capable of digesting complex nutrients after weaning. Due to contact of feces with the environment, this transmission method can also be considered partly horizontal. In humans, we consider the fetus sterile until birth (Lauder et al., 2016), after which it gets inoculated depending on the manner of birth with maternal vaginal or skin microbiota (vertical transmission). Moreover, human milk contains thousands of bacteria which will be (vertically) transmitted to the child during breastfeeding (Mueller et al., 2015). Examples of horizontally transmitted microbes are nitrogen-fixing rhizobia in legumes (Remigi et al., 2016) and the bioluminescent bacterium *Vibrio fischeri* colonizing the light organ in bobtail squids (McFall-Ngai, 2014). In most situations however, inheritance is not strictly vertical or horizontal, but a combination of both. Both transmission methods each have their advantages. Vertical transmission allows for precise transfer of the microbiome and therefore, promotes the maintenance of reciprocal metabolic processes within the holobiont, while horizontal transmission improves the chances of the holobiont in acquiring new genetic material from the environment (Rosenberg and Zilber-Rosenberg, 2018).

2.1.c. The holobiont in speciation

Another important point recognized by Darwin is that the formation of new species is associated with reproductive isolation (Kottler, 1978). Brucker and Bordenstein showed that the gut microbiome of closely related species of *Nasonia* form species-specific associations that cause lethality in interspecific hybrids, which could be undone with antibiotic treatment (Brucker and Bordenstein, 2013). From this they concluded that the hologenome breaks down during hybridization, originating from a mismatch between host genome and microbiome, promoting hybrid lethality and assisting speciation. A study analyzing the gut microbiota of two house mouse subspecies, *Mus musculus musculus* and *Mus musculus domesticus*, showed that hybrids display a variety of transgressive bacterial phenotypes combined with abnormal immune gene expression and increased intestinal pathology (Wang et al., 2015). Together, both studies suggest that microbiomes could contribute to reproductive isolation of the host, resulting in the formation of new species.

2.1.d. Controversy

The hologenome concept's usefulness, definition, and consequences have been a source of debate (Moran and Sloan, 2015; Bordenstein and Theis, 2015; Theis et al., 2016; Douglas and Werren, 2016). Both Moran and Sloan, and Douglas and Werren argue that the hologenome is most likely not the primary unit of selection, as here perfect agreement of selective interests among the microbiota and between microbiome and host is necessary (Moran and Sloan, 2015; Douglas and Werren, 2016). This is often not the case, even vertically transmitted microbes such as *Wolbachia*, result in fitness conflicts between the nuclear and maternally inherited genomes. *Wolbachia*, for example, is known to alter the sex ratio in favor of female offspring as they carry *Wolbachia* (Werren et al., 2008). Another argument opposing the hologenome concept of evolution is the lack of partner fidelity as this restricts the scope of the hologenome. As only a fraction of the microbiome is heritable (Org et al., 2015; van Opstal and Bordenstein, 2015;; Chen et al., 2018; Xu et al., 2020; Ishida et al., 2020) and thus, satisfies the necessary criteria, it is challenging to envision how the complete microbiome should be regarded as part of a 'hologenome' alongside its host (Douglas and Werren, 2016). However, a recent study in baboons showed that the gut microbiome heritability is nearly universal when using deep longitudinal sampling methods (Grieneisen et al., 2021). In conclusion, the hologenome concept of evolution is an interesting view on the evolution of host-microbe interactions, but whether this is an important unit of selection may be case-dependent.

3. Methods for studying the microbiome

3.1. The 16S rRNA gene as a phylogenetic marker

To date, bacterial 16S rRNA gene profiles have accounted for the lion's share of data obtained for the microbial component of holobionts. 16S ribosomal RNA (16S rRNA) is the RNA component of the 30S small subunit of a prokaryotic ribosome. Woese and Fox pioneered in using the 16S rRNA gene as a means for taxonomic classification to define the primary phylogenetic structure of the prokaryotic domain in 1977 (Woese and Fox, 1977). The ~1500 bp long conserved gene contains nine hypervariable regions, called V1-V9. Today's sequencing methods allow for amplifying 250 to 300bp long sequences, which makes the V1-V2 and the V3-V4 the most popular regions as a phylogenetic marker. However, PCR-based methods are sensitive to biases through sample preparation and amplification. Not all bacteria bind with the same efficiency to each of the primer sets for the different regions of the 16S rRNA, which results in an already biased sequences composition before sequencing (Johnson et al., 2019). Moreover, not all hypervariable regions are as variable for all taxa, which will result in a better discrimination in certain taxa for the V1-V2 region, while other taxa are better classified using the V3-V4 region (Rausch et al., 2019). Additionally, taxonomic classification is usually limited to the genus level depending on the database and classifiers used (Mizrahi-Man et al., 2013). Two common approaches are used to process the raw reads into an abundance table. The first one involves clustering sequences with 97% sequence similarity or more together into operational taxonomic units (OTUs), as pragmatic proxies for 'species' (Blaxter et al., 2005). In a second and more recent approach, erroneous sequences generated through PCR and sequencing are removed based on a error-learning algorithm and exact amplicon sequence variants (ASVs) are identified (Eren et al., 2013; Eren et al., 2015; Edgar and Flyvbjerg, 2015). This results in a higher resolution, where variation by single nucleotide change can be distinguished. This method has the additional advantage that ASVs inferred from different dataset can be validly compared, as ASVs are consistent labels representing a biological reality existing outside the data being analyzed (Callahan et al., 2017), making this approach into the current method of choice. Using the 16S rRNA gene sequencing will not provide any direct functional information, although this can be imputed with PICRUSt (Douglas et al., 2018).

3.2. Shotgun metagenome sequencing

Shotgun metagenomic sequencing involves the sequencing of all available DNA in the sample, instead of a particular marker gene. This limits the choice of samples to only samples with a low content of host DNA. Shotgun metagenome sequencing can offer species- and strain-level classification of bacteria (Li et al., 2020). It has the advantage of its relative lack of bias and allows the examination of the functional content of the microbiota (Heintz-Buschart and Wilmes, 2018). Furthermore, not yet classified bacteria can also be discovered through de novo genome binning if the sequencing coverage is high enough. Shotgun sequencing comes however with a relatively high cost and demands more computational power for analysis.

3.3. Quantitative trait locus analysis and genome-wide association studies

In quantitative trait locus (QTL) analysis and genome-wide association studies (GWAS), two types of data are associated (phenotypic data or traits, and genotypic data) in order to explain the genetic basis of variation in complex traits (Falconer, 1996; Lynch and Walsh, 1998; Walsh and Lynch, 2018). The term QTL mapping is frequently used when performing association studies in biparental populations, while GWAS is applied on unrelated or multi-parental individuals. Microbial QTL in inbred mice and GWAS in mice and humans have been successful in determining quantitative trait loci (QTLs) associated with quantitative measures of the microbiome. In mice, Benson et al. performed QTL mapping on a large murine advanced intercross population (Benson et al., 2010). They identified 18 host QTLs that showed significant or suggestive genome-wide associations with relative abundances of specific microbial taxa and thus providing strong evidence of the importance of host genetic regulation in shaping the composition of the mammalian microbiome. McKnite et al. and Org et al. both discovered a significant association of a bacterial taxon with a locus containing the *Irak4* gene using inbred mice strains (McKnite et al., 2012; Org et al., 2015). In humans, the first microbial GWAS was performed in a small sample set of 93 individuals from the Human Microbiome Project (Blekhman et al., 2015). Here, they found 83 associations between genetic variation in host coding sequence and abundance of specific microbial taxa, among them was a significant correlation between SNPs in the LCT gene and the abundance of *Bifidobacterium* in the gastrointestinal tract. This association was later confirmed in other studies (Bonder et al., 2016; Wang et al., 2016; Kato et al., 2018; Kurilshikov et al., 2020; Hughes et al., 2020; Rühlemann et al., 2021). Together, these studies show a clear genetic basis for microbial phenotypes, although results are very population-dependent making replication of results arduous.

3.4. From identification to function

Microbial GWAS and QTL studies are very helpful in identifying regions associated with bacterial abundances, however these are only correlations and not causal relationships. This calls for follow-up studies to determine the underlying functional basis and the mechanisms through which these variants act. As many of the GWAS loci are in regulatory regions, gene expression data can be used to scan for expression quantitive trait loci (eQTL) and identify genetic variants linked to changes in transcript abundance across individuals (Majewski and Pastinen, 2011). Colocalisation of GWAS loci with eQTL is one statistical method used to detect signals mediated by the same causative variations (Broekema et al., 2020). However, to advance from identification to function of genomic regions associated with bacterial abundances, we have to utilize genome edited model organisms, such as a knock-out (KO) or overexpression (OE) animal model for the gene of interest. Only using such models can we test the influence of the gene expression on the bacterial abundances. Moreover, we can use germ-free (GF) models to test the reciprocal effect of the bacterial species on gene expression.

Results

Chapter 1: Identifying host genomic regions influencing bacterial traits

This chapter has been submitted to eLIFE and bioRxiv (doi: https://doi.org/10.1101/2021.09.28.462095) as "Key features of the genetic architecture and evolution of host-microbe interactions revealed by high-resolution genetic mapping of the mucosa-associated gut microbiome in hybrid mice".

1. Introduction

The recent widespread recognition of the gut microbiome's importance to host health and fitness represents a critical advancement of biomedicine. Host phenotypes affected by the gut microbiome are documented in humans (Ley et al., 2006; Turnbaugh et al., 2009; Lynch and Pedersen, 2016), laboratory animals (Backhed et al., 2004; Turnbaugh et al., 2008; Rolig et al., 2015; Rosshart et al., 2017; Gould et al., 2018), and wild populations (Suzuki, 2017; Roth et al., 2019; Suzuki et al., 2020; Hua et al., 2020), and include critical traits such as aiding digestion and energy uptake (Rowland et al., 2018), and the development and regulation of the immune system (Davenport, 2020).

Despite the importance of gut microbiome, community composition varies significantly among host species, populations, and individuals (Benson et al., 2010; Yatsunenko et al., 2012; Brooks et al., 2016; Rehman et al., 2016; Amato et al., 2019). While a portion of this variation is expected to be selectively neutral, alterations of the gut microbiome are on the one hand linked to numerous human diseases (Carding et al., 2015; Lynch and Pedersen, 2016), such as diabetes (Qin et al., 2012), inflammatory bowel disease (IBD) (Ott et al., 2004; Gevers et al., 2014) and mental disorders (Clapp et al., 2017). On the other hand, there is evidence that the gut microbiome can play an important role in adaptation on both recent- (Hehemann et al., 2010; Suzuki and Ley, 2020) and ancient evolutionary timescales (Rausch et al., 2019). Collectively, these phenomena suggest that it would be evolutionarily advantageous for hosts to influence their microbiome.

An intriguing observation made in comparative microbiome research in the last decade is that the pattern of diversification among gut microbiomes appears to mirror host phylogeny (Ochman et al., 2010). This phenomenon, coined "phylosymbiosis"

(Brucker and Bordenstein, 2012a; Brucker and Bordenstein, 2012b; Lim and Bordenstein, 2020), is documented in a number of diverse host taxa (Brooks et al., 2016) and also extends to the level of the phageome (Gogarten et al., 2021). Several non-mutually exclusive hypotheses are proposed to explain phylosymbiosis (Moran and Sloan, 2015). However, it is likely that vertical inheritance is important for at least a subset of taxa, as signatures of co-speciation/-diversification are present among numerous mammalian associated gut microbes (Moeller et al., 2016; Groussin et al., 2017; Moeller et al., 2019), which could also set the stage for potential coevolutionary processes. Importantly, experiments involving interspecific fecal microbiota transplants indeed provide evidence of host adaptation to their conspecific microbial communities (Brooks et al., 2016; Moeller et al., 2019). Further, cospeciating taxa were observed to be significantly enriched among the bacterial species depleted in early onset IBD, an immune-related disorder, suggesting a greater evolved dependency on such taxa (Papa et al., 2012; Groussin et al., 2017). However, the nature of genetic changes involving host-microbe interactions that take place as new host species diverge remains under-explored.

House mice are an excellent model system for evolutionary microbiome research, as studies of both natural populations and laboratory experiments are possible (Suzuki, 2017; Suzuki et al., 2019). In particular, the house mouse species complex is comprised of subspecies that hybridize in nature, enabling the potential early stages of codiversification to be studied. We previously analyzed the gut microbiome across the central European hybrid zone of *Mus musculus musculus* and *M. m. domesticus* (Wang et al., 2015), which share a common ancestor ~ 0.5 million years ago (Geraldes et al., 2008). Importantly, transgressive phenotypes (i.e. exceeding or falling short of parental values) among gut microbial traits as well as increased intestinal histopathology scores were common in hybrids, suggesting that the genetic basis of host control over microbes has diverged (Wang et al., 2015). The same study performed an F2 cross between wild-derived inbred strains of *M. m. domesticus* and *M. m. musculus* and identified 14 quantitative trait loci (QTL) influencing 29 microbial traits. However, like classical laboratory mice, these strains had a history of rederivation and reconstitution of their gut microbiome, thus leading to deviations from the native microbial populations found in nature (Rosshart et al., 2017; Org and Lusis, 2018), and the genomic intervals were too large to identify individual genes.

In this study, we employed a powerful genetic mapping approach using inbred lines directly derived from the *M. m. musculus* - *M. m. domesticus* hybrid zone, and further focus on the mucosa-associated microbiota due to its more direct interaction with host cells (Fukata and Arditi, 2013; Chu and Mazmanian, 2013), distinct functions compared to the luminal microbiota (Wang et al., 2010; Vaga et al., 2020), and greater dependence on host genetics (Spor et al., 2011; Linnenbrink et al., 2013). Previous mapping studies using hybrids raised in a laboratory environment showed that high mapping resolution is possible due to the hundreds of generations of natural admixture between parental genomes in the hybrid zone (Turner and Harr, 2014; Pallares et al., 2014; Škrabar et al., 2018). Accordingly, we here identify 443 loci contributing to variation in 123 taxa, whose narrow genomic intervals (median <2Mb) enable many individual candidate genes and pathways to be pinpointed. We identify a high proportion of bacterial taxa with significant heritability estimates, and find that bacterial phenotyping based on 16S rRNA transcript compared to gene copy-based profiling yields an even higher proportion. Further, these heritability estimates also significantly positively correlate with cospeciation rate estimates, suggesting a more extensive host genetic architecture for cospeciating taxa. Finally, we identify numerous enriched functional pathways, whose role in host-microbe interactions may be particularly important as new species form.

2. Results

2.1. Microbial community composition

To obtain microbial traits for genetic mapping in the G2 mapping population, we sequenced the 16S rRNA gene from caecal mucosa samples of 320 hybrid male mice based on DNA and RNA (cDNA), which reflect bacterial cell number and activity, respectively. After applying quality filtering and subsampling 10,000 reads per sample, we identified a total of 4684 amplicon sequence variants (ASVs). For further analyses, we established a "core microbiome" (defined in Methods), such that analyses were limited to those taxa common and abundant enough to reveal potential genetic signal. The core microbiome is composed of four phyla, five classes, five orders, eleven families, 27 genera, and 90 ASVs for RNA, and four phyla, five classes, six orders, twelve families, 28 genera and 46 ASVs for DNA. A combined total of 98 unique ASVs belong to the core, of which 38 were shared between DNA and RNA (Suppl. Fig. 1). The most abundant genus in our core microbiome is *Helicobacter* (Suppl. Fig. 2), consistent with a previous study of the wild hybrid *M. m. musculus/M. m. domesticus* mucosa-associated microbiome (Wang et al., 2015).

2.2. Correlation between host genetic relatedness and microbiome structure

To gain a broad sense of the contribution of genetic factors to the variability of microbial phenotypes in our mapping population, we compared the kinship matrix based on genotypes to an equivalent based on gut microbial composition, whereby ASV abundances were used as equivalents of gene dosage. We found a significant correlation between these matrices (P = .001, R^2=0.03, Suppl. Fig. 3), indicating a host genetic effect on the diversity of the gut microbiota.

2.3. SNP-based heritability

Next, we used a SNP-based approach to estimate the proportion of variance explained (PVE) of the relative abundance of taxa, also called the narrow-sense heritability (h^2) or SNP-based heritability. Out of the 153 total core taxa, we identified 46 taxa for DNA and 69 taxa for RNA with significant heritability estimates (P_{RLRT} < .05), with estimates ranging between 29 and 91% (see Fig. 5A-B and Suppl. Table 1). An unclassified genus belonging to the phylum Bacteroidetes followed by ASV7 (genus *Paraprevotella*), *Paraprevotella* and Paraprevotellaceae showed the highest heritability among DNA-based traits (91.8%, 88.8%, 88.8%, and 87.1%, respectively; Fig. 5A), while ASV97 (genus *Oscillibacter*), followed by

Prevotellaceae, *Paraprevotella* and ASV7 (*Paraprevotella*) had the highest heritability among RNA-based traits (86.6%, 85.7%, 85.7%, and 85.6%, resp.; Fig. 5B). The heritability estimates for DNA- and RNA-based measurements of the same taxa are significantly correlated (*P* = 5.013 x 10-8, R2=0.58, Suppl. Fig. 4), and neither measure appears to be systematically more heritable than another, i.e. some taxa display higher RNA-based heritability estimates and others higher DNA-based estimates.

2.4. Heritability estimates are correlated with predicted co-speciation rates

In an important meta-analysis of the gut microbiome across diverse mammalian taxa, Groussin et al. (2017) estimated co-speciation rates of individual bacterial taxa by measuring the congruence of host and bacteria phylogenetic trees relative to the number of host-swap events. We reasoned that taxa with higher co-speciation rates might also demonstrate higher heritability, as these more intimate evolutionary relationships would provide a greater opportunity for genetic aspects to evolve. Intriguingly, we observe a significant positive correlation for RNA-based traits (*P*= .008, R2=.46, Fig. 5D) and a similar trend for DNA (*P*= 0.1; Fig. 5C). These results support the notion that cospeciating taxa evolved a greater dependency on host genes, and further suggest that bacterial activity may better reflect the underlying biological interactions.

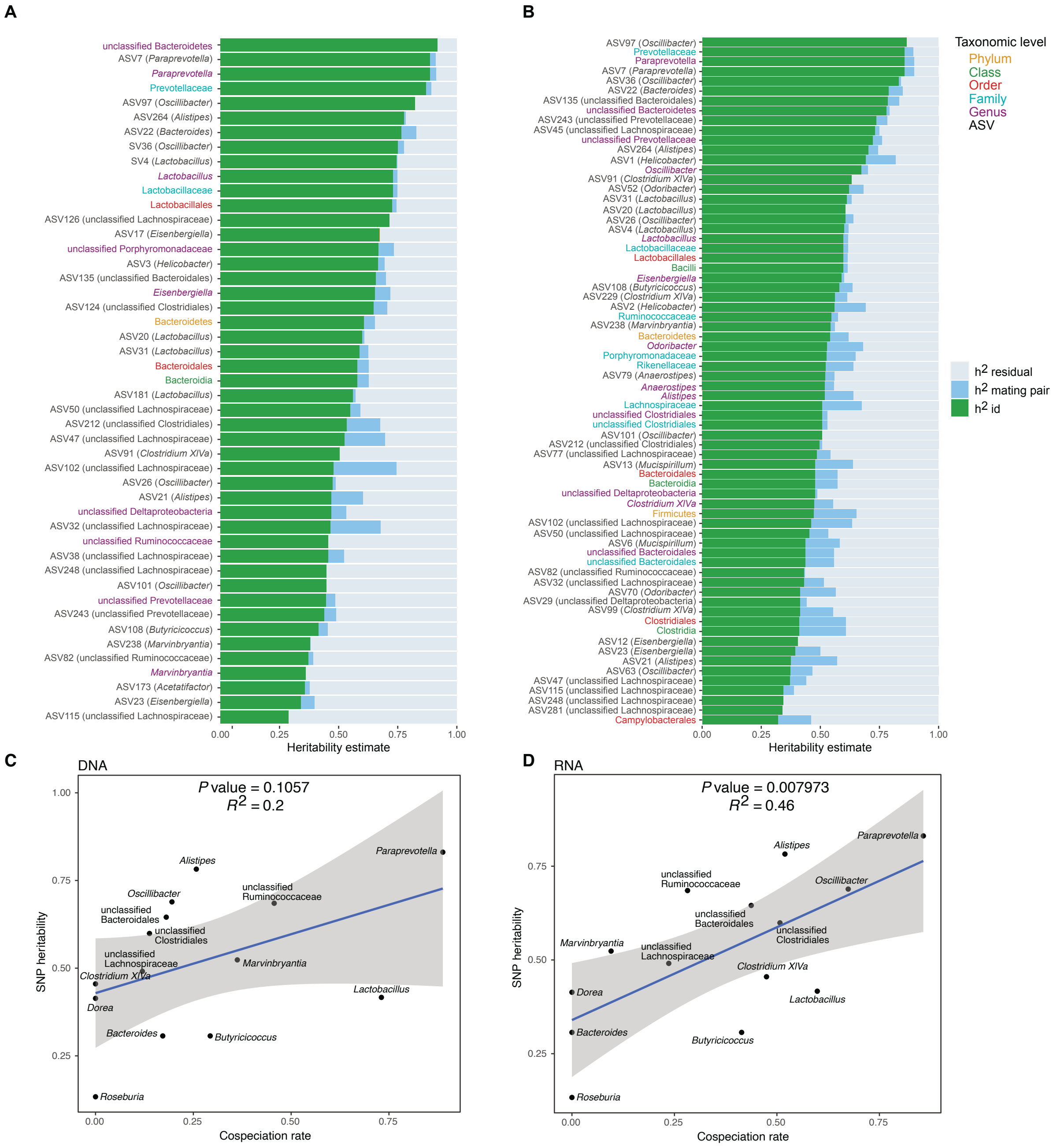


Figure 5: (A-B) Heritability estimates for the relative abundance of bacterial taxa. Proportion of variance explained for each taxon on DNA level (A), and RNA level (B) for all SNPs (GRM) in green, mating pair identifier in blue and residual variance in grey. Only significant heritability estimates are shown (P < .05). The text labels on the y-axis are colored according to taxonomic level: ASV in black, genus in purple, family in light blue, order in red, class in green, and phylum in yellow. (C-D) Relationship between the heritability estimates for the relative abundance of bacterial taxa and co-speciation rate for the same genus calculated by Groussin et al. (2017). DNA level (C), and RNA level (D). The blue line represents a linear regression fit to the data and the grey area the corresponding confidence interval.

2.5. Genetic mapping of host loci determining microbiome composition

Next, we performed genome-wide association mapping of the relative abundances of core taxa, in addition to two alpha-diversity measures (Shannon and Chao1 indices), based on 32,625 SNPs. We included both additive and dominance terms in the model to enable the identification of under- and over-dominance (see Methods). While we found no significant associations for alpha diversity at either the DNA or RNA level ($P > 1.53 \times 10^{-6}$), a total of 1099 genome-wide significant associations were identified for individual taxa ($P < 1.53 \times 10^{-6}$, Suppl. Table 2), of which 443 achieved study-wide significance ($P < 1.29 \times 10^{-8}$). Apart from the X chromosome, all autosomal chromosomes contained study-wide significant associations (Fig. 6). Out of the 153 mapped taxa, 123 had at least one significant association (Table 1). For the remainder of our analyses, we focus on the results using the more stringent study-wide threshold, and combined significant SNPs within 10 Mb into significant regions (Suppl. Table 3). The median size of significant regions is 1.91 Mb, which harbor a median of 14 protein-coding genes. On average, we observe 10 significant mouse genomic regions per bacterial taxon.

Of the significant loci with estimated interval sizes, we find 73 intervals (16.5%) that are smaller than one Mb (Suppl. Table 4). The smallest interval is only 231 bases and associated with the RNA-based abundance of an unclassified genus belonging to Deltaproteobacteria. It is situated in an intron of the C3 gene, a complement component playing a central role in the activation of the complement system, which modulates inflammation and contributes to antimicrobial activity (Ricklin et al., 2016).

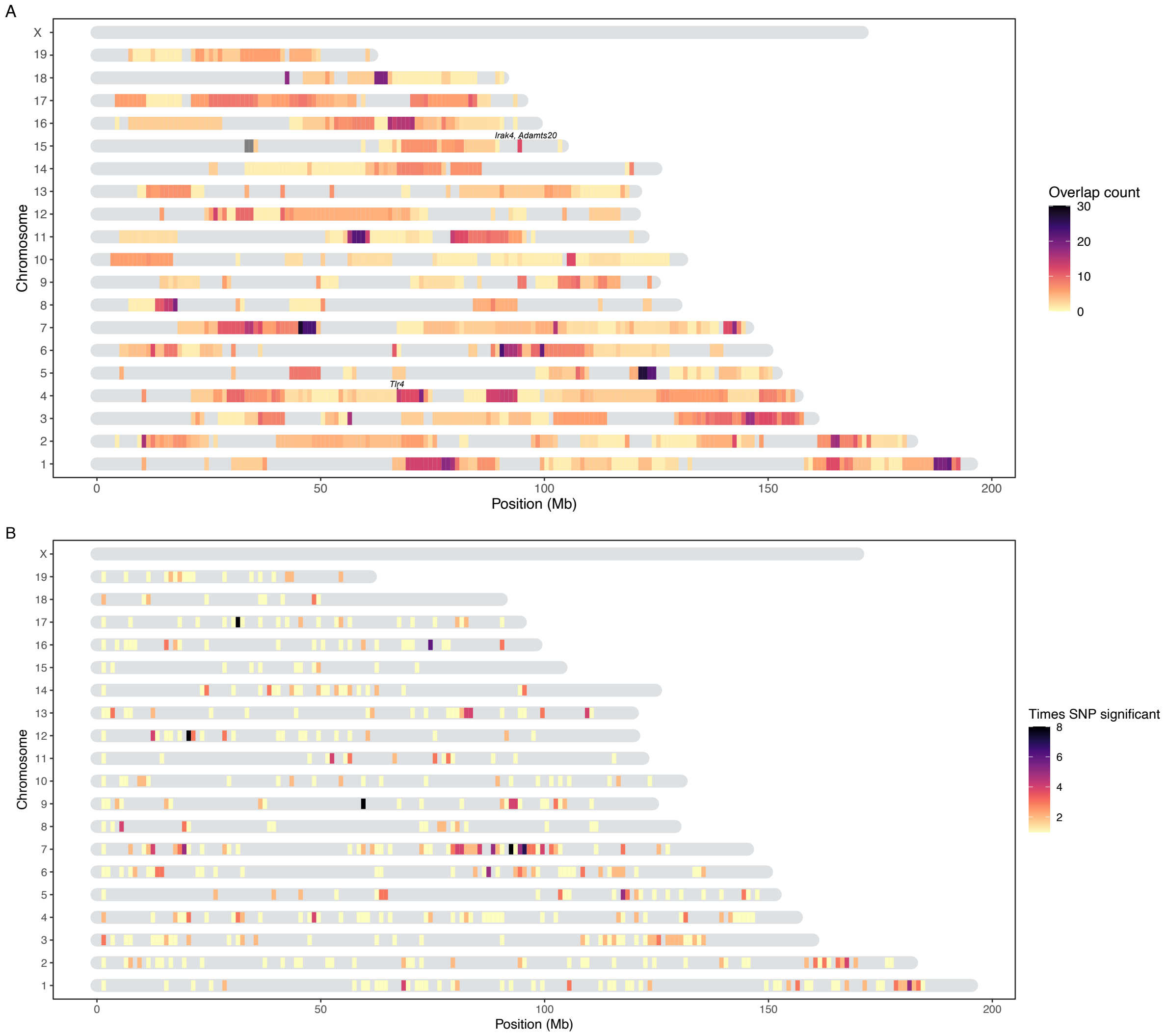


Figure 6: Heatmap of significant host loci from association mapping of bacterial abundances. Karotype plot showing the number of significant loci found using a study-wide threshold, where (A) plots the significance intervals, and (B) the significant SNP markers on the chromosomes.

The significant genomic regions and SNPs are displayed in Figure 6A and 3B, respectively. Individual SNPs were associated with up to 12 taxa, and significant intervals with up to 30 taxa. The SNPs with the lowest *P* values were associated with the genus *Dorea* and two ASVs belonging to *Dorea* (ASV184 and ASV293; Suppl. Fig. 5). At the RNA level this involves two loci: mm10-chr4: 67.07 Mb, where the peak SNP is 13 kb downstream of the closest gene *Tlr4* (UNC7414459, *P*=2.31 × 10^{-69}, additive *P*= 4.48 × 10^{-118}, dominance *P*= 1.37 × 10^{-111}), and mm10-chr15: 94.4 Mb, where the peak SNP is found within the *Adamts20* gene (UNC26145702, *P*=4.51 × 10^{-65}, additive *P*= 1.87 × 10^{-113}, dominance *P*= 1.56 × 10^{-105}; Suppl. Fig. 5). Interestingly, the *Irak4* gene, whose protein product is rapidly recruited after TLR4 activation, is also located 181 kb upstream of *Adamts20*. The five taxa displaying the most associations were ASV19 (*Bacteroides*), *Dorea*, ASV36 (*Oscillibacter*), ASV35 (*Bacteroides*), and ASV98 (unclassified Lachnospiraceae) (Suppl. Fig. 6).

Table 1: Overview of mapping statistics.

	DNA	RNA	Total
Mapped taxa	101	142	153
Taxa with significant loci	67	96	123
Median interval size (Mb)	1.52	2.29	1.91
Total significant loci	478	791	1269
Unique significant loci	179	313	443
Significant loci total *P*	91	167	233
Significant loci additive *P*	155	260	377
Significant loci dominance *P*	95	166	231
Median significant loci per trait	5	6	8
Median unique significant loci per trait	3	3	4
Median unique significant SNPs per locus	2	2.5	2
Median number of genes per locus	31	52	43
Median protein coding genes per locus	11	15	14

2.6. Ancestry, dominance, and effect sizes

A total of 435 significant SNPs were ancestry informative between *M. m. musculus* and *M. m. domesticus* (*i.e.* represent fixed differences between subspecies). To gain further insight on the genetic architecture of microbial trait abundances, we estimated the degree of dominance at each significant locus using the *d/a* ratio (Falconer, 1996), where alleles with strictly recessive, additive, and dominant effects have *d/a* values of -1, 0, and 1, respectively. As half of the SNPs were not ancestry informative (Fig. 7A), it was not possible to consistently have *a* associated with one parent/subspecies, hence we report *d/|a|* such that it can be interpreted with respect to bacterial abundance. For the vast majority of loci (83.53%), the allele associated with lower abundance is dominant or partially dominant (-1.25 < *d/|a|* < -0.75; Fig. 7B). On the basis of the arbitrary cutoffs we used to classify dominance, only a small proportion of alleles are underdominant (0.22%; *d/|a|* < -1.25) or overdominant (0.15%; *d/|a|*> 1.25). However for one-third of the significant SNPs, the heterozygotes display transgressive phenotypes, i.e. mean abundances that are either significantly lower (31% of SNPs)- or higher (2% of SNPs) than those of both homozygous genotypes. Interestingly, the *domesticus* allele was associated with higher bacterial abundance in two-thirds of this subset (33.2% vs 16.3% *musculus* allele; Fig. 7A).

Next, we estimated phenotypic effect sizes by calculating the percentage variance explained (PVE) by the peak SNP of each significant region. Peak SNPs explain between 3% and 64% of the variance in bacterial abundance, with a median effect size of 9.3% (Fig. 7C). The combined effects of all significant loci for each taxon ranged from 4.9% to 259%, with a median of 41.8% (Fig. 7D). Note, combined effects for many taxa (33 out of 59) exceed SNP-heritability estimates (Fig. 5). While exceeding 100% explained variance is biologically possible, as loci can have opposite phenotypic effects, many of these are likely inflated due to the Beavis effect (Beavis, 1994).

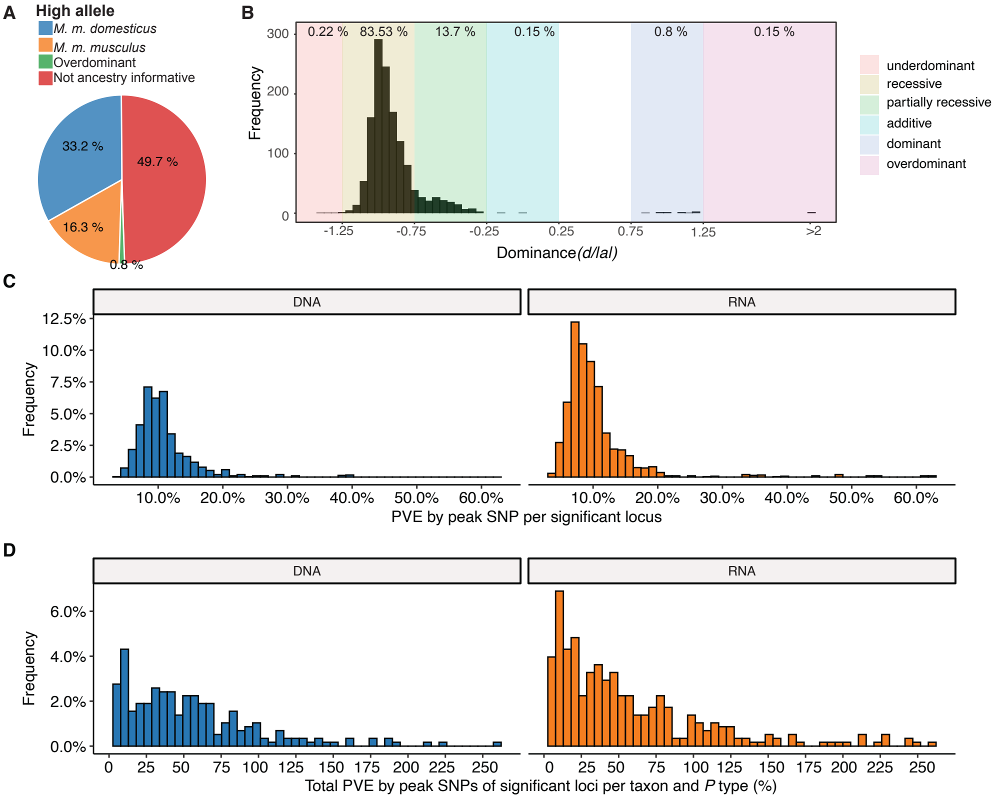


Figure 7: Genetic architecture of significant loci. A) Source of the allele with the highest phenotypic value. B) Histogram of dominance values d/a of significant loci reveals a majority of loci acting recessive or partially recessive. C) Histogram showing the percentage of variance explained (PVE) by the peak SNP for DNA (blue, left) and RNA (orange, right). D) Collective PVE by lead SNPs of significant loci within a taxon. Values are calculated separately for each P value type (total, additive, and dominance).

2.7. Functional annotation of candidate genes

In order to reveal potential higher level biological phenomena among the identified loci, we performed pathway analysis to identify interactions and functional categories enriched among the genes in significant intervals. We used STRING (Szklarczyk et al., 2019) to calculate a protein-protein interaction (PPI) network of 925 protein-coding genes nearest to significant SNPs (upstream and downstream). A total of 768 genes were represented in the STRING database, and the maximal network is highly significant (PPI enrichment *P* value: 2.15×10^{-14}) displaying 668 nodes connected by 1797 edges and an average node degree of 4.68. After retaining only the edges with the highest confidence (interaction score > 0.9), this results in one large network with 233 nodes, 692 edges and ten smaller networks (Fig. 8).

Next, we functionally annotated clusters using STRING's functional enrichment plugin. The genes of the largest cluster are part of the G protein-coupled receptor (GPCR) ligand binding pathway. GPCRs are the largest receptor superfamily and also the largest class of drug targets (Sriram and Insel, 2018). We then calculated the top ten hub proteins from the network based on Maximal Clique Centrality (MCC) algorithm with CytoHubba to predict important nodes that can function as 'master switches' (Suppl. Fig. 7). The top ten proteins contributing to the PPI network were GNG12, MCHR1, NMUR2, PROK2, OXTR, XCR1, TACR3, CHRM3, PTGFR, and C3, which are all involved in the GPCR signaling pathway.

Further, we performed enrichment analysis on the 925 closest genes using the *clusterprofiler* R package. We found 14 KEGG pathways to be over-represented: circadian entrainment, oxytocin signaling pathway, axon guidance, calcium signaling, cAMP signaling, cortisol synthesis and secretion, cushing syndrome, gastric acid secretion, glutamatergic synapse, mucin type O-glycan biosynthesis, inflammatory mediator regulation of TRP channels, PD-L1 expression and the PD-1 checkpoint pathway in cancer, tight junction, and the *Wnt* signaling pathway (Suppl. Table 5, Suppl. Fig. 8-9). Finally, genes involved in five human diseases are enriched, among them mental disorders (Suppl. Fig. 10).

Finally, due to the observation of a significant enrichment of cospeciating taxa among the bacterial species depleted in early onset IBD (Groussin et al., 2017) and the evidence that IBD is especially associated with a dysbiosis in mucosa-associated communities (Yang et al., 2020a; Daniel et al., 2021), we specifically examined possible over-representation of genes involved in IBD (Khan et al., 2021) among the 925 genes neighboring significant SNPs. We found 14 out of the 289 IBD genes, which was significantly more than expected by chance (10 000 times permuted mean: 2.7, simulated *P* = .0001; Suppl. Table 6). Interestingly, SNPs in five out of the 14 genes are associated with ASVs belonging to the genus *Oscillibacter*, a cospeciating taxon known to decrease during the active state of IBD (Metwaly et al., 2020).

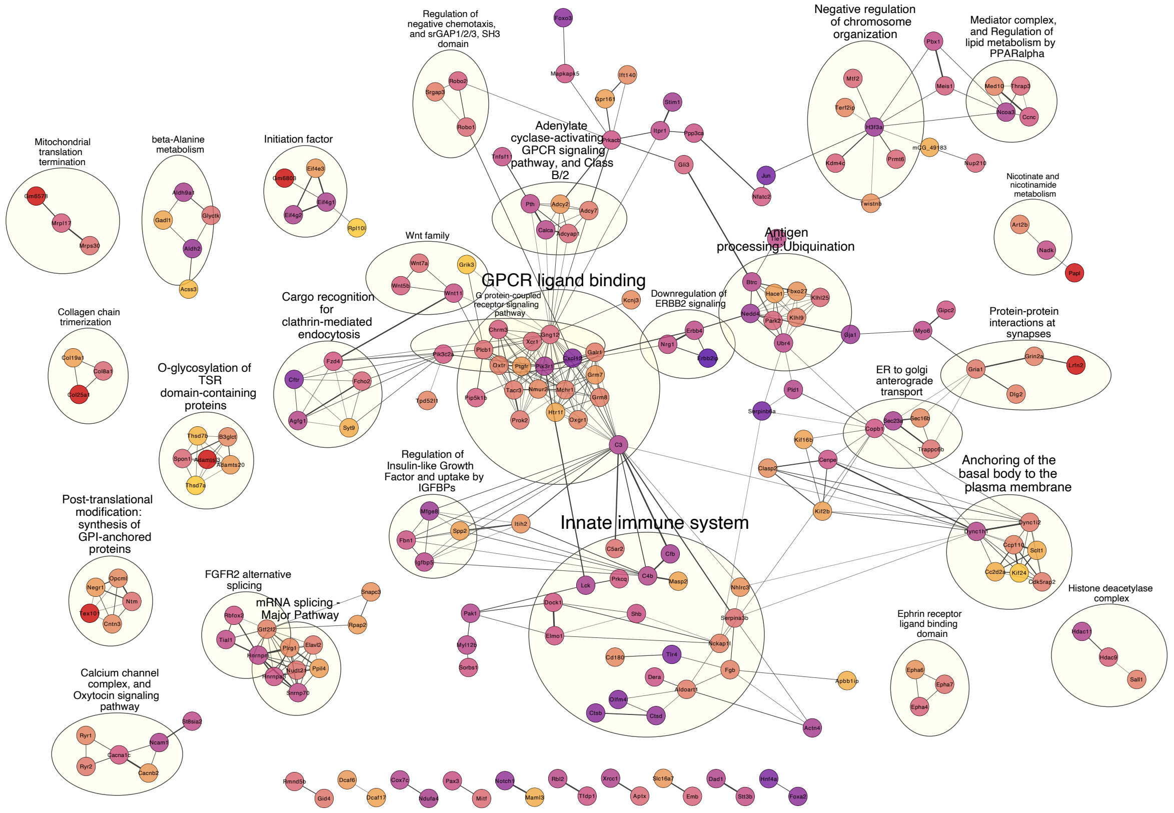


Figure 8: High confidence protein-protein interaction network of genes closest to significant SNPs. Network clusters are annotated using STRING's functional enrichment. Nodes represent proteins and edges their respective interactions. Only edges with an interaction score higher than 0.9 are retained. The width of the edge line expresses the interaction score calculated by STRING. The color of the nodes describe the expression of the protein in the intestine where yellow is not expressed and purple is highly expressed. Diamond shaped nodes are proteins coded by genes that contain a significant SNP. Round nodes are proteins from genes closest to a significant marker.

