Ecological and Evolutionary Forces Shaping Variation in the Wild Mouse Gut Microbiome

Dissertation
in fulfilment of the requirements for the degree

Doctor rerum naturalium

of the Faculty of Mathematics and Natural Sciences
at Kiel University

Jun Wang

Institute for Experimental Medicine
Christian-Albrechts-University Kiel
&
Evolutionary Genomics
Max Planck Institute for Evolutionary Biology

Table of Contents

Kurzfassung

Die Gesamtheit der Mikroorganismen (auch Mikrobiota oder Mikrobiom gennant), die Säugetiere und anderen Lebewesen bewohnen, sind essenziell für die Biologie ihrer Wirte. Das Mikrobiom kann den Stoffwechsel oder auch das Immunsystem der jeweiligen Wirte verändern, und die Verschiebung der Zusammensetzung der bakteriellen Gemeinschaften (Dysbiosis) kann zu Störungen von Struktur und/oder auch der Funktion dieser führen und Entzündungen oder Krankheiten zur Folge haben. Dies wurde bereits in vielen Studien in Menschen oder der Labormaus als Modellorganismus in verschiedenen Gesichtspunkten beleuchtet. Mäusen, die aus einer natürlichen Umgebung stammen (wilde Hausmäuse) wurde bisher erst wenig Beachtung geschenkt, obwohl diese dem Menschen in Ökologie und Biologie ähnlicher sind als Labormäuse. In meiner Doktorarbeit stellt die wilde europäische Hausmaus den Modellorganismus dar, anhand dessen einige wichtige Aspekte, die einen Einfluss auf das Darmmikrobiom haben, untersucht wurden. Mäuse von acht verschiedenen Regionen in Westeuropa wurden hinsichtlich verschiedener Faktoren, die das Darmmikrobiom beeinflussen, analysiert. Geographie stellte in diesem Zusammenhang den wichtigsten Parameter dar, der sowohl das Mikrobiom des Darmgewebes als auch des Darminhalts beeinflusst, aber auch Populationsstruktur und die genetische Distanz zwischen den Populationen haben einen Einfluss, wenn auch schwächer. Außerdem zeigt sich dieser Einfluss nur bei den Darmgewebe – assoziierten Gemeinschaften. Desweiteren umfasste diese Studie eine Analyse, die sich auf das Mikrobiom des Darminhalts bezieht. Die Ergebnisse sprechen dafür, dass es in wilden Mäusen sogenannte „Enterotypen“ gibt, was auch schon für Menschen und Schimpansen beschrieben wurde. Experimentell haben wir dies an neu gefangenen wilden Mäusen getestet und konnten feststellen, dass, nach Fütterung einheitlicher Nahrung, die Variation der Enterotypen verschwunden ist. Dass die Nahrung wesentlich zur Formierung der verschiedenen Enterotypen beiträgt, wurde anhand von metagenomischen Analysen und stabiler Isotopen Analyse, basierend auf der Rekonstruktion der Nahrung, festegestellt. Zusätzlich wurde in dieser Studie der Effekt von Hybridisierung des Wirts, in diesem Fall zweier Unterarten der Hausmaus (*Mus musculus musculus* und *M. m. domesticus*), auf das Darmmikrobiom untersucht. Die Mäuse stammen aus einem Bereich der Hybridzone in Bayern, zusätzlich wurden Kreuzungen verschiedener Laborstämme für diese Studie untersucht. Das Darmmikrobiom der Hybridmäuse unterscheidet sich deutlich von dem ihrer Elternarten, sowohl in natürlichen Hybriden als auch den gekreuzten Laborstämmen. Durch die Zuordnungen verschiedener Phänotypen zu verschiedenen Genloci (quantitative trait loci (QTL)) konnten transgressive Phäntoypen (Anzahl an Bakterien und die Diversität der Bakteriengemeinschaften), die das Hybrid- Mikrobiom

beeinflussen, identifiziert werden. Hinweise aus der Immun- und Histopathologie unterstützen noch unsere Vermutung, dass purifizierende Selektion auf den Erhalt einer funktionierenden Mikrobiota wirkt, während in Hybriden die gestörten Bakteriengemeinschaft Hinweise auf genetische Inkompatibilitäten geben und somit die Wirtsfitness beeinflussen. Abschließend wurde begonnen das Genom möglicher Wirts – assoziierter Bakterien zu sequenzieren. Bei *Staphylococcus epidermis*, einem bekannten Haut- assoziierten Bakterium konnte so eine genomische Variation in Bezug auf die Anpassung an den Wirt gezeigt werden.

Abstract

Microbial communities (*a.k.a*. the microbiota or microbiome) associated with mammals and other animals are essential to the biology of the host. The microbiome can modulate metabolism and immune functioning, and the disturbance of community structure and/or function can lead to dysbiosis, inflammation or disease. Extensive studies have been carried out in humans and lab mouse models with different emphases, while the study of mice in there natural environment (wild house mice) is yet under-represented, even though wild mice have a more similar ecology and biology to humans. In my thesis, the wild house mouse in Europe is used as a model for which numerous important factors influencing the gut microbiome are explored. First, we analyzed mice captured in eight different geographical regions of Western Europe and examined the relative importance of several components in shaping gut microbiome. We found geography to be the most significant factor influencing both the mucosal and luminal microbiome, with a comparatively weaker influence of host population structure and genetic distance, which was only significant in the mucosa-associated communities. Secondly, a separate analysis based on the luminal microbiome suggests the existence of "enterotypes" in wild mice, mirroring the findings in humans and chimpanzees. We experimentally tested newly captured wild mice and observed fast convergence to only one enterotype under a standardized diet. Additional functional metagenomic analysis and diet reconstruction based on stable isotope analysis both strongly support the determining role of diet in shaping enterotypes. In the third study, we focused on the effect of host hybridization on the gut microbiome between two subspecies of house mice (*Mus musculus musculus* and *M. m. domesticus*) from a transect of the hybrid zone in Bavaria, as well as artificial crosses between lab inbred strains. Hybrid mice have an intestinal microbiome distinct from their parental species in both setup, and with quantitative trait loci (QTL) mapping we identified transgressive phenotypes (bacterial abundances and comunity diversity) contributing to the distinct hybrid microbiome. Combined with additional immune and histopathological evidence, we propose that purifying selection acts to maintain a normal functioning microbiota, while in hybrids the distorted communities reflect genetic incompatibilities and may influence host fitness. Lastly, some primary efforts are devoted to studying candidate host-associated bacteria using a full genomic sequencing approach, where we begin to understand genomic variation of *Staphylococcus epidermidis*, a prominent skin-associated microbe, in the context of adaptation to the host.

Introduction

Prevalence and importance of microbes

Almost all known natural environments on earth are colonized by prokaryotes (bacteria and archaea), ranging from the most common (*e.g.* freshwater and marine environment, soil and air) to the extremities (*e.g.* hot springs, hydrothermal vents as well as acidic mines) (Ruff-Roberts *et al.*, 1994, Fife *et al.*, 2000, Dubilier *et al.*, 2001). The total number of these microorganisms is estimated to be at a magnitude of 10E30, and together they compose more biomass than animals and plants together (Hogan, 2010). Phylogenetically, bacteria and archaea form two out of the three kingdoms of life, occupying more than 90% of the branches in the tree of life, leaving only the remaining 10% to eukaryotes (McInerney *et al.*, 2008).

The importance of prokaryotes has become clear to researchers through extensive ecological studies: Cyanobacteria are the most important primary producers in oceans and freshwater systems (Ting *et al.*, 2002); bacteria compose the main final degraders in the terrestrial and aquatic food chain (Pomeroy *et al.*, 2007); archaea and bacteria are essential members for the biosynthesis and transformation of organic components (Whitman *et al.*, 1998), as well as the global cycling of fundamental elements like phosphate, sulfur and nitrogen (Zehr & Ward, 2002, Chu *et al.*, 2003, Dyhrman *et al.*, 2006). Recent investigations revealed even more biochemical potential of prokaryotes through genome sequencing of yet-uncultured environmental bacteria and archaea (Rinke *et al.*, 2013).

Besides studies focusing on environments at large, recent efforts also emphasized the prevalence and importance of microbes in host-associated micro-environments. It is well accepted that the organelles (chloroplasts and mitochondria) in eukaryotes have a bacterial origin and were acquired via endo-symbiosis (Thrash *et al.*, 2011, Chu *et al.*, 2004) at various stages of evolutionary history. Although higher plants do not usually have cavities for microbes to colonize as animals do, they nevertheless have specifically associated microbial communities in roots (Lundberg *et al.*, 2012) and external surfaces (phyllosphere) (Lopez-Velasco *et al.*, 2013, Ottesen *et al.*, 2013),

with an increasing recognition of their essential roles (Hirsch & Mauchline, 2012, Klein *et al.*, 2013).

For metazoans, specific associations with prokaryotes exist from simple organisms such as hydra (Franzenburg *et al.*, 2013) and sponges (Schmitt *et al.*, 2012) to complex vertebrates (Ley *et al.*, 2008) and contribute to a wide variety of biological functions. For example, symbionts can be primary producers for the host, such as symbiotic bacteria carrying out photosynthesis in corals (*Symbiodinium spp*.) (Sara *et al.*, 1998) or sponge symbionts performing chemoautotrophy in hydrothermal vents (Dubilier *et al.*, 2008). Bacteria can even contribute to host predatory behavior: *Vibrio spp.* produce bioluminescence regulated by quorum sensing system, which provides an advantage for predation by squids (Boettcher & Ruby, 1990). Further, gut microbes metabolize food to intermediate products to facilitate assimilation. Some fermentation products are of vital importance to the host (Egert *et al.*, 2006), whereas other animals such as termites solely rely on microbes to digest otherwise inedible food (Ohkuma *et al.*, 2007).

Not surprisingly, microbes also greatly influence the evolution of animals, especially in the sense of shaping the innate immune systems in most organisms (Bardoel & van Strijp, 2011) and adaptive immune systems in vertebrates (Pancer & Cooper, 2006). Finally, one major process of evolutionary importance, speciation, is directly influenced by microbes. The well-studied intracellular *Wolbachia* species cause cytoplasmic incompatibilities in *Drosophila* and other insects' zygotes (Zabalou *et al.*, 2004, Sinkins, 2004), and gut microbes modulate mate choice in *Drosophila* (Sharon *et al.*, 2010). A recent study carried out in *Nasonia* species reported the direct role of gut microbiome in preventing hybrid speciation, as the increase of certain bacterial genera in the hybrids greatly reduces hybrid fitness (Brucker & Bordenstein, 2013). A combination of immune deficiency in hybrids and consequent loss of control over bacteria, does not only lead to a distinct microbiome in hybrids, but also leads to higher mortality in the F1 generation.