2.8. Comparison of significant loci to published mouse gut microbiome studies

Next, we compiled a list of 648 unique confidence intervals of significant associations with gut bacterial taxa from seven previous mouse QTL studies (Benson et al., 2010; McKnite et al., 2012; Leamy et al., 2014; Wang et al., 2015; Org et al., 2015; Snijders et al., 2016; Kemis et al., 2019) and compared this list to our significance intervals for bacterial taxa at both the DNA and RNA level (346 unique intervals). Regions larger than 10Mb were removed from all studies. We found 434 overlapping intervals, which is significantly more than expected by chance (10 000 times permuted mean: 368, simulated *P*=.0073, see Methods). Several of our smaller significant loci overlapped with larger loci from previous studies and removing this redundancy left 186 significant loci with a median interval size of 0.78 Mb (Fig. 9). The most frequently identified locus is located on chromosome 2 169-171 Mb where protein coding genes *Gm11011, Znf217, Tshz2, Bcas1, Cyp24a1, Pfdn4, 4930470P17Rik, and Dok5* are situated.

Additionally, we collected genes within genome-wide significant regions reported in seven human microbiome GWAS (mGWAS) (Bonder et al., 2016; Turpin et al., 2016; Goodrich et al., 2016; Wang et al., 2016; Hughes et al., 2020; Rühlemann et al., 2021; Kurilshikov et al., 2021). However, no significant over-representation of genes was found within our significance intervals (*P* = .156*)*, nor within our list of genes closest to a significant SNP (*P* = .62).

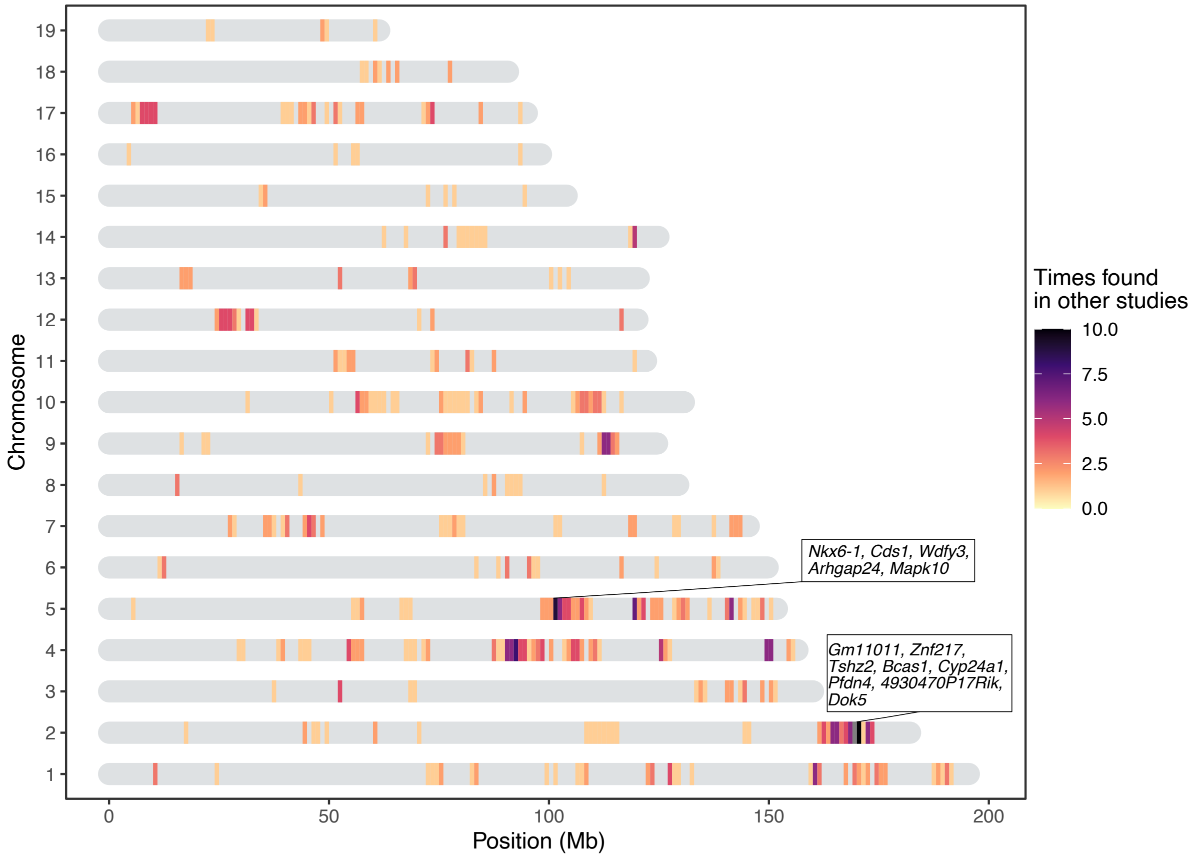


Figure 9: Heatmap showing the significant loci in this study that were previously found in other QTL studies of the mouse gut microbiome. The genes present in two repeatedly identified regions are depicted in boxes.

2.9. Proteins differentially expressed in germ-free vs conventional mice

To further validate our results, we compared the list of genes contained within intervals of our study to a list of differentially expressed protein between germ-free and conventionally raised mice (Mills et al., 2020). This comparison was made based on the general expectation that genes associated with variation in microbial abundances would be more likely to differ according to the colonization status of the host. Thus, we examined the intersection between genes identified in our study and the proteins identified as highly associated ($|\pi| > 1$) with the colonization state of the colon and the small intestine (Mills et al., 2020). Out of the 373 over- or under-expressed proteins according to colonization status, we find 198 of their coding genes to be among our significant loci, of which 17 are the closest genes to a significant marker (*Iyd, Nln, Slc26a3, Slc3a1, Myom2, Nebl, Tent5a, Fxr1, Cbr3, Chrodc1, Nucb2, Arhgef10l, Sucla2, Enpep, Prkcq, Aacs, and Cox7c*). This is significantly more than expected by chance (simulated *P*=.0156, 10 000 permutations). Further, analyzing the protein-protein interactions with STRING results in a significant network (*P*=1.73 × 10-14, and average node degree 2.4, Suppl. Fig. 11), with *Cyp2c65, Cyp2c55, Cyp2b10, Gpx2, Cth, Eif3k, Eif1, Sucla2,* and *Rpl17* identified as hub genes (Suppl. Fig. 12).

Subsequently, we merged the information from Mills et al. (2020) and the seven previous QTL mapping studies discussed above to further narrow down the most promising candidate genes, and found 30 genes overlapping with our study. Of these 30 genes, six are the closest gene to a significant SNP. These genes are myomesine 2 (*Myom2*), solute carrier family 3 member 1 (*Slc3a1*), solute carrier family 26 member 3 (*Slc26a3*), nebulette (*Nebl*), carbonyl reductase 3 (*Cbr3*), and acetoacetyl-coA synthetase (*Aacs*).

2.10. Candidate genes influencing bacterial abundance

Finally, all previously mentioned candidate genes were combined in one gene set of 304 genes and compiled in a highly significant PPI network (*P* < 1.0 × 10^{-16}, average node degree=4.85, see Methods 4.13). Guided by this network, we filtered out genes situated in the same genomic region and kept the gene with the highest connectivity and supporting information (original network see Suppl. Fig. 13). This gave a resulting gene set of 80 candidate genes (Fig. 10 and Suppl. Table 7). The G protein, GNG12 and the complement component 3 C3, are the proteins with the most edges in the network (30 and 25, respectively), followed by MCHR1, CXCL12, and NMUR2 with each 18 edges. Of these 80 highly connected genes, 66 are associated with bacteria that are either cospeciating (cospeciation rate > 0.5; Groussin et al., 2017) and/or have high heritability (> 0.5) suggesting a functionally important role for these bacterial taxa (Suppl. Table 7).

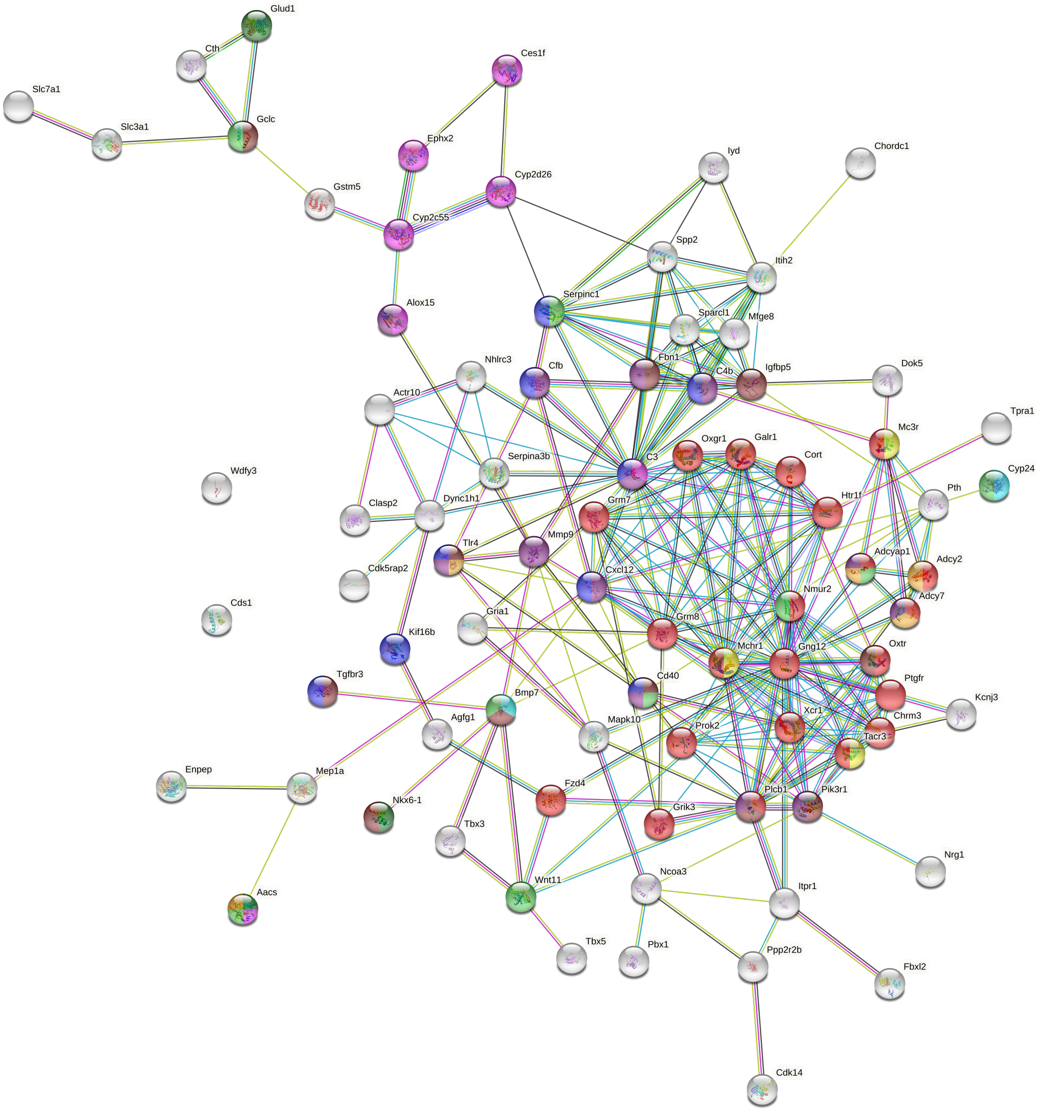


Figure 10: Network of host candidate genes influencing bacterial traits using STRING (https://string-db.org). The nodes represent proteins and are colored according to a selection of enriched GO terms and pathways: G protein coupled receptor (GPCR) signaling (red), regulation of the immune system process (blue), response to nutrient levels (light green), fatty acid metabolic process (pink), glucose homeostasis (purple), response to antibiotic (orange), regulation of feeding behavior (yellow), positive regulation of insulin secretion (dark green), circadian entrainment (brown), and response to vitamin D (turquoise). The color of the edges represents the interaction type: known interactions from curated databases (turquoise) or experimentally determined (pink); predicted interactions from gene neighborhood (green), gene fusions (red), gene co-occurrence (blue); other interactions from text-mining (light green), co-expression (black), and protein homology (purple).

3. Discussion

Understanding the forces that shape variation in host-associated bacterial communities within host species is key to understanding the evolution and maintenance of meta-organisms. Although numerous studies in mice and humans demonstrate that host genetics influences gut microbiota composition (McKnite et al., 2012; Leamy et al., 2014; Goodrich et al., 2014; Org et al., 2015; Davenport et al., 2015; Wang et al., 2016; Bonder et al., 2016; Goodrich et al., 2016; Kemis et al., 2019; Suzuki et al., 2019; Ishida et al., 2020; Hughes et al., 2020; Rühlemann et al., 2021), our study is unique in a number of important ways. First, the unique genetic resource of mice collected from a naturally occurring hybrid zone together with their native microbes yielded extremely high mapping resolution and the possibility to uncover ongoing evolutionary processes in nature. Second, our study is the first to perform genetic mapping of 16S rRNA transcripts in the gut environment, which was previously shown to be superior to DNA-based profiling in a genetic mapping study of the skin microbiota (Belheouane et al., 2017). Third, our study is one of the only to specifically examine the mucosa-associated community. It was previously reasoned that the mucosal environment may better reflect host genetic variation (Spor et al., 2011), and evidence for this hypothesis exists in nature (Linnenbrink et al., 2013). Finally, by cross-referencing our results with previous mapping studies and recently available proteomic data from germ-free versus conventional mice, we curated a more reliable list of candidate genes and pathways. Taken together, these results provide unique and unprecedented insight into the genetic basis for host-microbe interactions.

Importantly, by using wild-derived hybrid inbred strains to generate our mapping population, we gained insight into the evolutionary association between hosts and their microbiota at the transition from within species variation to between species divergence. Genetic relatedness in our mapping population significantly correlates with microbiome similarity, supporting a basis for codiversification at the early stages of speciation. A substantial proportion of microbial taxa are heritable, and heritability is correlated with cospeciation rates. This suggests that (i) vertical transmission could enable greater host adaptation to bacteria and/or (ii) the greater number of host genes associated with cospeciating taxa could indicate a greater dependency on the host, such that survival outside a specific host is reduced, making horizontal transmission less likely.

By performing 16S rRNA gene profiling at both the DNA and RNA level, we found that 30% (DNA-based) to 45% (RNA-based) of bacterial taxa are heritable, which is consistent with or higher than estimates reported in humans (~10%, Goodrich et al., 2016; ~21%, Turpin et al., 2016) and previous mouse studies (Kovacs et al., 2011; McKnite et al., 2012; Campbell et al., 2012; O'Connor et al., 2014; Carmody et al., 2015; Korach-Rechtman et al., 2019;). The high proportion of heritable taxa, with estimates of up to 91%, is likely explained in part by several factors of our study design. First, mice were raised in a controlled common environment, and heritability estimates in other mammals were shown to be contingent on the environment (Grieneisen et al., 2021). Further, bacterial communities were sampled from cecal tissue instead of fecal content (Linnenbrink et al., 2013), and genetic variation was higher than in a typical mapping study due to subspecies differences. For the RNA-based traits, heritability estimates were significantly correlated with previously reported cospeciation rates in mammals (Groussin et al., 2017). This pattern, as well as the higher proportion of heritable taxa in RNA-based traits, suggest that host genetic effects are more strongly reflected by bacterial activity than cell number.

Accordingly, we found a total of 179 and 313 unique significant loci for DNA-based and RNA-based bacterial abundance, respectively, passing the conservative study-wide significance threshold. Taxa had a median of eight significant loci, suggesting a complex and polygenic genetic architecture affecting bacterial abundances. We identify a higher number of loci in comparison to previous QTL and GWAS studies in mice (Benson et al., 2010; McKnite et al., 2012; Leamy et al., 2014; Wang et al., 2015; Org et al., 2015; Snijders et al., 2016; Kemis et al., 2019), which may be due to a number of factors. The parental strains of our study were never subjected to rederivation and subsequent reconstitution of their microbiota, and natural mouse gut microbiota are more variable than the artificial microbiota of laboratory strains (Kohl and Dearing, 2014; Weldon et al., 2015; Suzuki, 2017; Rosshart et al., 2017;). Furthermore, as noted above, our mapping population harbors both within- and between-subspecies genetic variation. We crossed incipient species sharing a common ancestor ~ 0.5 million years ago, hence we may also capture the effects of mutations that fixed rapidly between subspecies due to strong selection, which are typically not variable within species (Walsh, 1998; Barton and Keightley, 2002).

Importantly, our results also help to describe general features of the genetic architecture of bacterial taxon activity. For the majority of loci, the allele associated with lower relative abundance of the bacterial taxon was (partially) dominant. This suggests there is strong purifying selection against a high abundance of any particular taxon, which may help ensure high alpha diversity. The heterozygotes of one-third of significant SNPs displayed transgressive phenotypes. This is consistent with previous studies of hybrids (Turner et al., 2012; Turner and Harr, 2014; Wang et al., 2015;), for example, wild-caught hybrids showed broadly transgressive gut microbiome phenotypes. This pattern can be explained by over- or underdominance, or by epistasis (Rieseberg et al., 1999).

Notably, many loci significantly associated with bacterial abundance in this study were implicated in previous studies (Fig. 9). For example, chromosome 2 169-171 Mb is associated with ASV23 (*Eisenbergiella), Eisenbergiella* and ASV32 (unclassified

Lachnospiraceae) in this study, and overlaps with significant loci from three previous studies (Leamy et al., 2014; Snijders et al., 2016; Kemis et al., 2019). This region contains eight protein-coding genes: *Gm11011, Znf217, Tshz2, Bcas1, Cyp24a1, Pfdn4, 4930470P17Rik,* and *Dok5.* Another hotspot is on chromosome 5 101-103 Mb. This locus is significantly associated with four taxa in this study (Prevotellaceae, *Paraprevotella*, ASV7 genus *Paraprevotella* and *Acetatifactor*) and overlaps with associations for Clostridiales, Clostridiaceae, Lachnospiraceae, and Deferribacteriaceae (Snijders et al., 2016). Protein-coding genes in this region are: *Nkx6-1*, *Cds1*, *Wdfy3*, *Arhgap24*, and *Mapk10*. As previous studies were based on rederived mouse strains, identifying significant overlap in the identification of host loci suggests that some of the same genes and/or mechanisms influencing major members of gut microbial communities are conserved even in the face of community 'reset' in the context of re-derivation. The identity of the taxa is however not always the same, which suggests that functional redundancy may contribute to these observations, if *e.g.* several bacterial taxa fulfill the same function within the gut microbiome (Moya and Ferrer, 2016; Tian et al., 2020). Additionally, there is significant overlap of genes within loci identified in the current study and proteins differentially expressed in the intestine of germ-free mice compared to conventionally raised mice (Mills et al., 2020). Finally, by analyzing the functions of the genes closest to significant SNPs, we found that 12 of the 14 significantly enriched KEGG pathways were shown to be related to interactions with bacteria (Fonken et al., 2010; Thaiss et al., 2014; Neumann et al., 2014; Thaiss et al., 2015a; Thaiss et al., 2015b; Castoldi et al., 2015; Erdman and Poutahidis, 2016; Thaiss et al., 2016; Deaver et al., 2018; Wu et al., 2018; Peng et al., 2020; Nagpal et al., 2020; Hollander and Kaunitz, 2020; Suppl. Table 5).

To improve the robustness of our results, we combined multiple lines of evidence to prioritize candidates, resulting in a network of 80 genes (Suppl. Table 7). At the center of this network is a set of 22 proteins involved in G-protein coupled receptor signaling (Fig. 10, red nodes). MCHR1, NMUR2, and TACR3 (Fig. 10, yellow) are known to regulate feeding behavior (Saito et al., 1999; Cardoso et al., 2012; Smith et al., 2019), and CHRM3 to control digestion (Gautam et al., 2006; Tanahashi et al., 2009). Gut microbes can produce GPCR agonists to elicit host cellular responses (Cohen et al., 2017; Colosimo et al., 2019; Chen et al., 2019; Pandey et al., 2019). Thus, GPCRs may be key modulators of communication between the gut microbiota and host. Another interesting group of genes are those responding to nutrient levels (*Bmp7, Cd40, Aacs, Gclc, Nmur2, Cyp24a1, Adcyap1, Serpinc1,* and *Wnt11*) (Sethi and Vidal-Puig, 2008; Peier et al., 2009; Townsend et al., 2012; Yi and Bishop, 2015; Shi and Tu, 2015; Toderici et al., 2016; Yasuda et al., 2021; Gastelum et al., 2021;), as gut microbiota affect host nutrient uptake (Chung et al., 2018). In addition, CYP24A1, BMP7 and CD40 respond to vitamin D. Previous studies identified vitamin D/the vitamin D receptor to play a role in modulating the gut microbiota (Wang et al., 2016; Malaguarnera, 2020; Yang et al., 2020b; Singh et al., 2020), and

CD40 is known to induce a vitamin D dependent antimicrobial response through IFN-γ activation (Klug-Micu et al., 2013).

Another important category of candidate genes are those involved in immunity. Our most significant SNP was situated downstream of the *Tlr4* gene and was associated with the genus *Dorea* and several *Dorea* species. *Dorea* is a known short chain fatty acid producer (Taras et al., 2002; Reichardt et al., 2018) and interacts with tight junction proteins *Claudin-2* and *Occludin* (Alhasson et al., 2017). *Tlr4* is a member of the Toll-like receptor family, and has been linked with obesity, inflammation, and changes in the gut microbiota (Velloso et al., 2015). These combined results reflect an important role for *Dorea* in fatty acid harvesting and intestinal barrier integrity, both of which could act systemically to activate TLR4 and to promote metabolic inflammation (Cani et al., 2008; Delzenne et al., 2011; Nicholson et al., 2012). Moreover, the SNP with the second lowest *P* value was associated with the same taxa and situated 181 kb upstream of *Irak4.* IRAK4 is rapidly recruited after TLR4 activation to enable downstream activation of the NFκB immune pathway. *Irak4* has previously been associated with a change in bacterial abundance using inbred mice (McKnite et al., 2012; Org et al., 2015).

Finally, we identified notable links between candidate genes and five human diseases (mental disorders, blood pressure finding, systemic arterial pressure, substance-related disorders, and atrial septal deficits; Suppl. Fig. 10). The connection to mental disorders is intriguing as involvement of the gut microbiota is suspected (Kelly et al., 2015; Foster et al., 2017a; Cox and Weiner, 2018; Chen et al., 2019; Sarkar et al., 2020; Parker et al., 2020; Flux and Lowry, 2020). Taken together with our finding of an enriched set of GPCRs, this highlights the importance of host-microbial interplay along the gut-brain axis. Moreover, we also identify a significant over-representation of IBD genes (Khan et al., 2021) among the 925 genes nearest to significant SNPs (Suppl. Table 6). Interestingly, SNPs in five out of 14 genes are associated with ASVs belonging to the genus *Oscillibacter,* a highly cospeciating taxon known to decrease during the active state of IBD (Metwaly et al., 2020).

In summary, our study provides a number of novel insights into the importance of host genetic variation in shaping the gut microbiome, in particular for cospeciating bacterial taxa. These findings provide an exciting foundation for future studies of the precise mechanisms underlying host-gut microbiota interactions in the mammalian gut and should encourage future genetic mapping studies that extend analyses to the functional metagenomic sequence level.

4. Material and Methods

4.1. Intercross design

We generated a mapping population using partially inbred strains derived from mice mice captured in the *M. m. musculus - M. m. domesticus* hybrid zone around Freising in 2008 (Turner et al., 2012). Originally, four breeding stocks were derived from 8-9 ancestors captured from one (FS, HA, TU) or two sampling sites (HO), and maintained with four breeding pairs per generation using the HAN-rotation outbreeding scheme (Rapp, 1972). Eight inbred lines (two per breeding stock) were generated by brother/sister mating of the 8th generation lab-bred mice. We set up the cross when lines were at the 5th-9th generation of brother-sister meeting, with inbreeding coefficients of > 82%.

We first set up eight G1 crosses, each with one predominantly *domesticus* line (FS, HO - hybrid index <50%; see below) and one predominantly *musculus* line (HA, TU - hybrid index >50%); each line was represented as a dam in one cross and sire in another (Fig. 11). One line, FS5, had a higher hybrid index than expected, suggesting there was a misidentification during breeding (see genotyping below). Next, we set up G2 crosses in eight combinations (subcrosses), such that each G2 individual has one grandparent from each of the initial four breeding stocks. We included 40 males from each subcross in the mapping population.

This study was performed according to approved animal protocols and institutional guidelines of the Max Planck Institute. Mice were maintained and handled in accordance with FELASA guidelines and German animal welfare law (Tierschutzgesetz § 11, permit from Veterinäramt Kreis Plön: 1401-144/PLÖ-004697).

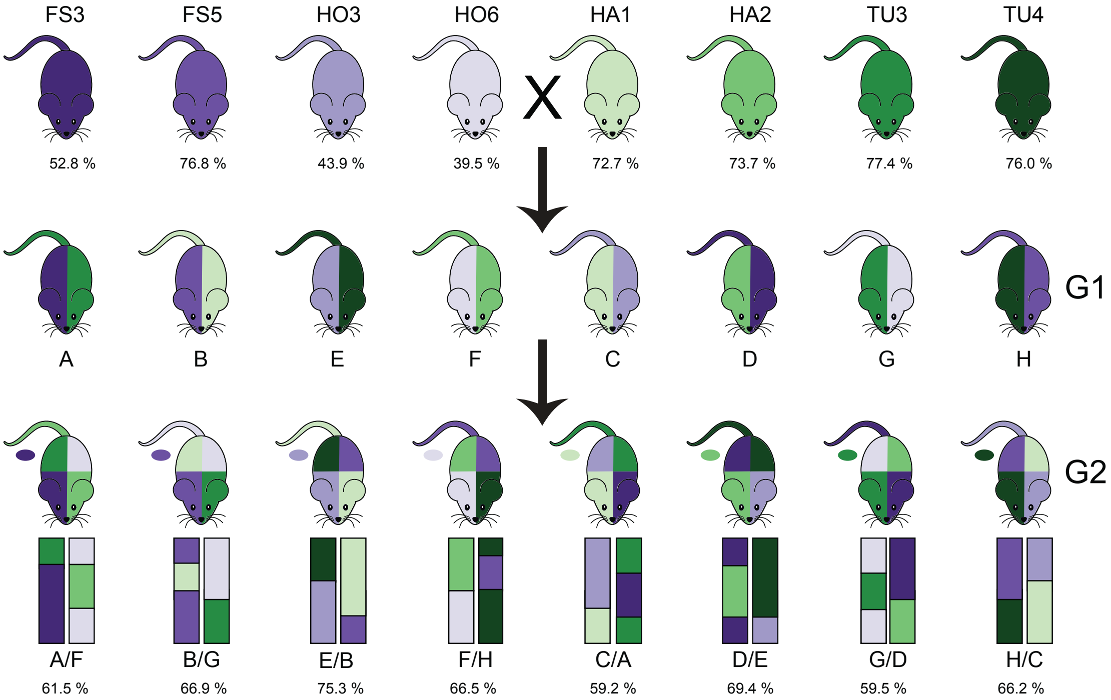


Figure 11: Overview of the intercross design. G0 mice are from eight partially inbred lines derived from mice wild-caught in four hybrid zone sites. Hybrid index - the percentage of musculus alleles - is reported as the mean for the G0 mice from each line, or mean of 40 G2s from each subcross at bottom. We performed eight G1 crosses with one line with hybrid index <50% (purple shades) and one line with hybrid index >50% (green shades); color on the left side of mouse diagram indicates dam line and right side indicates sire line. Next, G1 mice were crossed in eight combinations such that each G2 mouse had one grandparent from each of the four breeding stocks, indicated by colors of mouse diagram, and representative chromosomes below. Tail color indicates Y chromosome strain, and oval indicates mitochondrial strain.

4.2. Sample collection

Mice were sacrificed at 91 ± 5 days by CO2 asphyxiation. We recorded body weight, body length and tail length, and collected ear tissue for genotyping. The caecum was removed and gently separated from its contents through bisection and immersion in RNAlater (Thermo Fisher Scientific, Schwerte, Germany). After overnight storage in RNAlater at 4° C, the RNAlater was removed and tissue stored at -20° C.

4.3. DNA extraction and sequencing

We simultaneously extracted DNA and RNA from caecum tissue samples using Qiagen (Hilden, Germany) Allprep DNA/RNA 96-well kits. We followed the manufacturer's protocol, with the addition of an initial bead beating step using Lysing matrix E tubes (MP Biomedical, Eschwege) to increase cell lysis. We used caecum tissue because host genetics has a greater influence on the microbiota at this mucosal site than on the lumen contents (Linnenbrink et al., 2013). We performed reverse transcription of RNA with High-Capacity cDNA Transcription Kits from Applied Biosystems (Darmstadt, Germany). We amplified the V1-V2 hypervariable region of the 16S rRNA gene using barcoded primers (27F-338R) with fused MiSeq adapters and heterogeneity spacers following (Rausch et al., 2016) and sequenced amplicons with 250 bp paired-reads on the Illumina MiSeq platform.

4.4. 16S rRNA gene sequence analysis

We assigned sequences to samples by exact matches of MID (multiplex identifier, 10 nt) sequences and processed 16S rRNA gene sequences using the DADA2 pipeline, implemented in the DADA2 R package, version 1.16.0 (Callahan et al., 2016; Callahan, 2016). Briefly, raw sequences were trimmed and quality filtered with the maximum two 'expected errors' allowed in a read, paired sequences were merged and chimeras removed. For all downstream analyses, we rarefied samples to 10,000 reads each. Due to the quality filtering, we have phenotyping data for 286 individuals on DNA level, and 320 G2 individuals on RNA level. We classified taxonomy using the Ribosomal Database Project (RDP) training set 16 (Cole et al., 2014). Classifications with low confidence at the genus level (<0.8) were grouped in the arbitrary taxon 'unclassified_group'.

We used the phyloseq R package (version 1.32.0) to estimate alpha diversity using the Shannon index and Chao1 index, and beta diversity using Bray-Curtis distance (McMurdie and Holmes, 2013). We defined core microbiomes at the DNA- and RNA-level, including taxa present in > 25% of the samples and with median abundance of non-zero values > 0.2% for amplicon sequence variant (ASV) and genus; and >0.5% for family, order, class and phylum.

4.5. Genotyping

We extracted genomic DNA from ear samples using Qiagen Blood and Tissue 96 well kits (Hilden, Germany), according to the manufacturer's protocol. We sent DNA samples from 26 G0 mice and 320 G2 mice to GeneSeek (Neogen, Lincoln, NE) for genotyping using the Giga Mouse Universal Genotyping Array (GigaMUGA; Morgan et al., 2015), an Illumina Infinium II array containing 141,090 single nucleotide polymorphism (SNP) probes. We quality-filtered genotype data using plink 1.9 (Chang et al., 2015); we removed individuals with call rates <90% and SNPs that were: not bi-allelic, missing in >10% individuals, with minor allele frequency <5%, or Hardy-Weinberg equilibrium exact test *P* values <1e-10. A total of

64,103 SNPs and all but one G2 individual were retained. Prior to mapping, we LD-filtered SNPs with r2 >0.9 using a window of 5 SNPs and a step size of 1 SNP. We retain 32,625 SNPs for mapping.

4.6. Hybrid index calculation

For each G0 and G2 mouse, we estimated a hybrid index – defined as the percentage of *M. m. musculus* ancestry. We identified ancestry-informative SNP markers by comparing GigaMUGA data from ten individuals each from two wild-derived outbred stocks of *M. m. musculus* (Kazakhstan and Czech Republic) and two of *M. m. domesticus* (Germany and France) maintained at the Max Planck Institute for Evolutionary Biology (L.M. Turner and B. Payseur, unpublished data). We classified SNPs as ancestry informative if they had a minimum of 10 calls per subspecies, the major allele differed between *musculus* and *domesticus*, and the allele frequency difference between subspecies was > 0.3. A total of 48,361 quality-filtered SNPs from the G0/G2 genotype data were informative, including 8,775 SNPs with fixed differences between subspecies samples.

4.7. Correlation between host relatedness and microbiome structure

To investigate if host relatedness is correlated with individual variation in microbiome composition, we computed a centered relatedness matrix using the 32,625 filtered SNPs with GEMMA (v 0.98.1; Zhou and Stephens, 2012) and microbial composition-based kinship matrix among individuals based on relative bacterial abundances (Chen et al., 2018). The kinship matrix was calculated with the formula:

$$Kinship = 1/p \sum_{i=1}^{p} (x_i - 1_n \bar{x}_i)(x_i - 1_n \bar{x}_i)^T$$

where X denotes the $n \times p$ matrix of genotypes (or relative abundances), $\boldsymbol{x_i}$ as its *i*th column representing the genotypes of *i*th SNP (or the relative abundance of the *i*th ASV), $\bar{x}_i$ as the sample mean and $\mathbf{1}_n$ as a $n \times 1$ vector of 1's. We used a Mantel test to test for correlation between the host SNP-based kinship and microbial composition-based kinship.

4.8. SNP-based heritability of microbial abundances

We calculated SNP-based heritabilities for bacterial abundances using a linear mixed model implemented in the lme4qtl R package (Ziyatdinov et al., 2018). The SNP-based heritability is expressed as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_m^2 + \sigma_e^2}$$

where σ_g^2 is the genetic variance estimated by $K_{SNP,}$ σ_m^2 variance of the mating pair component, and σ_e^2 the variance due to residual environmental factors. We determined significance of the heritability estimates using exact likelihood ratio tests, following Supplementary Note 3 in (Ziyatdinov et al., 2018), using the exactLRT() function of the R package *RLRsim* (Fabian et al., 2008).

4.9. Genome-wide association mapping

Prior to mapping, we inverse logistic transformed bacterial abundances using the inv.logit function from the R package gtools (Gregory R. Warnes, 2020).

We performed association mapping in the R package *lme4qtl* (Ziyatdinov et al., 2018) with the following linear mixed model:

$$y_j = \mu + a_i X_{ij}^a + d_i X_{ij}^d + Wu + e$$

where y_j is the phenotypic value of the *j*th individual; μ is the mean, X^a_{ij} the additive and X^d_{ij} the dominance genotypic index values coded as for individual *j* at locus *i*. *a* and *d* indicate fixed additive and dominance effects, *W* indicates random effects mating pair and kinship matrix, plus residual error *e*.

We estimated additive and dominance effects separately because we expected to observe underdominance and overdominance in our hybrid mapping population, as well as additive effects, and aimed to estimate their relative importance. To model the additive effect (i.e. 1/2 distance between homozygous means), genotypes at each locus, *i*, were assigned additive index values (Xa ∊ 1, 0, −1) for AA, AB, BB, respectively, with A indicating the major allele and B the minor allele. To model dominance effects (i.e. heterozygote mean - midpoint of homozygote means), genotypes were assigned dominance index values (Xd ∊ 0, 1) for homozygotes and heterozygotes, respectively.

We included mating pair as a random effect to account for maternal effects and cage effects, as male litter mates are kept together in a cage after weaning. We included kinship coefficient as a random effect in the model to account for population and family structure. To avoid proximal contamination, we used a leave-one-chromosome-out approach, that is, when testing each single SNP association we used a relatedness matrix omitting markers from the same chromosome (Parker et al., 2014). Hence, for testing SNPs on each chromosome, we calculated a centered relatedness matrix using SNPs from all other chromosomes with GEMMA (v0.97; Zhou and Stephens, 2012).We calculated *P* values for single SNP associations by comparing the full model to a null model excluding fixed effects. Code for performing the mapping is available at https://github.com/sdoms/mapping_scripts.

We evaluated significance of SNP-trait associations using two thresholds; first, we used a genome-wide threshold for each trait, where we corrected for multiple testing across markers using the Bonferroni method (Abdi, 2007). Second, as bacteria interact with each other within the gut as members of a community, bacterial abundances are non-independent, so we calculated a study-wide thresholddividing the genome-wide threshold by the number of effective taxa included. We used *matSpDlite* (Nyholt, 2019; Li and Ji, 2005; Qin et al., 2020) to estimate the number of effective bacterial taxa based on eigenvalue variance.

To estimate the genomic interval represented by each significant LD-filtered SNP, we report significant regions defined by the most distant flanking SNPs in the full pre-LD-filtered genotype dataset showing $r_2 > 0.9$ with each significant SNP. We combined significant regions less than 10 Mb apart into a single region. Genes situated in significant regions were retrieved using *biomaRt* (Steffen Durinck, 2009), and the *mm10* mouse genome.

4.10. Dominance analyses

We classified dominance for SNPs with significant associations on the basis of the *d/a* ratio (Falconer, 1996) where *d* is the dominance effect, *a* the additive effect. As the expected value under purely additive effects is 0. As our mapping population is a multi-parental-line cross, and not all SNPs were ancestry-informative with respect to musculus/domesticus, the sign of *a* effects is defined by the major allele within our mapping population, which lacks clear biological interpretation. To provide more meaningful values, we report *d/|a|*, such that a value of 1 = complete dominance of the allele associated with higher bacterial abundance, and a value of -1 = complete dominance of the allele associated with lower bacterial abundance. Values above 1 or below -1 indicate over/underdominance. We classified effects of significant regions the following arbitrary *d/|a|* ranges to classify dominance of significant regions (Burke et al., 2002; Miller et al., 2014) : underdominant <-1.25, high abundance allele recessive between -1.25 and -0.75, partially recessive between -0.75 and -0.25, additive between -0.25 and 0.25, partially dominant between 0.25 and 0.75, dominant 0.75 and 1.25, and overdominant >1.25.

4.11. Gene ontology and network analysis

The nearest genes up- and downstream of the significant SNPs were identified using the locateVariants() function from the VariantAnnotation R package (version 1.34.0; Valerie et al., 2014) using the default parameters. A maximum of two genes per locus were included (one upstream, and one downstream of a given SNP).

To investigate functions and interactions of candidate genes, we calculated a a protein-protein interaction (PPI) network with STRING version 11 (Szklarczyk et al., 2019), on the basis of a list of the closest genes to all SNPs with significant trait associations. We included network edges with an interaction score >0.9, based on evidence from fusion, neighborhood, co-occurrence, experimental, text-mining, database, and co-expression. We exported this network to Cytoscape v 3.8.2 (Shannon et al., 2003) for identification of highly interconnected regions using the MCODE Cytoscape plugin (Bader and Hogue, 2003), and functional annotation of clusters using the stringApp Cytoscape plugin (Doncheva et al., 2019).

We identified overrepresented KEGG pathways and human diseases using the clusterprofiler R package (version 3.16.1; Yu et al., 2012). *P* values were corrected for multiple testing using the Benjamini-Hochberg method. Pathways and diseases with an adjusted *P* value < .05 were considered over-represented.

4.12. Calculating overlap with other studies

To test for significant overlap with loci identified in previous mapping studies and for over-representation of IBD genes, we used the tool *poverlap* (Brent Pedersen, 2013) to compare observed overlap to random expectations based on 10,000 permutations of significant regions. We identified genes within overlapping regions using the locateVariants() function from the *VariantAnnotation* R package (version 1.34.0; Valerie et al., 2014).

4.13. Combination of results

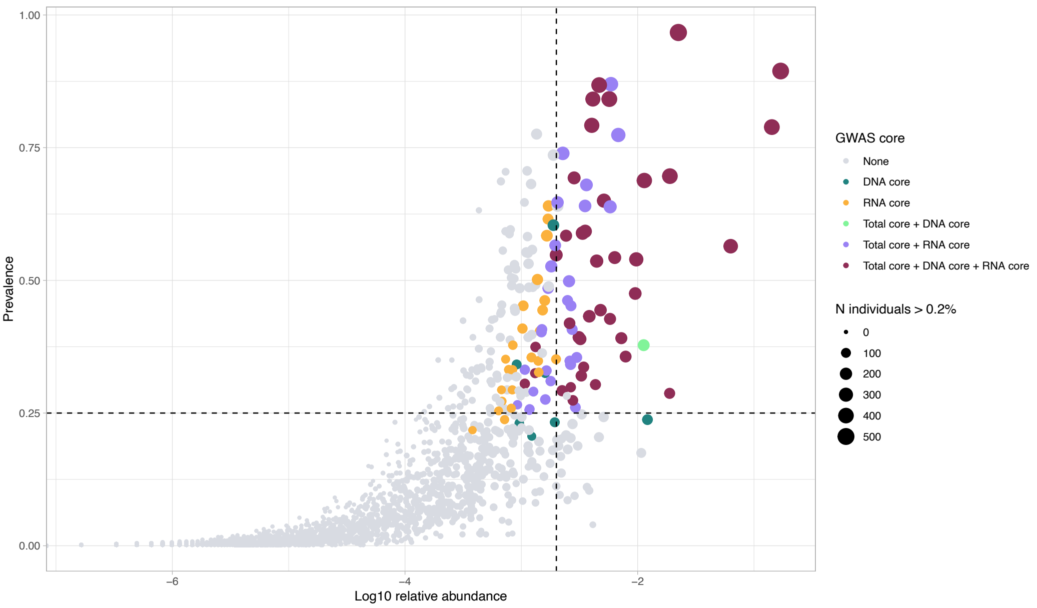
Hub genes SNP network and their first neighbors, the hub genes from the 'differentially expressed in GF mice'-network and their respective first neighbors, genes found in both Mills et al. (2020) and other mouse QTL studies, closest genes to a SNP found in Mills et al. (2020), genes situated in the 20 smallest intervals, six genes in the two intervals with the lowest P values, twenty genes in intervals found in most different taxa, genes situated in the region with most overlap within our study, and finally the genes situated in the intervals that most frequently overlapped with other studies were combined into on gene set and analyzed with STRING. Genes situated in the same genomic locus were curated according to the number of edges in the STRING network.

4.14. Data and code availability

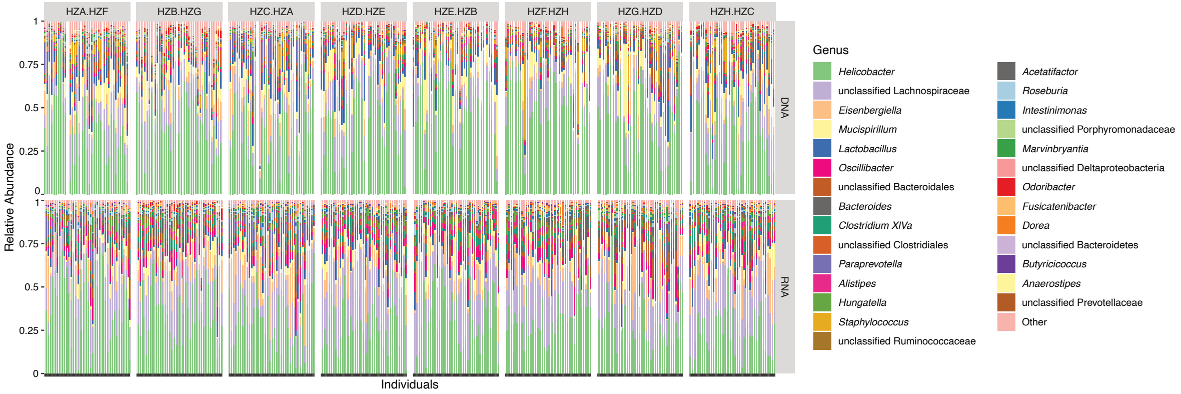
DNA- and RNA-based 16S rRNA gene sequences are available under project accession number PRJNA759194. Code is available at https://github.com/sdoms/mapping_scripts.

5. Supplementary material

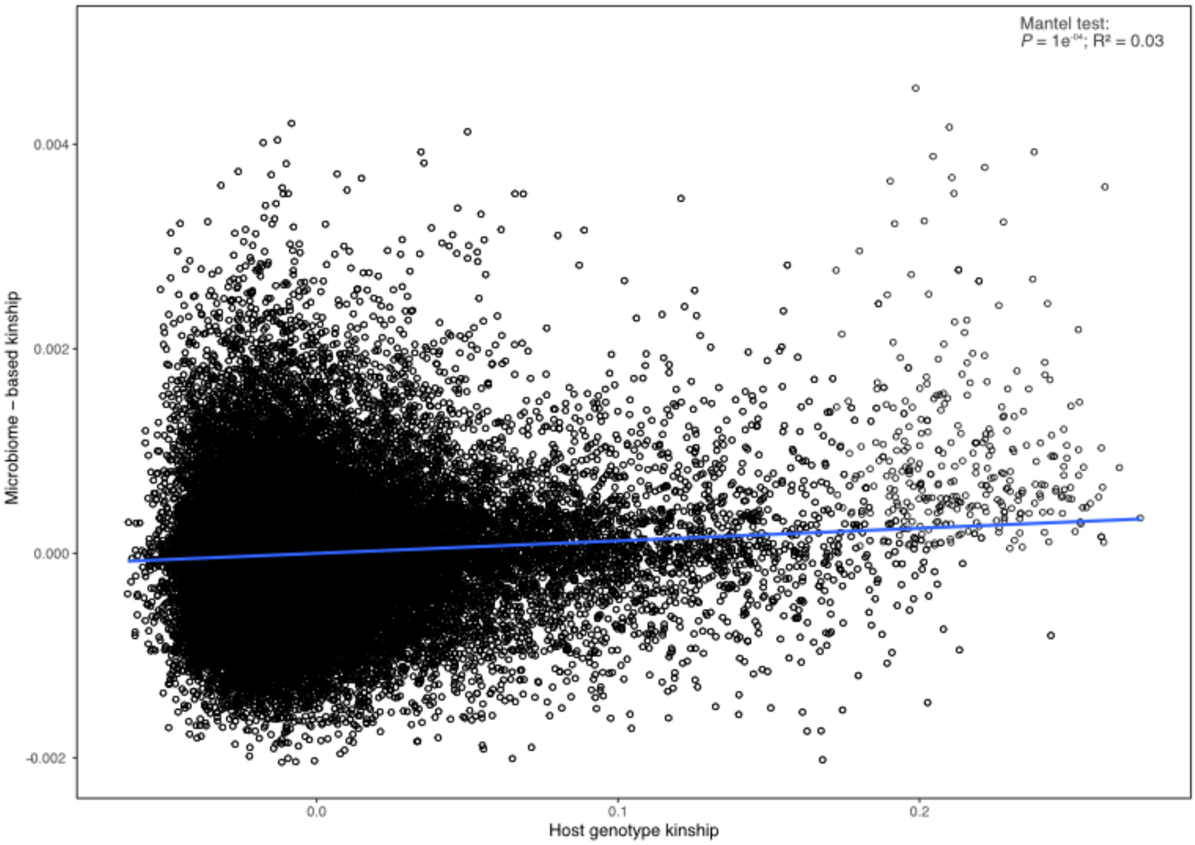
Suppl. Fig 1-14, **Suppl. Table 1**: Heritability estimates, **Suppl. Table 2**: Genome-wide significant associations, **Suppl. Table 3**: Study-wide significant associations, **Suppl. Table 4**: Intervals smaller than 1Mb, **Suppl. Table 5**: Overrepresented KEGG pathways, **Suppl. Table 6**: IBD genes, and **Suppl. Table 7**: Candidate genes.



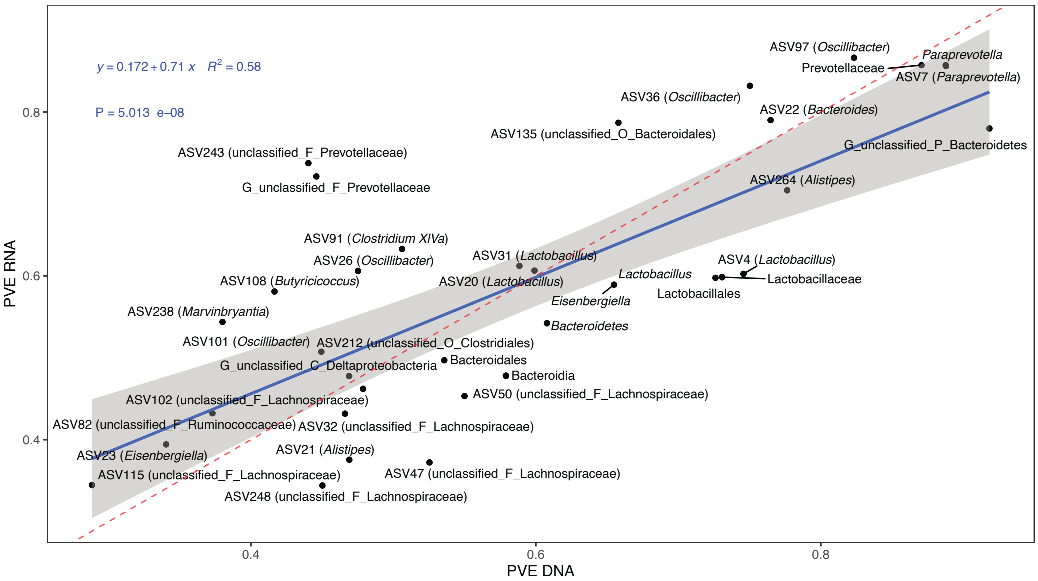
Suppl. Figure 1: Selection of taxa for mGWAS analysis. A scatter plot showing the association of average relative abundance of taxa with their prevalence in the G2 mapping population. Taxa retained for analysis are colored according to the originating core. The size of each dot represents the number of individuals that have a median abundance higher than 0.2% of the taxon. The dashed lines represent the thresholds of the core (vertical: median abundance>0.2% and horizontal prevalence of 25 %.



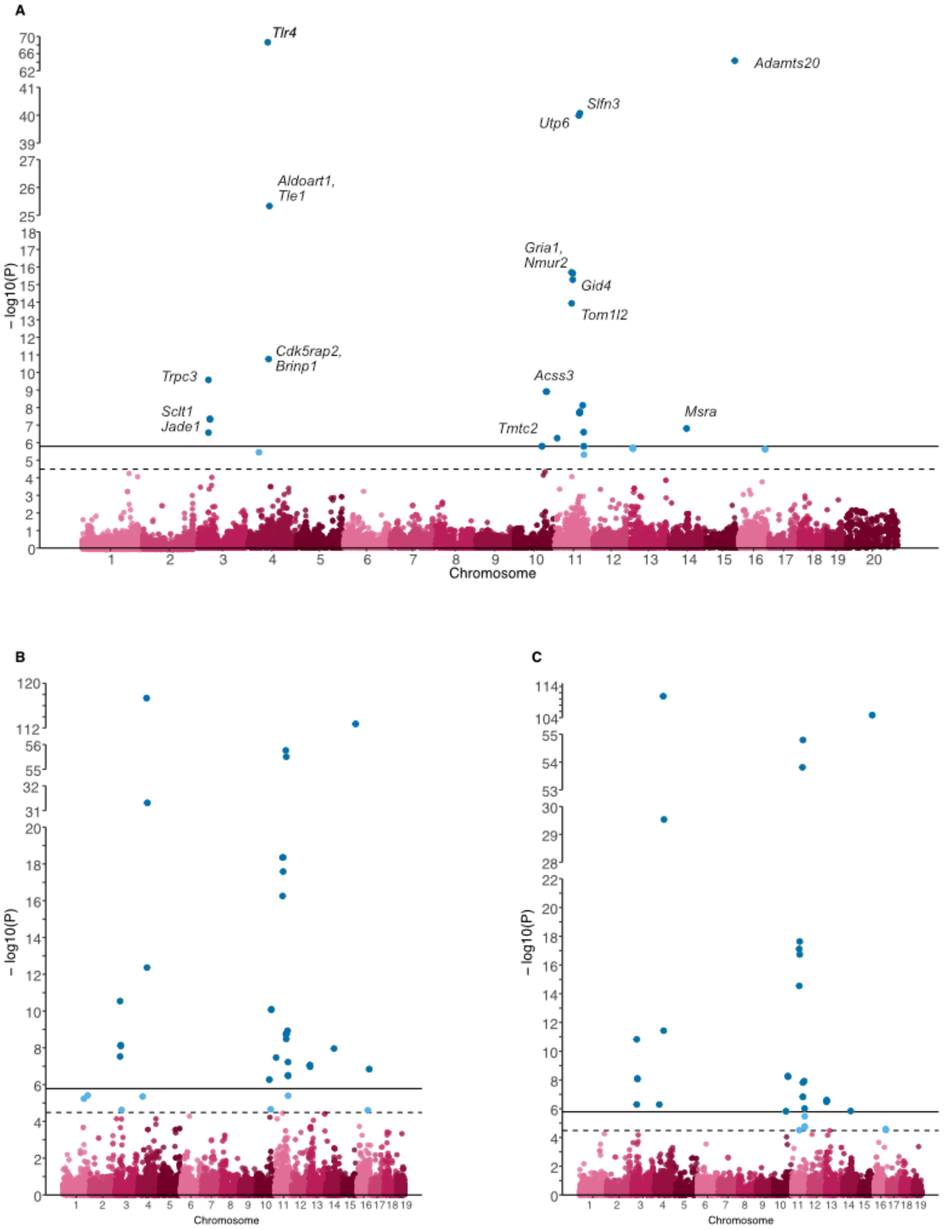
Suppl. Figure 2: Relative abundances of core genera in G2 mapping population. Each vertical line represents one individual. Subcross (see figure 11) is indicated at the top.



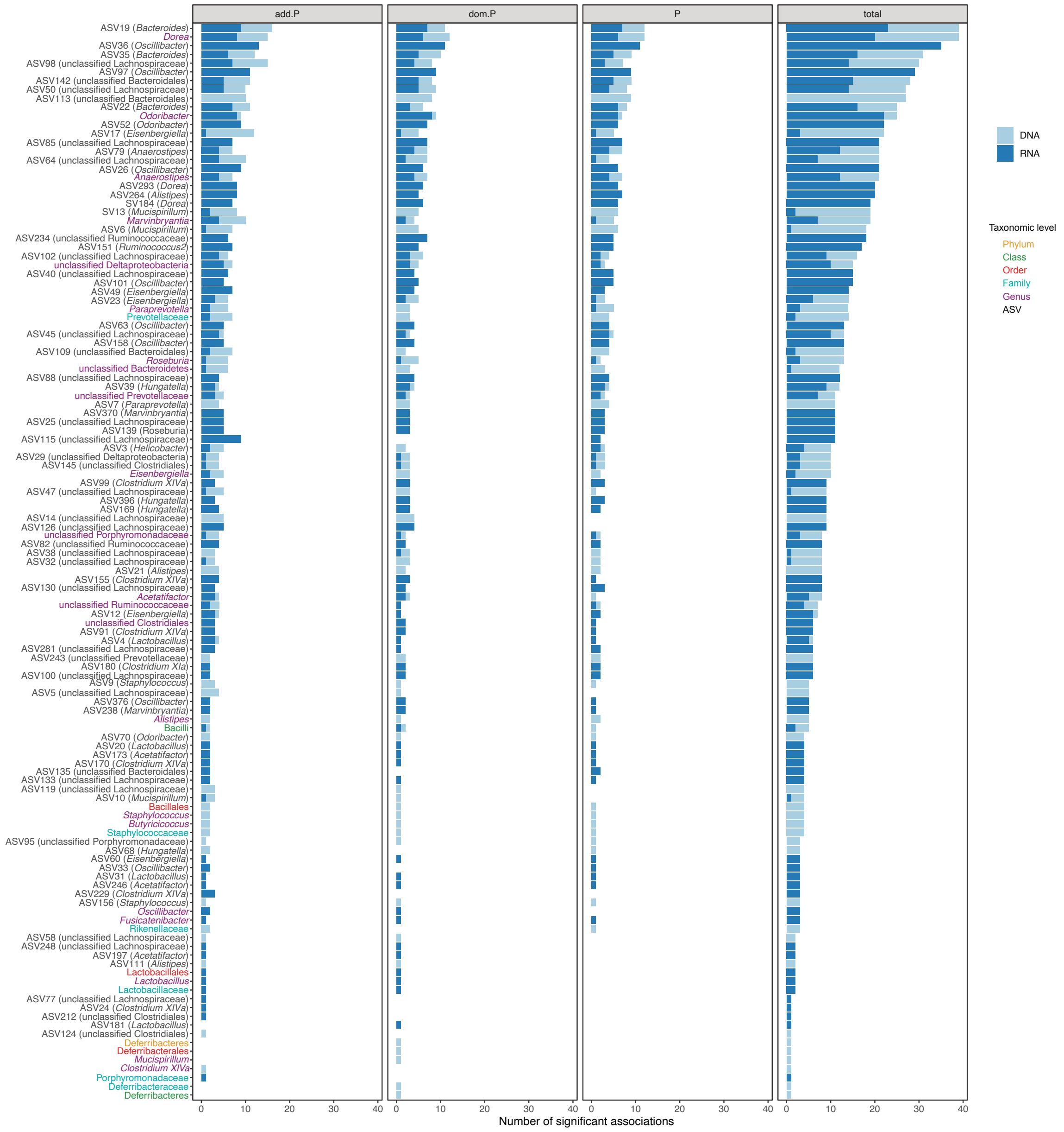
Suppl. Figure 3: Host genetic relatedness calculated from SNP data (x-axis) is correlated with microbial composition-based relatedness (y-axis) calculated from ASV abundances. The blue line represents a linear regression fit to the data.



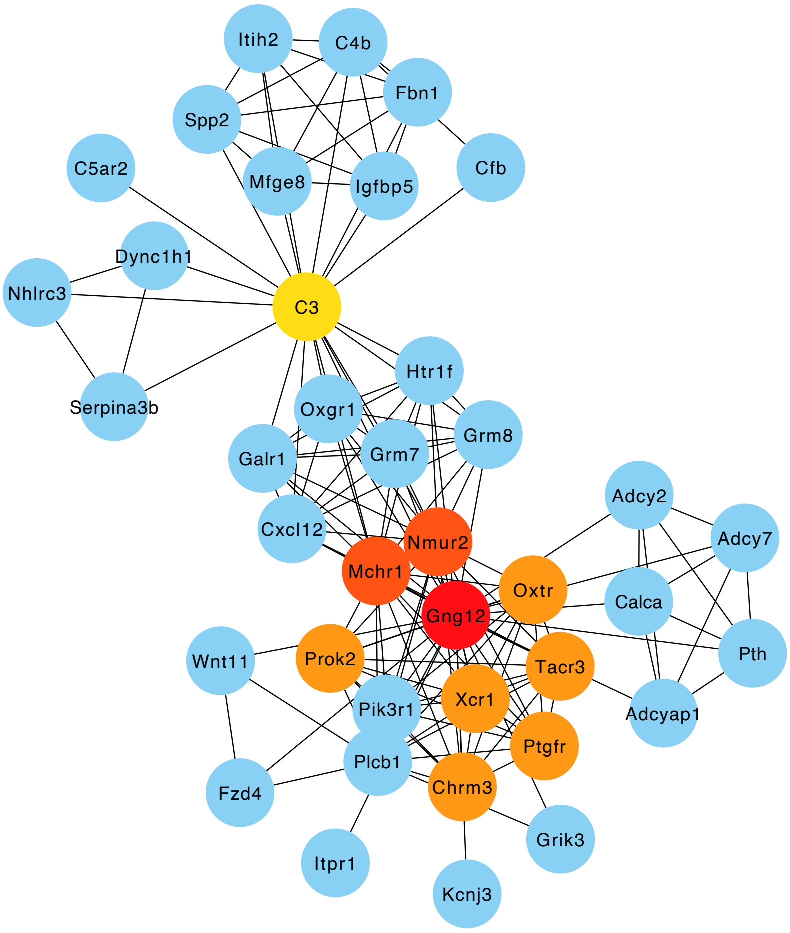
Suppl. Figure 4: Correlation of SNP-based heritability estimates based on DNA (x-axis) or RNA (y-axis). The blue line represents a linear regression fit to the data. Red dashed line represents the identity line with a slope of 1.



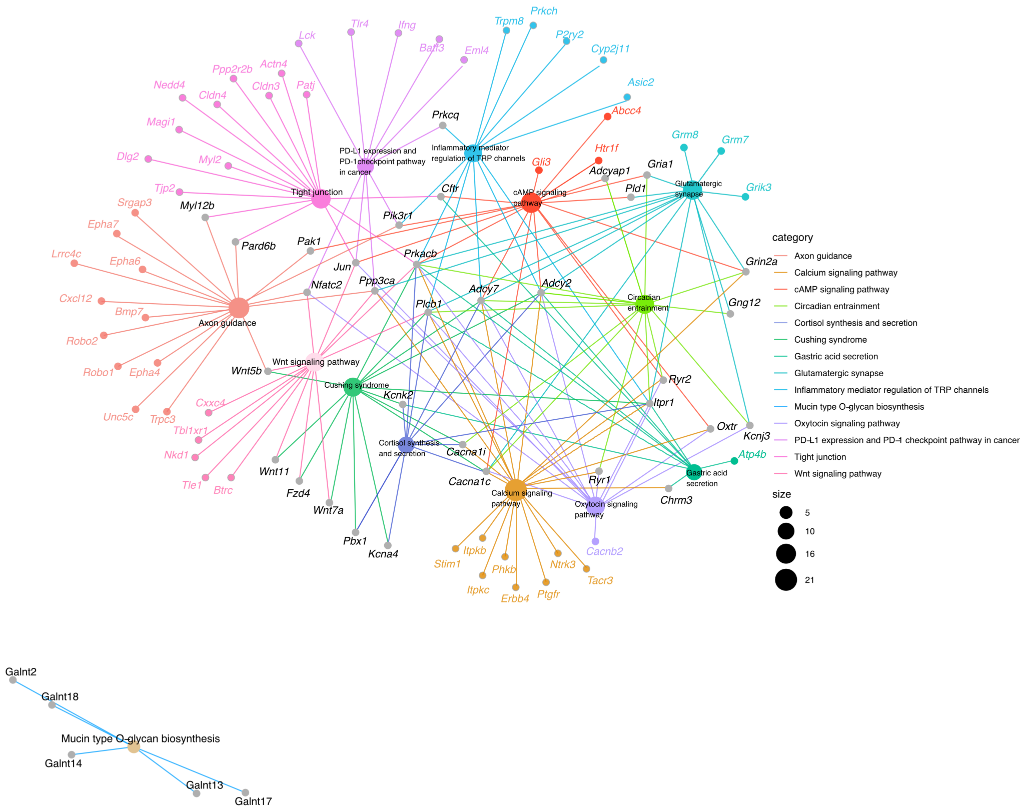
Suppl. Figure 5: Manhattan plots for ASV184 (Dorea) of the complete model (A), the additive effect (B) or the dominance effect (C). SNPs passing the study-wide significance threshold (solid line) are shown in dark blue, while genome-wide significant SNPs (dashed line) are shown in light blue. In panel A, the closest gene to the SNP is shown for a subset of significant SNPs.



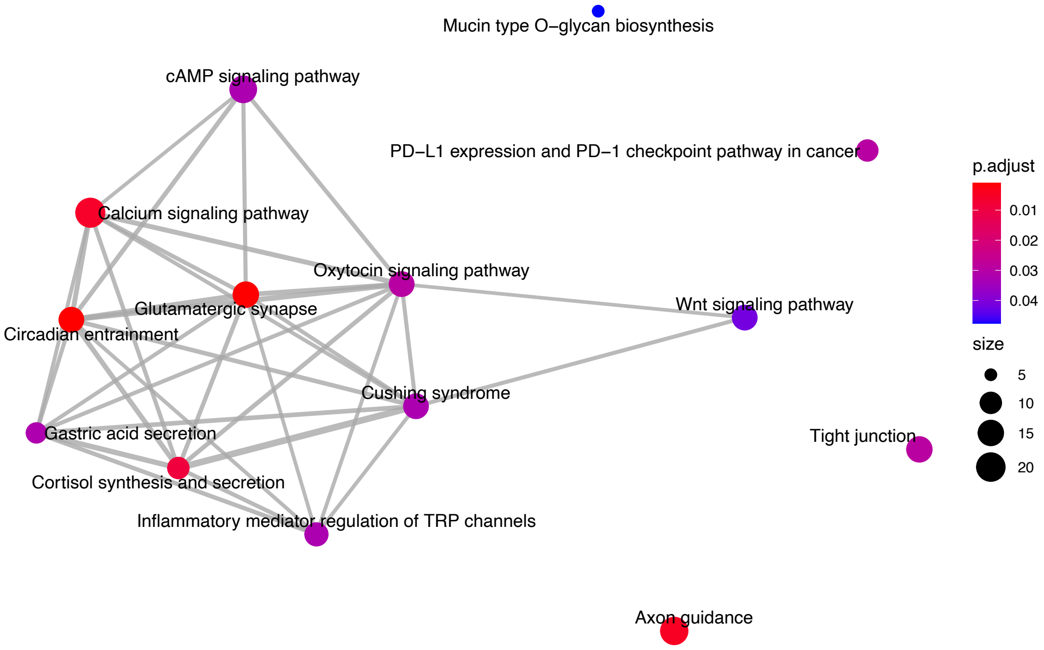
Suppl. Figure 6: Number of significantly associated loci per bacterial taxon. Loci with significant additive effects (add.P), dominance effects (dom.P) or effects in full model (P) are indicated.



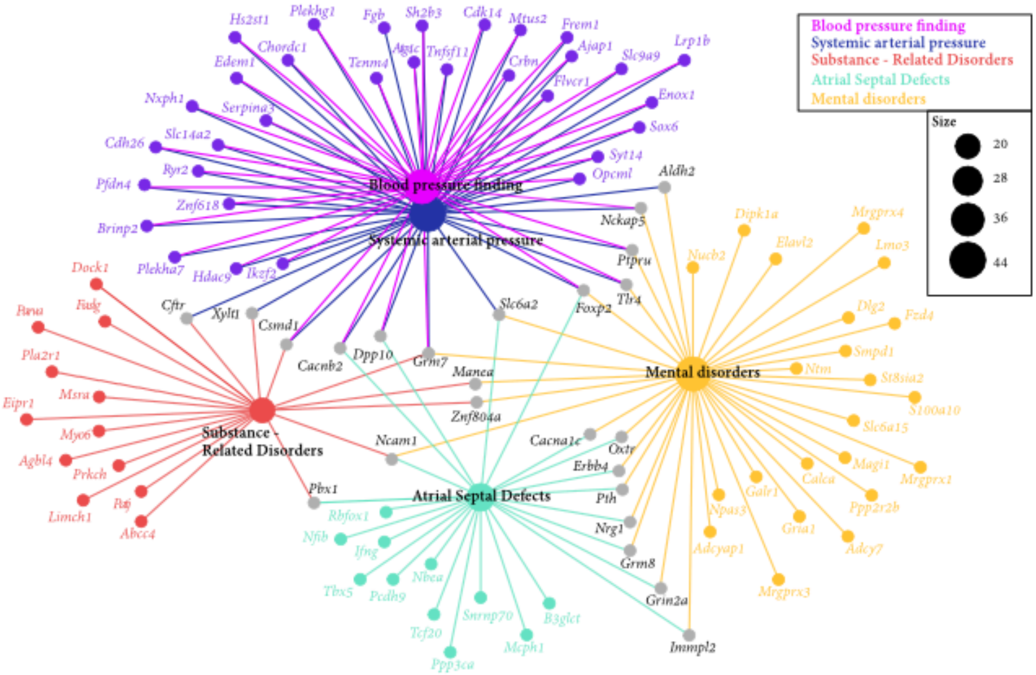
Suppl. Figure 7: Top ten hub genes of the protein-protein interaction (PPI) network with the closest genes to the host SNPs significantly associated with bacterial abundances. The nodes are colored according to hub gene rank from 1 (red) to 10 (yellow). Blue nodes are the first neighbors.


Tight junction
Axon guidance
PD-L1 expression and PD-1 checkpoint pathway in cancer
Inflammatory mediator regulation of TRP channels
cAMP signaling pathway
Glutamatergic synapse
Circadian entrainment
Wnt signaling pathway
Cushing syndrome
Cortisol synthesis and secretion
Calcium signaling pathway
Oxytocin signaling pathway
Gastric acid secretion
Mucin type O-glycan biosynthesis
Galnt2
Galnt18
Galnt14
Galnt13
Galnt17
category
Axon guidance
Calcium signaling pathway
cAMP signaling pathway
Circadian entrainment
Cortisol synthesis and secretion
Cushing syndrome
Gastric acid secretion
Glutamatergic synapse
Inflammatory mediator regulation of TRP channels
Mucin type O-glycan biosynthesis
Oxytocin signaling pathway
PD-L1 expression and PD-1 checkpoint pathway in cancer
Tight junction
Wnt signaling pathway
size
5
10
16
21

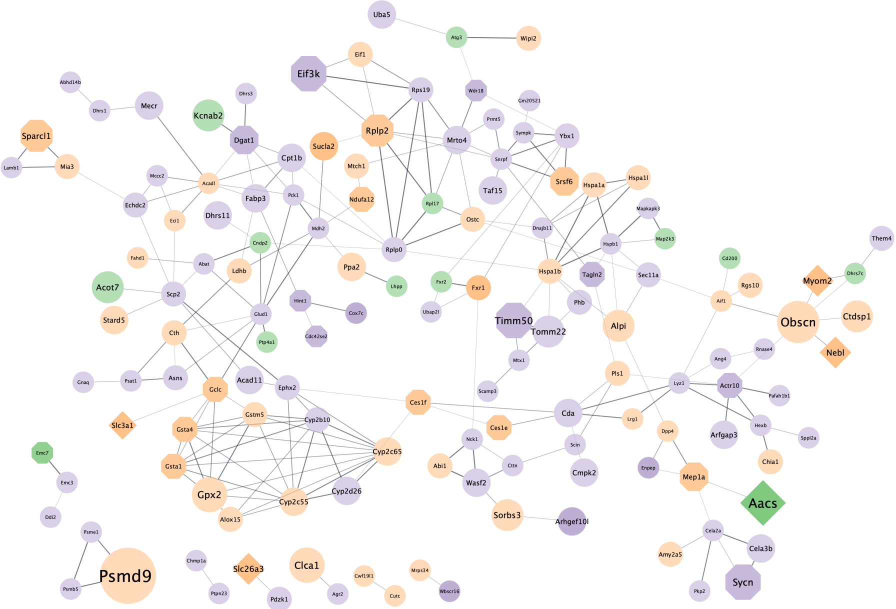
Suppl. Figure 8: Genes belonging to over-represented KEGG pathways within the host genes closest to significant SNPs from association analysis.


Mucin type O–glycan biosynthesis
cAMP signaling pathway
PD–L1 expression and PD–1 checkpoint pathway in cancer
Calcium signaling pathway
Oxytocin signaling pathway
Glutamatergic synapse
Wnt signaling pathway
Circadian entrainment
Cushing syndrome
Gastric acid secretion
Tight junction
Cortisol synthesis and secretion
Inflammatory mediator regulation of TRP channels
Axon guidance
p.adjust
0.01
0.02
0.03
0.04
size
5
10
15
20

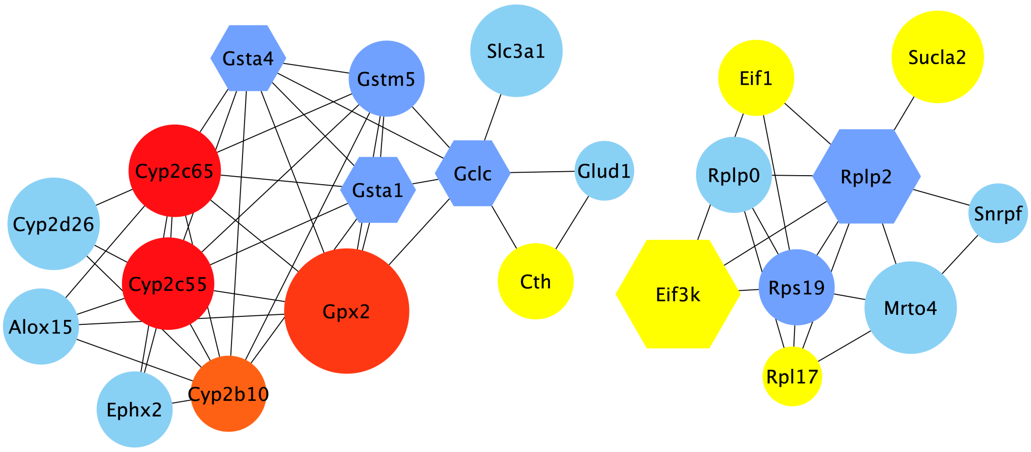
Suppl. Figure 9: Enriched KEGG pathways among closest genes to significant SNPs from association analysis. Node color indicates FDR-adjusted P value of enrichment and node size indicates number of candidate genes in pathway.



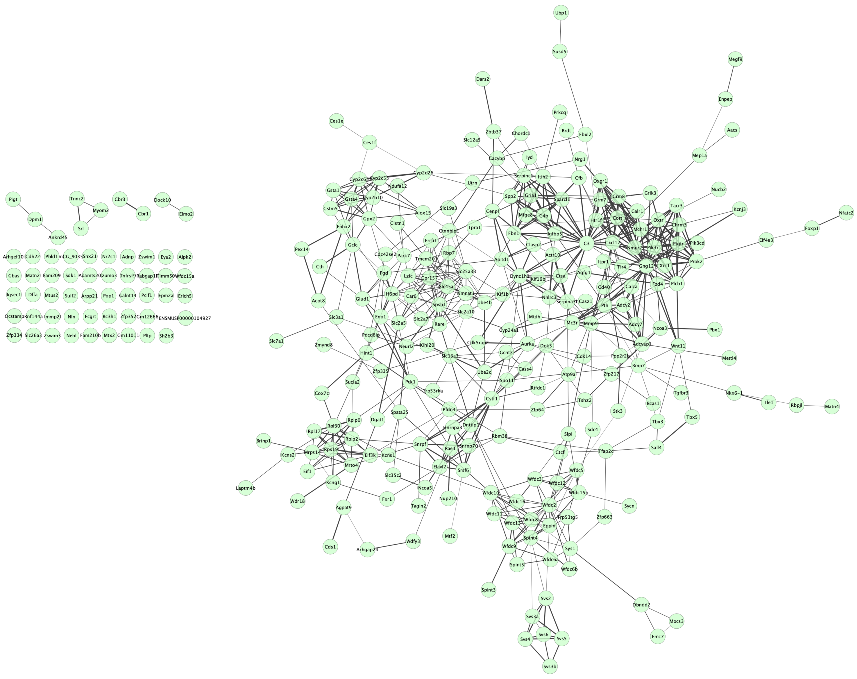
Suppl. Figure 10: Enriched human diseases among genes closest to significant SNPs from association analysis.



Suppl. Figure 11: STRING (Szklarczyk et al., 2019) protein-protein interaction network of proteins that are differentially expressed in the intestine (small intestine and colon) of germ-free (GF) mice compared to conventionally raised mice, found in the present study. The color of the network nodes indicates whether the QTL hit was found using the DNA abundances (green), RNA abundances (purple) or was found in both (orange). The shape represents if the gene of the protein was the closest gene to the significant SNP (rectangle), if the gene was also found in QTLs of other studies (octagon), a combination of both (diamond), or only differentially expressed in GF mice vs. conventionally raised mice. The node size expresses the number of taxa where the gene was found in a QTL. The edges represent protein-protein interactions, where the line thickness indicates the strength of the data support from text mining, experiments, databases, co-expression, gene-fusion, and co-occurrence.



Suppl. Figure 12: Visualization of the top hub genes calculated with the MCC algorithm and their first neighbors from the protein-protein interaction (PPI) network of genes found in intervals in present study that are also differentially expressed in germ-free versus conventionally raised mice. Edges represent the protein-protein associations. The red nodes represent genes with a high degree (= hub genes), and the yellow nodes with a low degree, while the blue nodes represent their first neighbors. All nodes shown are differentially expressed in GF mice. Hexagon shaped nodes are genes/proteins also found associated with gut microbiome abundances in other mouse QTL studies, and round nodes are 'only' differentially expressed in GF mice. The size of the node is an indication of the amount of taxa associated with the gene.



Suppl. Figure 13: Original protein protein interaction (PPI) network of 304 candidate genes closest to SNPs significantly associated with bacterial abundances. Generated in STRING (Szklarczyk et al., 2019) and Cytoscape (Shannon et al., 2003).

Chapter 2:
Characterizing candidate genes

1. Introduction

GWAS has proven itself very useful in identifying genetic loci contributing to many traits (Visscher et al., 2017). However, the approach does not, by definition, identify causal genes or genomic patterns, nor does it identify functional allelic variants that might contribute to the phenotype. Thus, the path from GWAS to biology is not clear-cut and follow-up studies are necessary to validate the results.

In this chapter, we follow-up on the association between ASV35 (putatively identified as *Bacteroides acidifaciens*) abundance and the host locus containing the *Sirt5* gene (peak SNP rs46570359) found in a previous version of the genome-wide association study (GWAS) of the gut microbial abundances presented in Chapter 1.

Evolutionary analysis provides clear evidence that all seven sirtuin families are highly-conserved in animal evolution, hence, a clearer understanding of the function of sirtuins may benefit from their analysis within the context of a simple biological system. Sirtuins are a family of highly conserved NAD$^+$-dependent protein deacylases with regulatory roles in numerous biological processes, such as genomic stability, metabolism and longevity (Haigis and Sinclair, 2010). Mammals have seven sirtuin proteins residing in different cellular locations. SIRT1, SIRT6 and SIRT7 are primarily situated in the nucleus, whereas SIRT2 dwells in the cytoplasm and SIRT3, SIRT4, and SIRT5 are localized in different compartments of the mitochondria (Michishita et al., 2005; Nakamura et al., 2008). The most important role of SIRT5 has been reported in the urea cycle in mitochondria. SIRT5 can deacylate, and thereby activate, carbamoyl phosphate synthetase (CPS1), leading to ammonia detoxification (Nakagawa et al., 2009). SIRT5 can also regulate ammonia production by controlling glutamine metabolism (Polletta et al., 2015). A succinylome study showed that SIRT5 desuccinylates a vast set of metabolic enzymes in the mitochondria involved in amino acid degradation, the tricarboxylic acid (TCA) cycle, and fatty acid metabolism (Park et al., 2013). Shortly thereafter, Nishida and colleagues identified glycolysis to be the top SIRT5-regulated pathway (Nishida et al., 2015). Taken together, these data suggest that SIRT5 may play an essential role in metabolic (nitrogen) homeostasis, whereby SIRT5 acts as a metabolic sensor under conditions of fasting, high-protein intake, or caloric restriction (CR), and that this response occurs in a circadian NAD+-dependent manner (Mauvoisin et al., 2017; Peek et al., 2012).

Emerging evidence suggests that gut microbes impact host metabolism (Cardona et al., 2016; Kuang et al., 2019; Lin and Zhang, 2017; Morowitz et al., 2011) and regulate nitrogen homeostasis through the de novo synthesis of amino acids and intestinal recycling of urea (Bergen and Wu, 2009; Stewart and Smith, 2005; Walpole et al., 2018). The *Sirt5*-associated bacterial species belongs to the phylum Bacteroidetes. This phylum is known to be linked with host metabolism, where it decreases upon excess energy intake and increases under conditions of CR (Johnson et al., 2017). Moreover, *B. acidifaciens* displays a cyclic diurnal oscillation that is dependent on feeding time (Thaiss et al., 2014). A recent study by Reese and colleagues proposes that the host can attenuate nitrogen limitation in the colon to upregulate preferred taxa, such as Bacteroidales, who increase in abundance with greater nitrogen supply (Reese et al., 2018). A possible interaction between *B. acidifaciens* and SIRT5 may therefore be mediated through the exchange of nitrogen.

Drosophila is an attractive organism to help understand the function of mitochondrial sirtuins and their association to the gut microbiome. The *Drosophila melanogaster* genome contains five sirtuins, *Sirt1, Sirt2, Sirt4, Sirt6*, and *Sirt7,* named for their closest mammalian orthologs. Of these, only *Sirt4* contains a predicted mitochondrial targeting sequence (Greiss and Gartner, 2009), suggesting that the *Drosophila Sirt4* (hereafter referred to as *dSirt4*) may act as the ancestral mitochondrial sirtuin that can perform many of the biological functions that are usually distributed across the three mitochondrial sirtuins within the mammalian system. Notably, *dSirt4* seems to also play a key role in metabolism and nitrogen regulation in *Drosophila*. Wood and colleagues found that *dSirt4* knockout flies display a number of phenotypes consistent with an inability to properly process and use energy stores (Wood et al., 2018) and have a markedly shorter lifespan compared to their wild-type controls, while overexpression increased the lifespan. Metabolomics results suggest that *dSirt4* may be involved in regulating both glycolysis and branched-chain amino acid oxidation.

Here, we present a follow-up case study of a gene-bacterial species association found in the genome-wide association study presented in Chapter 1. We make use of two different metaorganisms (house mouse and fruit fly) to determine the gene's role in association with bacteria in a circadian context. We hypothesize that a similar interaction between the *Drosophila* mitochondrial sirtuin, *dSirt4*, and a bacterial species functionally similar to mouse *B. acidifaciens* may exist. By combining the results of both model organisms, we provide evidence for a conserved role of sirtuins in regulation of bacterial abundance.

2. Results

2.1. Identification of *Sirt5* as a candidate for fine-scale characterization in mice

Within a preliminary run of the association mapping, we found an association between ASV35 (*Bacteroides acidifaciens/uniformis*) DNA-based abundance and a region on chromosome 13 containing two genes, *Gfod1* and *Sirt5* (Fig. 12). ASV35 represents an interesting candidate as it was found to be an indicator species for *Mus musculus musculus* in a geographic screen of mouse microbiomes (Fig. 13; Fokt, 2021). Moreover, inheritance deviates from Mendel's law of inheritance at the peak SNP rs46570359 with a lower number of heterozygotes compared to expected (X^2 (2, N=40)= 6.6, P = .037).

In the final GWAS, this association is no longer identified, due to the incorporation of the mating pair identifier as a random effect in the model which introduces collinearity with the genotype. However, there is a significant main effect of the genotype on the presence or absence of ASV35, F(2,297)= 28.64, P < .001, ω^2= 0.23. The Bonferroni and Tukey post hoc tests both revealed that the AA genotype had a higher presence of ASV35 compared to both the AG (P < .001) and the GG genotype (P < .001). Genotype GG also had a significantly higher presence of ASV35 compared to the AG (P < .001). There was a significant main effect of the genotype on the abundance of ASV35, F(2,297)= 4.29, P= .015, ω^2= 0.06 (Fig. 15A). The Bonferroni and Tukey post hoc test revealed that the abundance of ASV35 is significantly higher for genotype AA compared to genotype AG (P_{Bon} < .001 and P_{Tukey}= 0.048) and shows a similar trend for genotype GG compared to AG (P_{Bon} =0.097 and P_{Tukey}= 0.1).