Due to the widespread and still increasing recognition of microbial communities' involvement in host functioning and evolution, fundamental characteristics such as species richness (alpha diversity) and inter-individual variability (beta diversity) of

these communities have been addressed in a large range of studies, in particular those inhabiting the gut (Ley *et al.*, 2008). The development of high throughput sequencing including the 454 Roche and Illumina platforms greatly facilitated the understanding of taxonomical structure of the microbiome via 16S rRNA gene sequencing and functional genomic aspects by shot-gun metagenomic sequencing (Williams, 2011). Reports in basal organisms such as hydra and *Caenorhabditis elegans* have demonstrated effects on the microbiome from conserved cell signalling pathways (Franzenburg *et al.*, 2013, Portal-Celhay & Blaser, 2012). The natural ecology of gut bacteria are also widely studied in invertebrates, mainly in insects species, including honeybees (Martinson *et al.*, 2011, Fukatsu & Hosokawa, 2002), stinkbugs (Fukatsu & Hosokawa, 2002), wild *Drosophila* species (Chandler *et al.*, 2011) and European firebugs (Sudakaran *et al.*, 2012). In vertebrates, general patterns of gut bacteria are summarized and discussed in fish (Sullam *et al.*, 2012), pythons (Costello *et al.*, 2010), iguanas (Lankau *et al.*, 2012), chickens (Kim & Mundt, 2011), bats (Phillips *et al.*, 2012), cows (Jami & Mizrahi, 2012), sheep (Larue *et al.*, 2005) and even primates (Ochman *et al.*, 2010). While these studies provided individual insights into the microbial diversity of different animal groups, they are by and large observatory, and currently no consensus study has been performed. Meanwhile, the organisms with the highest effort ever devoted to microbiome research are humans (*Homo sapiens*) and house mice (*Mus musculus*).

Microbiota research in humans

Until the end of the last century, the majority of human-related microbiological studies were concerned with understanding disease and pathogens (Lederberg, 2000). The first hypothesis that microbes cause disease came from the 16th century when Fracastoro proposed that syphilis is caused tiny organisms (*Sive morbus Gallicus*), and in the 1930s the pathogenic agent, *Treponema pallidum*, was finally identified. The establishment of modern microbiology is largely attributed to Louis Pasteur and Robert Koch, and was greatly influenced by the onset of the antibiotic era with the discovery of penicillin by Alexander Fleming in 1928. This focus on medical microbiology is readily justified, as microbes also turned out to be some of the strongest selective forces in human history and greatly shaped the evolution and genomic landscape of humans. Repeated occurrences of Bubonic plague (black death)

reduced the European population by two thirds in recorded history (Bos *et al.*, 2011). *Mycobacterium* causes Tuberculosis and remains a major threat to human health and longevity in developing countries (Zaman, 2010). The plague and other diseases also drove a large proportion of Native Americans to extinction (Cook, 1973). This impact is recognized, among others, by the relatively new field of evolutionary medicine, which links vulnerability to disease to human evolution (Nesse & Williams, 1995), while other scholars have recognized that the epidemic of disease has even influenced the progress of social reform and history (Diamond, 1997).

The emergence of new pathogenic microbes in recent decades demonstrates the further need to understand pathogens, including the evolutionary processes that lead to their existence. Methicillin-resistant *Staphylococcus aureus* (MRSA) have spread throughout the globe beginning in the 1960s (Enright *et al.*, 2002), causing great casualties in hospitalized patients. It is recognized that the abuse of antibiotics has contributed to resistance evolving in MRSA (Wisplinghoff *et al.*, 2005, Deurenberg *et al.*, 2007), and the same conclusion is reached with regard to *Pseudomonas aeruginosa* (Mah *et al.*, 2003) and *Enterococcus faecalis* (Paulsen *et al.*, 2003). From an evolutionary genomic perspective, researchers also provided deeper insights into the processes contributing to the pathogenesis of these microbes. In the case of the EHEC (Enterohaemorrhagic *Escherichia coli*) outbreak in Germany in 2011, genome sequencing was performed within a few weeks (Mellmann *et al.*, 2011) and analysis demonstrated that horizontal gene transfer may have passed on lethal toxins to the originally commensal *E. coli* (Brzuszkiewicz *et al.*, 2011).

Meanwhile, the perception of pathogens has also changed more recently: due to new findings many bacteria have ceased to be recognized as solely pathogenic agents. *Helicobacter* was discovered to be the direct cause of stomach ulcers (Kroegel, 1994), but other studies revealed the anti-inflammatory properties of this species in the presence of other pathogenic bacteria (Every *et al.*, 2011). *Helicobacter* is also one the major bacteria that is maternally transmitted in humans, and their phylotypes have accordingly been used to track human migration routes (Falush *et al.*, 2003, Vale *et al.*, 2009, Yamaoka, 2009). *Staphylococcus epidermidis* is another important bacterial species colonizing animal skin, and in hospitals they are found in increasing cases of infections, facilitated by their abilities to attach to different surfaces including surgical

instruments (Conlan *et al.*, 2012, Widerstrom *et al.*, 2012). However, in human nasal cavities they can in fact inhibit infectious *Staphylococcus aureus* (Iwase *et al.*, 2010). More importantly, in increasing cases the expression of pathogenic factors and inflammation or infection by "pathogenic" microbes is shown to be context-related, whereby the whole microbial community (microbiome) determines the outcome of pathogens, or rather "potential/opportunistic pathogens" (Gilmore & Ferretti, 2003).

The human microbiome contains a collection of prokaryotes colonizing all of our exposed surfaces (skin, gastro-intestinal tract, respiratory system and reproductive organs). The skin is dominated by gram-positive bacteria (Actinobacteria and Firmicutes), and changes in its flora are associated with the health status of the host (Grice *et al.*, 2009, Kong *et al.*, 2012, Ravel *et al.*, 2011). Large-scale sampling has also been performed in the vaginal tract (White *et al.*, 2011a), which revealed the importance of the microbiome against diseases. Furthermore, during pregnancy the changes of the vaginal microbiome appear to be correlated with energy intake (Koren *et al.*, 2012). Lungs were thought to be a sterile environment except patients with cystic fibrosis (CF), whose lungs are colonized by dense biofilms primarily made up of *Pseudomonas* and other pathogens (Morris *et al.*, 2013). Recent findings, however, revealed that the lungs also have a relatively simple, but important microbial community, whose structures are affected by smoking (Morris *et al.*, 2013) or diseases such as chronic obstructive pulmonary disease (COPD)(Erb-Downward *et al.*, 2011).

The largest effort by far, however, has been directed towards the gut/fecal microbiome, the largest of microbial communities associated with humans. The total cell number is enormous (10E11) (Qin *et al.*, 2010), and functional metagenomics has revealed its gene content to be a magnitude greater than the host itself (Qin *et al.*, 2010). Large international consortia have been established including the Human Microbiome Project (Turnbaugh *et al.*, 2007) and Meta-Hit to name a few; and the taxonomical as well as genetic landscapes have been investigated by these combined efforts (Gevers *et al.*, 2012, Schloissnig *et al.*, 2013). Meanwhile, controversial conclusions were arrived at by several studies, for instance the three “universal clusters” of the human gut microbiome, termed "enterotypes" (Arumugam *et al.*, 2011). Inconclusive evidence of long-term diet influences on enterotypes exists (Wu

et al., 2011) but the concept continues to be debated by new studies and analysis (Moeller *et al.*, 2012, Koren *et al.*, 2013).

Through human microbiome studies, we have gained knowledge about the functional importance of the normal gut microbiome in human metabolism and immune defense (Maslowski & Mackay, 2011). Bacteria digest large molecules into smaller substances and facilitate assimilation, for instance long-chain- to short-chain fatty acids and complex carbohydrates into simpler sugars. Secondary metabolic products are also of vital importance, including vitamins (Tilg & Kaser, 2011). Microbes are vital during the early stages of infancy for the development of immune system, as studies suggest children born via Cesarean section (and thus not acquiring the same first inoculate of microbes compared to infants with a normal birth) tend to have compromised immune functioning in later in life (Azad *et al.*, 2013). Similarly, bacterial components are required for the development and homeostasis of a range of immune cells (Hill & Artis, 2010).

The delicate balance between microbes and host can be easily disturbed, and consequent dysbiosis can lead to deterioration of human health, infection or chronic inflammation. Dysbiosis of the intestinal microbiota has long been suspect in the etiology of inflammatory bowel diseases such as Crohn's disease (Ott *et al*., 2004, Erickson *et al.*, 2012). Despite the function of resisting pathogens in the context of a whole community (Stecher *et al.*, 2010), some of the microbes are by nature potential pathogens (*e.g. Helicobacter, Clostridium*). When the normal microbiota changes due to antibiotic usage or host factors, they can quickly replace the normal microbiota and express virulence factors (Lupp *et al.*, 2007, Qin *et al.*, 2012). Structural shifts in the community, even without detectable pathogens or pathology, could lead to detrimental human health conditions, *e.g*. obesity can be caused by shifts in the Firmicutes/Bacteroidetes ratio influencing energy intake (Turnbaugh *et al.*, 2008), and shifts in certain bacterial taxa can also lead to higher susceptibility to diabetes (Qin *et al.*, 2012).

Microbiota research in mice

While human microbiome studies are largely observational, experimental work has been carried out extensively in lab animals, including rats (*Rattus norvegicus*) (Manichanh *et al.*, 2010), guinea pigs (*Cavia porcellus*) (Hildebrand *et al.*, 2012) and pigs (*Sus scrofa*) (Lykke *et al.*, 2013), but with the vast majority being carried out in lab mice (*Mus musculus*, comprised predominantly of the *domesticus* subspecies). Gnotobiotic mice (mice that are made germ free or with known species of bacteria) or specific-pathogen free mice (mice that are devoid of selected pathogens) have provided a relatively simple system to study the different aspects of the microbiome (Backhed *et al.*, 2004). We have learned from lab mice that gut microbes are essential for digestion, as germ free mice have difficulties in food assimilation, while artificially inoculating a cocktail of a defined bacteria mixture (*e.g.* the altered Schaedler flora) greatly reduces such problems (Coates, 1975). The lack of essential substances in early life stages can greatly influence brain development and eventually affect behaviors in germ-free mice (Heijtza *et al.*, 2011). Also, it has been observed that the existence of bacteria is required for the development of intestinal mucosal tissues and their functioning, as germ free mice have deformed intestines (Smith *et al.*, 2007).

Mice serve as experimental models for a variety of diseases and infections by pathogens. *Salmonella typhi* (Mathur *et al.*, 2012), *Clostridium difficile* (Su *et al.*, 1986), and *Citrobacter rodentium* (D'Arienzo *et al.*, 2006) infection models were successfully established in germ-free mice or SPF mice. Not only the pathologic factors of bacteria were discovered and experimentally manipulated, the complex cellular or immunological pathways that are involved in infection/inflammation were also revealed under different scenarios (for instance TLR4 (Ubeda *et al.*, 2012) and NOD2 (Rehman *et al.*, 2011)). Our understanding is also improved for the *in vivo* functioning of previously neglected or newly discovered bacterial groups, for instance segmented filamentous bacteria (SFB) (Prakash *et al.*, 2011, Sczesnak *et al.*, 2011), *Fusobacterium* (Han et al., 2004), Lachnospriceae (Reeves *et al.*, 2012) and *Barnesiella* (Ubeda *et al.*, 2013), which can be either be pro-inflammatory or pro-biotic (preventing infection or inflammation).