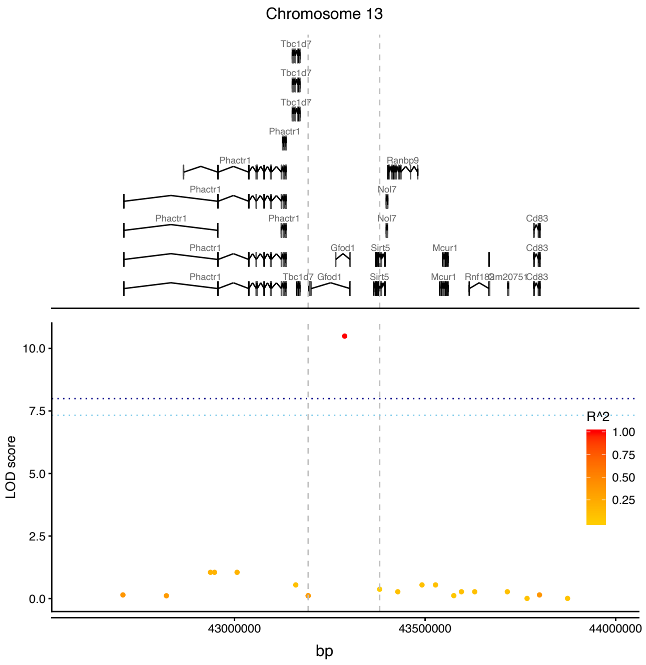


Figure 12: Region plot of the association of SNP rs46570359 with ASV35 (*Bacteroides acidifaciens/uniformis*) relative abundance on DNA level.

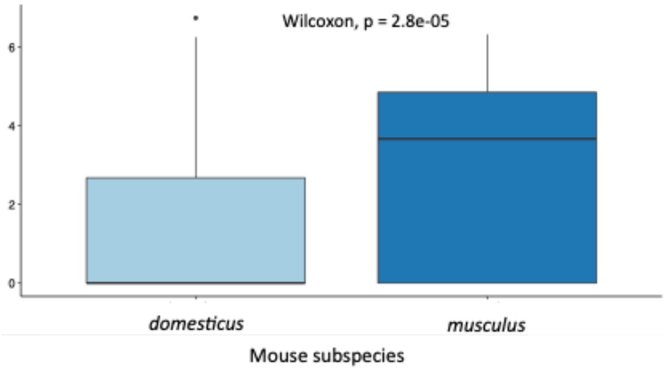


Figure 13: Relative abundance of ASV35 (*Bacteroides acidifaciens/uniformis*) in *Mus musculus musculus* and *Mus musculus domesticus* mice.

Using qRT-PCR analysis on the same cecal tissue used for microbial phenotyping for GWAS analysis, we explored whether the genotype at SNP rs46570359 is associated with a change in *Gfod1* and/or *Sirt5* expression in mice. The expression of *Gfod1* was significantly higher in mice homozygous for the *dom* allele (GG) compared to mice homozygous for the *mus* allele (AA; *P* = .016) or the heterozygotes (AG; *P* < .001; Fig. 14). However, there was no significant difference in expression of *Gfod1* between mice homozygous for the *mus* allele and the heterozygotes (Fig. 14).

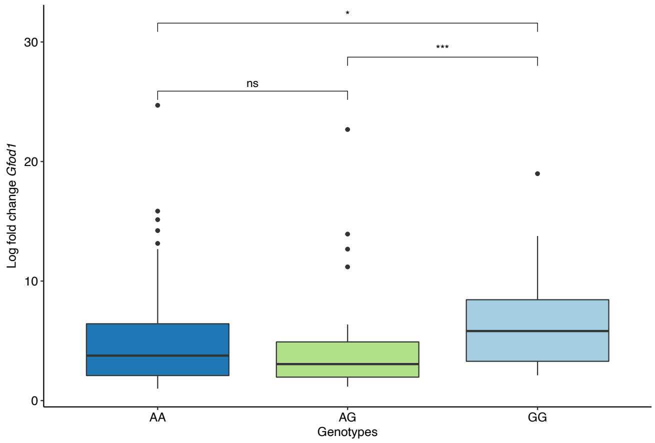


Figure 14: *Gfod1* expression with respect to genotype at SNP rs46570359. Fold change was determined by Δ-Δ Ct method relative to expression of *B2m* as a control gene. *P* values are calculated using a Wilcoxon rank sum test. ** *P* < .01; *** *P* < .001; **** *P* < .0001 .

Sirt5 expression was approximately three-fold greater in mice homozygous for the *mus* allele (AA) than mice homozygous for the *dom* allele (GG; *P* < .0001), while heterozygous animals had considerably lower expression than either homozygote (*P* < .0001 AA; *P* < .01 GG; Fig. 15B). A detailed look at the link between ASV35 abundance and the rs46570359 genotype revealed an interesting pattern: ASV35 was nearly completely absent in the heterozygous genotype (*P* < .0001 AA; *P* < .001 GG), but abundant in each of the homozygous parental genotypes (Fig. 15A). Additionally, the *mus* allele homozygotes had a greater abundance of ASV35 compared to the *dom* allele homozygotes (*P* < .01), mimicking the *Sirt5* expression pattern. This is in the same direction as seen in a geographical screen of gut microbiota of *mus* and *dom* mice, where *M. m. musculus* mice had a significantly greater abundance of ASV35 compared to *M. m. domesticus* mice (*P* = 2.8 × 10^{-5}; Fig. 13; Fokt, 2021). Due to the striking pattern observed between *Sirt5* expression and ASV35 abundance, we decided to focus on *Sirt5* for follow-up analyses.

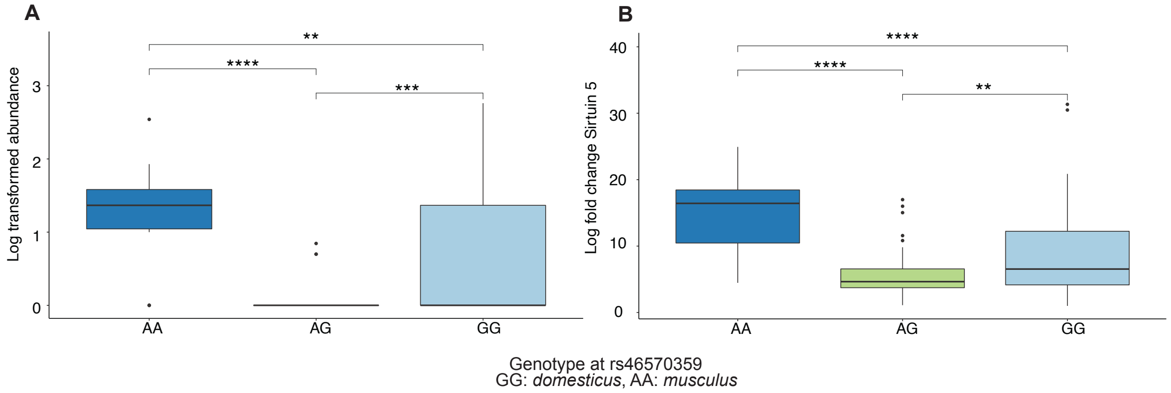


Figure 15: (A) ASV35 log10 transformed abundance according to genotype at SNP rs46570359 on chromosome 13. (B) *Sirt5* expression with respect to genotype at SNP. Fold change was determined by Δ-Δ Ct method relative to expression of *B2m* as a control gene. *P* values are calculated using a Wilcoxon rank sum test. ** *P* < .01; *** *P* < .001; **** *P* < .0001

Next, in a collaboration with Prof. Dr. Christoph Kaleta, we determined if there is metabolic interdependence between the host's *Sirt5* gene and *Bacteroides uniformis,* as this is the closest species to ASV35 with metabolic modeling data available. The metabolic metamodel included microbiome data, as well as host gene expression data of liver, brain and colon tissue. First, the metamodel containing all reactions found in *B. uniformis* was compared to a germ-free model and all metabolic reactions whose flow was decreased to less than 10% of their capability when *B. uniformis* reactions were enabled*,* were considered as *B. uniformis*-dependent reactions and their respective genes as *B. uniformis*-dependent genes. The model was initially validated by evaluating previously reported host gene expression data from GF mice mono-colonized with *B. uniformis* (Wang et al., 2019). This demonstrated a much larger response to *B. uniformis* colonization among genes anticipated to be dependent on *B. uniformis* products (Fig. 16A), lending credence to the model's validity.

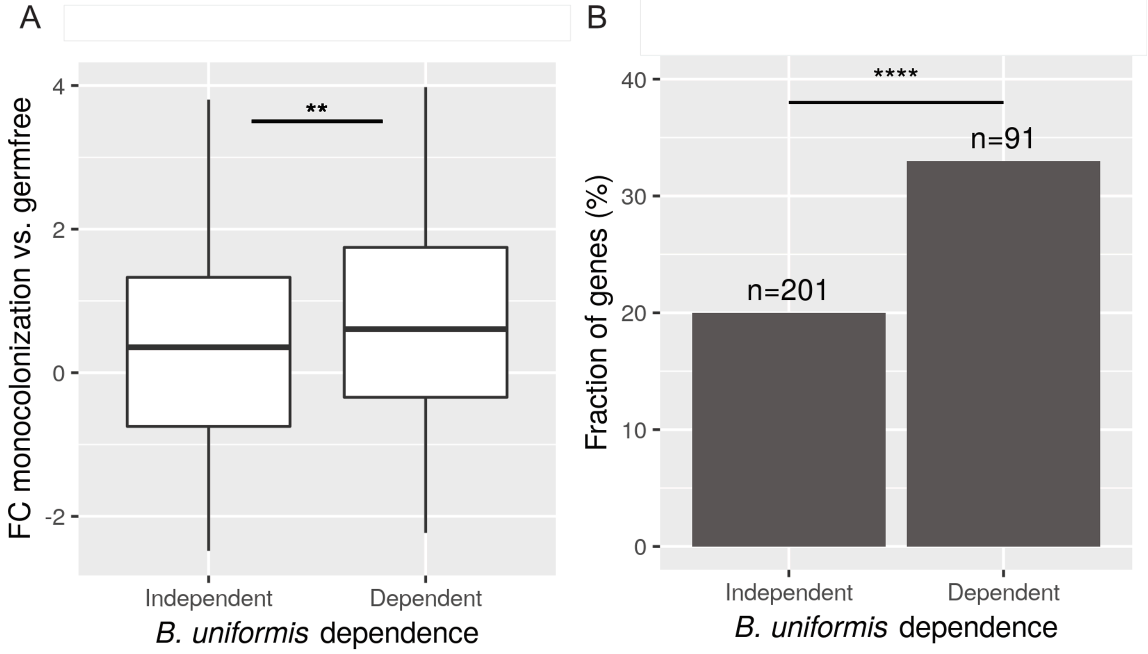


Figure 16: (A) Average fold-change of host genes in response to monocolonization with *B. uniformis*. Those host genes predicted to be dependent on *B. uniformis* metabolic products display on average a greater change in expression upon colonization. (B) Fraction of genes targeted by SIRT5 among *B. uniformis*-dependent vs. independent genes. *** *P* < .001, **** *P* < .0001.

Subsequently, we screened for overlap between *B. uniformis*-dependent genes and SIRT5 target genes, which are genes sensitive to SIRT5-mediated post-translational modification, mostly desuccinylation. In total, 647 SIRT5 targets have been discovered (Nakagawa and Guarente, 2009), 277 of which are engaged in metabolism. Notably, we find a highly significant enrichment of SIRT5 targets among *B. uniformis*-dependent genes (Fig. 16B), supporting the working assumption that ASV35's correlation with mouse genome variation is mediated by interacting with the *Sirt5* gene.

2.2. The influence of *Sirt5* on the bacterial community composition in a circadian context in mice

Within the original breeding stock from the animals used for mapping (Chapter 1), we determined their genotype at peak SNP rs46570359. Fecal samples were collected for ten mice per genotype (n_{tot}=30) every six hours over a period of 48 hours. The species richness (Chao1 Index) and evenness (Shannon Index) were significantly different between the homozygotes, AA and GG, where GG had a significantly higher alpha diversity (P_{Chao1} =.04 and $P_{Shannon}$ =.02, resp., Fig. 17A). Only time point 1pm has a significantly higher Shannon Index compared to the community of the 7pm time point, when comparing the alpha diversity measures within the genotypes according to sampling time point (*P =.03,* Fig. 17B). No time point showed a significantly different species richness using the Chao1 estimate (Fig. 17C). There is a significant difference in community composition according to genotype (*P* = .001, Fig. 18), but not according to time (*P* =.50), although the axes only explain about 1% of the variation. We found 38 ASVs and ten genera significantly more abundant in one genotype or a combination of genotypes (*P < .05*, Fig. 19), including ASV12, (corresponding to ASV35 from the mapping study), ASV7 (classifying to species level as *B. acidifaciens*), as well as several *Lactobacillus* species.

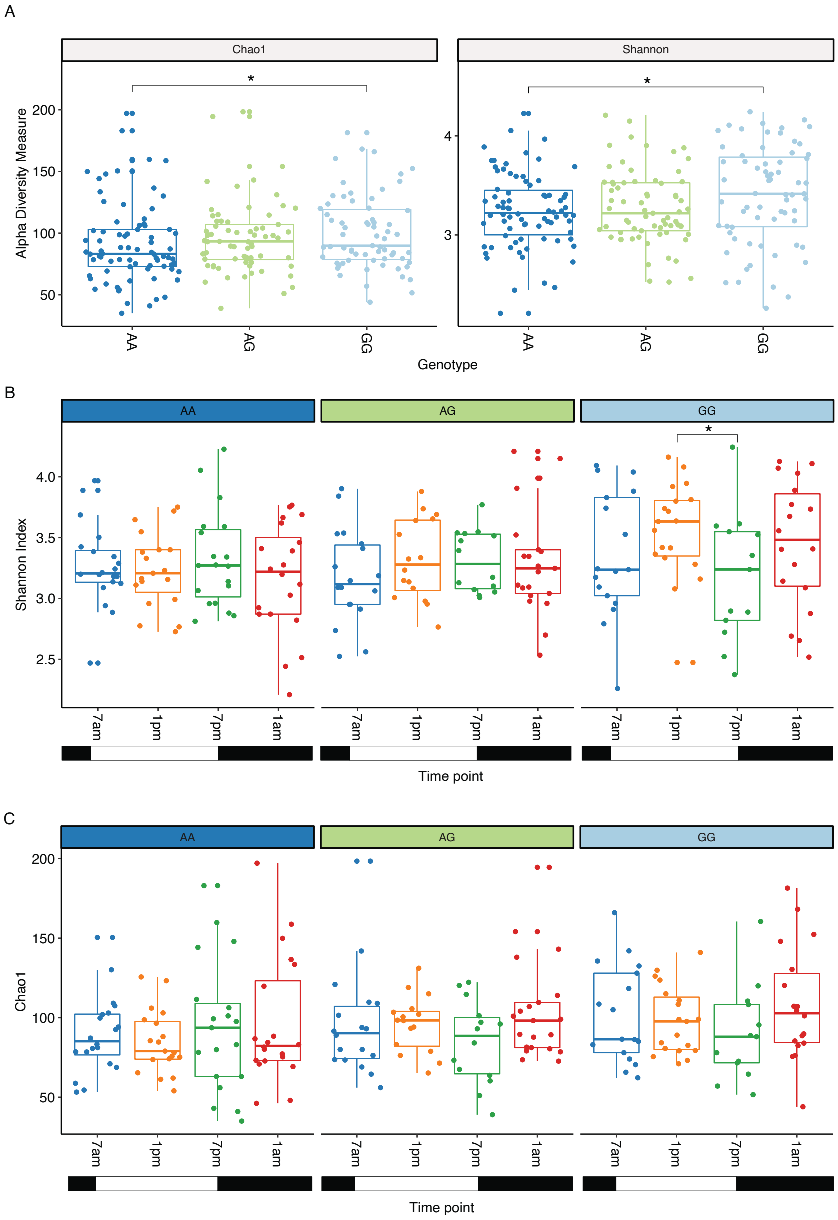


Figure 17: Species richness and evenness estimates according to genotype at rs46570359 on ASV level. (A) Chao1 and Shannon Index (left and right, resp.) according to genotype. (B) Shannon Index and (C) Chao1 estimate according to genotype and sampling time point. Significance is determined using the Wilcoxon Rank Sum test. Only significant comparisons are shown. * $P < .05$

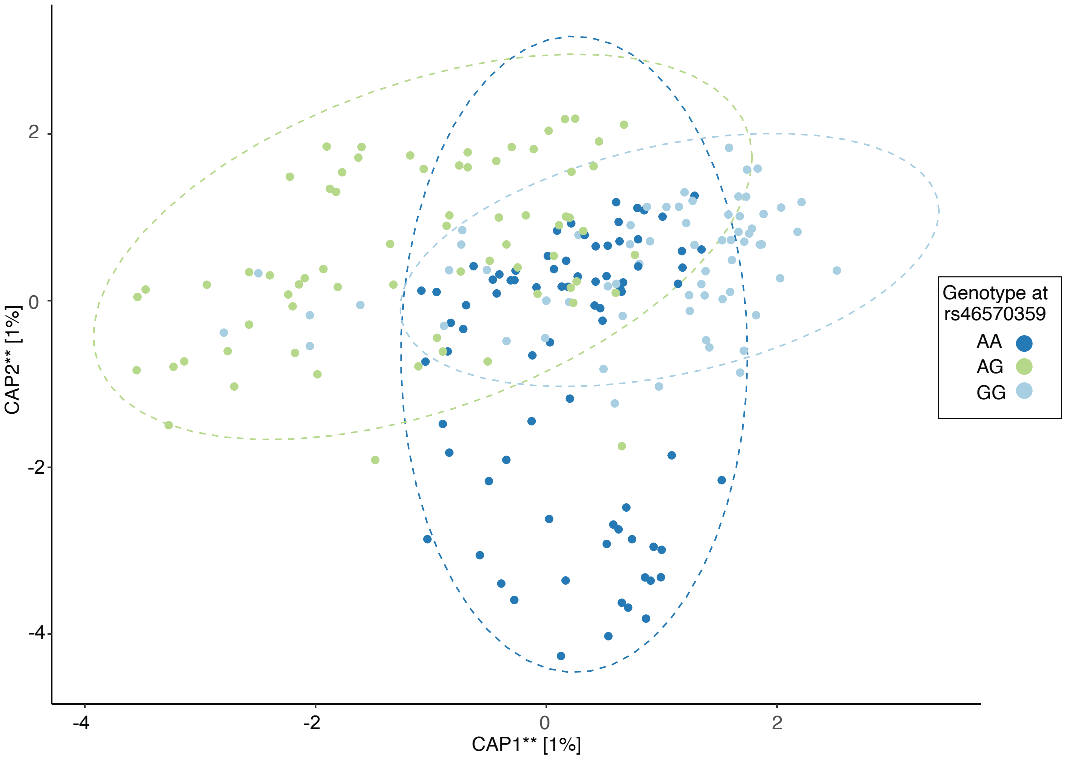


Figure 18: Constrained analysis of the Bray-Curtis dissimilarities according to genotype at rs46570359 at the ASV level. The originating mouse line is partialled out. ** *P* < .01

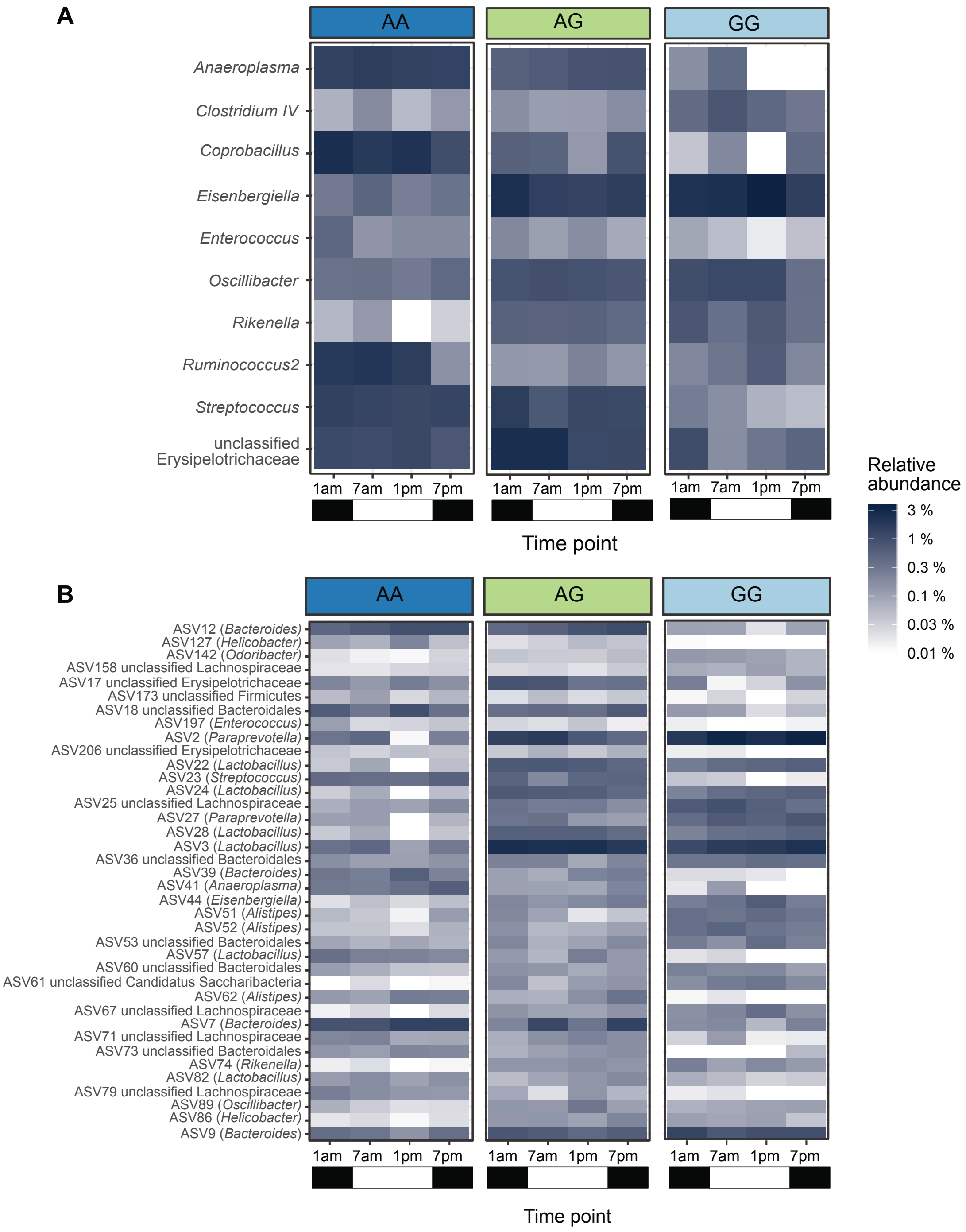


Figure 19: Indicator genera (A) and indicator ASVs (B) according to genotype or a combination of genotypes at rs46570359. The relative abundance is depicted by a color gradient. Only significant ASVs and genera are shown (10 000 permutations, FDR correction for multiple testing).

Next, we tested if the alpha diversity values and the individual bacterial abundances showed a circadian cycling pattern using JTK_CYCLE analysis (Hughes et al., 2010). Alpha diversity values do not show any significant cycling. Sixteen genera and 33 ASVs show cycling in one or more of the genotypes (Suppl. Tables 1 & 2, and Suppl.

Fig. 1 & 2). Interestingly, the heterozygote AG genotype has twice as many cycling bacteria compared to the homozygotes. Four of the 33 ASVs belong to the genus *Bacteroides*, including ASV7 and ASV12. The genus *Bacteroides* showed no significant cycling (Suppl. Table 2), while ASV7 showed significant cycling in the mice heterozygous at rs46570359 and ASV12 showed significant cycling in the heterozygous and AA genotype (Fig. 20 and Suppl. Table 1). Altogether, these results show that many bacterial taxa, including ASV35, exhibit a circadian pattern, however in a genotype dependent manner suggesting an effect of *Sirt5* expression on bacterial cycling.

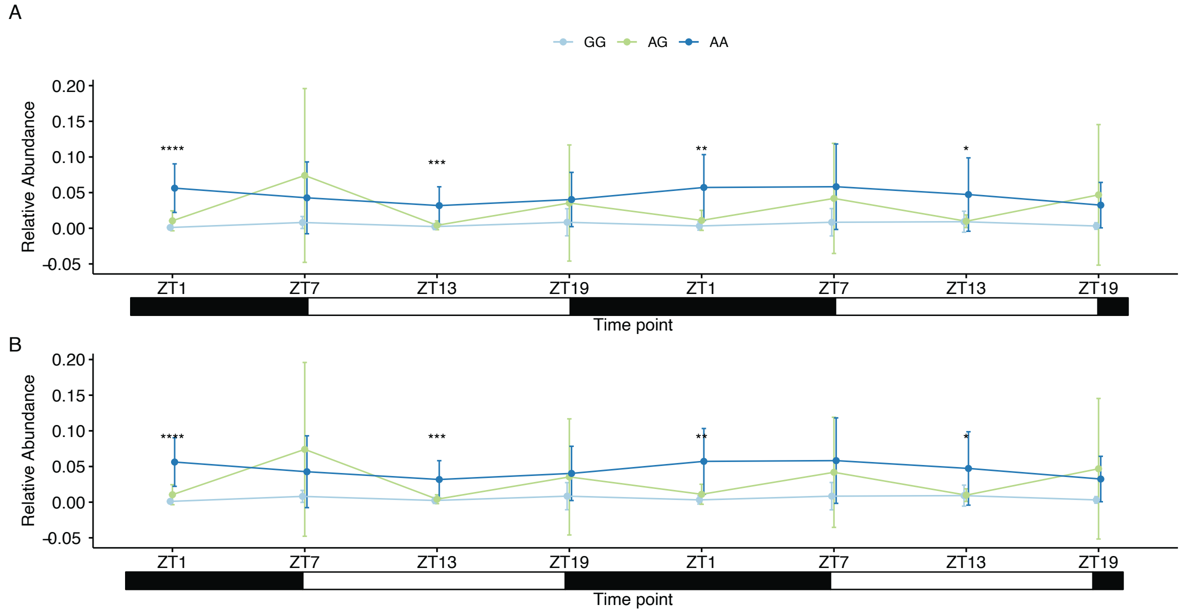


Figure 20: Relative abundance of ASV7 (*Bacteroides acidifaciens*) (A) and ASV12 (*Bacteroides* sp. ASV35) (B) at different time points over a 48 hour period for the different genotypes at rs46570359. Significant differences in relative abundances between genotypes within one time point are shown. ANOVA: **** *P*<.0001, *** *P* < .001, ** *P* < .01, * *P* < .05

2.3. Characterization of *dSirt4* and its relation to the gut microbiome and circadian rhythms in *Drosophila*

This section is based on joint work with Abdulgawaad Saboukh, who carried out the experiments, and Prof. Dr. Thomas Roeder and was funded with the CRC1182 Young Investigator 2019 Award, awarded to me.

2.3.a. Protein alignment of Sirt5 to dSirt4

Aligning the mouse *Sirt5* protein sequence to the *Drosophila Sirt4* (*dSirt4*) sequence reveals an overall sequence similarity of 23.7 %. However, the active sites, and the NAD+ and zinc binding sites and regions are conserved between both genes (Suppl. Fig. 3), as well as the folding of SIRT5 compared to dSIRT4 (Fig. 21) and of all three murine mitochondrial sirtuins (SIRT3, SIRT4, and SIRT5) (Suppl. Fig. 4).

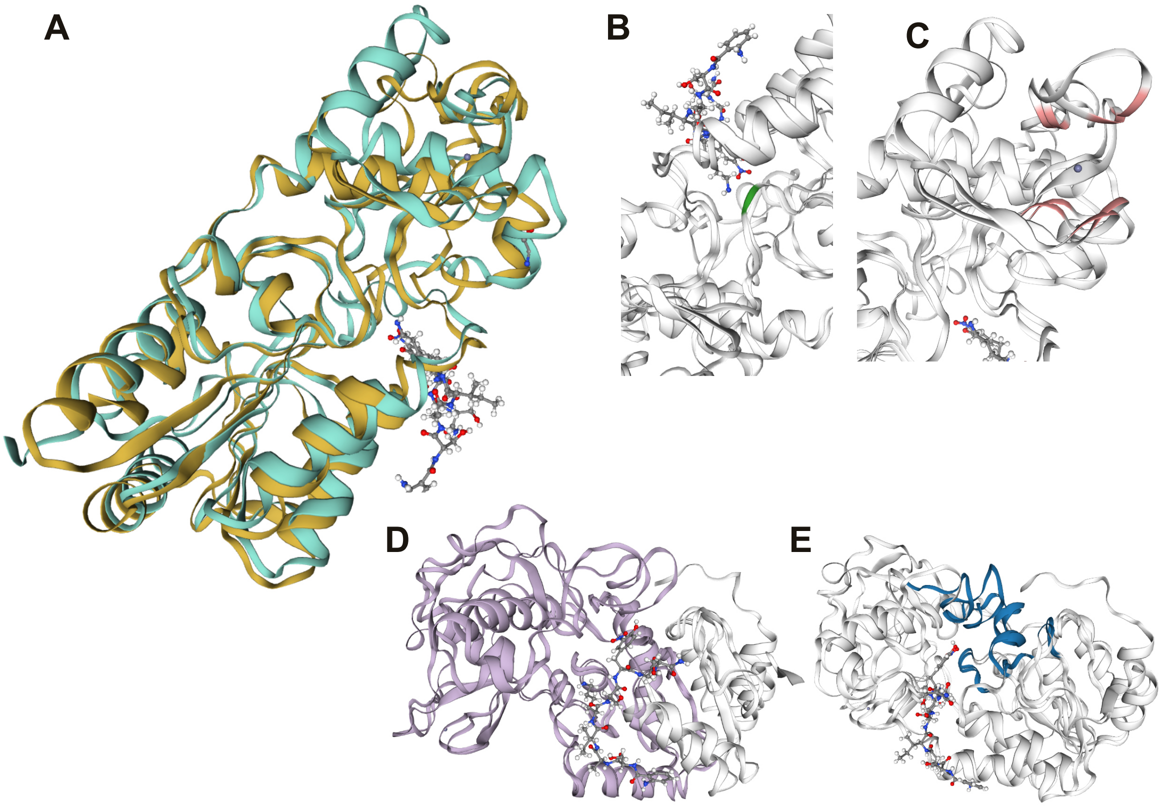
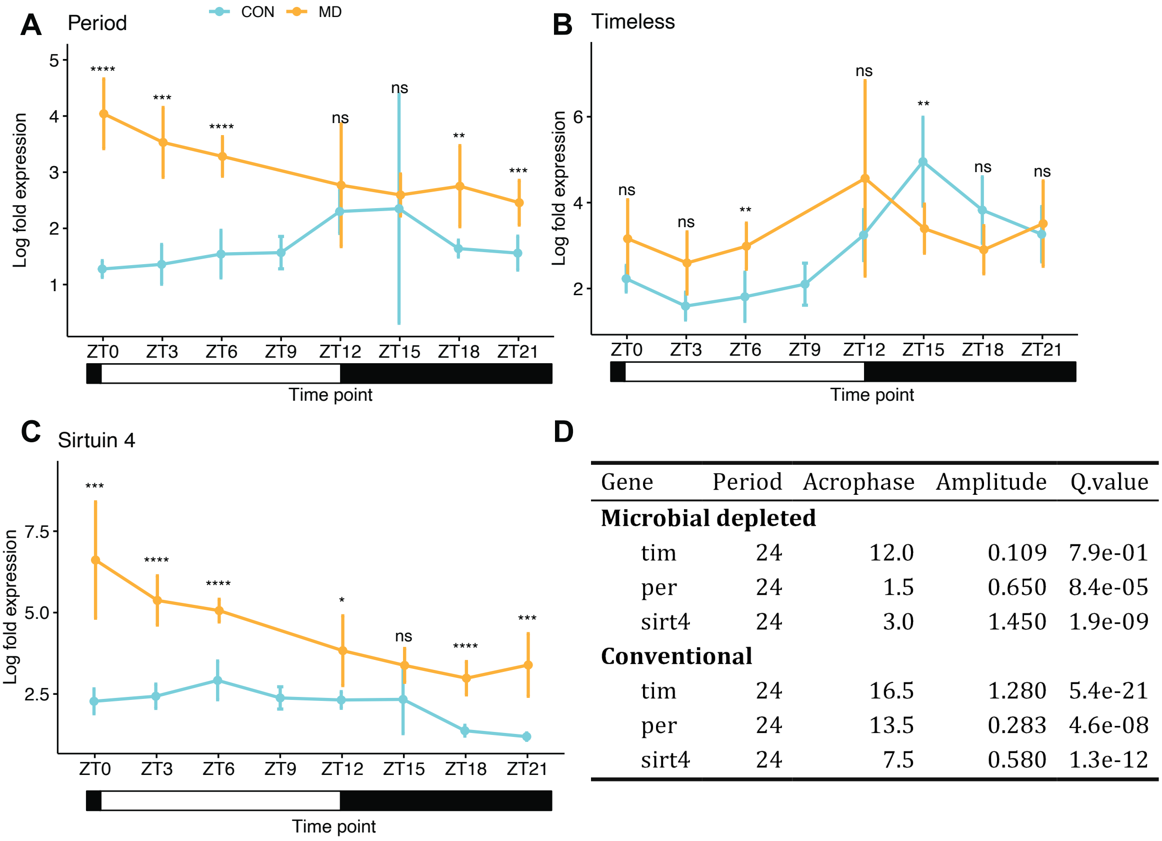


Figure 21: Superimposed predicted folded proteins dSIRT4 and SIRT5. (A) Protein structures of dSIRT4 (turquoise) and SIRT5 (ochre) bound to ligand (ball-stick structure). (B) Detail of the active site (green) interaction with ligand. (C) Detail of interaction of metal binding sites (pink) with zinc ion (purple ball). (D) Interpro sirtuin family domain colored in purple. (E) NAD+ binding regions are highlighted in blue.

2.3.b. Circadian cycling of dSirt4

To test if *dSirt4* displays a circadian pattern, we raised half of W^{1118} *Drosophila* flies conventionally (n=10 per time point), by inoculating them with six bacterial species (*Acetobacter Thailandi*, *Acetobacter pomorum*, *Lactobacillus brevis EW*, *Lactobacillus plantarum WJL*, *Enterococcus faecalis*, and *Commensalibacter intestini A911T*), and the other half were raised microbial depleted (n=7-10 per time point). We measured the expression level of *dSirt4* and two known cycling genes *Period* and *Timeless*. All three genes showed significant cycling patterns, however the acrophase (*i.e.* the time point with the highest expression) shifts depending on microbial status (Fig. 22). Interestingly, the expression levels of all three genes is higher in microbial depleted genes (Suppl. Fig. 5).



Gene	Period	Acrophase	Amplitude	Q.value
Microbial depleted				
tim	24	12.0	0.109	7.9e-01
per	24	1.5	0.650	8.4e-05
sirt4	24	3.0	1.450	1.9e-09
Conventional				
tim	24	16.5	1.280	5.4e-21
per	24	13.5	0.283	4.6e-08
sirt4	24	7.5	0.580	1.3e-12

Figure 22: Expression of Period (A), Timeless (B), and Sirtuin 4 (C) over a period of 24 hours within microbial depleted (MD; orange) and conventionally raised (CON; blue) flies. Expression levels are relative to housekeeping gene *Rpl32*. *P* values on plots A-C are calculated on the difference between CON and MD for a given time point using a Wilcoxon rank sum test. ns: *P* > .05; * : *P* < .05; ** : *P* < .01; *** : *P* < .001; **** : *P* < .0001 (D) Table showing the results from JTK_CYCLE analysis.

2.3.c. Influence of dSirt4 expression on bacterial load and microbiome composition

Next, we used four different groups of flies, *dSirt4* over-expression (OE), *dSirt4* knock-out (KO), *dSirt4* knock-down (KD), and wild type (WT) flies, to test whether dSirt4 expression has an influence on bacterial load and microbiome composition. Flies were sampled every three hours over a period of 24 hours. First, we evaluated the expression of *dSirt4* in the different groups. The expression of *dSirt4* in OE is 200 times higher compared to WT, while *dSirt4* in KD is on average 70 times lower expressed compared to WT.

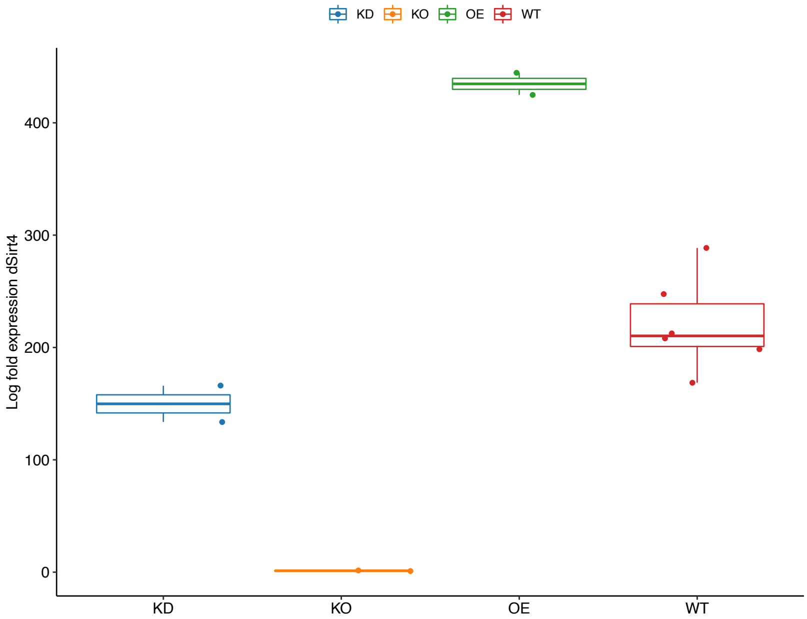


Figure 23: Log fold expression of *dSirt4* for OE (n=2), KD (n=2), KO (n=2), and WT (n=6) flies. Expression of *dSirt4* is relative to *Rpl32* and to the expression level in KO flies.

We measured the bacterial load in the KO, KD, OE, and WT flies and tested for cycling using JTK_CYCLE. We find that only OE flies show significant cycling in their bacterial load (*P* < .001, FDR corrected P values), however the other three groups show a trend (*P* = .01, FDR corrected *P* values; Table 2, and Fig. 24). Interestingly, the average expression of *dSirt4* shows a negative trend to the average bacterial load (*P* = .08, R^2 = -1, spearman correlation test; Suppl. Fig. 6).

Table 2: Results from JTK_CYCLE analysis on the bacterial load within the different fly groups.

Group	Period	Acrophase	Amplitude	P.value	Q.value
W1118	24	3.0	10100	7.57e-02	1.05e-01
KO	24	0.0	4330	7.90e-02	1.05e-01
KD	24	7.5	19800	7.90e-02	1.05e-01
OE	24	0.0	261	1.00e-11	8.00e-11

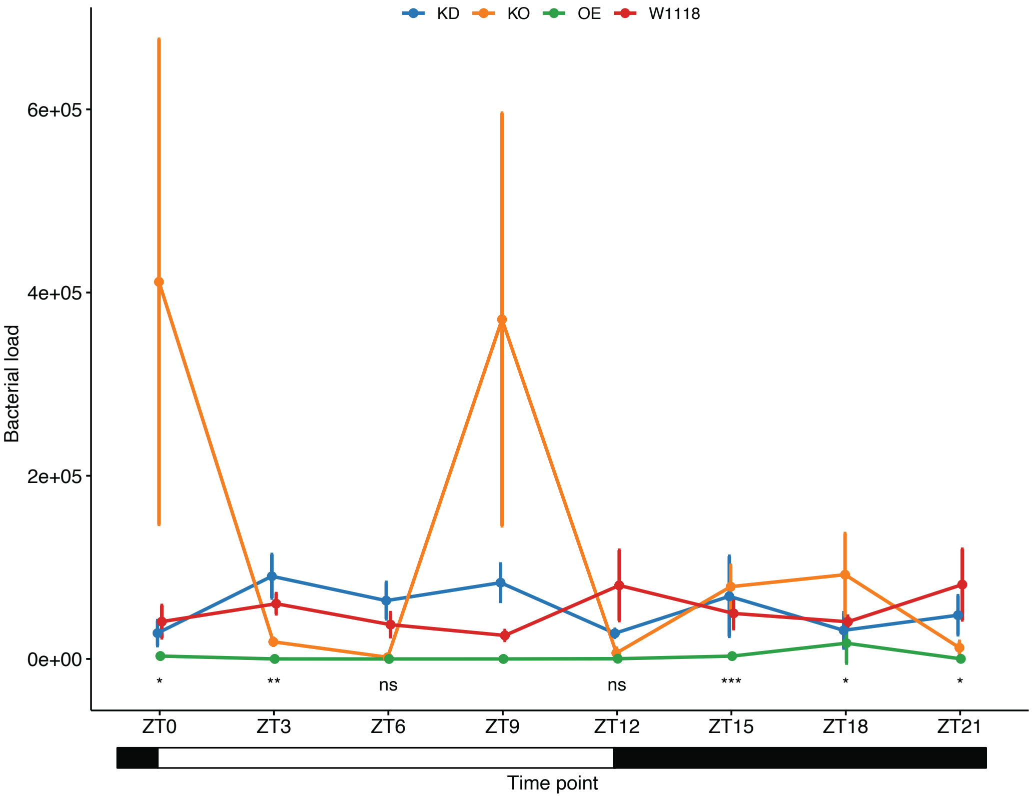


Figure 24: Bacterial load of the groups with different *dSirt4* expression measured every three hours over a period of 24 hours.