The dysbiosis of the human microbiome leading to deterioration of human health has been replicated in lab mice. The patterns of the microbiome between obese- and lean

mice mirror those in humans (Ley *et al.*, 2005, Turnbaugh *et al.*, 2006), where experimental transplantation confirmed a direct role of the microbiome (Koren *et al*., 2012). Changes in the mouse gut microbiome were shown to be correlated with susceptibility to diabetes (Boerner & Sarvetnick, 2011). Other studies lead to a deeper understanding of microbiome diversity and functionality in mice, as Stecher *et al*. (2010) showed that the diversity of the fecal community has a direct effect on the resistance against pathogens. Lastly, mice share a majority of taxa in their microbiome with humans (Ley *et al.,* 2005), as well as general patterns such as enterotypes (Hildebrand *et al.*, 2013), which refer to a number of discrete groups of individuals' intestinal bacterial communities characterized by differences in abundance of major bacterial taxa. However, the number of clusters in mice (two) is different than that of humans (Arumugam *et al.,* 2011) and chimpanzees (Moeller *et al.,* 2012) (in both cases three), and the leading cause of enterotypes in lab mice is suggested to be low-level inflammation rather than diet (Hildebrand *et al*., 2013).

The possibility of manipulating both the environment and genetics in mice enabled further tests of the factors structuring microbiomes. Dietary intervention suggests even minor nutritional components to influence the whole community, *e.g.* heme from red meat can change microbial communities and host health (IJssennagger *et al.*, 2012), while *e.g.* a calorie-restricted diet is correlated to the abundance of *Lactobacillus* and may influence mouse lifespan (Zhang *et al.*, 2013). At the same time, gene knock-out mice models have confirmed some observations made in human disease studies, that genetic mutations correlated to disease lead to differences in the microbiome, as in the examples of NOD2 (Rehman *et al.,* 2011) and TLR/MyD88 (Grivennikov *et al.*, 2012). Other studies revealed deeper and more complex interactions between host genetic and environmental factors (*e.g.* diet) that eventually affect the host health (Kashyap *et al.*, 2013).

Recent studies explored the genetic diversity among lab mice and its role in shaping the microbiome using a quantitative trait loci (QTL) approach. Benson *et al.* (2010) reported the first microbial QTL study in mice between the HR (an ICR-derived outbred line) and C57BL/6J strains, and identified 13 genomic regions correlated with the abundances of 26 core-measurable-microbiota (CMMs) as traits. Another mapping between the BL6 and DBA/2J strains revealed the correlation of six CMMs to five

genomic regions (McKnite *et al.*, 2012). Between the two studies, several bacterial taxa were identified to be genetically regulated, and the regions of overlap might be hotspots for genes with essential roles in host-microbe cross-talk. Interestingly, a different study primarily aimed at genome mapping of autoimmune skin disease also revealed several loci correlated to the abundances of major skin bacterial taxa (Srinivas *et al.*, 2013), and several genomic regions were discovered to regulate bacterial abundances in both the gut and skin.

Wild mice as an emerging model

A collection of questions, however, remains unanswered throughout the pioneering studies carried out in humans and lab mice. Spor *et al.* (2011) have summarized the great knowledge collected from available studies at the time, and pointed out the importance of "stochastic processes" influencing the structure and function of the microbiome in humans. A considerable proportion of such stochastic forces, however, can still be attributed to unstudied environmental factors, for instance geographical differences, dietary patterns and other lifestyle-related factors of the host. Thus far, single studies in humans have not generated enough data spanning a large enough region, and only a few studies have considered the effect of diet (Wu *et al.*, 2011) and age (Yatsunenko *et al.*, 2012), while the collection of other information (including genetic information) from participants often faces obstacles from practical and legal limitations. A recently published meta-study managed to organize a few thousand gut microbiome samples from humans, and reached the first general overview of geographical effects (Lozupone *et al.*, 2013), although leaving the other components largely unexplored.

Lab mice also have limited capabilities to address the hypotheses surrounding the complex structures/interactions of the microbiome with hosts like humans. Germ-free mice, as mentioned above, have difficulties in intestinal mucosa development, immune function and metabolism (Coates, 1975); and even though SPF- and conventionally reared mice are typically inoculated with altered Schaedler flora and further colonized by other bacteria over their lifetime, as well as over further generations of husbandry, the diversity of lab mice is still considerably lower than that of wild mice, as presented in the first chapter of this thesis (Linnenbrink *et al.*, 2013).

The synopsis by Martin *et al.* (2010) has summarized the inferior health status of lab mice and the important consequences of this to biomedical studies, which is suggested, in the first chapter of this thesis, to be at least partially due to a lack of important components of the microbiome (Linnenbrink *et al.,* 2013). Furthermore, most strains of lab mice have low genomic diversity due to inbreeding, while at the same time the most frequently used mouse strains (*e.g.* C57BL/6), represent genomic mosaics of multiple subspecies of the *Mus musculus* complex, contrasting the common thought of a single *M. m. domesticus* origin (Keane *et al.*, 2011). Such deviations from typical free-living mammalian hosts eventually pose problems for translating knowledge from model organisms to medical applications.

It has been brought to attention how little we actually know about the natural ecology of many model organisms, until some recent improvements were made in natural populations of *Drosophila melanogaster* (Nestel *et al.*, 2008), *Caenorhabditis elegans* (Chen *et al.*, 2006), *Daphania magna* (Coors *et al.*, 2009) and *Mus musculus* (Abolins *et al.*, 2011, Linnenbrink *et al.,* 2013). Only a particular few of studies have focused on the microbiome of natural counterparts of model organisms, for instance in *Drosophila* (Chandler et al., 2011, Staubach *et al.*, 2013). In the majority of this thesis (Chapters I-III), I focus on wild populations of *Mus musculus* to address several questions regarding the structure and functionality of the intestinal microbiome.

The *Mus musculus* species complex in nature is composed of five subspecies: *domesticus* occupying Western Europe, North Africa, Iran and the whole new world; *musculus* occupying Eastern Europe through the majority of Asia; *castaneus* in Southeast Asia and *bactrianus* in Central Asia, plus a later discovered subspecies *gentilulus* (Prager *et al.*, 1998). The *Mus* genus originated in the area of India-Pakistan and later migrated through the Eurasian continent, during which *Mus musculus* developed tight associations to humans, earning the name "house mouse" (Silver, 1995). To some extent, *Mus musculus*' migration (especially *domesticus* and *musculus*) reflects human migration history during the expansion of agriculture, and they have a high degree of similarity in natural ecology with humans (for instance diet, selective pressures from parasites and pathogens).

In the first two chapters, I demonstrate the use of natural populations of the western house mice (*domesticus*) to understand the forces shaping variation in the gut microbiome between individuals. Through the study of intestinal mucosa-associated and luminal content microbiota in eight wild mouse populations, I provide the first large-scale profiling of the gut microbiome of wild mice. By comparing beta-diversity (inter-individual variability) to other environmental factors as well as host genetics, I can conclude that geography is a major driving force of gut microbiome diversification in the wild, the effect of which is mainly confined to a continental scale of geographic isolation. Genetic distance plays a comparatively smaller role and is only detectable in the mucosal microbiome, not in the luminal contents. However, strong support for enterotypes in the luminal content microbiome of wild mice was identified. By comparing wild mice to those kept in the lab for a year, evidence of a dietary influence on forming enterotypes is revealed. Next, stable isotope analysis was conducted in wild-caught mice. Taken together, the data again supports a an effect of diet leading to enterotypes, particularly with respect to the ratio of protein versus carbohydrates, confirming the long-standing hypothesis proposed for human enterotypes (Wu *et al.,* 2011).

In the third chapter I focus on the naturally occurring hybrid zone between *M.m.domesticus* and *M.m.musculus* found through Central Europe. This hybrid zone has already provided insight into the process of speciation, as it is more representative of the early stages of speciation than other systems with complete reproductive isolation (Good *et al.*, 2008). The maintenance of the hybrid zone relies on the balance of migration of pure subspecies and selection against hybrids (Barton & Hewitt, 1985), leaving interesting questions open for the nature of the selective forces against hybrids, as reproductive isolation is not complete and can not maintain the hybrid zone alone. Different genetic components of hybrid sterility have been screened in wild populations (Good et al., 2008, Teeter *et al.*, 2010) and via QTL studies using crosses of wild-derived strains from both species (*musculus*: PWD; *domesticus*: WSB) (White *et al.*, 2011b, Dzur-Gejdosova *et al.*, 2012). Parasite load is also suggested to contribute to the lower fitness of hybrids (Sage *et al.*, 1986) and is considered to be the result of a compromised immune system (Moulia *et al.*, 1991), although there is inconsistency between studies (Baird *et al.*, 2012). In the third chapter I focus on the gut microbiome of mice in the hybrid zone as well as lab

crosses between PWD and WSB, where in both cases we found a distinct microbiome in hybrids compared to the parental subspecies. The hybrid gut microbiome has lower diversity in both systems, which is linked to altered immune function in hybrids and greater frequencies of intestinal pathology. Next, QTL analysis was performed to investigate the genetic architecture of hybrid microbiomes using an F2 intercross between the PWD and WSB inbred strains. This revealed numerous genome segments that lead to transgressive phenotypes, whereby inter-specific heterozygotic genotypes tend to have more extreme values compared to homozygotic genotypes. This suggests that purifying selection maintains a normal gut microbiome in both subspecies, while diverged genomes cause conflicts in hybrids that lead to a compromised immune system and microbiome, which may in turn lower fitness. This is the first study suggesting a role of the gut microbiome in a vertebrate speciation processes, which is until now only reported in invertebrates (Brucker & Bordenstein, 2013).

Bacterial genomic insights into microbe-host interaction

Investigation into the genomes of pathogens has gained extensive knowledge for medical studies. Some components are directly involved in infection or inflammation, like shiga toxins in *E. coli* (Tu *et al.*, 2007), hemeolysin and super toxin in *Staphylococcus* species (Gouaux *et al.*, 1997). Others mimic human molecules and avoid human immune defense, for instance *Francisella tularensis* (Champion, 2011) and *Neisseria gonorrhoeae / meningitidis* (Mandrell & Apicella, 1993), while *Porphyromonas gingivalis* modifies host proteins and eventually leads to the development of Rheumatoid arthritis (Mikuls *et al.*, 2012). Regulatory systems such as quorum sensing do not directly cause pathology, but rather modulate the expression of pathogenic factors and indirectly increase chances of successful invasions and infection (Bottomley *et al.*, 2007). Recent studies even go beyond the analysis of static genomes and report important dynamics in pathogenic genomes. De Paepe *et al.* (2011) report the rapid diversification of *E. coli* genomes within the mouse gut; and horizontal gene transfer (HGT), already known for its importance in acquiring pathogenic genes within bacteria (Vinatzer *et al.*, 2012, Anderson & Seifert, 2011), was shown to increase in frequency between pathogens and commensals during inflammation (Stecher *et al.*, 2012).

Neutral or probiotic components in bacterial genomes have also come to researchers' attention. *Helicobacter* is one of the major maternally transmitted gut bacteria, and accordingly, its patterns of genomic differentiation between strains sampled in different populations mirror host migration routes and population genetic structure (Falush *et al.,* 2003, Graham & Yamaoka, 2006). Antibacterial peptides are widely produced in gram-positive bacteria (de Vos *et al.*, 1995, Kovacs *et al.*, 2006), as well as other proteins like serine proteases (*esp* gene) produced by *Staphylococcus epidermidis* (Iwase *et al.*, 2010). Current genomic research not only focuses on long-known probiotic bacteria, for instance *Lactobacillus reuteri* (Mackos *et al.*, 2013, Eaton *et al.*, 2011), but also newly discovered potential probiotics (Chu *et al.*, 2011, Koenen *et al.*, 2004).

As described above, numerous bacterial infection models have been established in mice to study the details of pathology and bacteria-host interactions. However, the bacterial pathogens in these models are usually strains isolated from human patients (for instance *Staphylococcus aureus*), rather than the native strains inhabiting mice, and it has been shown that there are discrepancies between human and mouse strains regarding pathogenic host responses (Colton & Kosoy, 2013). Acquiring native strains of interest from mice and characterizing them by performing genome sequencing may lead to more realistic models of bacteria-host interactions, not only for pathogenic bacteria, but also beneficial bacterial strains. In the context of the studies of lab- as well as wild mice in the first three chapters, I identified several bacterial genera or species as interesting candidates warranting deeper analysis. In wild mice we identified the major members of the wild mouse gut microbiome: *Bacteroides*, *Robinsoniella* and *Helicobater*, which may be of essential importance for host homeostasis (Linnenbrink *et al.,* 2013). In the enterotype study (chapter II) we identified *Bacteroides* and *Robinsoniella* as signature genera for enterotypes. Lastly, we observed interesting patterns with respect to inflammation in the cecum, especially lower *Barnesiella* and higher *Mucisprillum* in wild and lab mice (chapter III), and we are indeed in the process of cultivating these candidate bacteria genera.

Chapter IV of this thesis will describe the genome sequencing and analysis of *Staphylococcus epidermidis*, as a primer for understanding host-microbiome interactions from bacterial genomic point of view. In a QTL study of autoimmune

blistering and the microbiota inhabiting mouse skin, we mapped genomic regions contributing to the disease while simultaneously considering the skin microbiome (supplement for this chapter, Srinivas *et al*., 2013). One major genus of the mouse skin microbiota, *Staphylococcus,* showed signs of anti-inflammatory activities, since mice with higher abundance of this genus have lower frequency of this disease, even though they have the genotypes that are susceptible to autoimmune skin blistering. I followed up on this study by isolating multiple *Staphylococcus* strains from the stock collection of wild mouse species and subspecies at the Max Planck Institute for Evolutionary Biology in Plön and carrying out genome sequencing and annotation of their genomes. These will serve as a basis for future comparative genomics studies to look for signatures of selection and identify genes related to adaptation to the host as possible reasons for the probiotic potential against skin disease in mice.

Chapter I Profiling gut microbiota in wild mice

Introduction

The microbial communities inhabiting the mammalian intestinal tract play an important role in diverse aspects of host biology and likely serve as an important target of selection. An example of a host gene that is subject to strong selective forces in nature and influences the gut microbiota is the *B4galnt2* gene (Linnenbrink *et al.,* 2011, Staubach *et al.,* 2012). Understanding the precise phenotype(s) under selection with regard to *B4galnt2* in the wild is an ongoing objective of the Baines lab. However, little is known about the forces shaping the gut microbiome of house mice in their natural habitats, thus, more fundamental description of these communities is necessary in order to disentangle *e.g*. selectively-relevant vs. neutral variation.

To establish a baseline for wild mice microbiome studies and shed light on the contributions of host genetics, transmission and geography to diversity in microbial communities between individuals, we performed a survey of intestinal microbial communities in a panel of 121 house mice derived from eight locations across western Europe using pyrosequencing of the bacterial 16S rRNA gene. The host factors studied included population structure estimated by microsatellite loci and mitochondrial DNA, genetic distance and geography. To determine whether host tissue (mucosa)-associated communities display properties distinct from those of the lumen, both the cecal mucosa and contents were examined.

We identified *Bacteroides*, *Robinsoniella* and *Helicobacter* as the most abundant genera in both the cecal content and mucosa-associated communities of wild house mice. Overall we found geography to be the most significant factor explaining patterns of diversity in the intestinal microbiota, with a comparatively weaker influence of host population structure and genetic distance. Furthermore, the influence of host genetic distance was limited to the mucosa communities, consistent with this environment being more intimately coupled to the host. This study provided insights

into essential questions regarding the gut microbiome in a model organism of critical importance.

Publication

Linnenbrink M*, Wang J*, Hardouin E, Künzel S, Metzler D, Baines JF, (2013). **The role of biogeography in shaping diversity of the intestinal microbiota in house mice.** Mol. Ecol. 22(7):1904-16. *Equal author contribution

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 1904–1916 doi: 10.1111/mec.12206

The role of biogeography in shaping diversity of the intestinal microbiota in house mice

MIRIAM LINNENBRINK,*[1] JUN WANG,*†[1] EMILIE A. HARDOUIN,‡ SVEN KÜNZEL,* DIRK METZLER§ and JOHN F. BAINES*†

*Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, D-24306 Plön, Germany, †Institute for Experimental Medicine, Christian-Albrechts-University (CAU) Kiel, Arnold-Heller-Str. 3, D-24105 Kiel, Germany, ‡School of Applied Sciences, Bournemouth University, Christchurch House, Talbot Campus, Poole, Dorset BH12 5BB, UK, §University of Munich, Grosshadernerstr. 2, 82152 Planegg-Martinsried, Germany

Abstract

The microbial communities inhabiting the mammalian intestinal tract play an important role in diverse aspects of host biology. However, little is known regarding the forces shaping variation in these communities and their influence on host fitness. To shed light on the contributions of host genetics, transmission and geography to diversity in microbial communities between individuals, we performed a survey of intestinal microbial communities in a panel of 121 house mice derived from eight locations across Western Europe using pyrosequencing of the bacterial 16S rRNA gene. The host factors studied included population structure estimated by microsatellite loci and mitochondrial DNA, genetic distance and geography. To determine whether host tissue (mucosa)-associated communities display properties distinct from those of the lumen, both the caecal mucosa and contents were examined. We identified *Bacteroides*, *Robinsoniella* and *Helicobacter* as the most abundant genera in both the caecal content and mucosa-associated communities of wild house mice. Overall, we found geography to be the most significant factor explaining patterns of diversity in the intestinal microbiota, with a comparatively weaker influence of host population structure and genetic distance. Furthermore, the influence of host genetic distance was limited to the mucosa communities, consistent with this environment being more intimately coupled to the host.

Keywords: biogeography, house mouse, intestinal microbiota, *Mus musculus* population structure

Received 9 March 2012; revision received 6 December 2012; accepted 11 December 2012

Introduction

Together with pioneering gnotobiotic animal studies (Bäckhed *et al.* 2004; Rawls *et al.* 2006; Turnbaugh *et al.* 2009), the recent expansion of culture-independent analyses of bacterial communities has highlighted the biological significance of those inhabiting vertebrate hosts, in particular the intestinal tract, leading some to consider the intestinal microbiota a 'forgotten organ' (O'Hara & Shanahan 2006) or a 'malleable third genome' (Carroll *et al.* 2009). Although the majority of the details surrounding host microbiota interactions have yet to be described, it is clear that the microbiota play an important role in diverse processes of host biology including the metabolism of nutrients and organic substrates (Hooper *et al.* 2002), the development of the intestinal epithelium (Falk *et al.* 1998), the detection of foreign pathogens (Stecher *et al.* 2010), the maturation and development of the immune system (Hooper 2001; Round & Mazmanian 2009) and even brain development and behaviour (Heijtz *et al.* 2011).

The intestinal microbiota are highly diverse, composed of *c.* 800 species and *c.* 10-fold more strains (Bäckhed *et al.* 2005), and despite serving important

Correspondence: John F. Baines, Fax: +49(0) 431 597 3730; E-mail: j.baines@iem.uni-kiel.de

[1]Authors contributed equally.

functions display considerable variation between individuals (Eckburg *et al.* 2005). While gross imbalances in these communities (dysbioses) are linked to inflammatory bowel disease (IBD) (Ott *et al.* 2004; Round & Mazmanian 2009), little is known regarding the impact of the 'normal' range of variation on host fitness, nor regarding the relative roles of the contributing factors to this variation, be they genetic and/or environmental (Ley *et al.* 2006). Comparative studies between humans differing by varying degrees of genetic relatedness suggest a strong influence of host genotype (Zoetendal *et al.* 2001), as does the codiversification of individual lineages together with their hosts (Falush *et al.* 2003a). However, individual phylotypes can also be found in multiple host species, suggesting they may be more promiscuous/environmentally acquired (Ley *et al.* 2008). Thus, it is likely that bacteria colonizing the intestinal environment exhibit different specificities, ranging from the highly specialized to the promiscuous. The relative contribution of these two strategies to overall community composition remains unclear, but a preponderance of specialized or promiscuous bacteria would generally be expected to display a greater influence of host lineage or the environment, respectively.

Previous studies of gut microbiota in various natural hosts demonstrate the importance of both environmental and host factors in shaping diversity. For instance, geographic isolation has been shown to contribute to variation in the intestinal communities of iguanas (Lankau *et al.* 2012), bats (Phillips *et al.* 2012) and European firebugs (Sudakaran *et al.* 2012). These studies also provide insight into the influence of host ecology and diet, which are known to be important in shaping the gut microbiota of mammals (Ley *et al.* 2008; Muegge *et al.* 2011) as well as fish (Sullam *et al.* 2012). Likewise, the significance of the host environment is emphasized by a recent study of humans and four species of great apes, which revealed a correspondence between host phylogeny based on mitochondrial DNA (mtDNA) and their faecal microbiota (Ochman *et al.* 2010). However, the mtDNA phylogeny explained only 25% of the variation in the microbial community tree and other factors such as geography within a host species were not thoroughly examined. Thus, while these studies highlight the importance of these individual factors in shaping diversity between individuals, the combined influence of host genetics and geography within a species remains largely unexplored.

In this study, we have performed a large biogeographical survey of natural *Mus musculus domesticus* populations to simultaneously assess the contribution of host genetics, maternal transmission and geography to the diversity in intestinal microbial communities *between* individuals, that is, beta-diversity. For each of 121 individual mice sampled from eight populations spanning a continuous area of Germany and France, we collected data from a panel of microsatellite loci, sequenced the mitochondrial D-loop and profiled the caecal mucosa-associated microbiota by performing barcoded 454 pyrosequencing of the bacterial 16S rRNA gene. Due to a more intimate association with the host, mucosal-associated communities are hypothesized to have a greater dependency on host genetics compared with faecal or caecal content communities (Spor *et al.* 2011). To test this possibility and enable a comparison of wild mouse communities to the greater number of previous studies focusing on faeces or caecal contents (Reviewed by Spor *et al.* (2011)), we additionally profiled the luminal microbiota for the subset of 80 mice that contained appreciable caecal contents. Analysis of host population structure and geography revealed both variables to be significantly correlated with the patterns of divergence in the composition and structure of the intestinal microbiota, with geography being a stronger determinant.