Next, we characterized the microbial composition of the different fly groups to identify cycling taxa that potentially have a homologous role to *B. acidifaciens* in mice. The relative abundances calculated from the 16S sequences were multiplied with the measured bacterial load in the samples in order to obtain an absolute abundance. First, we calculated the alpha diversity values, Shannon Index and Chao1 Index, to analyze species richness and evenness within bacterial communities and found that the Shannon Index significantly cycles in KD (P = 4.91 × 10^{-7}) and KO flies (P = 2.36 × 10^{-5}), while the Chao1 Index cycles in KO (P = 5.43 × 10^{-4}) and OE (P = 5.57 × 10^{-5}; Table 3, Fig. 25).

Table 3: Cycling analysis of alpha diversity values using JTK_CYCLE. Q.values are FDR corrected *P* values: . *P* < .01, * *P* < .05, ** *P* < .01, *** *P* < .001

Alpha.measure	Period	Acrophase	Amplitude	P.value	Q.value	Significance
Knock-down						
Shannon	24	0.0	0.182200	7.02e-08	4.91e-07	***
Chao1	0	NaN	0.000000	1.03e-01	1.80e-01	
Knock-out						
Shannon	24	13.5	0.008338	6.75e-06	2.36e-05	***
Chao1	24	13.5	0.707100	2.33e-04	5.43e-04	***
Overexpression						
Shannon	24	13.5	0.040710	5.17e-02	7.23e-02	.
Chao1	24	15.0	1.414000	7.95e-06	5.57e-05	***
Wild type W1118						
Shannon	0	NaN	0.000000	7.90e-02	2.77e-01	
Chao1	0	NaN	0.000000	2.04e-01	4.75e-01	

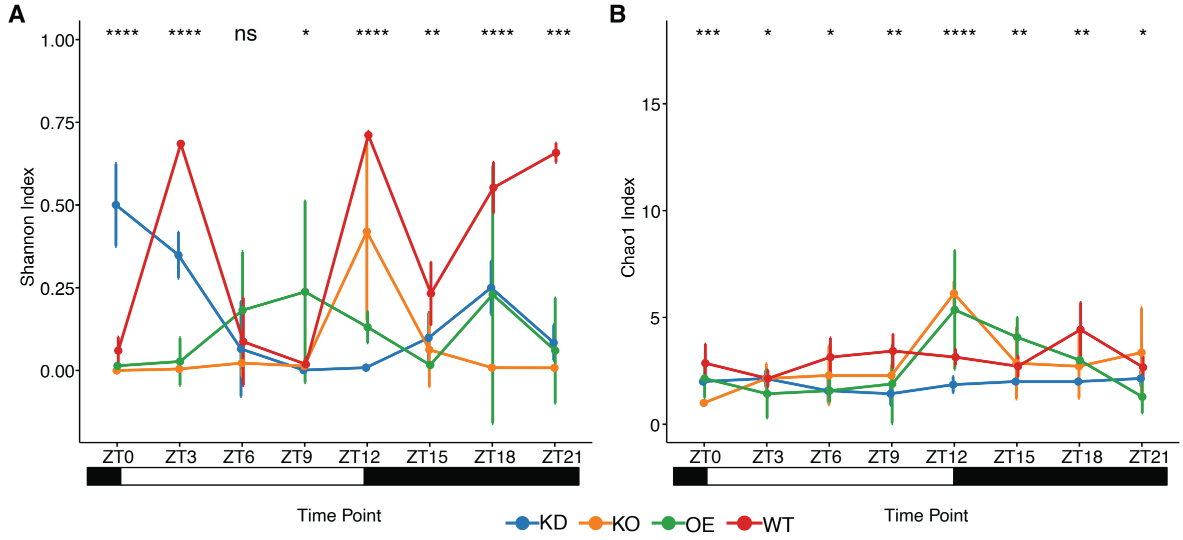


Figure 25: Alpha diversity measures across different time points for the different fly groups (OE, KO, KD and WT) using the Shannon Index (A) and Chao1 Index (B). Significant differences of alpha diversity between the *dSirt4* genotype within one time point is tested using ANOVA. ns *P* > 0.5; * *P* < .05; ** *P* < .01; *** *P* < .001; **** *P* < .0001

The flies of the different groups exhibit a distinct community composition, where most of the variation is explained by the *dSirt4* genotype, *i.e.* OE, KO, KD or WT (adonis of BC dissimilarity on ASV level: R^2_{sirt4}= 0.44, $R^2_{timepoint}$= 0.087, F_{sirt4}=65.44, $F_{timepoint}$=5.54, *P* < .001 for all; Fig. 26). Notably, KO and OE flies display a more similar community composition compared to WT and KD flies. In WT and KD flies, we see a high abundance of the genus *Wolbachia,* while the communities of KO and OE flies were consisting of mostly *Acetobacter (*Suppl. Fig. 7).

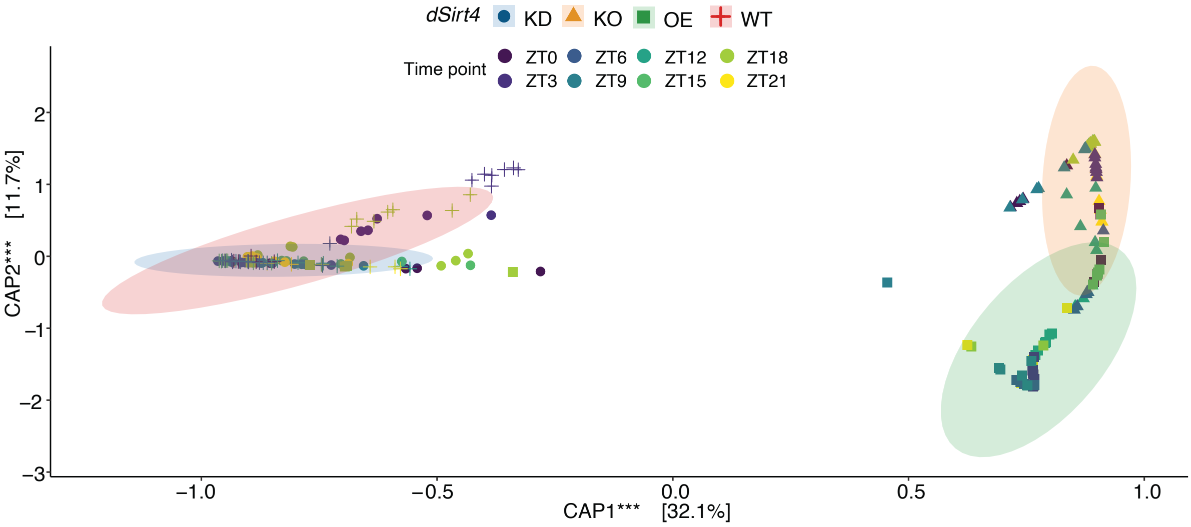
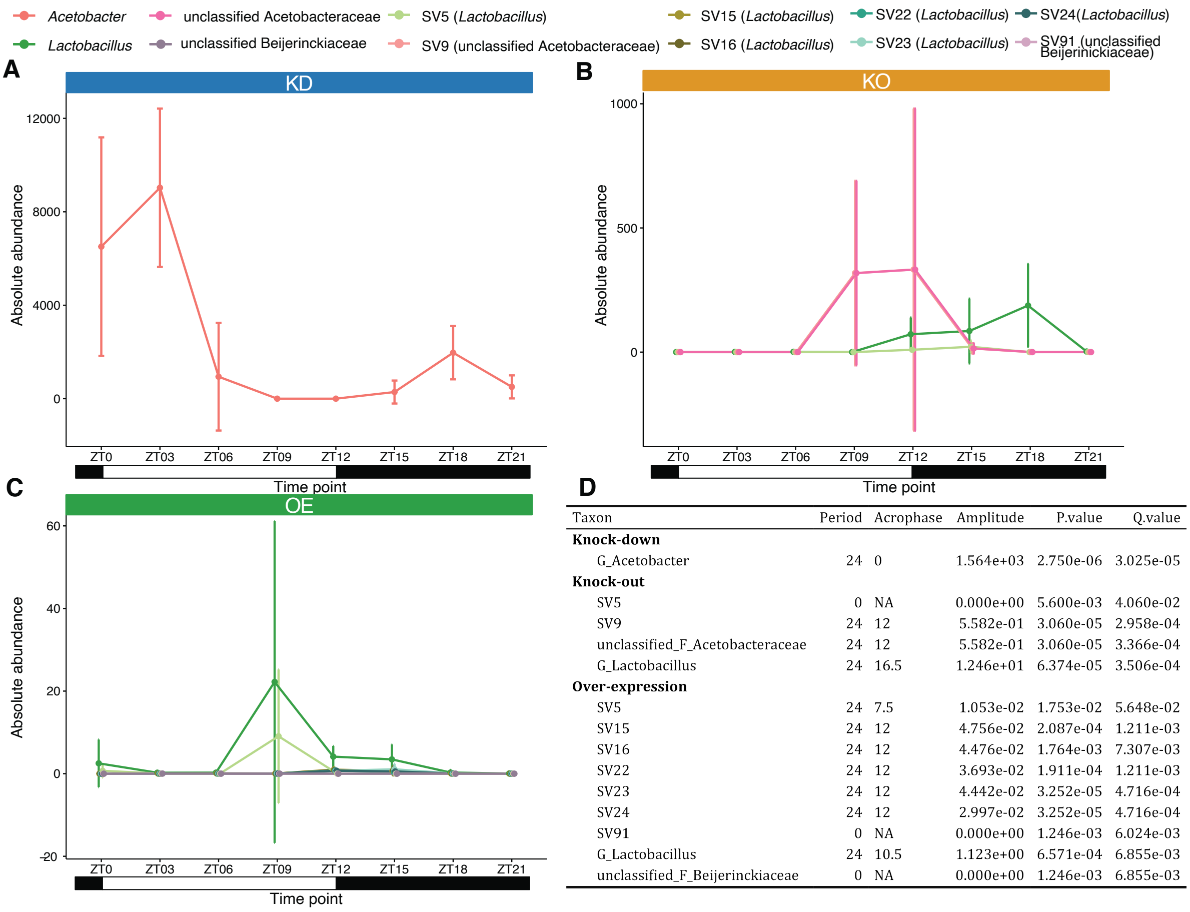


Figure 26: Constrained analysis of Bray-Curtis dissimilarity. Points are colored according to the sampling time points and shaped according to *dSirt4* status. *** *P* < .001, *P* values for axes are calculated using anova with 10,000 permutations.

Out of the 13 taxa tested (five genera and eight ASVs), none showed significant cycling in the wild type flies, while only one, annotated as *Acetobacter sp.*, showed significant cycling in the knock-down flies (Fig. 27A, D). Four taxa, ASV5 (*Lactobacillus* sp.), ASV9 (unclassified Acetobacteraceae), genus *Lactobacillus*, and an unclassified genus belonging to Acetobacteraceae, showed significant cycling in the knock-out flies (Fig. 27B, D) and nine in the over-expressed flies (ASV5 (*Lactobacillus* sp.*)*, ASV15 (*Lactobacillus* sp.), ASV16 (*Lactobacillus* sp.), ASV22 (*Lactobacillus* sp.), ASV23 (*Lactobacillus* sp.), ASV24 *(Lactobacillus* sp.), ASV91 (unclassified Beijerinickiaceae), genus *Lactobacillus*, an unclassified genus belonging to Beijerinickiaceae) (Fig. 27C, D).



Taxon	Period	Acrophase	Amplitude	P.value	Q.value
Knock-down					
G_Acetobacter	24	0	1.564e+03	2.750e-06	3.025e-05
Knock-out					
SV5	0	NA	0.000e+00	5.600e-03	4.060e-02
SV9	24	12	5.582e-01	3.060e-05	2.958e-04
unclassified_F_Acetobacteraceae	24	12	5.582e-01	3.060e-05	3.366e-04
G_Lactobacillus	24	16.5	1.246e+01	6.374e-05	3.506e-04
Over-expression					
SV5	24	7.5	1.053e-02	1.753e-02	5.648e-02
SV15	24	12	4.756e-02	2.087e-04	1.211e-03
SV16	24	12	4.476e-02	1.764e-03	7.307e-03
SV22	24	12	3.693e-02	1.911e-04	1.211e-03
SV23	24	12	4.442e-02	3.252e-05	4.716e-04
SV24	24	12	2.997e-02	3.252e-05	4.716e-04
SV91	0	NA	0.000e+00	1.246e-03	6.024e-03
G_Lactobacillus	24	10.5	1.123e+00	6.571e-04	6.855e-03
unclassified_F_Beijerinckiaceae	0	NA	0.000e+00	1.246e-03	6.855e-03

Figure 27: Taxa showing significant cycling patterns in *dSirt4* knock-down (A), knock-out (B), and over-expression flies. Panel D shows a table with the results of the JTK_CYCLE analysis.

3. Discussion

Genome-wide association studies provide numerous genetic loci associated with traits. However, the biology behind these interactions remains unknown, hence functional characterization of the candidate genes after GWAS is necessary. Here, we have demonstrated the use of two model organisms, *Mus musculus* and *Drosophila*, in characterizing the influence of candidate gene *Sirt5* on the bacterial composition. In both organisms, *Sirt5/dSirt4* expression had an effect on the community composition and on cycling of bacterial taxa suggesting a conserved role for the mitochondrial sirtuin in regulating the abundance of bacterial taxa.

In the GWAS mapping population, we noticed a significantly lower number of heterozygotes than expected, suggesting selection against heterozygotes for the SNP. A study characterizing *Sirt5*-deficient mice also noted a non-mendelian inheritance of *Sirt5*$^{-/-}$ mice with approximately 40% of prenatal loss of *Sirt5*$^{-/-}$ offspring (Yu et al., 2013). We evaluated whether the expression of *Sirt5* is associated with the genotype at SNP rs46570359. This revealed a significant nearly three-fold higher expression of *Sirt5* in the *mus* allele (AA) compared to the *dom* allele (GG). The expression of *Sirt5* in the heterozygote (AG) was lower than in both homozygotes. This pattern is consistent with previous findings in *dom*, *mus*, and hybrid mice, in which hybrid genetic backgrounds resulted in extensive misexpression of immune-related genes (Wang et al., 2015). Furthermore, the expression of *Sirt5* mirrored the abundance of ASV35 according to genotype, where the *mus* allele (AA) had a higher abundance of ASV35 (*Bacteroides acidifaciens/uniformis*) compared to the *dom* allele (GG), while ASV35 is absent in the heterozygote. This was also consistent with ASV35 abundance in a geographic screen of *mus* and *dom* mice, where *mus* mice had a higher abundance of ASV35. We used a metamodel to evaluate whether metabolic interdependency exists between the host and ASV35. This revealed a significant enrichment of SIRT5 targets among *B. uniformis*-dependent genes. This supports the assumption that the association between ASV35 and variation in the mouse genome is mediated by interacting with SIRT5. Taken together, these results provide compelling evidence of a host gene-microbe interaction involving the *Sirt5* gene and a taxon belonging to *B. acidifaciens* and/or *B. uniformis*, which has diverged since the common ancestor of the *dom* and *mus* subspecies roughly 0.5 MYA ago.

SIRT5 is an NAD+ dependent protein deacylase playing a role in maintaining metabolic homeostasis (Fischer et al., 2012; Chalkiadaki and Guarente, 2012; Cantó et al., 2015; Kumar and Lombard, 2018). NAD+ is its rate-limiting compound and nicotinamide (NAM), a product of the sirtuin reaction, inhibits SIRT5's desuccinylase activity without affecting its deacetylase activity (Du et al., 2011; Fischer et al., 2012; Madsen and Olsen, 2012). As NAD+ biosynthesis is tightly controlled by the circadian clock BMAL1:CLOCK complex (Nakahata et al., 2009;

Zhang and Sauve, 2018), we wanted to characterize the influence of the SNP rs46570359 near *Sirt5* on the bacterial community composition in mice in a circadian context. The *mus* allele (AA) showed a significant lower species richness compared to the *dom* allele (GG) and the whole bacterial community was significantly different according to genotype. Numerous ASVs and genera were more abundant in one genotype, including ASV7 (*B. acidifaciens*) and ASV12 (~ ASV35 in mapping population, *B. acidifaciens*), which were more abundant in the *mus* allele (AA). Nutrient availability, such as nitrogen, due to the expression levels of *Sirt5* in the different genotypes could cause the different microbiota composition. Reese et al. have shown that especially species belonging to the phylum *Bacteroidetes* increase in abundance with more nitrogen availability in the large intestine, suggesting resource limitation as way for hosts to control their gut communities (Reese et al., 2018).

Sixteen genera and 33 ASVs showed significant circadian cycling within different genotypes in the mouse model. These included *B. acidifaciens* sp. ASV7 (cycling in heterozygote) and ASV12 (cycling in AA and heterozygote), and *Lactobacillus* sp. ASV3 (cycling in AA and heterozygote) and ASV6 (cycling in heterozygote), all taxa previously shown to display a cyclic diurnal oscillation that is specifically related to feeding time, *i.e.* it can be shifted by forced 12 hour shifts of feeding times (Thaiss et al., 2014). Interestingly, the heterozygote, which is the allele with the lowest *Sirt5* expression, has twice as many cycling bacteria. A reduced *Sirt5* expression would result in a larger availability of ammonia, as SIRT5 activates carbamoyl phosphate synthetase 1 (CPS1), which starts the urea cycle. Several bacteria, such as *Lactobacillus* sp., have the capability to reduce ammonia (Singh et al., 2018; Wrong and Vince, 1984; Naidu et al., 2002), while others, such as *Bacteroides* sp. produce ammonia from amino acids and peptides (Vince and Burridge, 1980; Richardson et al., 2013). The opposite cycling pattern of *Lactobacillus* sp. (ASV3 and ASV6) and *Bacteroides* sp. (ASV7 and ASV12) suggests that bacteria could compensate for the expression of *Sirt5* by performing some of its metabolic functions and that their interaction could go through exchange of ammonia and urea. Short-chain fatty acids (SCFA) can be another point of interaction as SIRT5 desuccinylation regulates fatty acid oxidation (Du et al., 2018; Goetzman et al., 2020; Rardin et al., 2013) and SCFAs produced by bacteria induce circadian clock entrainment (Tahara et al., 2018), whereby *Bacteroides* is a known SCFA producer with a preference for propionate (Rios-Covian et al., 2017). Collectively, these result show a clear interaction of SIRT5 and several bacterial species through the exchange of nutrients.

In *Drosophila*, we first showed that *dSirt4* contains all active sites and binding regions of murine *Sirt5,* indicating that *dSirt4* might be the ancestral mitochondrial sirtuin fulfilling a similar role as the three murine mitochondrial sirtuins (*Sirt3, Sirt4, Sirt5*). *dSirt4* showed significant cycling in both conventionally raised as well as microbial depleted flies, where microbial depleted flies showed an overall higher expression of *dSirt4*. As *dSirt4* is highly expressed during caloric restriction and fasting (Wood et al., 2018), a higher expression in microbial depleted flies could be due to a missing cue normally provided by microbes to signal food intake (Han et al., 2021). The acrophase of *dSirt4* expression shifted 4.5 hours, from noon in conventionally raised flies to dawn in microbial depleted flies. *Timeless* also exhibited a 4.5 hour shift in acrophase, while the acrophase of *Period* shifted 12 hours. This would suggest missing signals from the microbiome disrupt the circadian rhythm of the flies, a phenomenon also seen in germ-free mice (Mukherji et al., 2013; Voigt et al., 2014; Ogawa et al., 2020).

Next, to determine the influence of *dSirt4* expression on the gut microbiome, we used flies with *dSirt4* over-expressed (OE), knocked down (KD), and knocked out (KO). The expression level of *dSirt4* showed a negative correlation with bacterial load. This is an extension of a study by Carneiro Dutra et al., who showed that the presence of the endosymbiont *Wolbachia* is associated with a decrease in *dSirt4* expression (Carneiro Dutra et al., 2020). Interestingly, the influence of *dSirt4* on the bacterial load goes in both directions: flies raised without bacteria show a higher *dSirt4* expression compared to WT and flies over-expressing *dSirt4* have a low bacterial load. This implies a reciprocal interaction between *dSirt4* and the bacterial community.

The bacterial load showed significant cycling in OE flies, while others showed a trend. The alpha diversity displayed significant circadian patterns in OE, KO, and KD, but not in WT flies. *dSirt4* expression had an influence on the community composition, where OE and KO showed a similar community composition, as well as KD and WT. We found no cycling taxa in WT flies, however one taxon cycled in KO flies, four taxa in KO, and nine taxa in OE flies, including several *Lactobacillus* sp. A study by Elya et al. (2016) showed that *Drosophila* exhibits very low host response to taxa-specific colonization of the gut (Elya et al., 2016). This could suggest that bacterial load is a more important factor, as an extreme bacterial load could be a burden for the host.

Wolbachia, present in WT, KD, and OE, but not KO flies, is an intracellular microbe that is known to alter host fitness and phenotypic analyses (Clark et al., 2005) and is therefore, a confounding factor in this study. Moreover, Carneiro Dutra et al. (2020) discovered that the presence of *Wolbachia* is associated with a decrease in *dSirt4* expression (Carneiro Dutra et al., 2020). Consequently, the interpretation of our results from the different fly groups is compromised, as we cannot prove that the effect (*e.g.* the higher bacterial load in KO flies) is independent of *Wolbachia* infection. Further studies in *Wolbachia*-free *dSirt4* KO, KD, OE, and WT flies are necessary to untangle the effect of *dSirt4* expression on the bacterial community composition in a circadian context.

In summary, our study showed that *Sirt5* and *dSirt4* have an influence on the gut microbiome composition of mice and flies, respectively. This suggest a conserved role across animals for a mitochondrial sirtuin in influencing bacterial abundance. This interaction most likely functions through an exchange of nutrients, as *Sirt5* and *dSirt4* play an important role in maintaining metabolic homeostasis (Kumar and Lombard, 2018; Wood et al., 2018). Further research into knock-out *Sirt5* mice will help elucidate the exact interaction between bacterial species and *Sirt5* expression.

4. Material and methods

4.1. qRT-PCR analysis of *Sirt5* expression

cDNA used for mapping (see section 1.4.3) was used for determining *Sirt5* expression. The following primers were used: *Sirt5* F 5'-GCAGACGGGTTGTGGTCAT-3', *Sirt5* R 5'-CTGGGCAGATCGGACTCCTA-3', *B2m* F 5'-GGTCTTTCTGGTGTTGTCTCA-3', *B2m* R 5'-GTTCGGCTTCCCATTCTCC-3'. qPCR was performed for 40 cycles using a BioRad qPCR machine and the Applied Biosystems SYBR Green Master Mix (Thermo Fisher Scientific, Schwerte, Germany). Log fold change expression of *Sirt5* was calculated using the ΔΔCt method relative to *B2m* expression levels.

4.2. *Sirt5* genotyping in mice

Ear punches were collected from mice of the original partially inbred mouse lines that were used for the mapping cross and were extracted using DNeasy Blood & Tissue kit (Qiagen, Hilden, Germany). The SNP situated at position 15177 of the ENSEMBL reference *Sirt5* gene was genotyped using forward primer 5'-CTGGTTCCTGGCTTCGACAT-3' and reverse 5'-TCTGCAAGAGATGGCCACAG-3'.

4.3. Murine feces sampling, extraction, and 16S rRNA gene sequencing

Ten mice of each genotype were chosen (AA, AG, GG) resulting in 30 mice in total. Mice were kept at a 12h:12h light:dark schedule, with light from 7am to 7pm. Feces was collected twice from the individuals at 7am, 1pm, 7pm, and 1am over several consecutive days. Bead beating using Lysing Matrix E tubes (MP biomedical) was used prior to extraction to ensure cell lysis. Feces samples were extracted using the 96 well DNA/RNA AllPrep kit (Qiagen, Hilden, Germany). The V1-V2 region of the 16S rRNA gene was amplified according to the conditions described (Rausch et al., 2016) and was sequenced with 250 bp paired-reads on the Illumina MiSeq platform. Sequences were assigned to each sample by exact matches to multiplex identifier (MID) sequences and processed with the dada2 R package (v 1.16.0) (Callahan et al., 2016). In brief, raw sequences were trimmed and quality filtered with a maximum of two 'expected errors' allowed in a read. Next, the paired sequences were merged, and chimeras removed before assigning taxonomy using the Ribosomal Database Project (RDP) training set 16. Samples were rarefied to a sequencing depth of 10,000 reads for all downstream analyses. Classifications with low confidence at the genus level

(< 0.8) were grouped in the arbitrary taxon 'unclassified_group'. Alpha (Shannon)- and beta (Bray-Curtis) diversity were analyzed using the phyloseq R package (v 1.32.0) (McMurdie and Holmes, 2013). The Vegan package in R (v 2.5-7) was used for analysis of dissimilarity using a constrained analysis of principal coordinates ('capscale'), a hypothesis-driven ordination that restricts the separation of the communities on the variable tested (Jari et al., 2020) for which the 'anova.cca' function was used to determine significance. Differentially abundant taxa between groups were determined with the IndVal.g function of the multipatt command in the IndicSpecies R package (Miquel and Pierre, 2009) with 10,000 permutations. Only taxa present in 25% of the samples were used for the IndicSpecies analysis. *P* values were corrected for multiple testing using FDR correction.

4.4. *dSirt4* and *Sirt5* sequence alignment and protein modelling

Protein sequences were downloaded from UniProt under the accession numbers Q8IRR5 for *dSirt4* and Q8K2C6 for *Sirt5*. Sequences were aligned using Geneious alignment with default parameters. The protein models were loaded in SWISS-MODEL, aligned and superimposed. Sequence similarity is calculated from a normalised BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992).

4.5. Fly strains and husbandry

Fly stocks were raised on standard cornmeal-molasses medium at 65% humidity, 25°C and a 12h-12h light:dark cycle. Media for microbial depleted flies were kept as germ-free, while a bacterial mixture containing *Acetobacter Thailandi ICUS* (OD_{600} =0.8) , *Acetobacter pomorum* (OD_{600} = 0.7), *Lactobacillus brevis EW* (OD_{600} = 8), *Lactobacillus plantarum WJL* (OD_{600} = 6), *Enterococcus faecalis* (OD_{600} = 0.8), and *Commensalibacter intestini A911T* (OD_{600} = 1.5), was added to the media of all conventially raised flies. Germ-free egg shells were generated using dechorionation before transferring them to the media. The following fly strains were used in the experiments: w^{1118} (Bloomington *Drosophila* stock center, USA) as wild type flies (WT), BDSC_8840 from Bloomington Drosophila stock as *dSirt4* full knock-out flies, *esg-Gal4* males were crossed to virgin female UAS-sirt4 (Jason Wood, private supplier) or *RNAi_dSirt4:UAS* (Bloomington Stock Center, USA) to generate *dSirt4* over-expression (OE) or knock-down (KD) in intestinal stem cells and enteroblasts (OE), respectively.

4.6. Fly tissue collection and extraction

Female flies were collected at ~12 days. Whole guts were collected every three hours over a period of 48 hours. Extra care was taken during night sampling points to sacrifice flies under red light before exposing them to light during dissections. Five independent biological replicates were samples per time point per condition and each replicate consisted of 6 individual fly guts. The samples were kept in RNAmagic (BioBudget, Krefeld, Germany) on ice before homogenization with a Bead Ruptor 24 (BioLab products, Bebensee, Germany) and stored at -80°C until extraction. RNA was isolated using a column-based phenol-chloroform phase segregation kit (Invitrogen™ PureLink™ RNA mini kit, Thermo Fisher Scientific, Schwerte, Germany) and purified using isopropanol precipitation.

4.7. qRT-PCR expression analysis of *dSirt4*, *Period*, and *Timeless* and bacterial load measurement

Gene expression of *dSirt4, Period,* and *Timeless* genes was measured using quantitative reverse transcriptase PCR. cDNA was generated using the SuperScript IV Reverse Transcriptase kit (Thermo Fisher Scientific, Schwerte, Germany). Random hexamer cDNA synthesis was performed using the High Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific, Schwerte, Germany) to measure bacterial load and for downstream 16S rRNA gene sequencing. qPCRBio Sygreen Mix Hi Rox (PCRBiosystems, London, UK) was used as master mix combined with the following primers:

Gene	**Forward primer**	**Reverse primer**
Sirt4	5’-CCGAAATGTTGTGGAGGTTC-3’	5’- ATTTAGCGACGCCAGTATGC-3’
Period	5’- TACCCGCATCCTTCGCTTTT-3’	5’- TTGTTGTACGCGGATTGGGA-3’
Timeless	5’- CCTCTGGTTCGAAGCCTCTC-3’	5’- CATTGCTGCCATTGTCCGAG -3’
Rpl32	5’- AAGCCGTAATGTCGTTTTTG -3’	5’- TGGGCAGTATCCATTGAGTT -3’
Bacterial load	5′-TTACCGCGGCKGCTGGCAC-3′	5′- AGAGTTTGATCMTGGCTCAG-3′

qPCR was performed for 40 cycles using a Thermo Fisher StepOne™ qPCR machine. Log fold change of expression was calculated relative to housekeeping gene *Rpl32* using ΔΔCt method.

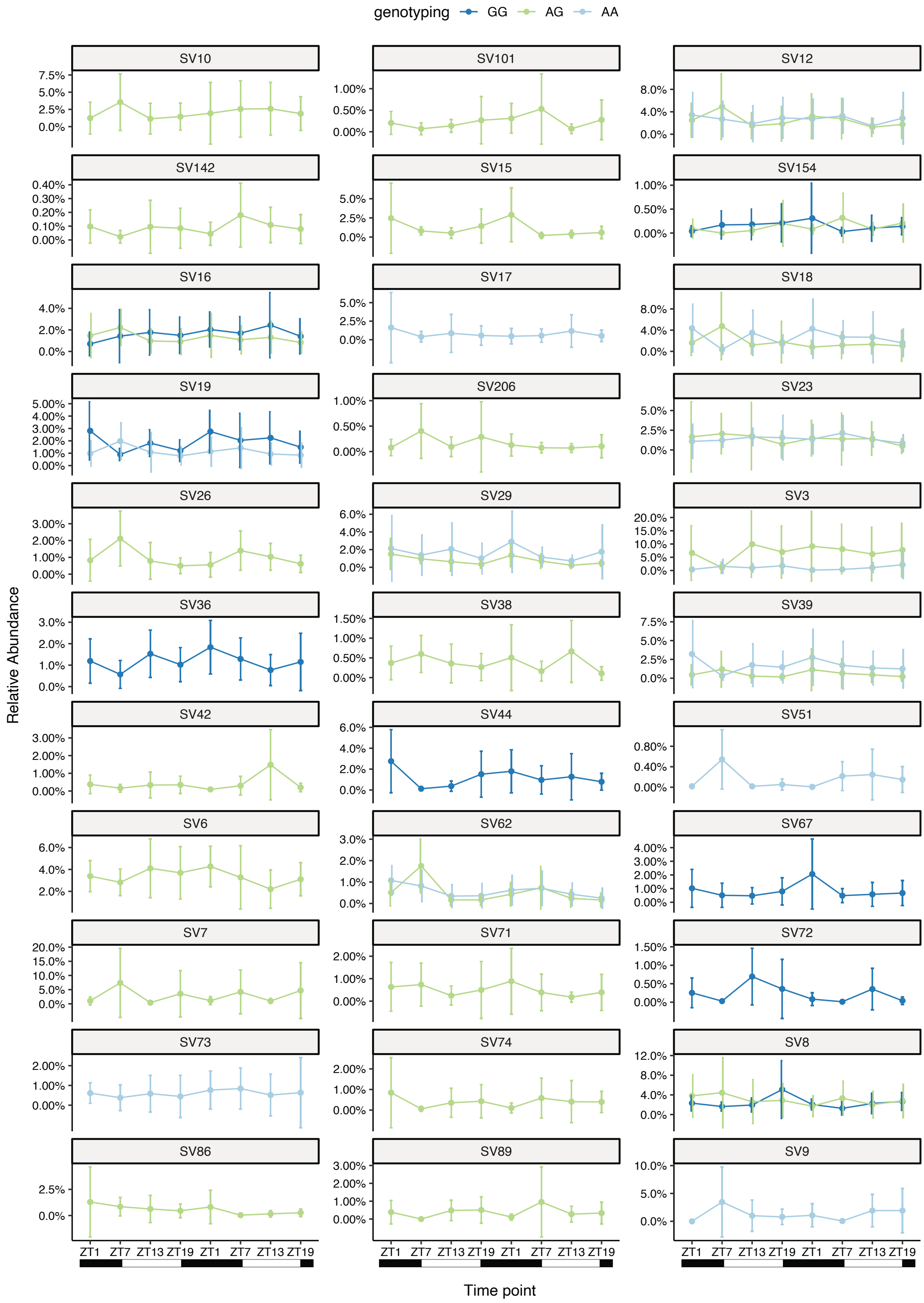
4.8. 16S rRNA gene sequencing and processing

Random hexamer cDNA was used for 16S rRNA amplicon sequencing. Libraries and sequencing was performed as described above, as well as processing the reads using dada2 (Callahan, 2016; Callahan et al., 2016; Methods 2.4.3). Before analysis, reads originating from contamination were filtered out using *decontam* R package (Nicole et al., 2017) with both the frequency and the prevalence methods. Next, taxa were filtered on abundance to protect against ASVs with a small mean and a trivially large coefficient of variation. Taxa with a minimum count of three in at least 20% of the samples were kept. Reads were rarefied at 6300 reads. This removed two samples (oeconzt9r7, wtconzt21r7). All statistical analyses were done in R v 4.0.2. JTK_CYCLE analysis was performed using the R package Discorhytm (Matthew et al., 2020).

5. Supplementary material

Suppl. Table 1: ASVs within the different genotypes showing significant cycling using JTK_CYCLE

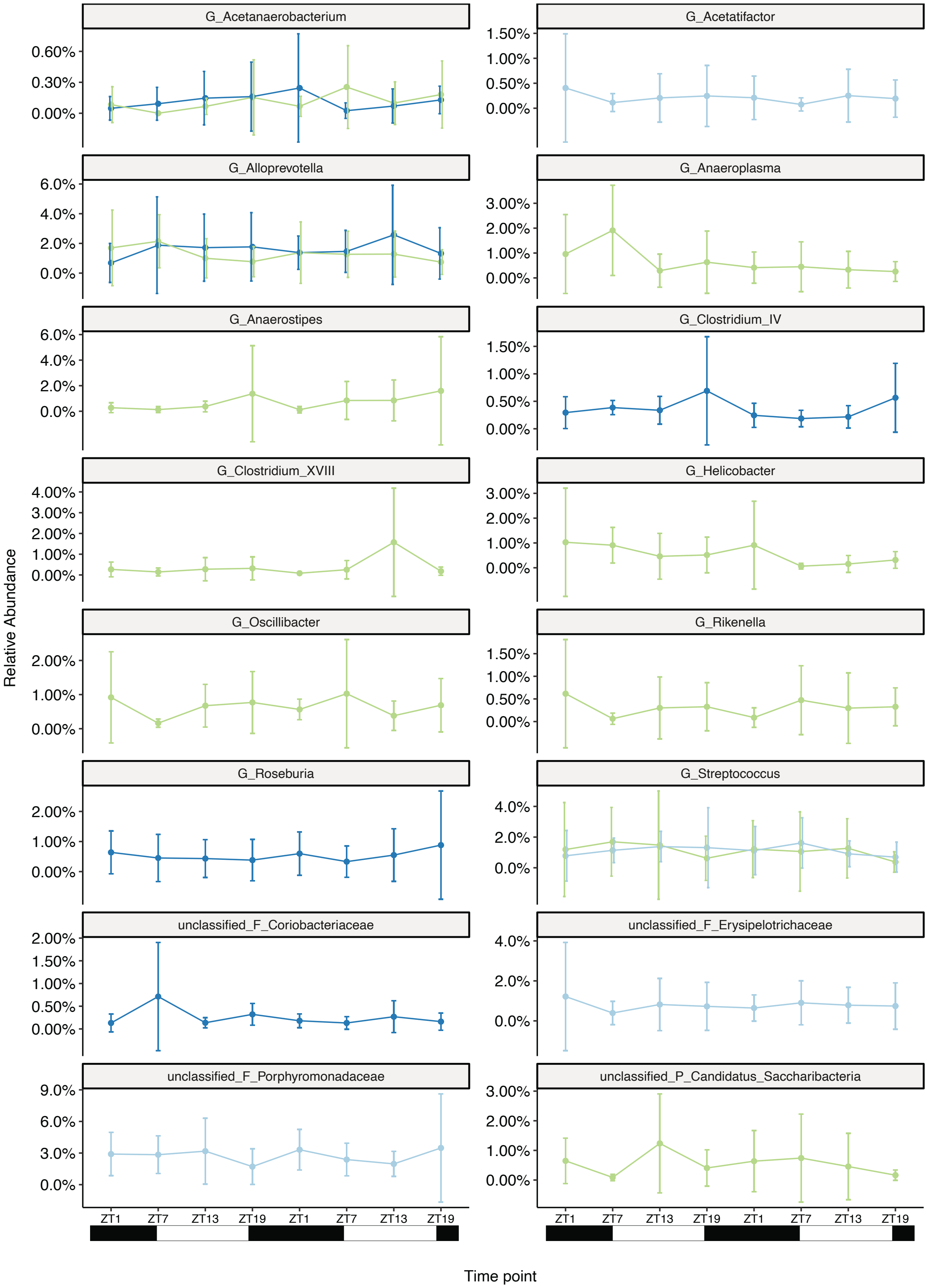
ASV	Genus	Species	genotype	acrophase	amplitude	pvalue	qvalue	period
SV3	G_Lactobacillus	NA	AG	4	0.0350254	0.0172486	0.0454857	24
SV3	G_Lactobacillus	NA	AA	19	0.0024571	0.0303238	0.0454857	24
SV6	G_Lactobacillus	NA	AG	22	0.0084696	0.0027259	0.0081778	24
SV7	G_Bacteroides	acidifaciens	AG	16	0.0054636	0.0037480	0.0112440	24
SV8	G_Alistipes	NA	GG	19	0.0058631	0.0047182	0.0141545	24
SV8	G_Alistipes	NA	AG	4	0.0116660	0.0128296	0.0192444	24
SV9	G_Bacteroides	NA	AA	16	0.0005940	0.0055586	0.0166758	24
SV10	unclassified_O_Bacteroidales	NA	AG	7	0.0146603	0.0053222	0.0159667	24
SV12	G_Bacteroides	NA	AG	16	0.0151648	0.0087537	0.0131306	24
SV12	G_Bacteroides	NA	AA	4	0.0223571	0.0000000	0.0000001	24
SV15	unclassified_F_Lachnospiraceae	NA	AG	16	0.0072833	0.0000210	0.0000629	24
SV16	G_Alloprevotella	NA	GG	4	0.0119247	0.0000175	0.0000525	24
SV16	G_Alloprevotella	NA	AG	16	0.0090837	0.0087537	0.0131306	24
SV17	unclassified_F_Erysipelotrichaceae	NA	AA	4	0.0018495	0.0021792	0.0065377	24
SV18	unclassified_O_Bacteroidales	NA	AG	16	0.0071301	0.0002096	0.0006287	24
SV18	unclassified_O_Bacteroidales	NA	AA	10	0.0210352	0.0063802	0.0095702	24
SV19	unclassified_O_Bacteroidales	NA	GG	7	0.0096190	0.0000082	0.0000245	24
SV19	unclassified_O_Bacteroidales	NA	AA	19	0.0059708	0.0048349	0.0072524	24
SV23	G_Streptococcus	NA	AG	13	0.0050906	0.0084190	0.0126285	24
SV23	G_Streptococcus	NA	AA	4	0.0041365	0.0007638	0.0022914	24
SV26	G_Paraprevotella	NA	AG	16	0.0036482	0.0026111	0.0078332	24
SV29	unclassified_F_Porphyromonadaceae	NA	AG	13	0.0039812	0.0138262	0.0207392	24
SV29	unclassified_F_Porphyromonadaceae	NA	AA	13	0.0079393	0.0012291	0.0036872	24
SV36	unclassified_O_Bacteroidales	NA	GG	7	0.0072194	0.0000829	0.0002487	24
SV38	G_Clostridium_XlVa	NA	AG	16	0.0016399	0.0029697	0.0089091	24
SV39	G_Bacteroides	NA	AG	NaN	0.0000000	0.0054325	0.0092475	0
SV39	G_Bacteroides	NA	AA	16	0.0093564	0.0061650	0.0092475	24
SV42	G_Clostridium_XVIII	cocleatum	AG	7	0.0018033	0.0063832	0.0191497	24
SV44	G_Eisenbergiella	NA	GG	10	0.0059382	0.0005913	0.0017740	24
SV51	G_Alistipes	NA	AA	22	0.0003981	0.0047501	0.0142503	24
SV62	G_Alistipes	NA	AG	13	0.0019029	0.0202929	0.0304394	24
SV62	G_Alistipes	NA	AA	7	0.0032398	0.0002365	0.0007096	24
SV67	unclassified_F_Lachnospiraceae	NA	GG	10	0.0030186	0.0019887	0.0059660	24
SV71	unclassified_F_Lachnospiraceae	NA	AG	10	0.0023585	0.0163257	0.0489772	24
SV72	G_Odoribacter	NA	GG	19	0.0007730	0.0041179	0.0123537	24
SV73	unclassified_O_Bacteroidales	NA	AA	1	0.0048525	0.0001361	0.0004084	24
SV74	G_Rikenella	NA	AG	1	0.0013248	0.0123560	0.0370681	24
SV86	G_Helicobacter	apodemus	AG	16	0.0020248	0.0000599	0.0001798	24
SV89	G_Oscillibacter	NA	AG	1	0.0025291	0.0002322	0.0006965	24
SV101	unclassified_F_Ruminococcaceae	NA	AG	19	0.0006288	0.0108186	0.0324559	24
SV142	G_Odoribacter	NA	AG	4	0.0006819	0.0026111	0.0078332	24
SV154	G_Acetanaerobacterium	NA	GG	1	0.0002318	0.0017608	0.0026411	24
SV154	G_Acetanaerobacterium	NA	AG	1	0.0007714	0.0008483	0.0025449	24
SV206	unclassified_F_Erysipelotrichaceae	NA	AG	13	0.0008038	0.0008889	0.0026667	24



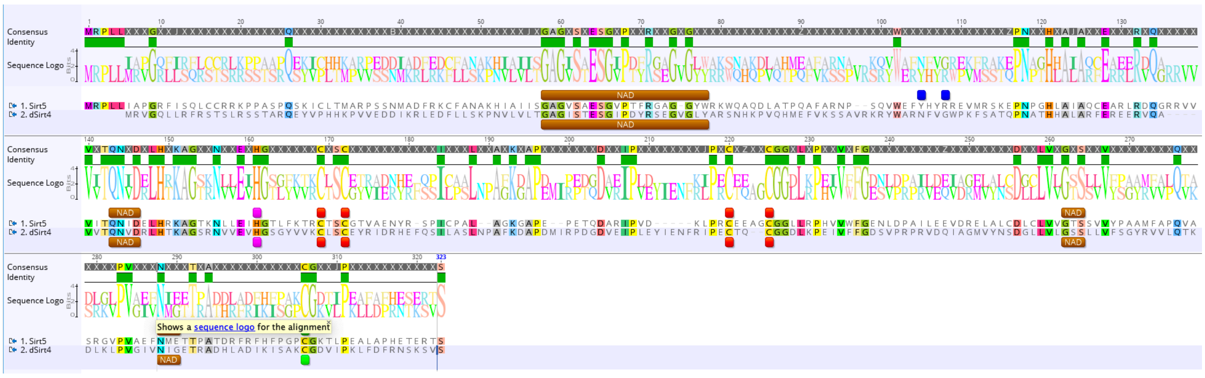
Suppl. Figure 1: ASVs occurring in the different genotypes showing significant cycling.