Materials and methods

Animal material and tissue sampling

In total, 121 adult mice were sampled in barns and stables in eight geographic locations throughout a continuous area of Germany and France in the summers of 2009 and 2010, although ultimately a subset of 80 mice were used for much of the analysis based on the presence of luminal contents in the caecum (See below; Table 1; Appendix S1). Analysis of gut microbiota composition was performed on all samples, while the analyses with respect to geography and host genetics were limited to the 80 mice with luminal contents to perform direct comparisons between the mucosa and contents. The sampling strategy described by Ihle *et al.* 2006 was followed by maintaining a minimum distance of one kilometre between sampling sites and using only one mouse from each sampling site for analysis. Thus, 121 distinct sampling sites were included.

The caecum was the location of the gastrointestinal tract chosen for study due to its large size, diverse microbial communities and important role in hindgut fermentation. Mice were dissected directly in the field. To avoid contamination, utensils were washed in 70% ethanol and flamed prior to each individual dissection, and separate sets of instruments were used for opening the abdominal cavity and removing the caecum. After removal, the caecum tissue was carefully flayed and separated from its contents by immersing and gently shaking the opened tissue in 4 mL of ice-cold RNALater (after which containing the contents), followed by transferring the tissue to a separate tube containing

Table 1 General information on sampling locations and molecular data

Abbreviation	Area	Year	Country	Location		Microsatellites and mitochondrial DNA	16S rRNA gene	
				Latitude	Longitude		Tissue	Content
CB	Cologne/Bonn	2010	Germany	51° 0′6.00″N	7° 2′18.00″E	15*	15	5
SL	Schömberg/Langenbrand	2010	Germany	48°47′31.80″N	8°38′7.38″E	12	12	12
MC	Severac le Château	2009	France	44°19′6.13″N	3° 3′57.00″E	18	18	8
ES	Espelette	2009	France	43°21′10.49″N	1°26′49.34″W	22	22	14
AN	Angers	2009	France	47°27′11.48″N	0°35′41.77″W	18	18	8
NA	Nancy	2010	France	48°39′32.39″N	6° 8′29.41″E	12	12	10
LO	Louan-Villegruis	2010	France	48°37′57.53″N	3°29′4.10″E	12	12	12
DB	Divonne les Bains	2010	France	46°22′35.04″N	6° 7′12.77″E	12	12	11

*Numbers indicate the number of individuals for which data was obtained.

1.5 mL ice-cold RNALater for future processing, according to the manufacturer's instructions. RNALater was removed from content samples by centrifuging at full speed (16 783 rcf) for 10 min followed by careful removal of the supernatant. DNA extraction, 16S rRNA gene amplification and sequencing were performed separately for caecal tissue and luminal contents, although visible content material was present only in a subset of 80 of the 121 mice captured. The same procedure was performed for all caecal tissue samples irrespective of the presence of contents.

DNA extraction

Mouse DNA was isolated from ear punches using the DNeasy Blood and Tissue Kit (QIAGEN). Bacterial DNA was extracted using a modification of the QIAmp DNA stool mini kit (QIAGEN). Approximately, half of the caecum tissue and 200 mg aliquots of contents were transferred to 2-mL screw-cap tubes containing 50 mg each of 0.1 mm, 0.5 mm and 1 mm glass beads (BioSpec Products). The tubes containing beads were treated with UV exposure for 2 h prior to the addition of tissue or contents and performing the extraction. After adding 1.4 mL ASL lysis buffer, samples were subjected to bead beating using the Precellys (Peqlab) for 3 × 15 s at 4723 *g*. Samples were then heated to 95 °C for 10 min, after which the manufacturer's protocol was followed.

Molecular analysis of mouse DNA

D-loop sequencing. An 885 bp portion of the mitochondrial D-loop was sequenced as described by Prager *et al.* (1993). Sequencing reactions were performed using ABI Big Dye v3.1 sequencing chemistry (Applied Biosystems, Foster City, CA, USA) and run on an ABI 3730 automated sequencer. Sequence chromatograms were edited using Seqman (DNASTAR, Inc., Madison, WI, USA) and aligned using CLUSTALW (Thompson *et al.* 1994) in the program MEGA 4.0.2 (Tamura *et al.* 2007).

Microsatellite genotyping. A set of 18 unlinked autosomal microsatellites described by Thomas *et al.* 2007 were chosen to perform analyses of population structure. These include: Chr01_25, Chr02_01, Chr03_21, Chr03_24, Chr04_31, Chr05_15, Chr05_45, Chr07_38, Chr08_11, Chr 09_20, Chr11_64, Chr12_05, Chr13_22, Chr14_16, Chr16_21, Chr17_09, Chr18_08 and Chr19_08 (Hardouin *et al.* 2010). PCRs containing forward primers labelled with either FAM or HEX were performed on 10 ng DNA template using a Multiplex PCR kit (QIAGEN). PCR products were then diluted 1:20 in water and 1 μL of diluted product was added to 10 μL HiDi formamide and 0.1 μL of 500 ROX size standard before running on an ABI 3730 automated sequencer (Applied Biosystems). The alleles were analysed using GENEMAPPER 4.0 (Applied Biosystems).

Population genetic analysis

For the analysis of the mtDNA, we used one sequence of *Mus musculus domesticus* (GenBank Accession no: AM182648) as a reference and one sequence each of *Mus spretus* (GenBank Accession no: U47539), *Mus spicilegus* (GenBank Accession no: U47536) and *Mus famulus* as outgroups. A NeighbourNet network was constructed using the program SPLITSTREE (Huson 1998; Huson & Bryant 2006). For microsatellites, the observed and expected heterozygosities and average number of alleles were calculated using ARLEQUIN (Excoffier *et al.* 2005) and the Cavalli-Sforza Chord distance (CAS) was calculated using Microsatellite Analyser (MSA version 4.05, Dieringer & Schlötterer 2003).

Population structure. Two measures of genetic differentiation were used: Slatkin's R_{ST} (Slatkin 1995) for

microsatellites and γ_{ST} for mitochondrial D-loop sequence. Slatkin's R_{ST} is a measure analogous to Wright's F_{ST} (Wright 1951) that is adapted to the high rate of stepwise mutations occurring at microsatellites. γ_{ST} is a direct measure of differentiation between populations based on the mean number of pairwise differences. Isolation by distance (Wright 1943) was tested for by applying a Mantel test to both data sets. Population substructure was investigated using the software STRUCTURE version 2.3.1 (Pritchard *et al.* 2000; Falush *et al.* 2003b). The parameters used were 500 000 burn-in period and 1 000 000 Markov chain Monte Carlo (MCMC) simulations with four iterations per number of clusters (K) for K equals 2 12. For the choice of K, we applied the criterion of Evanno *et al.* 2005.

Pyrosequencing of the 16S rRNA gene and sequence processing

The 27F-338R primer pair spanning the V1 and V2 hypervariable regions of the bacterial 16S rRNA gene sequence was used for PCR amplification and barcoded pyrosequencing on the 454 GS-FLX platform with Titanium sequencing chemistry as described by Rausch *et al.* (2011). Raw sequences were filtered using MOTHUR version 1.22.2 (Schloss *et al.* 2009) requiring sequences to have a mean quality score >20 and minimum length of 200 bp. Exact matches of 10 bp multiplex identifier (MID) barcodes were used to assign sequences to samples. Details on the reads per sample before and after quality filtering (including Chimera removal) are provided in Table S1 (Supporting information). Chimera detection and sequence clustering of operational taxonomic units (OTUs) at the species level (97% similarity threshold) were carried out using USEARCH/UCHIME version 5.2.32 (Edgar 2010; Edgar *et al.* 2011). The clustering procedure within USEARCH (UCLUST) was performed using the default parameters. Representative consensus sequences for each OTU were obtained from USEARCH. Sequence classification was performed using the version of RDP classifier (Wang *et al.* 2007) implemented in MOTHUR version 1.22.2, with a threshold bootstrap value of 80% for each taxonomic level from phylum to genus. This resulted in approximately 84% of all sequences in the data set being classified at this level. All sequences not meeting this criterion were binned into artificial unclassified genera as described on the RDP website, and the majority of these taxa (approximately 83%) were classifiable at the family level. To compare the bacterial composition of gut microbiota from our study to laboratory mice, we obtained caecum content data from wild type C57BL/6J from Staubach *et al.* (2012) and Ubeda *et al.* (2012) and applied the same sequence filtration, standardization and sequence classification as described previously.

Statistical analyses of the gut microbial community

To avoid potential biases in alpha- and beta-diversity estimates due to the sensitivity of these metrics to sampling effort (Schloss *et al.* 2011), a random subset of 1000 sequences was generated prior to OTU clustering by a perl script implementing random sampling with replacement to normalize the read distribution. This is close to the lowest number of sequences per sample (1028) in our data set after quality filtering and chimera removal and represents a reasonable sampling depth for studies of the microbiome (Hamady & Knight 2009). Bacterial community analyses including the Bray-Curtis and Jaccard distances, analysis of dissimilarity (*'adonis'*, which performs a multidimensional analysis of variance on distance matrices and was applied to Bray-Curtis, Jaccard and UniFrac distances) and constrained analysis of principal coordinates (*'capscale'*) were carried out using the 'VEGAN' R package (R Development Core Team 2011; Oksanen *et al.* 2011). Statistical significance for constrained analysis of principle coordinates was determined by an ANOVA-like permutation test function in Vegan with 1000 permutations (anova.cca). Phylogenetic-based analysis of beta-diversity (UNIFRAC, Lozupone & Knight 2005) was based on a phylogeny produced by FASTTREE (Price *et al.* 2009) implemented in MOTHUR (Schloss *et al.* 2009). To test the significance of factors contributing to the divergence in bacterial communities, we used the R package 'ncf' (Bjørnstad 2009) to perform Mantel and partial Mantel tests. The decay of bacterial community similarity with respect to geographic distance was tested using beta-diversity measures (Bray-Curtis, Jaccard, weighted and unweighted UniFrac distances) on the entire data set as well as separating within and between sampling regions following Martiny *et al.* (2011). To identify OTUs contributing to the Bray-Curtis distance, we applied the SIMPER method (Clarke 1993), which identifies the OTUs contributing to similarity within- and dissimilarity between groups and ranks their contribution.

To assess our ability to measure the contribution of environmental and genetic factors to beta-diversity, we performed three technical replicates (i.e. independent PCR amplification and sequencing of the same DNA template using different MID barcodes) on six independent samples. This revealed the correlation (r^2) of ln abundances for OTUs in the data set to be 0.915, while those OTUs with more than 20 reads in the data set reached 0.94. This indicates a high level of technical

Fig. 1 Sampling locations across Western Europe. More detailed information is provided in Table 1.

reproducibility and enables us to reliably measure the factors influencing beta-diversity.

Results

Host population structure

In total, we sampled 121 *Mus musculus domesticus* individuals from eight geographic locations throughout a continuous area of Germany and France in the summers of 2009 and 2010 (Fig. 1). To confirm the taxonomic status and evaluate population structure based on the maternal lineage, we sequenced 885 bp of the mitochondrial D-loop region of all 121 mice. A total of 51 haplotypes were present in the 121 individuals (Table S2, Supporting information). A NeighbourNet network of the sequences reveals the presence of six closely related haplogroups that correspond to a subset of the haplogroups described in a much larger survey of *M. m. domesticus* including 1313 mtDNA sequences (Bonhomme *et al.* 2011) (Fig. S1, Supporting information). An analysis of γ_{ST} (average over all comparisons = 0.258) with respect to geographic distance revealed no evidence of isolation by distance (Mantel test; $r = 0.04324$, $P = 0.38116$).

To evaluate population structure based on the nuclear genome, we analysed 18 unlinked microsatellite loci (Tables S2 and S3, Supporting information) using STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003b). After conducting four independent runs for each number of clusters (K) ranging from 2 to 12, we applied the criterion of Evanno *et al.* (2005) and chose $K = 5$ for the number of clusters (Fig. 2). The two southernmost localities in France (Severac le Château and Espelette) and Schömberg/Langenbrand displayed the most homogeneity in ancestry, being composed of nearly single clusters, while the other localities displayed multiple clusters and/or admixed individuals. A test for isolation by distance based on R_{ST} (average over all comparisons = 0.039) revealed no significant correlation with geographic distance (Mantel test; $r = 0.12$, $P = 0.28627$).

Composition of the gut microbiota

For the analysis of intestinal bacterial communities, we normalized the read number to 1000 reads for each of the 201 community profiles (121 mucosal plus 80 content) and compared the overall composition at the phylum and genus levels between the mucosa and contents. In total, 17 phyla were identified, with the three most abundant being Bacteroidetes (43.37% vs. 53.83%), Firmicutes (27.85% vs. 36.35%) and Proteobacteria (22.50% vs. 7.25%) in the mucosa vs. content, respectively (Fig. 3a). The differences in Firmicutes and Proteobacteria between the mucosa and contents are significant (ANOVA, $P = 0.0104$ and 3E-6, respectively) and compare well with previous observations in C57BL6/J laboratory mice (Staubach *et al.* 2012). On the genus level, the most abundant taxa were *Bacteroides* (30.67% vs. 36.44%), *Helicobacter* (22.50% vs. 4.25%) and *Robinsoniella* (4.91% vs. 6.95%), which were also, respectively, the most abundant members of the three major phyla Bacteroidetes, Proteobacteria and Firmicutes

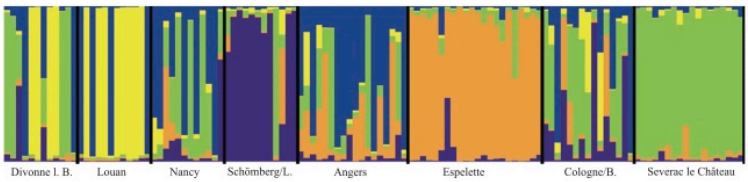


Fig. 2 Genetic clusters identified by STRUCTURE for $K = 5$. The program was run using the admixture model and the optimal number of clusters ($K = 5$) was determined using the criterion of Evanno *et al.* (2005). Each colour refers to one of the five clusters indentified. The ancestry of individuals is denoted by vertical bars, whereby the presence of multiple colours indicates a mixed ancestry. The individuals are sorted according to their geographic locations, which are separated by black vertical lines.

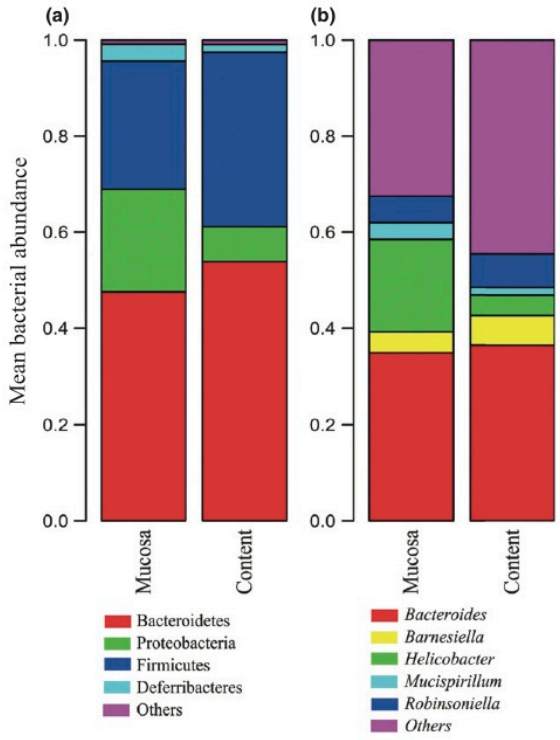


Fig. 3 Composition of mucosa vs. content bacterial communities at the phylum (a) and genus (b) levels. Only the four most abundant phyla and five most abundant genera are shown. The area within each bar represents the mean abundance across the 80 mice for which both mucosa and content data were obtained. Colours represent different phyla (a) or genera (b).

(Fig. 3b). The difference in *Helicobacter* abundance in the mucosa vs. content is significant (ANOVA, P = 6e-7). Other genera with >1% overall abundance which significantly differ between the mucosa vs. content are *Mucispirillum* (3.51% vs. 1.61%; ANOVA, $P = 0.043$) and *Oscillibacter* (1.69% vs. 2.74%; ANOVA, $P = 0.0156$).

In addition, we also detected significant differences in the abundances of these major taxa between the mucosa samples from mice containing caecal contents ($n = 80$) compared with those that were empty ($n = 41$) (Appendix S1, Fig. S2, Supporting information). This may reflect differences in the mucosa-associated communities due to lack of recent food intake and/or an influence of the luminal contents on mice *with* recent food intake. To make direct comparisons between the mucosal and luminal communities, subsequent analyses focus on the subset of individuals for which both data were obtained ($n = 80$).

To evaluate the level and pattern of diversity of intestinal communities within individual wild mice, we applied measures of alpha-diversity including the Chao1 measure of species richness and the Shannon index to measure species evenness. These analyses were applied to genus-level composition (Table S4, Supporting information), as defined by RDP classifier (Wang *et al.* 2007), because this level is less influenced by singletons and clustering artefacts compared with species-level OTUs. The values varied considerably, with mucosa-associated communities displaying overall significant differences with respect to geographic location (ANOVA, $P = 0.0166$ and 0.0406 for Chao1 and Shannon, respectively, Fig. S3, Supporting information), although the same was not observed for contents (ANOVA, $P = 0.363$ and 0.243 for Chao1 and Shannon, respectively). Significant differences in these measures between the mucosa and contents are not apparent (Pairwise t-test, $P = 0.2996$ and 0.2498 for Chao1 and Shannon, respectively). We also tested the possible contribution of body length, weight and their ratio as proxies for age and body mass index but found no significant contribution ($P > 0.05$ in all cases).

To provide insight into possible compositional differences between the wild mice investigated here and commonly used laboratory mice, we compared the genus-level composition of caecum contents between the current and two previously published studies containing data from wild-type laboratory mice (C57BL/6J) housed in different animal facilities (Staubach *et al.* 2012; Ubeda *et al.* 2012). This revealed an on average 1.3–1.5-fold greater number of genera per mouse in the wild mice compared with laboratory mice (Table S5, Supporting information). Although the studies are not directly comparable due to differences in sample size and experimental design, examination of the 359 genera detected among these studies revealed 61 shared, 103 unique to laboratory mice and 111 genera unique to wild mice (Table S5, Supporting information).

Influence of geography on the gut microbiota

To determine the relative contribution of environmental and host genetic factors to the overall pattern of diversity between individual wild mice, we employed both taxon (OTU)- and phylogenetic-based measures of beta-diversity. The Bray-Curtis and Jaccard distances, which are based on weighting taxon abundance and the presence/absence, respectively, were calculated based on OTUs at a 97% similarity threshold. The phylogenetic distances between communities based on weighting taxon abundance and the presence/absence were measured using the weighted and unweighted UniFrac indices, respectively, which are based on comparing the unique vs. shared branch lengths underlying different communities (Lozupone & Knight 2005).

Prior to biogeographical analysis, we compared the similarity between mucosa and content communities within and between mice. Overall, we observed significantly greater similarity between the mucosa and content communities within mice relative to the mucosa or content communities between mice using all four beta-diversity measures (Table 2). Thus, the mucosal communities of individual mice displayed more similarity to their own content communities than to the mucosal communities of other mice, and significant correlations between the mucosal and content communities are present (Mantel test; Bray-Curtis distance: $r = 0.4339$, $P = 0.001$, Jaccard distance: $r = 0.4178$, $P = 0.001$, weighted UniFrac distance: $r = 0.5847$, $P = 0.001$, unweighted UniFrac distance: $r = 0.7084$, $P = 0.001$). Nonetheless, as observed at the level of phyla and genera abundances, significant differences between the mucosa and content communities are apparent using three of four measures of beta-diversity (*adonis*; Bray-Curtis distance: $r^2 = 0.0206$, $P = 0.001$, Jaccard distance: $r^2 = 0.0143$, $P = 0.001$, weighted UniFrac distance: $r^2 = 0.0234$, $P = 0.023$, unweighted UniFrac distance: $r^2 = 0.0086$, $P = 0.210$; Fig. S4, Supporting information), and subsequent analyses were thus conducted separately for the mucosa and contents.

As a first assessment of the influence of geography, we analysed and visualized the degree to which communities separate according to geographic location using a constrained analysis of principal coordinates ('*capscale*') based on this categorical variable (Fig. 4). This analysis is a hypothesis-driven ordination that limits ('constrains') the separation of the communities only according to the variable being tested (in this case geographic location), rather than allowing communities to separate by all possible variables influencing the statistic (*i.e.* an 'unconstrained' ordination) (Anderson & Willis 2003). These results are supported by multivariate ANOVA performed using analysis of dissimilarity ('*adonis*') applied to beta-diversity measures. The mucosa samples displayed significant separation according to Bray-Curtis distance (*adonis*, $r^2 = 0.1552$, $P = 0.001$; Fig. 4) and Jaccard distance (*adonis*, $r^2 = 0.1275$, $P = 0.001$), but not weighted- (*adonis*, $r^2 = 0.1258$, $P = 0.107$) or unweighted UniFrac distance (*adonis*, $r^2 = 0.09586$, $P = 0.338$). The analysis based on caecal content displayed similar results, with significant separation based on Bray-Curtis distance (*adonis*, $r^2 = 0.1323$, $P = 0.001$), Jaccard distance (*adonis*, $r^2 = 0.1141$, $P = 0.001$), as well as weighted- (*adonis*, $r^2 = 0.1848$, $P = 0.008$), but not unweighted UniFrac distance (*adonis*, $r^2 = 0.0834$, $P = 0.543$).

We next investigated the influence of geographic distance on a continuous scale by applying Mantel tests and regression analyses to each of four measures of beta-diversity. Mantel tests revealed a significantly negative relationship for mucosa and content both for Bray-Curtis and Jaccard distances, but not for weighted or unweighted UniFrac distance (Table S6, Supporting information). To further investigate the influence of geography, we applied regression analyses of community similarities (1 dissimilarities) at different scales, namely within a sampling location (local scale, average distance between sites = 11.51 km) and between sampling locations (continental scale, average distance between sites = 530 km) following Martiny *et al.* (2011). Interestingly, this revealed significant similarity-distance decay for OTU-based measures at the continental, but not local scale (Table S6, Fig. S5, Supporting information). Lastly, we examined whether the climatic zone to which a sampling location belongs contributes to the influence of geography, but all but one location (Severac le Château) belongs to the same zone (Peel *et al.* 2007).