Suppl. Table 2: Genera showing significant cycling using JTK_CYCLE.

Genus	genotype	acrophase	amplitude	pvalue	qvalue	period
G_Alloprevotella	GG	4	0.0102556	0.0000144	0.0000432	24
G_Alloprevotella	AG	16	0.0081582	0.0094597	0.0141895	24
unclassified_F_Erysipelotrichaceae	AA	1	0.0017589	0.0152640	0.0457921	24
G_Streptococcus	AG	16	0.0042520	0.0014722	0.0044167	24
G_Streptococcus	AA	4	0.0025196	0.0041985	0.0062977	24
unclassified_F_Porphyromonadaceae	AA	13	0.0131229	0.0003632	0.0010895	24
G_Anaeroplasma	AG	16	0.0033604	0.0001089	0.0003267	24
G_Clostridium_XVIII	AG	7	0.0014209	0.0045180	0.0135539	24
G_Anaerostipes	AG	7	0.0007103	0.0125888	0.0377664	24
unclassified_P_Candidatus_Saccharibacteria	AG	22	0.0017807	0.0011206	0.0033617	24
G_Rikenella	AG	1	0.0010655	0.0094597	0.0283790	24
G_Helicobacter	AG	16	0.0023068	0.0000097	0.0000290	24
G_Oscillibacter	AG	1	0.0032189	0.0022927	0.0068782	24
G_Acetatifactor	AA	4	0.0003570	0.0087927	0.0263782	24
unclassified_F_Coriobacteriaceae	GG	1	0.0010586	0.0019887	0.0059660	24
G_Roseburia	GG	1	0.0017691	0.0043097	0.0129292	24
G_Acetanaerobacterium	GG	1	0.0003541	0.0040259	0.0060389	24
G_Acetanaerobacterium	AG	1	0.0007082	0.0004445	0.0013335	24
G_Clostridium_IV	GG	22	0.0010611	0.0159281	0.0477842	24

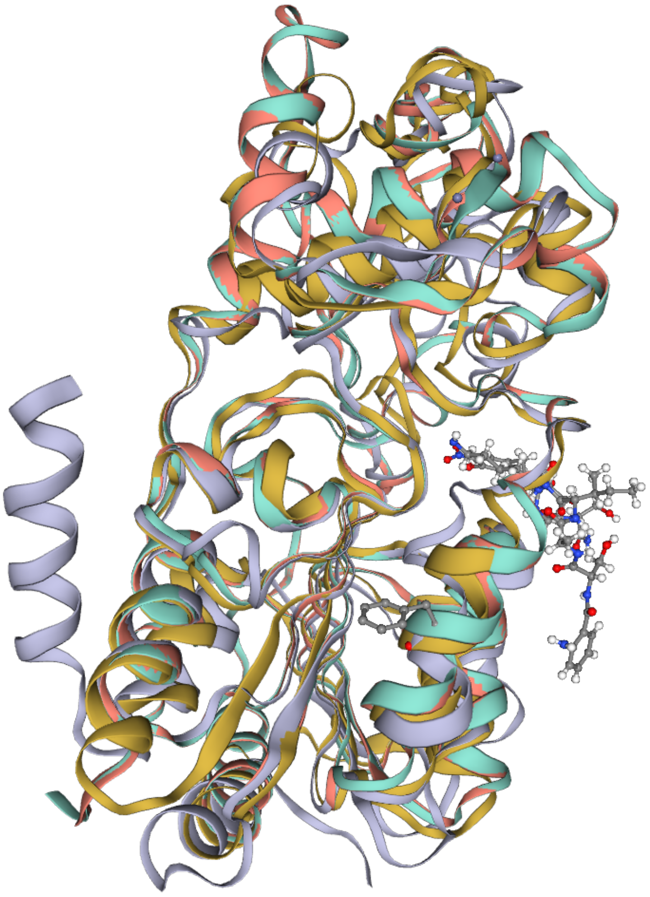


Suppl. Figure 2: Genera within the genotypes showing significant cycling using JTK_CYCLE

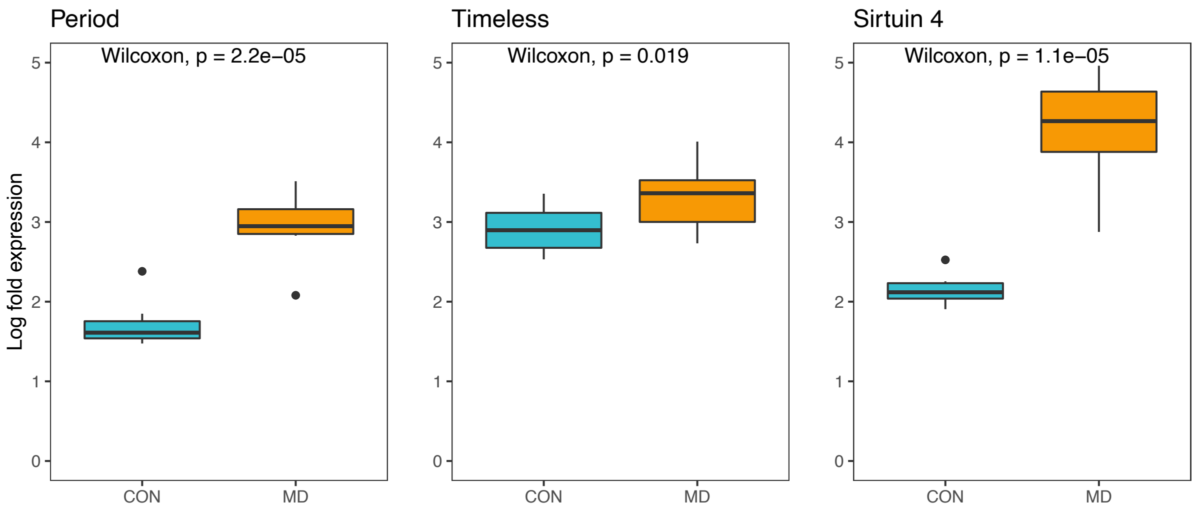


```
Consensus   MRPLLXXXGXXJXXXXXXXXXXXXXQXXXXXXXXXXXXBXXXXXXXXXXXXXXXXJXGAGXSXESGXPXXRXXGXGXXXXXXXXXXXXXZXXXXXXXXXXXWXXXXXXXXXXXXXZPNXXHXAJAXXEXXXRXQXXXXX
1. Sirt5    MRPLLIAPGRFISQLCCRRKPPASPQSKICLTMARPSSNMADFRKCFANAKHIAIISGAGVSAESGVPTFRGAG-GYWRKWQAQDLATPQAFARNP--SQVWEFYHYRREVMRSKEPNPGHLAIAQCEARLRDQGRRVV
2. dSirt4        MRVGQLLRFRSTSLRSSTARQEYVPHHKPVVEDDIKRLEDFLLSKPNVLVLTGAGISTESGIPDYRSEGVGLYARSNHKPVQHMEFVKSSAVRKRYWARNFVGWPKFSATQPNATHHALARFEREERVQA----

Consensus   VXTQNXDXLHXKAGXXNXXEXHGXXXXXXCXSCXXXXXXXXXXXIXXXLXXAXKXAPXXXXXXXDXXIPXXXXXXXXJPXCXZXXCGGXLXPXXVXFGXXXXXXXXXZXXXXXXXXDXXLVXGXSXXVXXXXXXXXQXX
1. Sirt5    VITQNIDELHRKAGTKNLLEIHGTLFKTRCTSCGTVAENYR-SPICPAL--AGKGAPE--PETQDARIPVD-----KLPRCEEAGCGGLLRPHVVWFGENLDPAILEEVDRELALCDLCLVVGTSSVVYPAAMFAPQVA
2. dSirt4   VVTQNVDRLHTKAGSRNVVEVHGSGYVVKCLSCEYRIDRHEFQSILASLNPAFKDAPDMIRPDGDVEIPLEYIENFRIPECTQ--CGGDLKPEIVFFGDSVPRPRVDQIAGMVYNSDGLLVLGSSLLVFSGYRVVLQTK

Consensus   XXXXPVXXXNXXXTXAXXXXXXXXXXXCGXXJPXXXXXXXXXXXS
1. Sirt5    SRGVPVAEFNMETTPATDRFRFHFPGPCGKTLPEALAPHETERTS
2. dSirt4   DLKLPVGIVNIGETRADHLADIKISAKCGDVIPKLFDFRNSKSVS
```

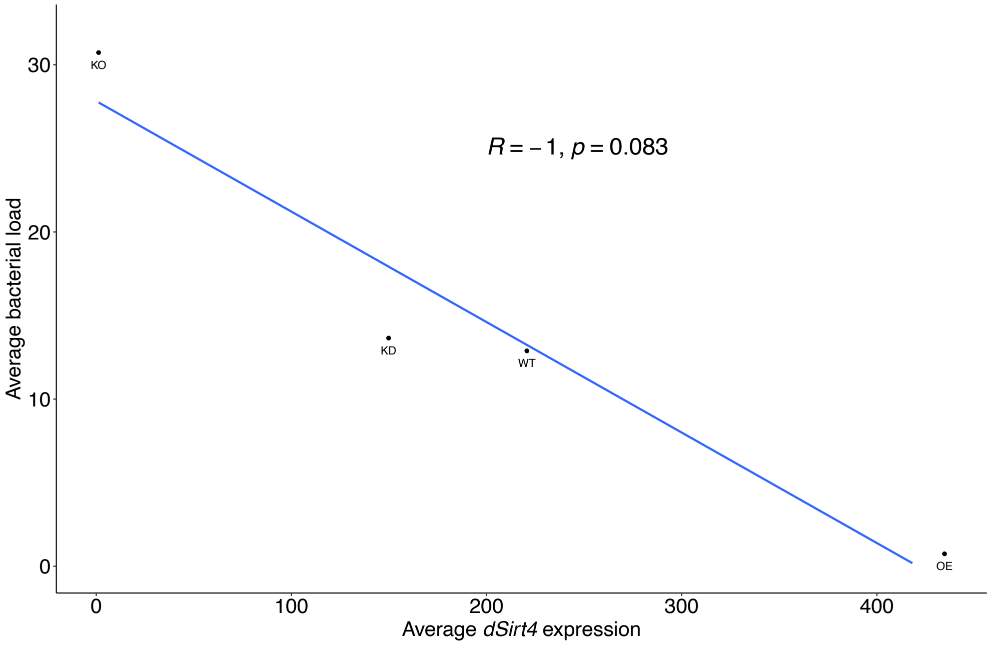
Suppl. Figure 3: Protein sequence alignment of mouse *Sirt5* (top) to *dSirt4* (bottom sequence). Active sites and binding sites/regions are annotated as colored blocks below the sequences: proton acceptor sites (pink), NAD+ binding regions (brown), substrate binding sites (blue), zinc binding sites (red) and binding site for NAD+ via amine nitrogen (green).



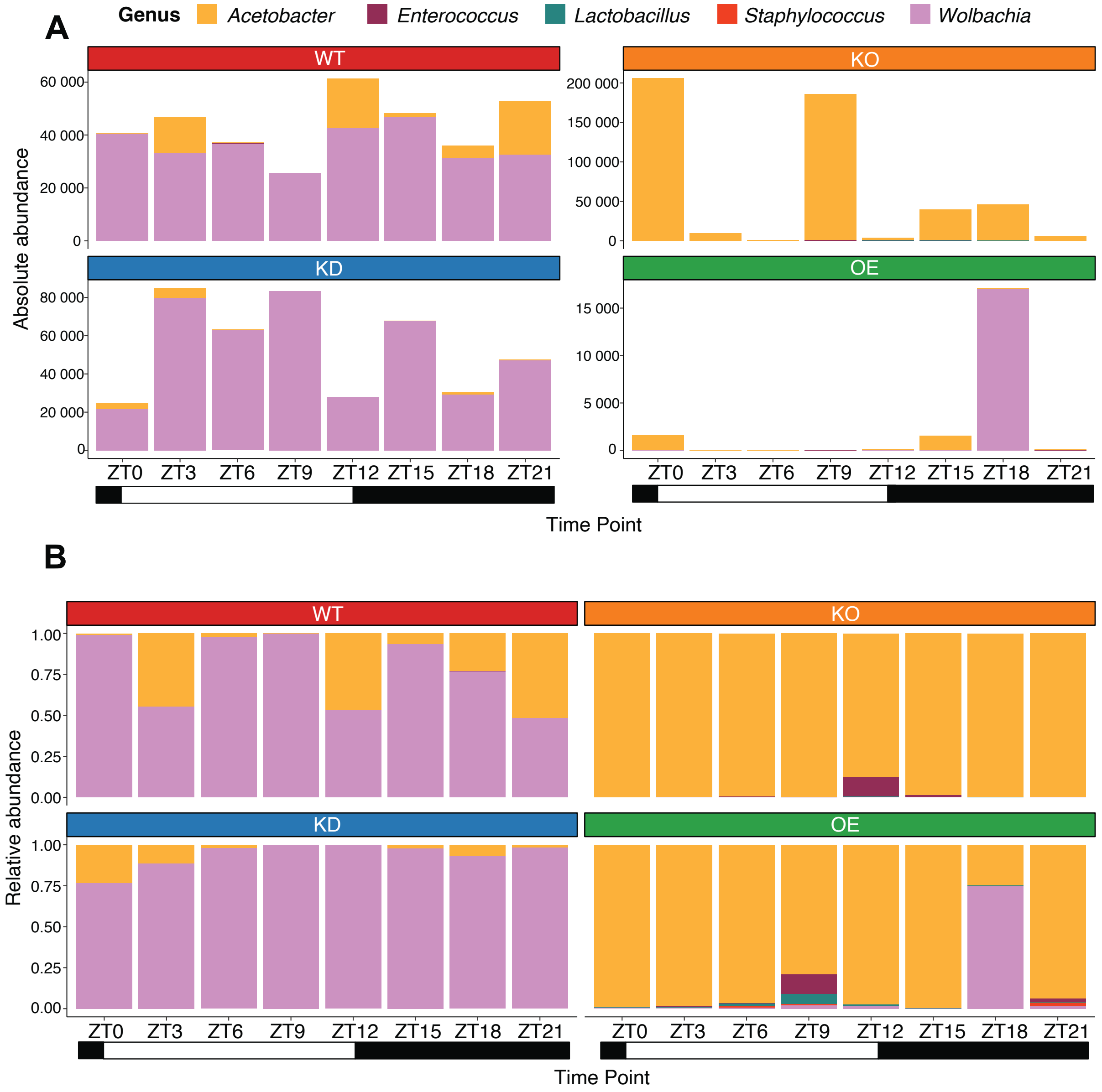
Suppl. Figure 4: Predicted 3D structure of Drosophila SIRT4 (turquoise) and murine mitochondrial sirtuins, SIRT3 (lavender), SIRT4 (salmon pink), and SIRT5 (ochre).



Suppl. Figure 5: Average gene expression per replicate for Period, Timeless, and Sirtuin 4 in microbial depleted (MD) and conventionally raised (CON) flies.



Suppl. Figure 6: The average bacterial load in function of the average dSirt4 expression.



Suppl. Figure 7: Absolute (A) and relative (B) abundances of genera over time in *Drosophila* differentially expressing *dSirt4*: wild type (WT, red), knock-down (KD, dark blue), knock-out (KO, orange), and over-expression (OE, green). Absolute abundances are calculated by multiplying the bacterial load (from qPCR) with the relative abundance in percentage (from 16S).

Chapter 3: Comparison of microbial DNA enrichment kits for shotgun metagenome sequencing

1. Introduction

Advancements in sequencing technologies have fundamentally altered how the scientific field designs and conceptualizes microbial ecology research in order to understand community complexity. The affordability of shotgun sequencing on a large-scale, coupled with advances in read length and throughput, has made it possible to apply on metagenome samples. Shotgun metagenome sequencing can offer species- and strain-level classification of bacteria (Li et al., 2020), with a relative lack of bias and allows the examination of the functional content of the microbiota (Heintz-Buschart and Wilmes, 2018). Furthermore, not yet classified bacteria can also be discovered through *de novo* genome binning if the sequencing coverage is high enough. As mucosal populations are more dependent on host genetics than luminal-associated bacteria (Linnenbrink et al., 2013; Spor et al., 2011), using shotgun metagenomic sequencing data as input for a genome-wide association study would allow us to not only map the gut community structure, but also their functional capabilities. However, a major hurdle remains. Without the ability to specifically target microbial DNA, much of the DNA sequenced will belong to the host rather than the microbiota. Several bioinformatic tools to remove host DNA exist (Bush et al., 2020), but this requires a greater sequencing depth to obtain enough microbial reads which quickly increases the sequencing cost.

To address this problem, several commercial kits have arrived on the market to enrich for microbial DNA in tissue and blood samples. One such kit, the Ultra-Deep Microbiome Prep kit (Molzym GmbH, Germany) selectively breaks down eukaryotic cells first using chaotropic reagents and degrade their DNA with DNases before lysing the bacterial cells to extract the DNA. This approach will also remove 'dead' microbial DNA. Two other methods, LOOXSTER® Enrichment kit (Analytik Jena, AG, Germany) and NEBNext Microbiome DNA Enrichment kit (New England Biolabs GmbH, Germany), take advantage of the difference in CpG methylation rates between microbial and host DNA. The LOOXSTER Enrichment kit uses the specific affinity of the CXXC-domain of the LOOXSTER®-protein, which is based on the human CXXC zinc finger protein 1 (CFP1), to form a stable complex with nonmethylated CpG-dinucleotides (Xu et al., 2011). Unbound (eukaryotic) DNA can be removed through a stringent washing step and the bound (microbial) DNA is released from LOOXSTER®-paramagnetic particles. The NEBNext kit uses the MBD2-Fc protein that binds to protein A on magnetic beads through the Fc domain and binds specifically and tightly to CpG methylated DNA (Feehery et al., 2013). Application of a magnetic field then pulls out the CpG-methylated (eukaryotic) DNA, leaving the non-CpG-methylated (microbial) DNA in the supernatant. These methods have been been tested with varying outcomes (Hansen et al., 2009; Feehery et al., 2013; Thoendel et al., 2016; Marotz et al., 2018; Yap et al., 2020; Heravi et al., 2020;), however no direct comparisons between these methods have been published using shotgun metagenome sequencing. Furthermore, these methods have not been used on host tissue samples.

In this study, we compared three different commercial kits designed to enhance the proportion of microbial DNA to host DNA. We processed eight cecum tissue samples to be processed in parallel with the three kits and then sequenced on a n Illumina NextSeq 500 run. One untreated (raw) aliquot of each sample was also sequenced at greater sequencing depth to act as a comparison control and five luminal samples and a mock community DNA standard served as additional controls.

2. Results

2.1. Sequencing and enrichment efficiency

Extracted DNA from eight murine cecum tissue samples was sequenced on one NextSeq run, while the DNA treated with the enrichment kits was sequenced on another NextSeq run (Fig. 28). As an additional controls, we included five luminal samples: four originating from the same mice as the cecal mucosa, plus one mock community standard containing DNA from eight bacterial species and two yeast species (ZYMOBIOMICS® microbial community DNA standard). After removal of the host reads, sequencing data was processed with the SqueezeMeta pipeline for functional and taxonomical classification (see Methods 3.4.3).

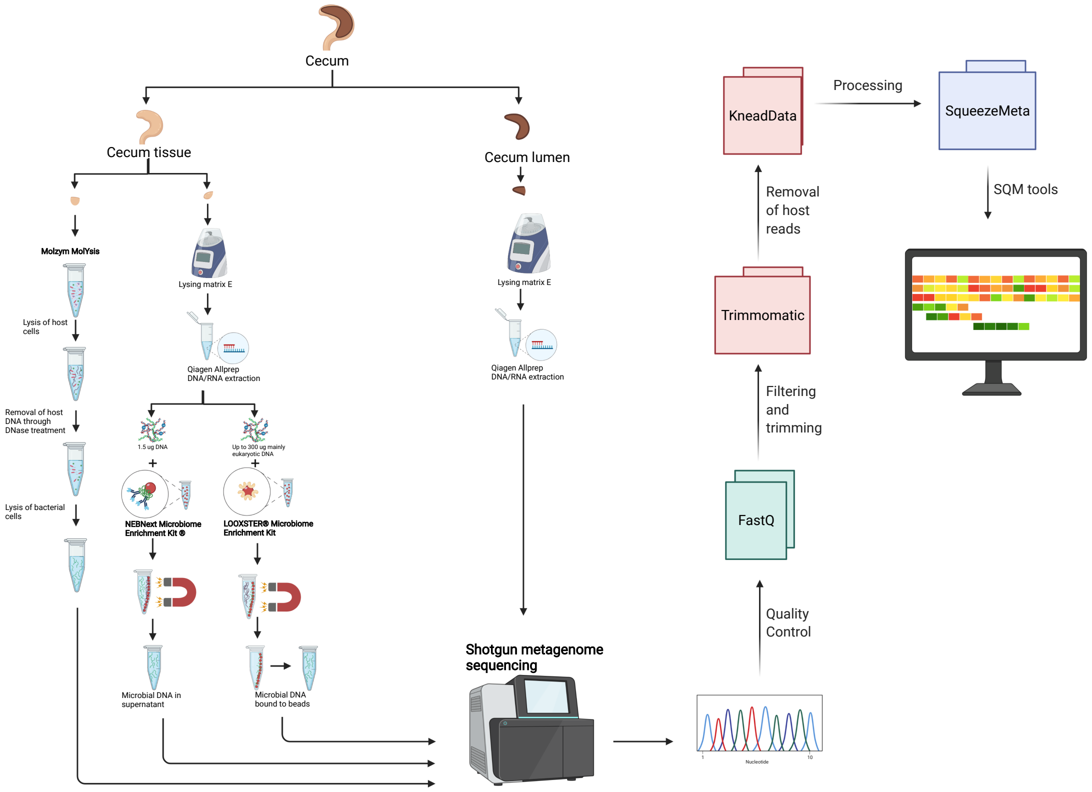

Cecum
Cecum tissue
Cecum lumen
Molzym MolYsis
Lysis of host cells
Removal of host DNA through DNase treatment
Lysis of bacterial cells
Lysing matrix E
Qiagen Allprep DNA/RNA extraction
1.5 ug DNA
Up to 300 ug mainly eukaryotic DNA
NEBNext Microbiome Enrichment Kit ®
LOOXSTER® Microbiome Enrichment Kit
Microbial DNA in supernatant
Microbial DNA bound to beads
Lysing matrix E
Qiagen Allprep DNA/RNA extraction
Shotgun metagenome sequencing
Quality Control
FastQ
Filtering and trimming
Trimmomatic
Removal of host reads
KneadData
Processing
SqueezeMeta
SQM tools

Figure 28: Workflow for comparing microbial DNA enrichment kits to full metagenome shotgun sequencing and 16S rRNA gene amplicon sequencing. Created with BioRender.com.

With the exception of *Bacillus subtilis,* all ten microorganisms of the mock community have been classified to species level; *B. subtilis* is classified to the genus level (unclassified *Bacillus*; Suppl. Fig. 1). The average number of raw reads sequenced ranged from 57.5 million reads (untreated DNA) to 2.26 million reads (luminal samples; Fig. 29A), while the number of bacterial reads left after filtering ranged from 0.8 million (untreated DNA) to 4.5 million (mock; Fig. 29B). The percentages of microbial DNA present in the raw reads were highly variable: untreated DNA (1.50% ± 0.32%), NEBNext enrichment (3.55% ± 1.12%), LOOXSTER enrichment (5.05% ± 2.03%), Molzym (8.96% ± 9.04%), lumen (96.46% ± 4.11%), and mock (99.11% ± 0.44%). There was no significant difference in microbial enrichment efficiency between the different kits (*P* > .05, Wilcoxon rank sum test; Fig. 29D). To obtain a 5X coverage of the fourth most abundant species, we estimated we would need 1.96Gb of sequencing data. We only achieved this coverage for two samples that were processed with the Molzym enrichment kit (Fig. 29C).

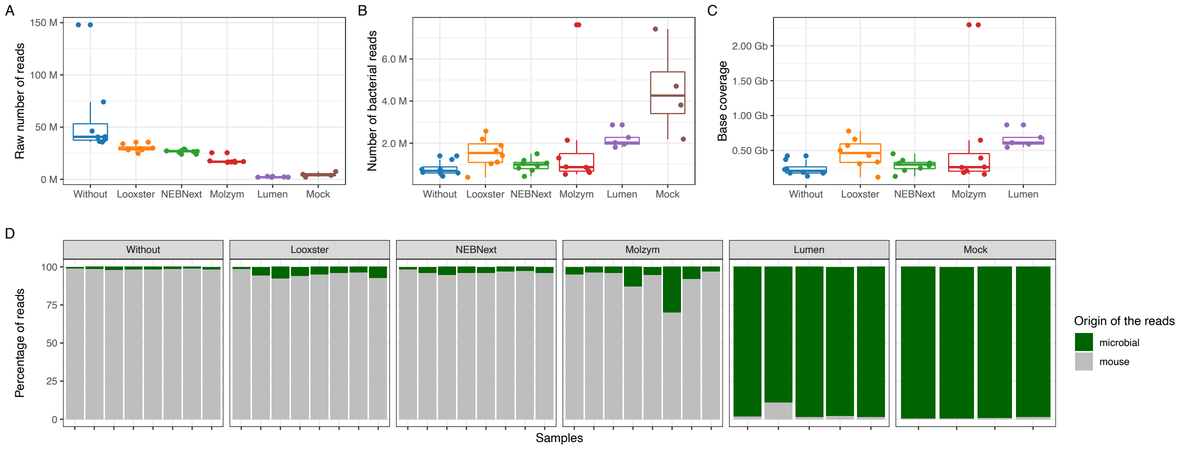
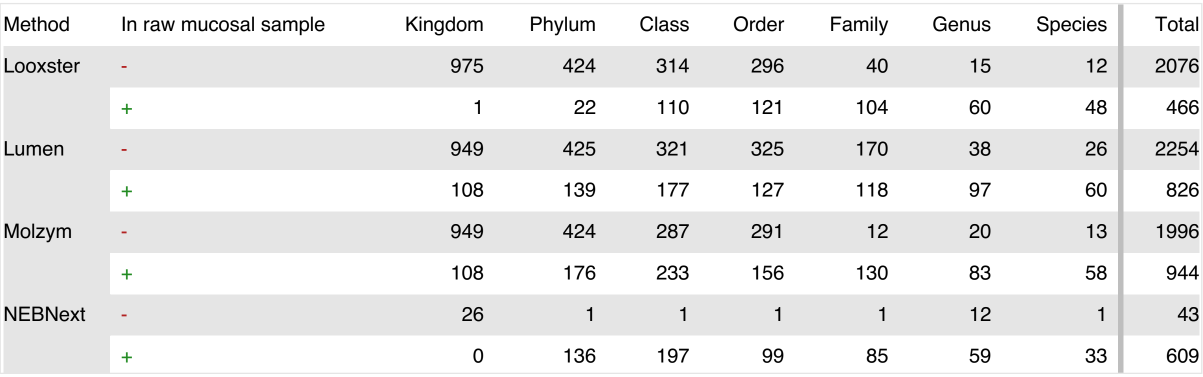


Figure 29: Efficiency of methods for improving the percentage of microbial reads. (A) Raw number of reads sequenced. (B) Number of bacterial reads. (C) Base coverage of microbial reads. (D) Percentage of mouse versus microbial reads. .

2.2. Influence of microbial-enrichment methods on taxonomic composition

Each sample exhibited a specific pattern of microbial genera that changed upon host depletion (Fig. 30). The untreated mucosal samples contain the most unclassified and unmapped reads. LOOXSTER enriched samples have a significantly higher species richness and evenness compared to all other methods (Fig. 31A). In a principal coordinate analysis (PCoA) samples cluster by method and by individual (Fig. 31B-C). The pairwise Bray-Curtis (BC) dissimilarity metric is smaller between individuals within methods compared to within individuals between enrichment methods (*P* < .001, Fig. 31D). This suggests that the enrichment method had a larger influence on the microbial composition than the inter-individual variability (adonis of BC dissimilarity $R^2{}_{\text{method}}$= 0.566, $R^2_{\text{individual}}$= 0.333, F_{method}=33.54, $F_{\text{individual}}$=9.874, *P* < .001 for all). We then compared the BC dissimilarities of each of the microbial-enriched samples to the corresponding raw samples (mucosa and lumen, Fig. 31E and F, resp.). The microbiota composition of the samples enriched with the NEBNext kit (0.11 ± 0.022) were the most similar to the mucosal raw samples compared to LOOXSTER (0.29 ± 0.042), Molzym (0.31 ± 0.122) and luminal raw samples (0.47 ± 0.09) (Fig. 31E). Interestingly, mucosal samples enriched with Molzym (0.20 ± 0.064) and LOOXSTER (0.21 ± 0.103) show more similarity to the luminal sample than the mucosal sample (Fig. 31F). To determine if this effect is independent of sequencing depth, we subsampled all datasets to 100,000 reads and found similar results (adonis of BC dissimilarity R^2_{method}= 0.564, $R^2_{individual}$= 0.332, F_{method}=32.86, $F_{individual}$=9.67, P < .001 for all, Suppl. Fig. 2). Similar patterns were also found using the functional content instead of the taxonomic composition (Suppl. Fig. 3, COG; Suppl. Fig. 4, KEGG). Finally, we investigated which taxa were most affected by microbial enrichment methods using the Wilcoxon rank sum test on the differences in median abundances of the samples (Fig. 32). Molzym enriched samples miss the most taxa (944 taxa) compared to raw samples and LOOXSTER samples the least (466 taxa) (Table 4).

Table 4: Amount of significantly differentially abundant taxa compared to raw mucosa sample. '+' in 'In raw mucosal sample' means taxa are more abundant in the untreated sample compared to the microbial enriched sample, '-' means taxa are less abundant in the untreated sample.



Method	In raw mucosal sample	Kingdom	Phylum	Class	Order	Family	Genus	Species	Total
Looxster	-	975	424	314	296	40	15	12	2076
	+	1	22	110	121	104	60	48	466
Lumen	-	949	425	321	325	170	38	26	2254
	+	108	139	177	127	118	97	60	826
Molzym	-	949	424	287	291	12	20	13	1996
	+	108	176	233	156	130	83	58	944
NEBNext	-	26	1	1	1	1	12	1	43
	+	0	136	197	99	85	59	33	609

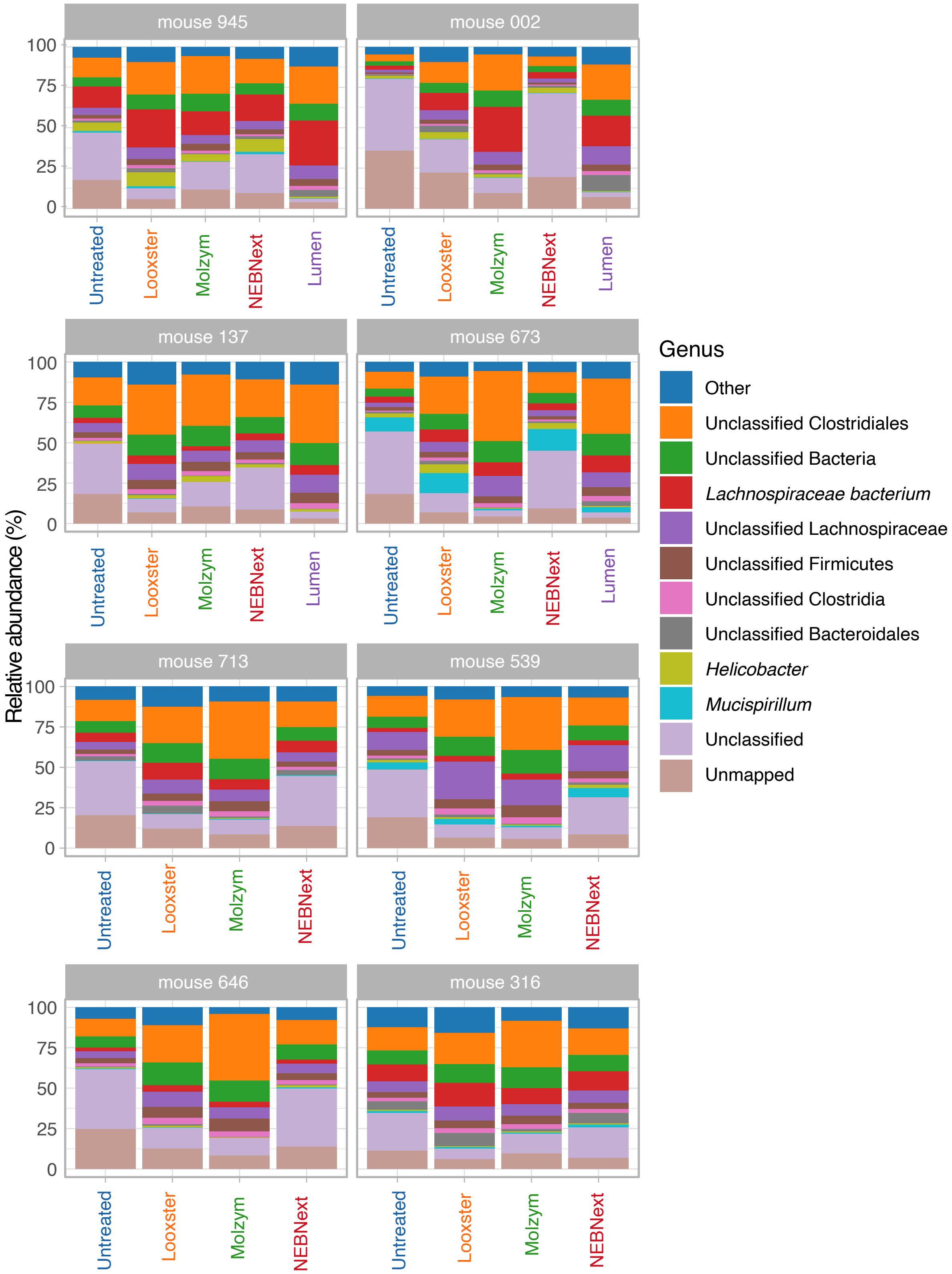


Figure 30: Taxonomic composition of each sample ordered by individual at the genus level.

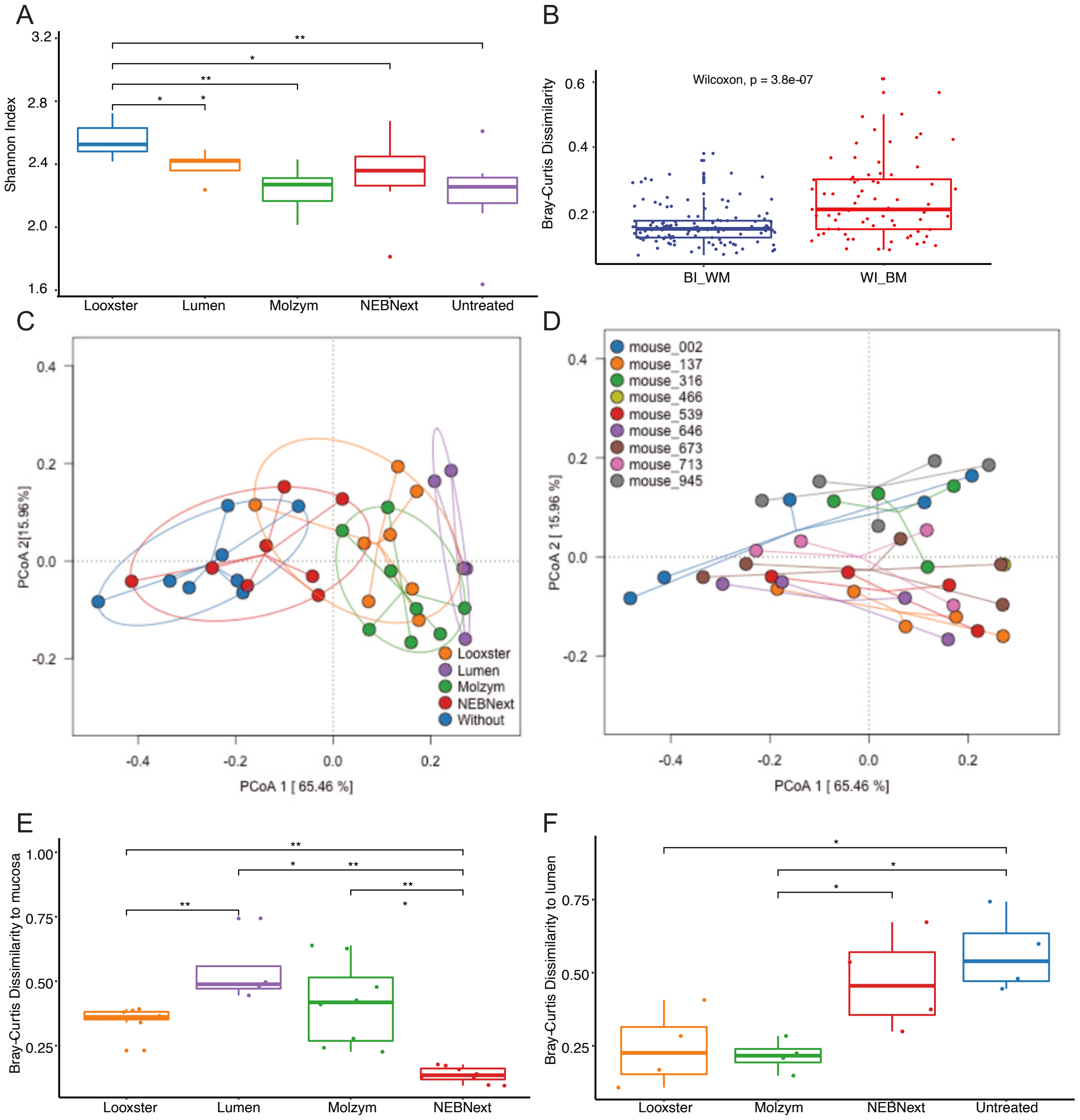


Figure 31: Alpha and beta diversity analyses. (A) Species richness and evenness in the samples processed with different enrichments methods. (B) Pairwise Bray-Curtis Dissimilarity within individual between methods (WI_BM) and between individuals within methods (BI_WM). (C-D) PCoA of samples using Bray-Curtis Dissimilarities colored by method (C) or by individual (D). (E-F) Pairwise Bray-Curtis dissimilarity to untreated mucosal sample (E) or to untreated luminal sample (F). Only significant comparisons are shown. * *P* < .05; ** *P* < .01; *** *P* < .001

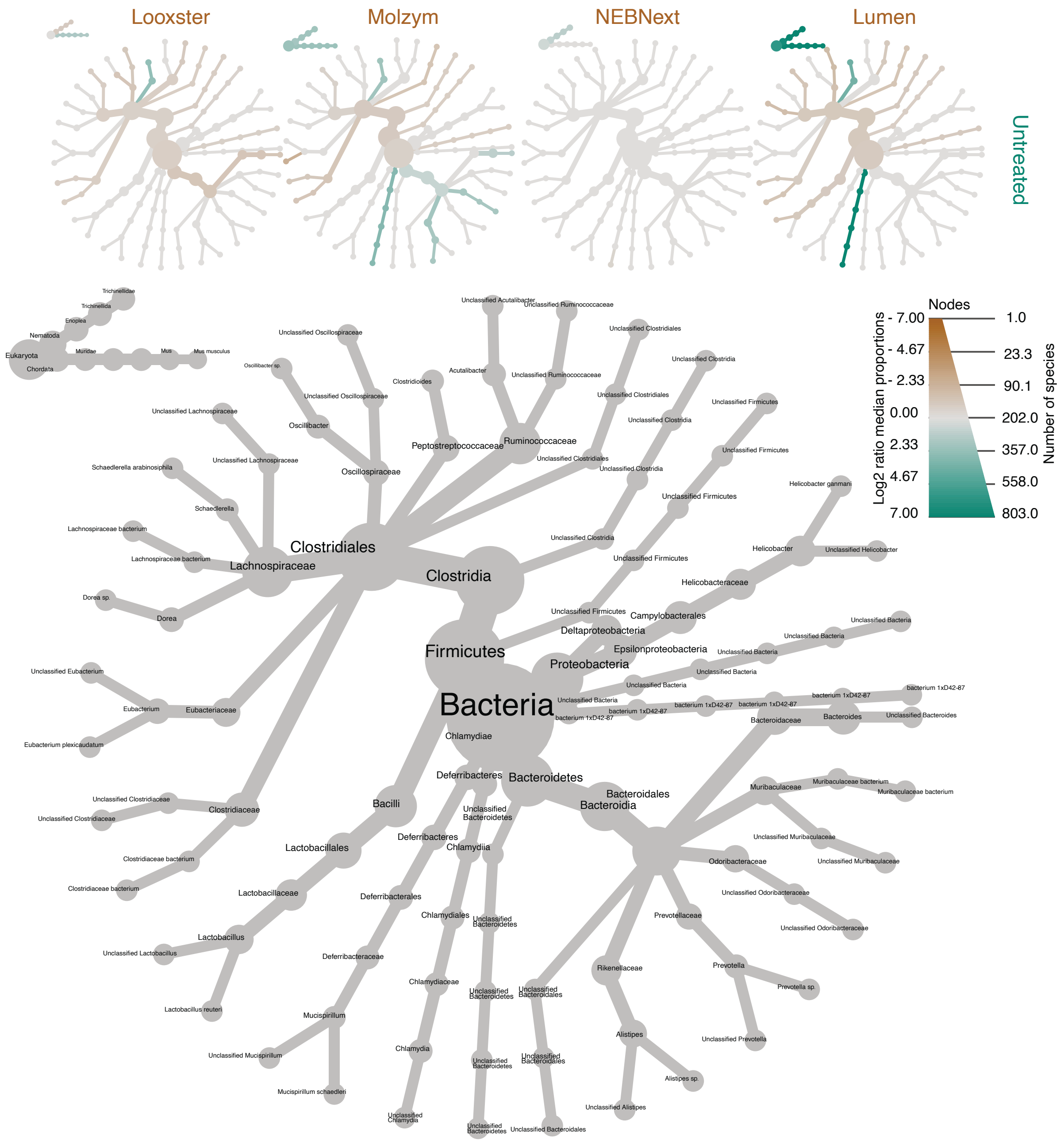


Figure 32: Differentially abundant taxa between the raw samples (Without, green), and the microbial-enriched samples and luminal samples (brown). The size of the nodes shows the number of species it contains. The color of the node represents the log2 ratio of the median relative abundances. Only significantly differentially abundant taxa are shown (Wilcoxon rank sum test and FDR correction for multiple testing).

3. Discussion

Shotgun metagenomic sequencing is an effective tool for studying a wide range of microbiome sites and to determine their taxonomic composition and functional capabilities with a relative lack of bias compared to 16S rRNA gene sequencing. The relative lack of bias is particularly important in quantitative studies, such as microbiome QTL mapping or GWAS, which rely on biologically correct microbial abundances. This cannot be achieved with 16S rRNA gene sequencing, as for example, different primers preferentially amplify different sets of taxa (Schmalenberger et al., 2001; D'Amore et al., 2016; Rausch et al., 2019). To use shotgun metagenomic sequencing to its full potential, substantial sequencing coverage needs to be achieved in order to refine classification to the species or strain level and to have reliable functional information. This is, however, a problem in tissue samples, which largely consist of host DNA. Thus, to maximize tissue sample use, host DNA must be removed prior to sequencing.

In this study, we evaluated the effectiveness of three different commercially available microbial DNA enrichment kits (LOOXSTER, NEBNext, Molzym) in removing host DNA and their influence on the taxonomic and functional composition. Results show that enrichment with the Molzym kit resulted in the highest amount of microbial reads, although this was not significantly different from the other methods and the efficiency of the Molzym kit was variable between samples. Contrary to previous studies, we find a significant effect of the microbial enrichment method on the taxonomic and functional composition of the samples and we also find specific taxa that are missing in the different methods (Feehery et al., 2013; Thoendel et al., 2016; Yap et al., 2020; Hansen et al., 2009; Heravi et al., 2020; Marotz et al., 2018). The Bray-Curtis dissimilarity to the raw untreated samples was lowest for the NEBNext kit and the BC dissimilarity was smaller between individuals within methods than within individuals between methods, indicating that the enrichment methods have a larger influence on the community composition compared to the inter-individual variability.

In the NEBNext kit, the MBD2-Fc protein binds to CpG methylated DNA of the host to remove it (Feehery et al., 2013) and thus, it is the only method tested that only manipulates the host DNA. This can explain why the taxonomic composition is the least different compared to the raw samples. The NEBNext kit however, needs at least 15 kb long sequences for optimal performance and is limited to 1.5 µg of input DNA. As the method relies on binding to the methylated eukaryotic DNA, the enrichment is limited to the amount of eukaryotic DNA binding protein. Thus, tissue samples with predominately host DNA might have such a high host:microbial DNA ratio that there is not enough protein to bind all of the host DNA, which could explain the low enrichment efficiency (3.6% of microbial reads) combined with the presence of short sequences. To conclude, the taxonomic and functional composition of samples processed with the NEBNext kit are the closest to the untreated samples, however the enrichment efficiency is low in samples with a high concentration of host DNA.

The LOOXSTER kit binds the nonmethylated CpG-dinucleotides of the microbial DNA to remove the methylated host DNA (Xu et al., 2011). It is easier to use as it can take up to 300 µg of input DNA to produce a maximum of 3µg enriched DNA. Its separating power is highest for microbial DNA with high genomic GC content. As this can be highly variable between microbes (20-80%), this could explain the difference in taxonomic composition in comparison to the untreated samples. Samples enriched with LOOXSTER contained on average 5% of microbial reads and had a significantly higher alpha diversity compared to the other samples. Only 466 taxa were differentially less abundant compared to the untreated samples. The LOOXSTER kit was the most user friendly, but is not efficient and produces a biased taxonomic community composition.

Lastly, the Molzym kit relies on selective lysis of the host tissue first followed by a DNase treatment. Thus, the condition of the samples is critical and premature lysis of microbial cells should be prevented. Due to the selective lysis procedure, a different piece of tissue was extracted, which makes the results not perfectly comparable to the other methods. The Molzym kit is more labour intensive compared to the NEBNext and the LOOXSTER kit and does not allow for simultaneous extraction of DNA and RNA. The enrichment of microbial reads was the most efficient with the Molzym kit (8.96% ± 9.04% microbial reads), but also was the most variable. We believe that the variation in the efficiency of this kit relies on our use of RNAlater as a stabilizer for the RNA and DNA, which could have interfered with the disruption and homogenization of the tissue with an incomplete lysis of the host cells as a result. The community composition was the least similar to the untreated mucosal samples and was closer to the luminal samples. This is partly due to the biological variation present between two pieces of the same tissue sample, but mostly due to the selective lysis procedure. Some microbial cells can lyse during the host cell lysis step, the DNase treatment step removes DNA from prematurely lysed cells (e.g. due to freezing), and the extraction protocol does not include a bead beating step, although this has become a standard procedure in bacterial cell extraction for lysing Gram-positive bacteria (Velásquez-Mejía et al., 2018; Zhang et al., 2020). To summarize, the Molzym kit was the most efficient method tested in microbial DNA enrichment, but the selective lysis procedure causes a large change in taxonomic and functional community composition.

Even though the study used a modest number of samples, the present sample size was adequate to identify significant differences across techniques. Our study also included adequate controls, including untreated mucosal DNA, luminal samples and a mock community. However, some limitations remain. We did not quantify the concentration of microbial or host DNA, which might give further insight into the ratios of host to microbial DNA observed in the cecum tissue samples. More studies are needed to investigate the influence of stabilizers such as RNAlater on the host tissue when using the Molzym Ultra Deep Microbiome Prep kit.

Overall, this evaluation can only conclude that currently available microbial DNA enrichment methods are not efficient enough to sufficiently remove host DNA from tissue samples in order to reach adequate sequencing depth. Moreover, the bias of the enrichment methods tested on the community composition makes them not suitable for quantitative studies, such as QTL mapping or GWAS from microbiome taxonomic/functional abundances.

4. Material and Methods

4.1. Sample collection, extraction and enrichment

Eight mice originating from a wild derived *Mus musculus musculus* and *Mus musculus domesticus* hybrid mouse breeding stock (11th lab generation at the sampling time point) were sacrificed according to the German animal welfare law and Federation of European Laboratory Animal Science Associations guidelines with CO_2 followed by cervical dislocation. Cecal tissues and content were conserved in 1ml RNAlater for 24h at 5°C after which the RNAlater was removed, and the sample stored at -20°C until extraction. Figure 28 shows the overall workflow for comparing the microbial DNA enrichment kits. One piece of cecum tissue (max 30 mg) disrupted and homogenized using a bead-beating step (3 x 15 s at speed 6500 with Precellys 24) with Lysing Matrix E tubes (MP Biomedicals) before extraction using the DNA/RNA AllPrep kit (Qiagen) according to the manufacturer's protocol. Extracted DNA was divided between the NEBNext (New England Biolabs) and the LOOXSTER enrichment kits (Analytik Jena). The NEBNext kit has a restriction of maximum 1.5 µg of input DNA. The leftover DNA was used for the LOOXSTER enrichment kit (9.2-16.8 µg). Both methods were performed as specified by the manufacturer. Another piece of cecum tissue (max 30 mg) was extracted according to manufacturer's protocol using the Ultra Deep Microbiome Prep kit (Molzym).

4.2. Library preparation and shotgun sequencing

The library for shotgun sequencing was prepared using the NexteraXT kit (Illumina) according to manufacturer's protocol. Sequencing was performed on two runs on an Illumina NextSeq 500 platform via 2 X 150 bp Mid Output Kit at the sequencing center of the Max Planck Institute for Evolutionary Biology (Plön, Germany). One run was filled with 8 raw, untreated DNA samples and the other run was filled with eight samples treated with the three different microbial DNA enrichment kits (24 samples in total).

4.3. Processing of shotgun sequences

Reads were pre-processed with Kneaddata (v0.10.0). It integrates FastQC (v0.11.9), Trimmomatic (v0.39) (Bolger et al., 2014) and Bowtie2 (v2.3.5.1) (Langmead and Salzberg, 2012) to perform quality control, quality filtering, and *in silico* separation of host reads from microbial reads, respectively. First, adapters and the first 10 bp were removed from the reads. Next, reads were trimmed using a sliding window approach, where the average base Phred quality score over four bases cannot be less than 20.

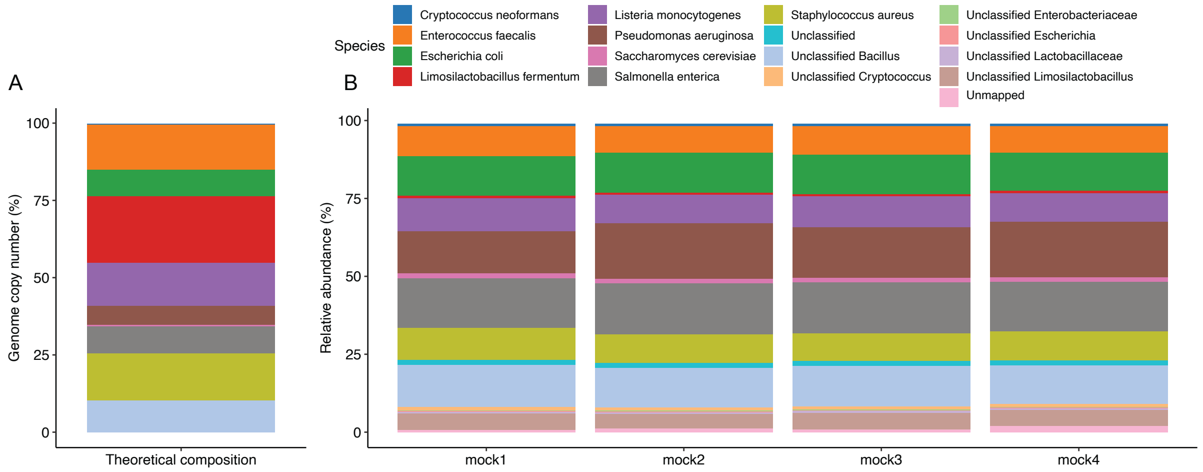
Reads with a length less than 50 bp were discarded. After quality filtering, mouse contaminant reads were identified with BowTie2, by mapping against the C57BL/6 reference genome in the very sensitive mode. Only the microbial reads were kept for subsequent analyses.

The processing of the shotgun sequences was done with *SqueezeMeta* (v1.3.1) (Tamames and Puente-Sánchez, 2018). Assembly was done using *Megahit* (Li et al., 2015). Short contigs (<200 bps) were removed and contig statistic calculated using *prinseq* (Schmieder and Edwards, 2011). RNAs were predicted using *Barrnap* (Seemann, 2014). 16S rRNA gene sequences were taxonomically classified using the *RDP classifier* (Wang et al., 2007) and tRNA/tmRNA sequences were predicted using *Aragorn* (Laslett and Canback, 2004). ORFs were predicted using *Prodigal* (Hyatt et al., 2010). Similarity searches for GenBank (Clark et al., 2016), eggNOG (Huerta-Cepas et al., 2016), and KEGG (Kanehisa and Goto, 2000) were conducted using *Diamond* (Buchfink et al., 2015). HMM homology searches were done by *HMMER3* (Eddy, 2009) for the Pfam database (Piovesan et al., 2019). Read mapping against contigs was performed using *Bowtie2* (). Binning was done using *MaxBin2* (Wu et al., 2016) and *Metabat2* (Kang et al., 2019). Binning results were combined using *DAS Tool* (Sieber et al., 2018) and bin statistics were computed using *CheckM* (Parks et al., 2015). Pathway prediction for KEGG (Kanehisa, 2002) and MetaCyc (Caspi et al., 2020) databases was completed using *MinPath* (Ye and Doak, 2009). SqueezeMeta implements a fast LCA algorithm for taxonomic assignment of genes that looks for the last common ancestor of the hits for each query gene using the results of the Diamond search against Genbank nr database based on (Luo et al., 2014).

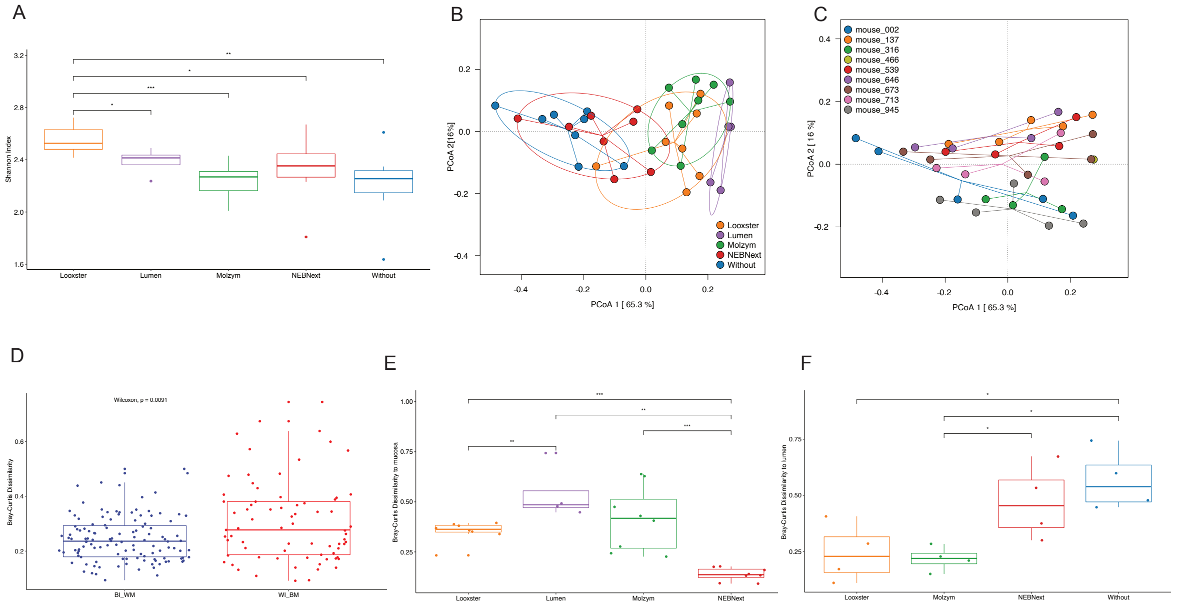
4.4. Analysis of shotgun sequences

All analyses were performed in R v 4.0.2. The output from *SqueezeMeta* was handled with the SQMtools R package (Puente-Sánchez et al., 2020) to load into R and to plot the taxonomic and functional compositions. Alpha and beta diversity measures where calculated with the *vegan* R package (Jari Oksanen, 2020). The *ape* package (Paradis and Schliep, 2019) was used for the principal coordinate analysis. Differential abundance analysis was performed with the *metacoder* package (Foster et al., 2017b). Plots were made use base R, *ggplot2* (Wickham and Chang, 2016), *ggsci* (Nan Xiao, 2018), and *ggpubr* (Kassambara and Kassambara, 2020) packages.

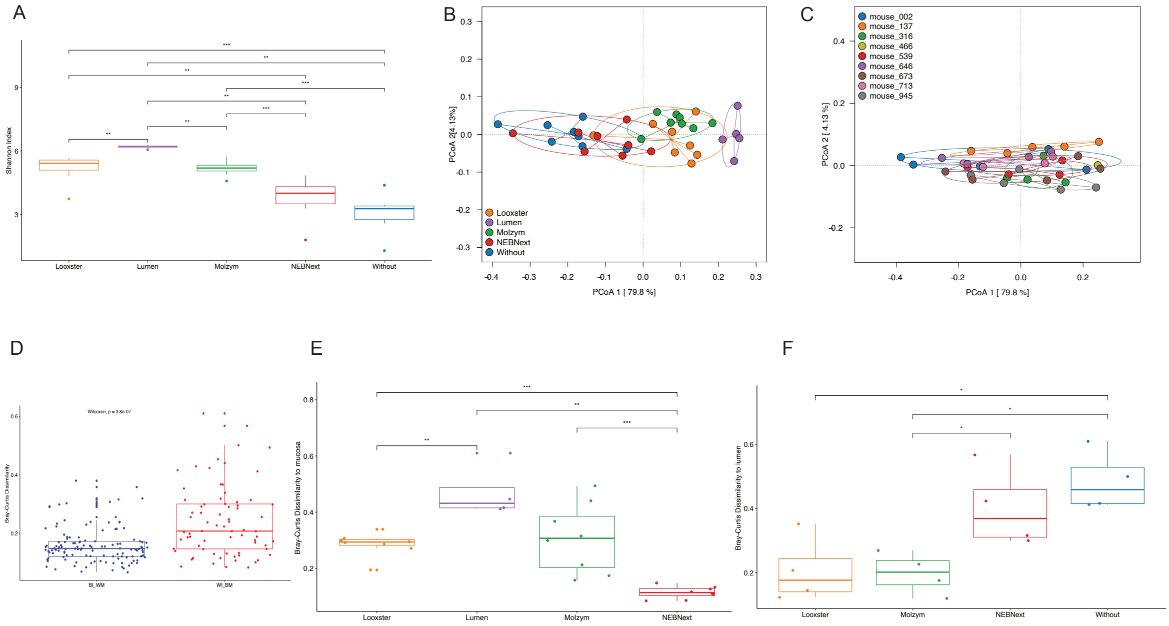
5. Supplementary figures



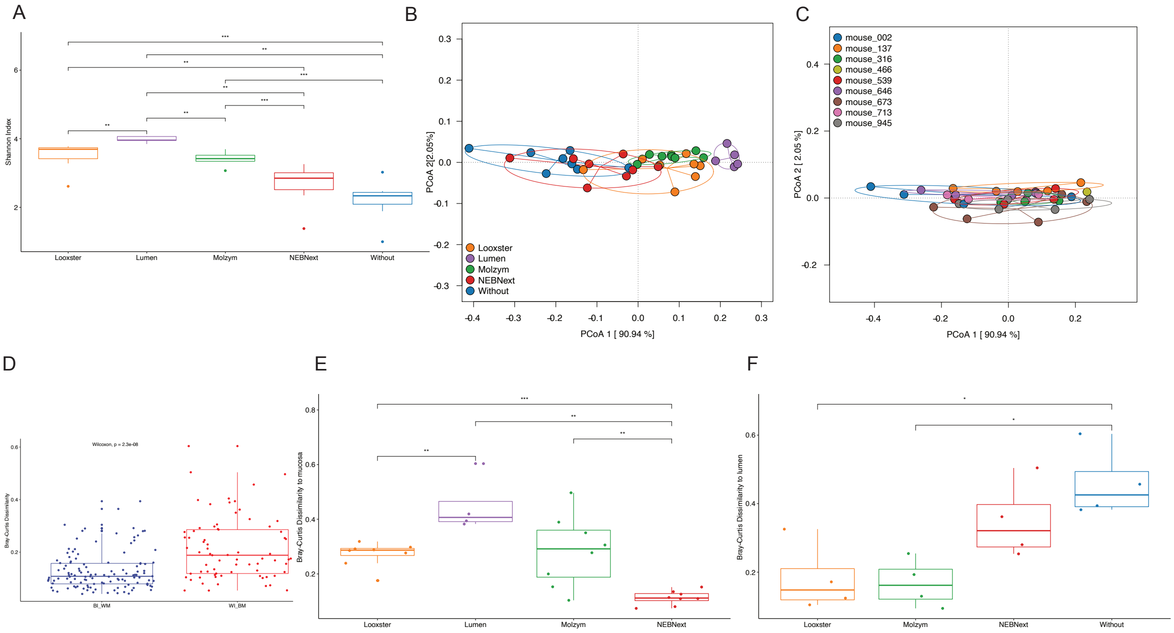
Suppl. Figure 1: Mock community composition. (A) Theoretical composition in terms of genome copy number. The genome copy number is calculated from the theoretical genomic DNA composition using the following formula: genome copy number = total genomic DNA (g) X unit conversion constat (bp/g) / genome size (bp). *Bacillus subtilis* is colored light blue (Unclassified *Bacillus* in the legend). (B) Observed community composition in terms of relative abundance of reads.



Suppl. Figure 2: Alpha and beta diversity on 100,000 subsampled reads. (A) Species richness and evenness in the samples processed with different enrichments methods. (B-C) PCoA of samples using Bray-Curtis Dissimilarities colored by method (B) or by individual (C). (D) Pairwise Bray-Curtis Dissimilarity within individual between methods (WI_BM) and between individuals within methods (BI_WM). (E-F) Pairwise Bray-Curtis dissimilarity to untreated mucosal sample (E) or to untreated luminal sample (F). Only significant comparisons are shown. * P < .05; ** P < .01; *** P < .001



Suppl. Figure 3: Alpha and beta diversity analyses on COG functions. (A) Species richness and evenness in the samples processed with different enrichments methods. (B-C) PCoA of samples using Bray-Curtis Dissimilarities colored by method (B) or by individual (C). (D) Pairwise Bray-Curtis Dissimilarity within individual between methods (WI_BM) and between individuals within methods (BI_WM). (E-F) Pairwise Bray-Curtis dissimilarity to untreated mucosal sample (E) or to untreated luminal sample (F). Only significant comparisons are shown. * P < .05; ** P < .01; *** P < .001



Suppl. Figure 4: Alpha and beta diversity analyses on KEGG functions. (A) Species richness and evenness in the samples processed with different enrichments methods. (B-C) PCoA of samples using Bray-Curtis Dissimilarities colored by method (B) or by individual (C). (D) Pairwise Bray-Curtis Dissimilarity within individual between methods (WI_BM) and between individuals within methods (BI_WM). (E-F) Pairwise Bray-Curtis dissimilarity to untreated mucosal sample (E) or to untreated luminal sample (F). Only significant comparisons are shown. * P < .05; ** P < .01; *** P < .001

Conclusion

The present work provides a better understanding of the factors that shape diversity in host-associated bacterial populations. We employed a genome-wide association mapping approach to identify genomic regions influencing bacterial abundances. Next, we pursued a candidate gene-bacteria association identified in an earlier version of the association study in different model organisms. Finally, we explored the possibility of employing shotgun metagenome sequencing on the same cecum tissue samples used for mapping in Chapter 1.

In **Chapter 1**, we acquired insight into the evolutionary connection between hosts and their microbiota at the transition from within species variation to between species divergence by utilizing wild-derived hybrid inbred strains to construct our mapping population. Our mapping population's genetic relatedness was substantially associated with microbiome similarity, indicating phylosymbiosis during the early phases of speciation. Heritability estimates were associated with cospeciation rates in a significant fraction of microbial species, suggesting that vertical transmission may foster the evolution of greater host genetic effects for highly cospeciating taxa. We went on to investigate general features of host loci, including candidate genes and pathways. Taken together, these findings show that host genetics plays a crucial role in shaping microbiome composition and pinpoint particular underlying processes.

Chapter 2 presented a follow-up study of an association between ASV35 (*Bacteroides acidifaciens/uniformis*) DNA-based abundance and a region on chromosome 13 containing two genes, *Gfod1* and *Sirt5*. ASV35 is a promising candidate since it was identified to be an indicator species for *Mus musculus musculus* in a geographic screen of mouse microbiomes (Fokt, 2021). We used two model organisms, *Mus musculus* and *Drosophila melanogaster,* to characterize the influence of the candidate gene *Sirt5* on the bacterial composition. *Sirt5/dSirt4* expression had an influence on the community composition and cycling of bacterial taxa in both species, indicating a conserved role for the mitochondrial sirtuin in controlling bacterial taxa abundance. Moreover, this approach to pursue a candidate gene-bacteria association can be adopted on the most promising associations found in the current version of the GWAS.

Shotgun metagenome sequencing would be useful to determine the functional capacity of the microbiome in order that we can proceed from 'who is the singer' to 'what is their song' and to tackle the problem of functional redundancy between bacterial species. Moreover, the hologenome concept of evolution assumes that selection happens on the genomic content of the bacteria as part of the hologenome instead of the taxonomic composition. In addition, as the tissue-associated bacterial community and more heritable, we would be able to determine reliable heritability estimates of specific metabolic functions and pathways present in the community. In **Chapter 3**, I assessed the feasibility of using cecum tissue with shotgun metagenome sequencing. We evaluated the performance of three different commercially available microbial DNA enrichment kits (LOOXSTER, NEBNext, Molzym) in removing host DNA and their influence on taxonomic and functional composition. Current microbial DNA enrichment techniques are inefficient in removing enough host DNA from tissue samples to achieve appropriate sequencing depth. Furthermore, the enrichment methods evaluated on community composition are biased, making them unsuitable for quantitative investigations like QTL mapping or GWAS using microbiome taxonomic/functional abundances. Further research is needed on optimizing microbial DNA enrichment methods to make shotgun metagenome sequencing possible on samples containing host DNA.

In conclusion, these findings indicate that host genetics has a significant influence on the variance in the mouse gut microbiota. We demonstrated that host genes may impact bacterial abundances via metabolic (nitrogen) homeostasis, and that this relationship is likely conserved from *Drosophila* to mice. The precise functioning of the mechanism, however, remains unclear. These results may pave the path for more in-depth research into various host loci associated with certain bacterial taxa in order to understand the functional pathways involved. Further research on functionally characterizing the association of bacterial species with host genomic regions will reveal the structural mechanisms of the interactions opening the door to therapeutics that attempt to replicate, emulate, or boost natural protective genetic diversity.

Acknowledgements

I would like to express my most sincere appreciation to my supervisor Prof. Dr. John Baines for his patient guidance, enthusiastic encouragement and useful critiques of this research work. I am also particularly grateful for the invaluable support of Dr. Leslie Turner. Our weekly meetings were always a moment I would look forward to discussing our collaborative project combined with some chitchat.

I would like to thank Hanna Fokt and Cecilia Chung for their assistance with the collection of the data and performing all those numerous dissections. My thanks also extends to all the other members of our lab group for our everyday conversations and support. Special thanks goes to Ms Katja Cloppenborg-Schmidt for her assistance in the lab.

I am also thankful to the Collaborative Research Centre 1182 through which the funding for the research project was acquired to provide the structure to develop as a scientist and who awarded me with a Young Investigator Award, which allowed me to fund an additional research project. I would also like to thank Abdulgawaad Saboukh, who executed this project.

In addition, I would like to thank my parents for their continuous support, love and sympathetic ear. Mama, Papa, ik zie jullie graag! You are always there for me. Also I express my thanks to my uncle Prof. Dr. Joery De Kock for fueling my interests in science and encouraging me to pursue a Ph.D. and to my brother, my sister and the rest of the family for their love and support.

Some special words of gratitude go to my friends who have always been a major source of support when things would get a bit discouraging and have provided me with many memorable days and evenings to counter it: Evelien Beirnaert, Malte Rühlemann, Danielle Harris, Nadia Andreani, Meghna Basu, Aleksa Čepić, Marek Ehlers, Nina Tonn, and Britt Marie Hermes. Thank you for your friendship!

Finally, I could not have completed this dissertation without the endless support, love and encouragement from my partner Robin Wolke. It was a great comfort to have you by my side. My heartfelt thanks. This would not have been possible without you.

References

Abdi, Hervé (2007), 'The Bonferonni and Šidák Corrections for Multiple Comparisons', in Salkind, Neil J. (ed.), (Encyclopedia of Measurement and Statistics, SAGE), 9.

Adiyodi, KG and Rita G Adiyodi (1983), *Reproductive biology of invertebrates*, ((592.016 R4);).

Agosta, SJ and JA Klemens (2008), 'Ecological fitting by phenotypically flexible genotypes: implications for species associations, community assembly and evolution.', *Ecol Lett*, 11 1123-34.

Alhasson, Firas, et al. (2017), 'Altered gut microbiome in a mouse model of Gulf War Illness causes neuroinflammation and intestinal injury via leaky gut and TLR4 activation', *PLoS One*, 12 (3), e0172914.

Amato, Katherine R, et al. (2019), 'Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes', *The ISME journal*, 13 (3), 576-87.

Backhed, F., et al. (2004), 'The gut microbiota as an environmental factor that regulates fat storage', *Proceedings of the National Academy of Sciences*, 101 (44), 15718-23.

Bader, Gary D. and Christopher WV Hogue (2003), 'An automated method for finding molecular complexes in large protein interaction networks', *BMC Bioinformatics*, 4 (1), 2.

Barton, Nicholas H. and Peter D. Keightley (2002), 'Understanding quantitative genetic variation', *Nat. Rev. Genet.*, 3 (1), 11-21.

Baumann, Paul, et al. (1995), 'Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids', *Annu. Rev. Microbiol.*, 49 (1), 55-94.

Beavis, WD (1994), 'The power and deceit of QTL experiments: lessons from comparative QTL studies', Proceedings of the forty-ninth annual corn and sorghum industry research conference 250 266.

Belheouane, Meriem, et al. (2017), 'Improved detection of gene-microbe interactions in the mouse skin microbiota using high-resolution QTL mapping of 16S rRNA transcripts', *Microbiome*, 5 (1), 1-17.

Benson, Andrew K., et al. (2010), 'Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors', *Proceedings of the National Academy of Sciences of the United States of America*

Proc. Natl. Acad. Sci. U.S.A., 107 (44), 18933-38.

Berg, Gabriele, et al. (2020), 'Microbiome definition re-visited: old concepts and new challenges', *Microbiome*, 8 (1), Bergen, Werner G and Guoyao Wu (2009), 'Intestinal nitrogen recycling and utilization in health and disease', *The Journal of nutrition*, 139 (5), 821-25.

Blaxter, Mark, et al. (2005), 'Defining operational taxonomic units using DNA barcode data', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360 (1462), 1935-43.

Blekhman, Ran, et al. (2015), 'Host genetic variation impacts microbiome composition across human body sites', *Genome Biology*, 16 (1), 191.

Bolger, AM, M Lohse, and B Usadel (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data.', *Bioinformatics*, 30 (15), 2114-20.

Bonder, Marc Jan, et al. (2016), 'The effect of host genetics on the gut microbiome', *Nat. Genet.*, 48 (11), 1407-12.

Bordenstein, Seth R. and Kevin R. Theis (2015), 'Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes', *PLoS Biol.*, 13 (8),

Brent Pedersen, Joe Brown (2013), 'poverlap: significance testing over interval overlaps',

Broekema, RV, OB Bakker, and IH Jonkers (2020), 'A practical view of fine-mapping and gene prioritization in the post-genome-wide association era.', *Open Biol*, 10 (1), 190221.

Brooks, AW, et al. (2016), 'Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History.', *PLoS Biol.*, 14 (11), e2000225.

Brucker, RM and SR Bordenstein (2012a), 'Speciation by symbiosis.', *Trends Ecol Evol*, 27 (8), 443-51.

——— (2013), 'The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus Nasonia.', *Science*, 341 (6146), 667-69.

Brucker, Robert M and Seth R Bordenstein (2012b), 'The roles of host evolutionary relationships (genus: Nasonia) and development in structuring microbial communities', *Evolution: International Journal of Organic Evolution*, 66 (2), 349-62.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2015), 'Fast and sensitive protein alignment using DIAMOND', *Nature methods*, 12 (1), 59-60.

Burke, John M., et al. (2002), 'Genetic Analysis of Sunflower Domestication', *Genetics*, 161 (3), 1257-67.

Bush, SJ, et al. (2020), 'Evaluation of methods for detecting human reads in microbial sequencing datasets.', *Microb Genom*, 6 (7),

Callahan, Benjamin J (2016), 'DADA2 pipeline', *DADA2*,

Callahan, Benjamin J, et al. (2016), 'DADA2: High resolution sample inference from Illumina amplicon data', *Nature methods*

Nat Methods, 13 (7), 581-83.

Callahan, Benjamin J, Paul J McMurdie, and Susan P Holmes (2017), 'Exact sequence variants should replace operational taxonomic units in marker-gene data analysis', *The ISME Journal*, 11 (12), 2639-43.

Campbell, JH, et al. (2012), 'Host genetic and environmental effects on mouse intestinal microbiota.', *ISME J*, 6 (11), 2033-44.

Cani, Patrice D., et al. (2008), 'Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice', *Diabetes*, 57 (6), 1470-81.

Cantó, Carles, Keir J. Menzies, and Johan Auwerx (2015), 'NAD+ Metabolism and the Control of Energy Homeostasis: A Balancing Act between Mitochondria and the Nucleus', *Cell Metabolism*, 22 (1), 31-53.

Carding, Simon, et al. (2015), 'Dysbiosis of the gut microbiota in disease', *Microb. Ecol. Health Dis.*, 26

Cardona, C, et al. (2016), 'Network-based metabolic analysis and microbial community modeling.', *Curr Opin Microbiol*, 31 124-31.

Cardoso, JC, et al. (2012), 'Feeding and the rhodopsin family g-protein coupled receptors in nematodes and arthropods.', *Front Endocrinol (Lausanne)*, 3 157.

Carmody, RN, et al. (2015), 'Diet dominates host genotype in shaping the murine gut microbiota.', *Cell Host Microbe*, 17 (1), 72-84.

Carneiro Dutra, Heverton Leandro, Mark Anthony Deehan, and Horacio Frydman (2020), 'Wolbachia and Sirtuin-4 interaction is associated with alterations in host glucose metabolism and bacterial titer', *PLoS Pathog.*, 16 (10), e1008996.

Caspi, R, et al. (2020), 'The MetaCyc database of metabolic pathways and enzymes-a 2019 update', *Nucleic acids …*,

Castoldi, Angela, et al. (2015), 'They Must Hold Tight: Junction Proteins, Microbiota And Immunity In Intestinal Mucosa', *Current Protein & Peptide Science*

Curr Protein Pept Sci, 16 (7), 655-71.

Chalkiadaki, Angeliki and Leonard Guarente (2012), 'Sirtuins mediate mammalian metabolic responses to nutrient availability', *Nature Reviews Endocrinology*, 8 (5), 287-96.

Chang, Christopher C, et al. (2015), 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience*

GigaSci, 4 (1), 7.

Chen, Congying, et al. (2018), 'Contribution of Host Genetics to the Variation of Microbial Composition of Cecum Lumen and Feces in Pigs', *Frontiers in Microbiology*

Front. Microbiol., 9

Chen, Haiwei, et al. (2019), 'A forward chemical genetic screen reveals gut microbiota metabolites that modulate host physiology', *Cell*, 177 (5), 1217-1231.e18.

Chu, Hiutung and Sarkis K Mazmanian (2013), 'Innate immune recognition of the microbiota promotes host-microbial symbiosis', *Nat. Immunol.*, 14 (7), 668-75.

Chung, HJ, et al. (2018), 'Gut Microbiota as a Missing Link Between Nutrients and Traits of Human.', *Front Microbiol*, 9 1510.

Clapp, M, et al. (2017), 'Gut microbiota's effect on mental health: The gut-brain axis.', *Clin Pract*, 7 (4), 987.

Clark, Karen, et al. (2016), 'GenBank', *Nucleic Acids Res.*, 44 (D1), D67-72.

Clark, ME, et al. (2005), 'Widespread prevalence of wolbachia in laboratory stocks and the implications for Drosophila research.', *Genetics*, 170 (4), 1667-75.

Cohen, Louis J., et al. (2017), 'Commensal bacteria make GPCR ligands that mimic human signalling molecules', *Nature*, 549 (7670), 48-53.

Cole, JR, et al. (2014), 'Ribosomal Database Project: data and tools for high throughput rRNA analysis.', *Nucleic Acids Res.*, 42 (Database issue), D633-42.

Colosimo, Dominic A., et al. (2019), 'Mapping Interactions of Microbial Metabolites with Human G-Protein-Coupled Receptors', *Cell Host & Microbe*, 26 (2), 273-282.e7.

Cox, Laura M. and Howard L. Weiner (2018), 'Microbiota Signaling Pathways that Influence Neurologic Disease', *Neurotherapeutics*, 15 (1), 135-45.

D'Amore, Rosalinda, et al. (2016), 'A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling', *BMC Genomics*, 17 (1),

Daniel, Noëmie, Emelyne Lécuyer, and Benoit Chassaing (2021), 'Host/microbiota interactions in health and diseases—Time for mucosal microbiology', *Mucosal Immunology*, 1-11.

Davenport, Emily R., et al. (2015), 'Genome-Wide Association Studies of the Human Gut Microbiota', *PLoS One*, 10 (11), e0140301.

Davenport, Emily R. (2020), 'Genetic Variation Shapes Murine Gut Microbiota via Immunity', *Trends in Immunology*, 41 (1), 1-3.

Deaver, Jessica A., Sung Y. Eum, and Michal Toborek (2018), 'Circadian Disruption Changes Gut Microbiome Taxa and Functional Gene Composition', *Frontiers in Microbiology*

Front Microbiol, 9 737.

Delzenne, Nathalie M., et al. (2011), 'Targeting gut microbiota in obesity: effects of prebiotics and probiotics', *Nature Reviews. Endocrinology*

Nat Rev Endocrinol, 7 (11), 639-46.

Doncheva, Nadezhda T., et al. (2019), 'Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data', *Journal of Proteome Research*

J Proteome Res, 18 (2), 623-32.

Douglas, AE and JH Werren (2016), 'Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts.', *mBio*, 7 (2), e02099.

Douglas, GM, RG Beiko, and MGI Langille (2018), 'Predicting the functional potential of the microbiome from marker genes using PICRUSt', *Microbiome Analysis*,

Du, J, et al. (2011), 'Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase.', *Science*, 334 (6057), 806-9.

Du, Y, et al. (2018), 'SIRT5 deacylates metabolism-related proteins and attenuates hepatic steatosis in ob/ob mice.', *EBioMedicine*, 36 347-57.