Influence of host genetics on the gut microbiota

To evaluate the contribution of host genetics to the pattern of diversity between individuals, we first examined the genetic clusters identified by STRUCTURE as a categorical variable. Because a portion of the individuals contained a mixture of genetic backgrounds (i.e. were admixed), we limited the analysis of genetic clusters to those individuals with $\geq 80\%$ of their ancestry being assigned to a given cluster. This resulted in 61 mice belonging to five defined genetic clusters. For the mucosa communities, we found Bray-Curtis and Jaccard distances to be significantly influenced by genetic clusters (*adonis*; $r^2 = 0.1578$, $P = 0.001$ and $r^2 = 0.1401$,

Table 2 Comparison of beta-diversity measures between mucosa and content communities within single mice and among mucosa/content from different mice

Beta-diversity measure	Average between mucosa and content within mice	Average among mucosa between mice	Average among content between mice	ANOVA *P* value
Bray-Curtis	0.820	0.938	0.936	2.1E-16
Jaccard	0.897	0.967	0.966	2.3E-16
Weighted UniFrac	0.263	0.494	0.503	2.2E-16
Unweighted UniFrac	0.618	0.786	0.784	2.9E-16

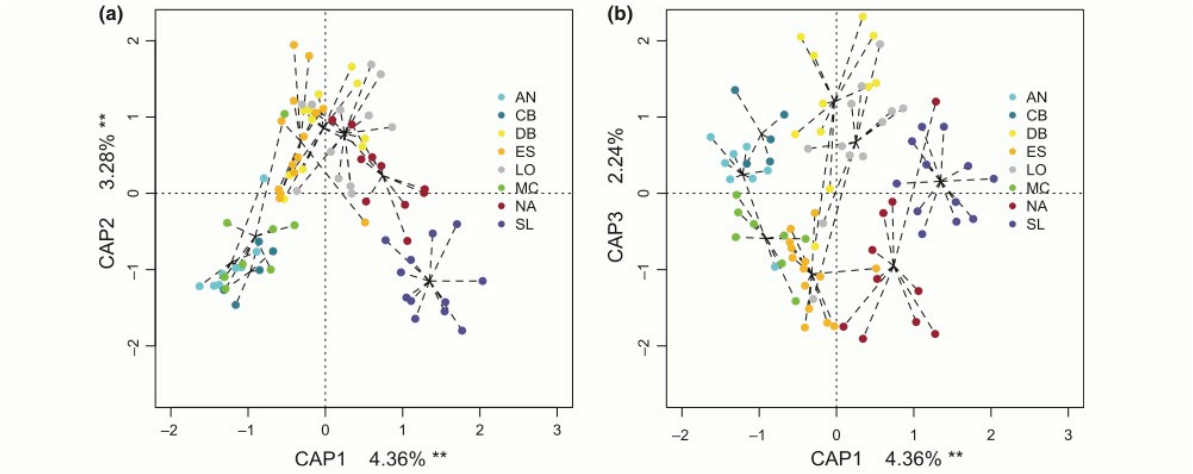


Fig. 4 Constrained analysis of principle coordinates of Bray-Curtis distance based on geographic location for mucosa samples. CAP1, CAP2 and CAP3 are the axes from the constrained analysis of principle coordinates ('*capscale*', see Materials and methods) and account for 8.96% of the total variation. Panel (a) displays the first and second axes (CAP1, CAP2) and (b) displays the first and third axes (CAP1, CAP3). Each of the eight geographic sampling locations is indicated by a unique colour; abbreviations for the sampling locations (e.g. 'AN' for Angers) are given in Table 1. **Represents significance from the 'anova.cca' test with respect to geographic location as a categorical variable with 1000 permutations (see Materials and methods; $P < 0.01$).

$P = 0.002$, respectively), but not weighted or unweighted UniFrac distance (*adonis*; $r^2 = 0.05413$, $P = 0.858$ and $r^2 = 0.0896$, $P = 0.290$, respectively). Similarly, content communities also displayed a significant influence of genetic clusters for Bray-Curtis and Jaccard distances (*adonis*; $r^2 = 0.1132$, $P = 0.002$ and $r^2 = 0.1052$, $P = 0.001$, respectively), but not weighted or unweighted UniFrac distance (*adonis*; $r^2 = 0.1092$, $P = 0.255$ and $r^2 = 0.0749$, $P = 0.627$, respectively).

We next investigated the influence of genetics on a continuous scale by applying Mantel tests between genetic distance at the individual level [Cavalli-Sforza Chord distance, (CAS)] and each of four measures of beta-diversity. For the mucosa, we found a significant contribution of genetic distance for three of four beta-diversity measures (Table 3). The analysis of caecal contents yielded similar results but were significant only for the OTU-based measures (Table 3). However, because we found individual level genetic distances to be correlated with geographic distance (Mantel test; $r = 0.264$, $P = 0.002$), partial Mantel tests are necessary to distinguish the contribution of each single factor while controlling for the effect of the other. For the mucosa communities, a significant effect of genetic distance remained for Bray-Curtis, Jaccard and unweighted UniFrac distances, while no significant effects remained for content communities (Table 3). In contrast, the influence of geography remained significant for both OTU-based measures in the mucosa and content communities. Furthermore, the correlation coefficient for geography is on average approximately threefold higher than that of genetic distance for OTU-based measures.

Lastly, we also examined the potential contribution of sex to patterns of beta-diversity. However, we identified no significant effect on any of the four measures used for both mucosa and content communities, consistent with the findings of other recent studies (McKnite *et al.* 2012).

Mitochondrial DNA as a proxy for maternal transmission

A final important factor to consider in our data set is maternal transmission, which if present may be reflected by differences with respect to mtDNA haplogroups. We treated each of the six previously defined haplogroups (Bonhomme *et al.* 2011) detected as categorical variables and detected a significant signal of differentiation using OTU-based measures for both mucosa (*adonis*; Bray-Curtis distance: $r^2 = 0.0728$, $P = 0.022$; Jaccard distance: $r^2 = 0.0688$, $P = 0.031$; weighted UniFrac distance: $r^2 = 0.07788$, $P = 0.184$; unweighted UniFrac distance: $r^2 = 0.06715$, $P = 0.131$) and content communities (*adonis*: Bray-Curtis distance: $r^2 = 0.0758$, $P = 0.004$; Jaccard distance: $r^2 = 0.0702$, $P = 0.006$; weighted UniFrac distance: $r^2 = 0.0834$, $P = 0.098$; unweighted UniFrac distance: $r^2 = 0.0675$, $P = 0.125$).

Comparison of beta-diversity metrics

Interestingly, the majority of significant relationships with respect to beta-diversity were identified using

Table 3 Summary of Mantel and partial Mantel tests for geographic and genetic distances

Community location	Beta-diversity measure	Geographic distance*		Cavalli-Sforza Chord distance (CAS) genetic distance		Geographic distance		CAS genetic distance	
		Mantel *r*	Mantel *P*	Mantel *r*	Mantel *P*	Partial Mantel *r*	Partial Mantel *P*	Partial Mantel *r*	Partial Mantel *P*
Mucosa	Bray-Curtis	0.179	**0.001**†	0.1041	**0.001**	0.158	**0.001**	0.0597	**0.013**
Mucosa	Jaccard	0.172	**0.001**	0.1016	**0.001**	0.151	**0.001**	0.0589	**0.011**
Mucosa	Weighted UniFrac	0.016	0.304	0.0389	0.077	0.005	0.402	0.036	0.113
Mucosa	Unweighted UniFrac	0.033	0.197	0.0699	**0.026**	0.014	0.335	0.064	**0.023**
Content	Bray-Curtis	0.128	**0.001**	0.0676	**0.002**	0.113	**0.001**	0.0354	0.091
Content	Jaccard	0.125	**0.001**	0.0675	**0.007**	0.111	**0.001**	0.0359	0.077
Content	Weighted UniFrac	−0.008	0.601	0.0204	0.276	−0.015	0.361	0.0238	0.233
Content	Unweighted UniFrac	0.061	0.064	0.0366	0.124	0.052	0.087	0.0214	0.243

*Results for geographic distance are the same as those reported in Table 3 but are included here for comparison.
†Bold values indicate significant ($P < 0.05$) comparisons.

OTU-based metrics (Bray-Curtis and Jaccard distances) as opposed to those based on underlying phylogeny (weighted- and unweighted UniFrac distance). To investigate the possible reasons for these discrepancies, we applied the SIMPER method (Clarke 1993) to identify the major taxa contributing to the Bray-Curtis index with respect to geographic location. Using this method, we quantified the contribution of every OTU to pairwise comparisons and found that OTUs belonging to only three genera (515 OTUs belonging to *Bacteroides*, 66 belonging to *Helicobacter* and 334 belonging to *Robinsoniella*) contribute on average 55%, 18% and 13%, respectively, to the variation in the Bray-Curtis distance with respect to geographic location. However, the sequences belonging to these genera contribute only 6%, 1% and 9% to the branch length of the maximum-likelihood tree we used to calculate the UniFrac distance. Thus, divergence in community structure predominately among closely related taxa might lead to limited resolution for phylogenetic-based methods.

Discussion

Determining the relative role of the environment and genetics on the gut microbiota is a challenging undertaking due to the complexity of these communities and the interaction of factors contributing to diversity. Several different categories of studies addressing this question have been performed including human twin studies, comparison and experimental manipulation of mouse lines, quantitative trait loci (QTL) mapping and analyses of single host genes (reviewed by Spor *et al.* 2011). Although different twin studies have offered some conflicting results (*e.g.* fingerprinting based studies previously revealed host genetic influences (Zoetendal *et al.* 2001), while more recent sequence-based methods did not (Turnbaugh *et al.* 2009)), by and large each of these approaches yields support for a role of host genetics. Using a different comparative evolutionary approach, Ochman *et al.* (2010) revealed a correspondence between the gut microbiota and host phylogeny among closely related hominid species, providing evidence for vertical inheritance among genetically differentiated hosts. Our study complements these previous studies by analysing a large number of mice in their natural environment, with varying degrees of relatedness and spanning a large geographic area. In addition, we focused on both the caecal mucosa as well as the luminal content to test the hypothesis that the mucosal community displays a stronger dependence on host genetics due to its more intimate association with the host (Spor *et al.* 2011).

Overall, we determined the most significant contributing factor to microbial diversity between individuals

to be geography, as measured by the distance between sampling sites. This effect was found both in the mucosa-associated and in content microbial communities, where a partial Mantel test (controlling for genetic distance) revealed geographic distance to explain approximately 16% and 11% of the variation in OTU-based measures in the mucosa and content communities, respectively (Table 3). The same test revealed genetic distance to explain only 6% of the variation for the mucosa, while no significant influence was observed for the content communities. Interestingly, further regression analysis detected an influence of geographic distance at the continental (between sampling locations), but not local (within a sampling location) scale.