Dworkin, M (2012), 'Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist.', *FEMS Microbiol Rev*, 36 (2), 364-79.

Eddy, Sean R (2009), 'A new generation of homology search tools based on probabilistic inference', *Genome Informatics 2009: Genome Informatics Series Vol. 23* (World Scientific), 205-11.

Edgar, Robert C. and Henrik Flyvbjerg (2015), 'Error filtering, pair assembly and error correction for next-generation sequencing reads', *Bioinformatics*, 31 (21), 3476-82.

Elya, C, et al. (2016), 'Stable Host Gene Expression in the Gut of Adult Drosophila melanogaster with Different Bacterial Mono-Associations.', *PLoS One*, 11 e0167357.

Erdman, S.E. and T. Poutahidis (2016), 'Microbes and Oxytocin', *131* (Int. Rev. Neurobiol., Elsevier), 91-126.

Eren, A Murat, et al. (2015), 'Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences', *The ISME Journal*, 9 (4), 968-79.

Eren, A. Murat, et al. (2013), 'Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data', *Methods in Ecology and Evolution*, 4 (12), 1111-19.

Fabian, Scheipl, Greven Sonja, and Kuechenhoff Helmut (2008), 'Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models.', *Computational Statistics & Data Analysis*, 52 (7), 3283-99.

Falconer, D. S (1996), *Introduction to quantitative genetics*, (Harlow, England: Prentice Hall).

Feehery, GR, et al. (2013), 'A method for selectively enriching microbial DNA from contaminating vertebrate host DNA.', *PLoS One*, 8 (10), e76096.

Fischer, Frank, et al. (2012), 'Sirt5 deacylation activities show differential sensitivities to nicotinamide inhibition', Flux, M. C. and Christopher A. Lowry (2020), 'Finding intestinal fortitude: Integrating the microbiome into a holistic view of depression mechanisms, treatment, and resilience', *Neurobiology of Disease*

Microbiome in neurological and psychiatric disease

Neurobiology of Disease, 135 104578.

Fokt, Hanna (2021), 'The Evolution of the Genus Bacteroides in the House Mouse Species Complex',

Fonken, Laura K., et al. (2010), 'Light at night increases body mass by shifting the time of food intake', *Proceedings of the National Academy of Sciences*

PNAS, 107 (43), 18664-69.

Foster, Jane A., Linda Rinaman, and John F. Cryan (2017a), 'Stress & the gut-brain axis: Regulation by the microbiome', *Neurobiology of Stress*, 7 124-36.

Foster, ZS, TJ Sharpton, and NJ Grünwald (2017b), 'Metacoder: An R package for visualization and manipulation of community taxonomic diversity data.', *PLoS Comput. Biol.*, 13 (2), e1005404.

Fukata, Masayuki and Moshe Arditi (2013), 'The role of pattern recognition receptors in intestinal inflammation', *Mucosal immunology*, 6 (3), 451-63.

Futuyma, Douglas J. (1983), *Coevolution*, (Sunderland, Mass: Sinauer Associates Inc) 555.

Gastelum, C, et al. (2021), 'Adaptive Changes in the Central Control of Energy Homeostasis Occur in Response to Variations in Energy Status.', *Int J Mol Sci*, 22 (5), 2728.

Gaulke, Christopher A., et al. (2018), 'Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota', *mBio*, 9 (5),

Gautam, D, et al. (2006), 'A critical role for beta cell M3 muscarinic acetylcholine receptors in regulating insulin release and blood glucose homeostasis in vivo.', *Cell Metab*, 3 (6), 449-61.

Geraldes, A, et al. (2008), 'Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes.', *Mol Ecol*, 17 (24), 5349-63.

Gevers, Dirk, et al. (2014), 'The treatment-naive microbiome in new-onset Crohn's disease', *Cell host & microbe*, 15 (3), 382-92.

Goetzman, ES, et al. (2020), 'Impaired mitochondrial medium-chain fatty acid oxidation drives periportal macrovesicular steatosis in sirtuin-5 knockout mice.', *Sci Rep*, 10 (1), 18367.

Gogarten, Jan F, et al. (2021), 'Primate phageomes are structured by superhost phylogeny and environment', *Proceedings of the National Academy of Sciences*, 118 (15),

Goodrich, Julia K., et al. (2014), 'Human genetics shape the gut microbiome', *Cell*, 159 (4), 789-99.

Goodrich, Julia K., et al. (2016), 'Genetic Determinants of the Gut Microbiome in UK Twins', *Cell host & microbe*,

Gould, AL, et al. (2018), 'Microbiome interactions shape host fitness.', *Proc. Natl. Acad. Sci. U S A*, 115 (51), E11951-60.

Gregory R. Warnes, Ben Bolker and Thomas Lumley (2020), 'gtools: Various R Programming Tools',

Greiss, S and A Gartner (2009), 'Sirtuin/Sir2 phylogeny, evolutionary considerations and structural conservation.', *Mol Cells*, 28 (5), 407-15.

Grieneisen, L, et al. (2021), 'Gut microbiome heritability is nearly universal but environmentally contingent.', *Science*, 373 (6551), 181-86.

Groussin, Mathieu, et al. (2017), 'Unraveling the processes shaping mammalian gut microbiomes over evolutionary time', *Nature Comm.*, 8 (1), 14319.

Haigis, Marcia C. and David A. Sinclair (2010), 'Mammalian sirtuins: biological insights and disease relevance', *Annual Review of Pathology*

Annu Rev Pathol, 5 253-95.

Han, H, et al. (2021), 'From gut microbiota to host appetite: gut microbiota-derived metabolites as key regulators.', *Microbiome*, 9 (1), 162.

Hansen, WL, CA Bruggeman, and PF Wolffs (2009), 'Evaluation of new preanalysis sample treatment tools and DNA isolation protocols to improve bacterial pathogen detection in whole blood.', *J. Clin. Microbiol.*, 47 (8), 2629-31.

Hart, Michael W (2002), 'Life history evolution and comparative developmental biology of echinoderms', *Evolution & Development*, 4 (1), 62-71.

Hehemann, Jan-Hendrik, et al. (2010), 'Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota', *Nature*, 464 (7290), 908-12.

Heintz-Buschart, A and P Wilmes (2018), 'Human Gut Microbiome: Function Matters.', *Trends Microbiol.*, 26 (7), 563-74.

Henikoff, Steven and Jorja G Henikoff (1992), 'Amino acid substitution matrices from protein blocks', *Proceedings of the National Academy of Sciences*, 89 (22), 10915-19.

Heravi, Fatemah Sadeghpour, et al. (2020), 'Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples', *J. Microbiol. Methods*, 170 105856.

Hollander, Daniel and Jonathan D. Kaunitz (2020), 'The "Leaky Gut": Tight Junctions but Loose Associations', *Digestive Diseases and Sciences*

Dig Dis Sci, 65 (5), 1277-87.

Hua, Yan, et al. (2020), 'Gut microbiota and fecal metabolites in captive and wild North

China leopard (Panthera pardus japonensis) by comparsion using 16 s rRNA gene sequencing and LC/MS-based metabolomics', *BMC Veterinary Research*, 16 (1),

Huerta-Cepas, Jaime, et al. (2016), 'eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Res.*, 44 (D1), D286-93.

Hughes, David A., et al. (2020), 'Genome-wide associations of human gut microbiome variation and implications for causal inference analyses', *Nature Microbiology*, 5 (9), 1079-87.

Hughes, ME, JB 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), 372-80.

Hurst, GDD (2017), 'Extended genomes: symbiosis and evolution.', *Interface Focus*, 7 (5), 20170001.

Hyatt, Doug, et al. (2010), 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC bioinformatics*, 11 (1), 1-11.

Ishida, Sachiko, et al. (2020), 'Genome-wide association studies and heritability analysis reveal the involvement of host genetics in the Japanese gut microbiota', *Communications Biology*

Commun Biol, 3

Janzen, D. H. (1985), 'On Ecological Fitting', *Oikos*, 45 (3), 308.

Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter

R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene

Wagner (2020), 'vegan: Community Ecology Package. R package version 2.5-7.',

Jari, Oksanen, et al. (2020), 'vegan: Community Ecology Package',

Johnson, EL, et al. (2017), 'Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes.', *J Mol Med (Berl)*, 95 (1), 1-8.

Johnson, Jethro S., et al. (2019), 'Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis', *Nature Comm.*, 10 (1),

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

Kanehisa, Minoru (2002), 'The KEGG database', Novartis Foundation Symposium 91-100.

Kang, DD, et al. (2019), 'MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*,

Kassambara, A and MA Kassambara (2020), 'Package 'ggpubr', *mran.microsoft.com*, Kato, Kumiko, et al. (2018), 'Association between functional lactase variants and a high abundance of Bifidobacterium in the gut of healthy Japanese people', *PLoS One*, 13 (10),

Kelly, John R., et al. (2015), 'Breaking Down the Barriers: The Gut Microbiome, Intestinal Permeability and Stress-related Psychiatric Disorders', *Frontiers in Cellular Neuroscience*

Front. Cell. Neurosci., 9

Kemis, Julia H., et al. (2019), 'Genetic determinants of gut microbiota composition and bile acid profiles in mice', *PLoS Genet.*, 15 (8), e1008073.

Khan, Farhat, et al. (2021), 'IBDDB: a manually curated and text-mining-enhanced database of genes involved in inflammatory bowel disease', *Database*, 2021

Klug-Micu, GM, et al. (2013), 'CD40 ligand and interferon-γ induce an antimicrobial response against Mycobacterium tuberculosis in human monocytes.', *Immunology*, 139 (1), 121-28.

Knoll, Andrew H (2003), 'Life on a young planet, Princeton and Oxford',

Kohl, KD and MD Dearing (2014), 'Wild-caught rodents retain a majority of their natural gut microbiota upon entrance into captivity.', *Environ Microbiol Rep*, 6 (2), 191-95.

Korach-Rechtman, H, et al. (2019), 'Murine Genetic Background Has a Stronger Impact on the Composition of the Gut Microbiota than Maternal Inoculation or Exposure to Unlike Exogenous Microbiota.', *Appl. Environ. Microbiol.*, 85 (18), e00826-19.

Kottler, Malcolm J (1978), 'Charles Darwin's biological species concept and theory of geographic speciation: the transmutation notebooks', *Annals of Science*, 35 (3), 275-97.

Kovacs, Amir, et al. (2011), 'Genotype is a stronger determinant than sex of the mouse gut microbiota', *Microbial ecology*, 61 (2), 423-28.

Kuang, Z, et al. (2019), 'The intestinal microbiota programs diurnal rhythms in host metabolism through histone deacetylase 3.', *Science*, 365 (6460), 1428-34.

Kumar, S and DB Lombard (2018), 'Functions of the sirtuin deacylase SIRT5 in normal physiology and pathobiology.', *Crit. Rev. Biochem. Mol. Biol.*, 53 (3), 311-34.

Kurilshikov, Alexander, et al. (2020), 'Genetics of human gut microbiome composition',

Kurilshikov, Alexander, et al. (2021), 'Large-scale association analyses identify host factors influencing human gut microbiome composition', *Nat. Genet.*, 53 (2), 156-65.

Langmead, B and SL Salzberg (2012), 'Fast gapped-read alignment with Bowtie 2.', *Nat Methods*, 9 (4), 357-59.

Laslett, Dean and Bjorn Canback (2004), 'ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences', *Nucleic Acids Res.*, 32 (1), 11-16.

Lauder, Abigail P, et al. (2016), 'Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota', *Microbiome*, 4 (1), 1-11.

Leamy, Larry J, et al. (2014), 'Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice', *Genome Biology*

Genome Biol, 15 (12),

Lederberg, Joshua and Alexa T McCray (2001), 'Ome SweetOmics--A genealogical treasury of words', *The scientist*, 15 (7), 8-8.

Ley, RE, et al. (2006), 'Microbial ecology: human gut microbes associated with obesity.', *Nature*, 444 (7122), 1022-23.

Li, Dinghua, et al. (2015), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31 (10), 1674-76.

Li, J. and L. Ji (2005), 'Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix', *Heredity*, 95 (3), 221-27.

Li, Shenghui, et al. (2020), 'Analysis of metagenome-assembled genomes from the mouse gut microbiota reveals distinctive strain-level characteristics',

Lim, SJ and SR Bordenstein (2020), 'An introduction to phylosymbiosis.', *Proc Biol Sci*, 287 (1922), 20192900.

Lin, L and J Zhang (2017), 'Role of intestinal microbiota and metabolites on gut homeostasis and human diseases.', *BMC Immunol*, 18 (1), 2.

Linnenbrink, Miriam, et al. (2013), 'The role of biogeography in shaping diversity of the intestinal microbiota in house mice', *Molecular Ecology*, 22 (7), 1904-16.

Lockyer, N. (1869), *Nature*, ((v. 315); Macmillan Journals Limited).

Luo, C, LM Rodriguez-R, and KT Konstantinidis (2014), 'MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences.', *Nucleic Acids Res.*, 42 (8), e73.

Lynch, M and B Walsh (1998), 'Genetics and analysis of quantitative traits', *invemar.org.co*,

Lynch, SV and O Pedersen (2016), 'The Human Intestinal Microbiome in Health and Disease.', *N. Engl. J. Med.*, 375 (24), 2369-79.

Madsen, AS and CA Olsen (2012), 'Substrates for efficient fluorometric screening employing the NAD-dependent sirtuin 5 lysine deacylase (KDAC) enzyme.', *J. Med. Chem.*, 55 (11), 5582-90.

Majewski, J and T Pastinen (2011), 'The study of eQTL variations by RNA-seq: from SNPs to phenotypes.', *Trends Genet.*, 27 (2), 72-79.

Malaguarnera, L (2020), 'Vitamin D and microbiota: Two sides of the same coin in the immunomodulatory aspects.', *Int Immunopharmacol*, 79 106112.

Marchesi, Julian R. and Jacques Ravel (2015), 'The vocabulary of microbiome research: a proposal', *Microbiome*, 3 (1),

Margulis, Lynn (1991), 'Symbiosis as a Source of Evolutionary Innovation', in Margulis, Lynn and Rene Fester (eds.), (Symbiosis as a Source of Evolutionary Innovation, MIT Press),

Marotz, CA, et al. (2018), 'Improving saliva shotgun metagenomics by chemical host DNA depletion.', *Microbiome*, 6 (1), 42.

Matthew, Carlucci, et al. (2020), 'DiscoRhythm: an easy-to-use web application and R package for discovering rhythmicity', *Bioinformatics*, 36 (6), 1952-54.

Mauvoisin, Daniel, et al. (2017), 'Circadian and Feeding Rhythms Orchestrate the Diurnal Liver Acetylome', *Cell Reports*, 20 (7), 1729-43.

Mazel, F, et al. (2018), 'Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life', *mSystems*, 3 (5), e00097-18.

McFall-Ngai, M, et al. (2013), 'Animals in a bacterial world, a new imperative for the life sciences.', *Proc. Natl. Acad. Sci. U S A*, 110 (9), 3229-36.

McFall-Ngai, MJ (2014), 'The importance of microbes in animal development: lessons from the squid-vibrio symbiosis.', *Annu. Rev. Microbiol.*, 68 177-94.

McKnite, Autumn M., et al. (2012), 'Murine Gut Microbiota Is Defined by Host Genetics and Modulates Variation of Metabolic Traits', *PLoS One*, 7 (6),

McMurdie, Paul J and Susan Holmes (2013), 'phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data', *PLoS One*, 8 (4), e61217.

Metwaly, Amira, et al. (2020), 'Integrated microbiota and metabolite profiles link Crohn's disease to sulfur metabolism', *Nature Comm.*, 11 (1),

Meyer-Abich, A. (1943), 'Beiträge zur Theorie der Evolution der Organismen. I. Das typologische Grundgesetz und seine Folgerungen für Phylogenie und Entwicklungsphysiologie [Contributions to the evolutionary theory of organisms: I. The basic typological law and its implications for phylogeny and developmental physiology].', *Acta Biotheoretica*, 7 1-80.

Walsh, Michael Lynch and Bruce (1998), *Genetics and Analysis of Quantitative Traits*, (Sunderland, MA: Sinauer).

Michishita, Eriko, et al. (2005), 'Evolutionarily conserved and nonconserved cellular localizations and functions of human SIRT proteins', *Molecular Biology of the Cell*

Mol. Biol. Cell, 16 (10), 4623-35.

Miller, Craig T., et al. (2014), 'Modular Skeletal Evolution in Sticklebacks Is Controlled by Additive and Clustered Quantitative Trait Loci', *Genetics*, 197 (1), 405-20.

Mills, Robert H., et al. (2020), 'Organ-level protein networks as a reference for the host effects of the microbiome', *Genome Research*

Genome Res., 30 (2), 276-86.

Miquel, De Caceres and Legendre Pierre (2009), *Associations between species and groups of sites: indices and statistical inference*, (Ecology).

Mizrahi-Man, O, ER Davenport, and Y Gilad (2013), 'Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs.', *PLoS One*, 8 (1), e53608.

Moeller, Andrew H., et al. (2016), 'Cospeciation of gut microbiota with hominids', *Science (New York, N.Y.)*

Science, 353 (6297), 380-82.

Moeller, Andrew H., et al. (2019), 'Experimental Evidence for Adaptation to Species-Specific Gut Microbiota in House Mice', *mSphere*, 4

Moran, Nancy A. and Daniel B. Sloan (2015), 'The Hologenome Concept: Helpful or Hollow', *PLoS Biol.*, 13 (12), e1002311.

Morgan, Andrew P., et al. (2015), 'The Mouse Universal Genotyping Array: From Substrains to Subspecies', *G3: Genes$_{Genomes}$Genetics*

G3 (Bethesda), 6 (2), 263-79.

Morowitz, MJ, EM Carlisle, and JC Alverdy (2011), 'Contributions of intestinal bacteria to nutrition and metabolism in the critically ill.', *Surg Clin North Am*, 91 (4), 771-85, viii.

Moya, Andrés and Manuel Ferrer (2016), 'Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance', *Trends Microbiol.*, 24 (5), 402-13.

Mueller, Noel T, et al. (2015), 'The infant microbiome development: mom matters', *Trends Mol. Med.*, 21 (2), 109-17.

Mukherji, Atish, et al. (2013), 'Homeostasis in Intestinal Epithelium Is Orchestrated by the Circadian Clock and Microbiota Cues Transduced by TLRs', *Cell*, 153 (4), 812-27.

Nagpal, Ravinder, et al. (2020), 'Role of TRP Channels in Shaping the Gut Microbiome', *Pathogens*, 9

Naidu, AS, et al. (2002), 'Reduction of sulfide, ammonia compounds, and adhesion properties of Lactobacillus casei strain KE99 in vitro.', *Curr Microbiol*, 44 (3), 196-205.

Nakagawa, Takashi and Leonard Guarente (2009), 'Urea cycle regulation by mitochondrial sirtuin, SIRT5', *Aging*

Aging (Albany NY), 1 (6), 578-81.

Nakagawa, Takashi, et al. (2009), 'SIRT5 Deacetylates carbamoyl phosphate synthetase 1 and regulates the urea cycle', *Cell*, 137 (3), 560-70.

Nakahata, Y, et al. (2009), 'Circadian control of the NAD+ salvage pathway by CLOCK-SIRT1.', *Science*, 324 (5927), 654-57.

Nakamura, Yasuhiko, et al. (2008), 'Localization of mouse mitochondrial SIRT proteins: Shift of SIRT3 to nucleus by co-expression with SIRT5', *Biochem. Biophys. Res. Commun.*, 366 (1), 174-79.

Nan Xiao, Joshua Cook, Miaozhu Li (2018), 'Package 'ggsci'',

Neumann, Philipp-Alexander, et al. (2014), 'Gut Commensal Bacteria and Regional Wnt Gene Expression in the Proximal Versus Distal Colon', *The American Journal of Pathology*

Am J Pathol, 184 (3), 592-99.

Nicholson, Jeremy K., et al. (2012), 'Host-gut microbiota metabolic interactions', *Science (New York, N.Y.)*

Science, 336 (6086), 1262-67.

Nicole, M Davis, et al. (2017), 'Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data', *bioRxiv*, 221499.

Nishida, Yuya, et al. (2015), 'SIRT5 Regulates both Cytosolic and Mitochondrial Protein Malonylation with Glycolysis as a Major Target', *Mol. Cell.*, 59 (2), 321-32.

Nyholt, Dale R. (2019), 'matSpD local version - Statistical and Genomic Epidemiology Laboratory (SGEL)',

O'Connor, Annalouise, et al. (2014), 'Responsiveness of cardiometabolic-related microbiota to diet is influenced by host genetics', *Mammalian Genome*, 25 (11), 583-99.

Ochman, Howard, et al. (2010), 'Evolutionary relationships of wild hominids recapitulated by gut microbial communities', *PLoS Biol.*, 8 (11), e1000546.

Ogawa, Yukino, et al. (2020), 'Gut microbiota depletion by chronic antibiotic treatment alters the sleep/wake architecture and sleep EEG power spectra in mice', *Scientific Reports*

Sci Rep, 10

Org, Elin, et al. (2015), 'Genetic and environmental control of host-gut microbiota interactions', *Genome Research*

Genome Res., 25 (10), 1558-69.

Org, Elin and Aldons J. Lusis (2018), 'Using the natural variation of mouse populations to understand host-gut microbiome interactions', *Drug discovery today. Disease models*

Drug Discov Today Dis Models, 28 61-71.

Ott, SJ, et al. (2004), 'Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease', *Gut*, 53 (5), 685-93.

Page, Roderic D. M. (2003), *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, (University of Chicago Press) 364.

Page, Roderic DM (2006), 'Cospeciation',

Pallares, LF, et al. (2014), 'Use of a natural hybrid zone for genomewide association mapping of craniofacial traits in the house mouse.', *Mol Ecol*, 23 5756-70.

Pandey, Shubhi, Jagannath Maharana, and Arun K. Shukla (2019), 'The Gut Feeling: GPCRs Enlighten the Way', *Cell Host & Microbe*, 26 (2), 160-62.

Papa, Eliseo, et al. (2012), 'Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease', *PLoS One*, 7 (6), e39242.

Paradis, E and K Schliep (2019), 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.', *Bioinformatics*, 35 (3), 526-28.

Park, Jeongsoon, et al. (2013), 'SIRT5-Mediated Lysine Desuccinylation Impacts Diverse Metabolic Pathways', *Mol. Cell.*, 50 (6), 919-30.

Parker, Bianca J., et al. (2020), 'The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health', *Frontiers in Immunology*

Front. Immunol., 11

Parker, CC, et al. (2014), 'High-resolution genetic mapping of complex traits from a combined analysis of F2 and advanced intercross mice.', *Genetics*, 198 (1), 103-16.

Parks, DH, M Imelfort, and CT Skennerton... (2015), 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome ...*,

Peek, Clara B., et al. (2012), 'Nutrient sensing and the circadian clock', *Trends in Endocrinology & Metabolism*, 23 (7), 312-18.

Peier, Andrea, et al. (2009), 'The Antiobesity Effects of Centrally Administered Neuromedin U and Neuromedin S Are Mediated Predominantly by the Neuromedin U Receptor 2 (NMUR2)', *Endocrinology*, 150 (7), 3101-9.

Peng, Zhi, et al. (2020), 'The Gut Microbiome Is Associated with Clinical Response to Anti–PD-1/PD-L1 Immunotherapy in Gastrointestinal Cancer', *Cancer Immunology Research*

Cancer Immunol Res, 8 (10), 1251-61.

Pickard, JM, et al. (2017), 'Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease.', *Immunol. Rev.*, 279 (1), 70-89.

Piovesan, D, SCE Tosatto, and RD Finn (2019), 'The Pfam protein families database in 2019', *Nucleic acids ...*,

Polletta, L, et al. (2015), 'SIRT5 regulation of ammonia-induced autophagy and mitophagy.', *Autophagy*, 11 (2), 253-70.

Puente-Sánchez, F, N García-García, and J Tamames (2020), 'SQMtools: automated processing and visual analysis of 'omics data with R and anvi'o.', *BMC Bioinformatics*, 21 (1), 358.

Qin, Junjie, et al. (2012), 'A metagenome-wide association study of gut microbiota in type 2 diabetes', *Nature*, 490 (7418), 55-60.

Qin, Youwen, et al. (2020), 'Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort',

Rapp, K (1972), 'HAN-rotation, a new system for rigorous outbreeding', *Z. Versuchstierk.*, 14 133-42.

Rardin, M.J., et al. (2013), 'SIRT5 regulates the mitochondrial lysine succinylome and metabolic networks', *Cell Metabolism*, 18 (6), 920-33.

Rausch, P, et al. (2016), 'Analysis of factors contributing to variation in the C57BL/6J fecal microbiota across German animal facilities.', *Int. J. Med. Microbiol.*, 306 (5), 343-55.

Rausch, Philipp, et al. (2019), 'Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms', *Microbiome*, 7 (1),

Reese, Aspen T., et al. (2018), 'Microbial nitrogen limitation in the mammalian large intestine', *Nature microbiology*

Nat Microbiol, 3 (12), 1441-50.

Rehman, A, et al. (2016), 'Geographical patterns of the standing and active human gut microbiome in health and IBD.', *Gut*, 65 (2), 238-48.

Reichardt, Nicole, et al. (2018), 'Specific substrate-driven changes in human faecal microbiota composition contrast with functional redundancy in short-chain fatty acid production', *The ISME Journal*, 12 (2), 610-22.

Remigi, P, et al. (2016), 'Symbiosis within Symbiosis: Evolving Nitrogen-Fixing Legume Symbionts.', *Trends Microbiol.*, 24 (1), 63-75.

Richardson, Anthony J, Nest McKain, and R John Wallace (2013), 'Ammonia production by human faecal bacteria, and the enumeration, isolation and characterization of bacteria capable of growth on peptides and amino acids', *BMC Microbiol.*, 13 (1), 6.

Ricklin, D, et al. (2016), 'Complement component C3 - The "Swiss Army Knife" of innate immunity and host defense.', *Immunol. Rev.*, 274 (1), 33-58.

Rieseberg, Loren H, Margaret A Archer, and Robert K Wayne (1999), 'Transgressive segregation, adaptation and speciation', *Heredity*, 83 (4), 363-72.

Rios-Covian, D, et al. (2017), 'Shaping the Metabolism of Intestinal Bacteroides Population through Diet to Improve Human Health.', *Front Microbiol*, 8 376.

Robert C. King, William D. Stansfield, and Pamela K. Mulligan (2007), *A Dictionary of*

Genetics, (7 edn., Oxford University Press).

Rolig, AS, et al. (2015), 'Individual Members of the Microbiota Disproportionately Modulate Host Innate Immune Responses.', *Cell Host Microbe*, 18 (5), 613-20.

Rosenberg, E and I Zilber-Rosenberg (2018), 'The hologenome concept of evolution after 10 years.', *Microbiome*, 6 (1), 78.

Rosshart, Stephan P., et al. (2017), 'Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance', *Cell*, 171 (5), 1015-1028.e13.

Roth, TL, et al. (2019), 'Reduced Gut Microbiome Diversity and Metabolome Differences in Rhinoceros Species at Risk for Iron Overload Disorder.', *Front Microbiol*, 10 2291.

Rowland, Ian, et al. (2018), 'Gut microbiota functions: metabolism of nutrients and other food components', *European Journal of Nutrition*

Eur J Nutr, 57 (1), 1-24.

Rühlemann, Malte Christoph, et al. (2021), 'Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome', *Nat. Genet.*, 1-9.

Saito, Yumiko, et al. (1999), 'Molecular characterization of the melanin-concentrating-hormone receptor', *Nature*, 400 (6741), 265-69.

Sarkar, Amar, et al. (2020), 'The role of the microbiome in the neurobiology of social behaviour', *Biol. Rev.*, 95 (5), 1131-66.

Schmalenberger, A, F Schwieger, and CC Tebbe (2001), 'Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling.', *Appl. Environ. Microbiol.*, 67 (8), 3557-63.

Schmieder, Robert and Robert Edwards (2011), 'Quality control and preprocessing of metagenomic datasets', *Bioinformatics*, 27 (6), 863-64.

Seemann, T (2014), 'Prokka: rapid prokaryotic genome annotation.', *Bioinformatics*, 30 (14), 2068-69.

Sethi, JK and AJ Vidal-Puig (2008), 'Wnt signalling at the crossroads of nutritional regulation.', *Biochem. J.*, 416 (2), e11-3.

Shannon, Paul, et al. (2003), 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research*

Genome Res, 13 (11), 2498-504.

Shi, L and BP Tu (2015), 'Acetyl-CoA and the regulation of metabolism: mechanisms and consequences.', *Curr. Opin. Cell Biol.*, 33 125-31.

Sieber, CMK, et al. (2018), 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nature ...*,

Singh, Parul, et al. (2018), 'Elucidation of the anti-hyperammonemic mechanism of Lactobacillus amylovorus JBD401 by comparative genomic analysis', *BMC Genomics*, 19 (1),

Singh, Parul, et al. (2020), 'The potential role of vitamin D supplementation as a gut microbiota modifier in healthy individuals', *Scientific Reports*, 10 (1), 21641.

Škrabar, N, et al. (2018), 'Using the Mus musculus hybrid zone to assess covariation and genetic architecture of limb bone lengths.', *Mol Ecol Resour*, 18 (4), 908-21.

Smith, Ashley E., et al. (2019), 'Binge-Type Eating in Rats is Facilitated by Neuromedin U Receptor 2 in the Nucleus Accumbens and Ventral Tegmental Area', *Nutrients*, 11 (2), 327.

Snijders, Antoine M., et al. (2016), 'Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome', *Nature Microbiology*, 2 16221.

Spor, Aymé, Omry Koren, and Ruth Ley (2011), 'Unravelling the effects of the environment and host genotype on the gut microbiome', *Nature Reviews Microbiology*, 9 (4), 279-90.

Sriram, K and PA Insel (2018), 'G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs', *Mol. Pharmacol.*, 93 (4), 251-58.

Steffen Durinck, Paul T. Spellman, Ewan

Birney and Wolfgang Huber (2009), 'Mapping identifiers for the integration of genomic datasets with the

R/Bioconductor package biomaRt.', *Nature Protocols*, 4 1184-91.

Stewart, GS and CP Smith (2005), 'Urea nitrogen salvage mechanisms and their relevance to ruminants, non-ruminants and man.', *Nutr Res Rev*, 18 (1), 49-62.

Suzuki, TA (2017), 'Links between Natural Variation in the Microbiome and Host Fitness in Wild Mammals.', *Integr Comp Biol*, 57 (4), 756-69.

Suzuki, TA, et al. (2020), 'The gut microbiota and Bergmann's rule in wild house mice.', *Mol Ecol*, 29 (12), 2300-11.

Suzuki, Taichi A and Ruth E Ley (2020), 'The role of the microbiota in human genetic adaptation', *Science*, 370 (6521),

Suzuki, Taichi A., et al. (2019), 'Host genetic determinants of the gut microbiota of wild mice', *Molecular Ecology*, 28 (13), 3197-207.

Szklarczyk, Damian, et al. (2019), 'STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', *Nucleic Acids Research*

Nucleic Acids Res, 47 (D1), D607-13.

Tahara, Yu, et al. (2018), 'Gut microbiota-derived short chain fatty acids induce circadian clock entrainment in mouse peripheral tissue', *Scientific reports*, 8 (1), 1-12.

Tamames, J and F Puente-Sánchez (2018), 'SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline.', *Front Microbiol*, 9 3349.

Tanahashi, Yasuyuki, et al. (2009), 'Multiple muscarinic pathways mediate the suppression of voltage-gated Ca2+ channels in mouse intestinal smooth muscle cells', *Br. J. Pharmacol.*, 158 (8), 1874-83.

Taras, David, et al. (2002), 'Reclassification of Eubacterium formicigenerans Holdeman and Moore 1974 as Dorea formicigenerans gen. nov., comb. nov., and description of Dorea longicatena sp. nov., isolated from human faeces.', *International Journal of Systematic and Evolutionary Microbiology*, 52 (2), 423-28.

Thaiss, Christoph A., et al. (2014), 'Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis', *Cell*, 159 (3), 514-29.

Thaiss, Christoph A., Maayan Levy, and Eran Elinav (2015a), 'Chronobiomics: The Biological Clock as a New Principle in Host–Microbial Interactions', *PLoS Pathog.*, 11 (10), e1005113.

Thaiss, Christoph A., et al. (2015b), 'A day in the life of the meta-organism: diurnal rhythms of the intestinal microbiome and its host', *Gut Microbes*, 6 (2), 137-42.

Thaiss, Christoph A., et al. (2016), 'Microbiota Diurnal Rhythmicity Programs Host Transcriptome Oscillations', *Cell*, 167 (6), 1495-1510.e12.

Theis, KR, et al. (2016), 'Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes.', *mSystems*, 1 (2), e00028-16.

Thoendel, Matthew, et al. (2016), 'Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing', *J. Microbiol. Methods*, 127 141-45.

Thompson, J. N. (1994), *The Coevolutionary Process*, (University of Chicago Press).

Tian, Liang, et al. (2020), 'Deciphering functional redundancy in the human microbiome', *Nature Comm.*, 11 (1),

Toderici, M, et al. (2016), 'Identification of Regulatory Mutations in SERPINC1 Affecting Vitamin D Response Elements Associated with Antithrombin Deficiency.', *PLoS One*, 11 (3), e0152159.

Townsend, KL, et al. (2012), 'Bone morphogenetic protein 7 (BMP7) reverses obesity and regulates appetite through a central mTOR pathway.', *FASEB J.*, 26 (5), 2187-96.

Turnbaugh, Peter J., et al. (2009), 'A core gut microbiome in obese and lean twins', *Nature*, 457 (7228), 480-84.

Turnbaugh, PJ, et al. (2008), 'Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome.', *Cell Host Microbe*, 3 (4), 213-23.

Turner, Leslie M., Denise J. Schwahn, and Bettina Harr (2012), 'Reduced Male Fertility Is

Common but Highly Variable in Form and Severity in a Natural House Mouse Hybrid Zone', *Evolution*, 66 (2), 443-58.

Turner, Leslie M. and Bettina Harr (2014), 'Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions', *eLife*, 3 e02504.

Turpin, W., et al. (2016), 'Association of host genome with intestinal microbial composition in a large healthy cohort', *Nat. Genet.*, 48 (11), 1413-17.

Vaga, Stefania, et al. (2020), 'Compositional and functional differences of the mucosal microbiota along the intestine of healthy individuals', *Scientific Reports*, 10 (1),

Valerie, Obenchain, et al. (2014), 'VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants', *Bioinformatics*, 30 (14), 2076-78.

van Opstal, Edward J. and Seth R. Bordenstein (2015), 'Rethinking heritability of the microbiome', *Science*, 349 (6253), 1172-73.

Vaughn, Dawn (2010), 'Why run and hide when you can divide? Evidence for larval cloning and reduced larval size as an adaptive inducible defense', *Marine Biology*, 157 (6), 1301-12.

Velásquez-Mejía, EP, J de la Cuesta-Zuluaga, and JS Escobar (2018), 'Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces.', *Appl Microbiol Biotechnol*, 102 (1), 403-11.

Velloso, Licio A., Franco Folli, and Mario J. Saad (2015), 'TLR4 at the Crossroads of Nutrients, Gut Microbiota, and Metabolic Inflammation', *Endocrine Reviews*

Endocr Rev, 36 (3), 245-71.

Veneti, Zoe, et al. (2004), 'Heads or tails: host-parasite interactions in the Drosophila-Wolbachia system', *Appl. Environ. Microbiol.*, 70 (9), 5366-72.

Vince, AJ and SM Burridge (1980), 'Ammonia production by intestinal bacteria: the effects of lactose, lactulose and glucose.', *J. Med. Microbiol.*, 13 (2), 177-91.

Visscher, PM, et al. (2017), '10 Years of GWAS Discovery: Biology, Function, and Translation.', *Am. J. Hum. Genet.*, 101 (1), 5-22.

Voigt, Robin M., et al. (2014), 'Circadian Disorganization Alters Intestinal Microbiota', *PLoS One*, 9 (5), e97500.

Walpole, C, et al. (2018), 'Investigation of facilitative urea transporters in the human gastrointestinal tract.', *Physiol Rep*, 6 (15), e13826.

Walsh, Bruce and Michael Lynch (2018), *Evolution and selection of quantitative traits*, (Oxford University Press).

Wang, Jun, et al. (2015), 'Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome', *Nature Communications*

Nat Commun, 6

Wang, Jun, et al. (2016), 'Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota', *Nat. Genet.*, 48 (11), 1396-406.

Wang, Qiong, et al. (2007), 'Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy', *Appl. Environ. Microbiol.*, 73 (16), 5261-67.

Wang, Y, et al. (2010), 'Regional mucosa-associated microbiota determine physiological expression of TLR2 and TLR4 in murine colon.', *PLoS One*, 5 (10), e13607.

Wang, Yijie, et al. (2019), 'An overview of Sirtuins as potential therapeutic target: Structure, function and modulators', *European journal of medicinal chemistry*, 161 48-77.

Weldon, L, et al. (2015), 'The Gut Microbiota of Wild Mice.', *PLoS One*, 10 (8), e0134643.

Werren, John H, Laura Baldo, and Michael E Clark (2008), 'Wolbachia: master manipulators of invertebrate biology', *Nature Reviews Microbiology*, 6 (10), 741-51.

Whipps, JM, K Lewis, and RC Cooke (1988), 'Mycoparasitism and plant disease control', *Fungi in biological control systems*, 161-87.

Wickham, H and W Chang (2016), 'Package 'ggplot2', *Create Elegant Data ...*,

Winogradsky, Sergej Nikolaevič (1949), *Microbiologie du sol: problèmes et méthodes*, (Paris : Masson).

Woese, CR and GE Fox (1977), 'Phylogenetic structure of the prokaryotic domain: the primary kingdoms.', *Proc. Natl. Acad. Sci. U S A*, 74 (11), 5088-90.

Wood, Jason G., et al. (2018), 'Sirt4 is a mitochondrial regulator of metabolism and lifespan in <i>Drosophila melanogaster</i', *Proceedings of the National Academy of Sciences*, 115 (7), 1564-69.

Wrong, Oliver M. and Angela Vince (1984), 'Urea and ammonia metabolism in the human large intestine', *Proc. Nutr. Soc.*, 43 (1), 77-86.

Wu, Guangyan, et al. (2018), 'Light exposure influences the diurnal oscillation of gut microbiota in mice', *Biochemical and Biophysical Research Communications*

Biochem Biophys Res Commun, 501 (1), 16-23.

Wu, YW, BA Simmons, and SW Singer (2016), 'MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets.', *Bioinformatics*, 32 (4), 605-7.

Xu, C, et al. (2011), 'The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain.', *Nat Commun*, 2 227.

Xu, Fengzhe, et al. (2020), 'The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases', *Microbiome*, 8

Yang, M, et al. (2020a), 'Mucosal-Associated Microbiota Other Than Luminal Microbiota Has a Close Relationship With Diarrhea-Predominant Irritable Bowel Syndrome.', *Front. Cell. Infect. Microbiol.*, 10 515614.

Yang, Q, et al. (2020b), 'Role of Dietary Nutrients in the Modulation of Gut Microbiota: A Narrative Review.', *Nutrients*, 12 (2), E381.

Yap, Min, et al. (2020), 'Evaluation of methods for the reduction of contaminating host reads when performing shotgun metagenomic sequencing of the milk microbiome', *Scientific Reports*, 10 (1),

Yasuda, K, et al. (2021), 'Elucidation of metabolic pathways of 25-hydroxyvitamin D3 mediated by CYP24A1 and CYP3A using Cyp24a1 knockout rats generated by CRISPR/Cas9 system.', *J. Biol. Chem.*, 296 100668.

Yatsunenko, Tanya, et al. (2012), 'Human gut microbiome viewed across age and geography', *Nature*, 486 (7402), 222-27.

Ye, Y and TG Doak (2009), 'A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes', *PLoS Comput. Biol.*,

Yi, Z and GA Bishop (2015), 'Regulatory role of CD40 in obesity-induced insulin resistance.', *Adipocyte*, 4 (1), 65-69.

Youngblut, Nicholas D., et al. (2019), 'Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades', *Nature Comm.*, 10 (1), 2200.

Yu, Guangchuang, et al. (2012), 'clusterProfiler: an R package for comparing biological themes among gene clusters', *Omics: a journal of integrative biology*, 16 (5), 284-87.

Yu, Jiujiu, et al. (2013), 'Metabolic Characterization of a Sirt5 deficient mouse model', *Scientific Reports*

Sci Rep, 3

Zeevi, David, et al. (2019), 'Structural variation in the gut microbiome associates with host health', *Nature*, 568 (7750), 43-48.

Zhang, Bo, et al. (2020), 'Impact of bead-beating intensity on microbiome recovery in mouse and human stool: Optimization of DNA extraction',

Zhang, Ning and Anthony A Sauve (2018), 'Regulatory effects of NAD+ metabolic pathways on sirtuin activity', *Progress in molecular biology and translational science*, 154 71-104.

Zhou, Xiang and Matthew Stephens (2012), 'Genome-wide efficient mixed-model analysis for association studies', *Nat. Genet.*, 44 (7), 821-24.

Zilber-Rosenberg, Ilana and Eugene Rosenberg (2008), 'Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution', *FEMS Microbiology*

Reviews, 32 (5), 723-35.

Ziyatdinov, Andrey, et al. (2018), 'lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals', *BMC Bioinformatics*, 19 (1), 1-5.