While geography might be viewed as an approximation for the sum of environmental effects such as local weather patterns, availability of food sources, *etc.*, we note that traps were placed in and around artificial structures on a mixture of arable, livestock and mixed farms throughout the sampling range. Thus, we expect no systematic bias in human-related dietary availability. Furthermore, the Köppen climatic classification system (Peel *et al.* 2007), which incorporates average annual and monthly temperatures and precipitation as well as the seasonality of precipitation, was largely invariant throughout our sampling range (seven of the eight locations sampled belong to the same zone). Thus, while we cannot rule out the possible contribution of local environmental variables or dietary availability, an alternative and perhaps likely explanation for the signature of geographic distance we observe is neutral dispersal limitations among microbes. Moreover, the analysis of OTUs contributing to the Bray-Curtis distance would suggest this phenomenon to occur largely among closely related taxa.

Compared with geography, the influence of host genetic distance identified by our analyses is less substantial. However, the signal we do observe is clearly limited to the mucosa, where three of four measures of beta-diversity are significant compared with none for content communities after controlling for the influence of geography. This supports the hypothesis that the mucosal environment may be more strongly determined by host genetics (Spor *et al.* 2011).

In addition to the analysis of geographic vs. genetic distance on the individual level, we also analysed the structuring of our host samples using the clustering procedure provided by STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003b), which assigns individuals to populations without a priori knowledge of the population units. However, in contrast to absolute measures of genetic distance such as the Chord distance, this method provides no information on the relationship of populations to one another. Thus, it is not certain how much the association is due to genetic differences between these clusters vs. an increase in community similarity within groups of interbreeding individuals, which would presumably arise due to greater transmission (e.g. via corprophagy) within such groups. At a minimum, maternal transmission is an important factor to consider as indicated by the significant clustering we observe with respect to mtDNA haplogroups. Finally, it should be noted that house mice colonized Europe approximately 3000 years ago (Cucchi *et al.* 2005), and this and other studies that analysed *Mus musculus domesticus* populations in Western Europe (Salcedo *et al.* 2007) report little genetic differentiation among populations. Thus, the scale of genetic divergence among the populations included in our data set may be limited with regard to the potential influence on microbial communities.

In addition to providing a unique perspective on the contributions of geography, host genetics and transmission to diversity between individuals, our study also offers insight into the communities of free-living, wild house mice compared with those kept in laboratory environments. Typically, laboratory mice from commercial providers are initially inoculated with the Altered Schaedler Flora (ASF), which contains a mixture of eight bacterial strains representing seven genera (Dewhirst *et al.* 1999), but can increase in diversity over time depending on differences in housing practices such as the use of conventional vs. individually ventilated cages (IVCs) (Stecher *et al.* 2010). As expected the wild mice in our study are very similar to laboratory mice at the bacterial phyla level (i.e. the three most abundant being Bacteroidetes, Firmicutes and Proteobacteria, followed by Deferribacteres). However, they display more substantial differences at lower taxonomic levels (e.g. genus level). Previous studies of wild *M. musculus* revealed them to have greater immune function than their laboratory-reared counterparts (Abolins *et al.* 2011) but did not examine their microbiota. Thus, future studies may shed light on whether aspects of wild mouse microbial communities might contribute to this important fitness-related trait.

Our analysis of the caecal mucosal and content communities in natural house mouse populations provides several important points of information for understanding the ecology and evolution of host microbiota interactions. The influence of geography on a scale of only hundreds of kilometres within mostly one climate zone illustrates the importance of considering these effects in comparative studies, which are also likely relevant to human hosts (Yatsunenko *et al.* 2012). Further analyses including variation in intestinal bacterial communities both within and *between* mouse host subspecies may shed additional light on the nature of co-evolution in this important model organism.

Acknowledgements

We wish to thank Hermann Autengruber, Urs Benedikt Müller, Knut Albrecht, Philipp Rausch and Ann Kathrin Jarms for assistance in fieldwork, Iris Fischer, Ateequr Rehman and Aurélien Tellier for helpful discussion, Heinke Buhtz, Katja Cloppenborg-Schmidt and Silke Carstensen for excellent technical assistance and the editor and anonymous reviewers for helpful comments that improved the manuscript. This manuscript was supported by DFG Grants ME3134/3 to D.M. and BA 2863/2-1, BA 2863/2-2 and the Excellence Cluster 'Inflammation at Interfaces' to J.F.B.

References

Abolins SR, Pocock MJO, Hafalla JCR, *et al.* (2011) Measures of immune function of wild mice, *Mus musculus*. *Molecular Ecology*, **20**, 881–892.

Anderson MJ, Willis TJ (2003) Canonical analysis of principle coordinates: a useful method of constrained ordination for ecology. *Ecology*, **84**, 511–525.

Bäckhed F, Ding H, Wang T, *et al.* (2004) The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15718–15723.

Bäckhed F, Ley RE, Sonnenburg JL, *et al.* (2005) Host-bacterial mutualism in the human intestine. *Science*, **307**, 1915–1920.

Bjørnstad ON (2009) *ncf: Spatial Nonparametric Covariance Functions. R Package Version 1.1-3*. Available from http://CRAN.R-project.org/package=ncf.

Carroll IM, Threadgill DW, Threadgill DS (2009) The gastrointestinal microbiome: a malleable, third genome of mammals. *Mammalian Genome*, **20**, 395–403.

Clarke KR (1993) Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, **18**, 117–143.

Cucchi T, Vigne J-D, Auffray J-C (2005) The genus Mus as a model for evolutionary studies first occurrence of the house mouse (*Mus musculus domesticus* Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. *Biological Journal of the Linnean Society*, **84**, 429–445.

Dewhirst FE, Chien CC, Paster BJ, *et al.* (1999) Phylogeny of the defined murine microbiota: altered Schaedler flora. *Applied and Environmental Microbiology*, **65**, 3287–3292.

Dieringer D, Schlötterer C (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, **3**, 167–169.

Eckburg PB, Bik EM, Bernstein CN, *et al.* (2005) Diversity of the human intestinal microbial flora. *Science*, **308**, 1635–1638.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar RC, Haas BJ, Clemente JC, *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Falk PG, Hooper LV, Midtvedt T, *et al.* (1998) Creating and maintaining the gastrointestinal ecosystem: what we know and need to know from gnotobiology. *Microbiology and Molecular Biology Reviews*, **62**, 1157–1170.

Falush D, Wirth T, Linz B, *et al.* (2003a) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.

Falush D, Stephens M, Pritchard JK (2003b) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects tools, techniques, and challenges. *Genome Research*, **19**, 1141–1152.

Hardouin EA, Chapuis J-L, Stevens MI, *et al.* (2010) House mouse colonization patterns on the sub-Antarctic Kerguelen Archipelago suggest singular primary invasions and resilience against re-invasion. *BMC Evolutionary Biology*, **10**, 325.

Heijtz RD, Wang S, Anuar F, *et al.* (2011) Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 3047–3052.

Hooper LV (2001) Commensal host–bacterial relationships in the gut. *Science*, **292**, 1115–1118.

Hooper LV, Midtvedt T, Gordon JI (2002) How host–microbial interactions shape the nutrient environment of the mammalian intestine. *Annual Review of Nutrition*, **22**, 283–307.

Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Ihle S, Ravaoarimanana I, Thomas M, *et al.* (2006) An analysis of signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution*, **23**, 790–797.

Lankau EW, Hong PY, Mackie RI (2012) Ecological drift and local exposures drive enteric bacterial community differences within species of *Galapagos iguanas*. *Molecular Ecology*, **21**, 1779–1788.

Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, **124**, 837–848.

Ley RE, Lozupone CA, Hamady M, *et al.* (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews. Microbiology*, **6**, 776–788.

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.

Martiny JBH, Eisen JA, Penn K, *et al.* (2011) Drivers of bacterial beta-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7850–7854.

McKnite AM, Perez-Munoz ME, Lu L, *et al.* (2012) Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS ONE*, **6**, e39191.

Muegge BD, Kuczynski J, Knights D, *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.

Ochman H, Worobey M, Kuo C-H, *et al.* (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology*, **8**, e1000546.

O'Hara AM, Shanahan F (2006) The gut flora as a forgotten organ. *EMBO Reports*, **7**, 688–693.

Oksanen J, Blanchet FG, Kindt R, *et al.* (2011) *vegan: Community Ecology Package. R Package Version 2.0-2*. Available from http://CRAN.R-project.org/package=vegan.

Ott SJ, Musfeldt M, Wenderoth D, *et al.* (2004) Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, **53**, 685–693.

Peel MC, Finlayson BL, McMahon TA, *et al.* (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, **11**, 1633–1644.

Phillips CD, Phelan G, Dowd SE, *et al.* (2012) Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular Ecology*, **21**, 2617–2627.

Prager EM, Sage RD, Gyllenstein U, *et al.* (1993) Mitochondrial-DNA sequence diversity and the colonization of Scandinavia by house mice from east Holstein. *Biological Journal of the Linnean Society*, **50**, 85–122.

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26**, 1641–1650.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2011) *R: A Language and Environment Computing*. R Development Core Team, Vienna, Austria.

Rausch P, Rehman A, Künzel S, *et al.* (2011) Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and *FUT2* (Secretor) genotype. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 19030–19035.

Rawls JF, Mahowald MA, Ley RE, *et al.* (2006) Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell*, **127**, 423–433.

Round JL, Mazmanian SK (2009) The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology*, **9**, 313–323.

Salcedo T, Geraldes A, Nachman MW (2007) Nucleotide variation in wild and inbred mice. *Genetics*, **177**, 2277–2291.

Schloss PD, Westcott SL (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology*, **77**, 3219–3226.

Schloss PD, Westcott SL, Ryabin T, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Spor A, Koren O, Ley RE (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews. Microbiology*, **9**, 279–290.

Staubach F, Kunzel S, Baines AC, *et al.* (2012) Expression of the blood-group-related glycosyltransferase *B4galnt2* influences the intestinal microbiota in mice. *The ISME Journal*, **6**, 1345–1355.

Stecher B, Chaffron S, Käppeli R, *et al.* (2010) Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathogens*, **6**, e1000711.

Sudakaran S, Salem H, Kost C, *et al.* (2012) Geographical and ecological stability of the symbiotic mid-gut microbiota in European firebugs, *Pyrrhocoris apterus* (Hemiptera, Pyrrhocoridae). *Molecular Ecology*, **21**, 6134–6151.

Sullam KE, Essinger SD, Lozupone CA, *et al.* (2012) Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Molecular Ecology*, **21**, 3363–3378.

Tamura K, Dudley J, Nei M, *et al.* (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Thomas M, Möller F, Wiehe T, *et al.* (2007) A pooling approach to detect signatures of selective sweeps in genome scans using microsatellites. *Molecular Ecology*, **7**, 400–403.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Turnbaugh PJ, Hamady M, Yatsunenko T, *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Ubeda C, Lipuma L, Gobourne A, *et al.* (2012) Familial transmission rather than defective innate immunity shapes the distinct intestinal microbiota of TLR-deficient mice. *Journal of Experimental Medicine*, **209**, 1445–1456.

Wang Q, Garrity GM, Tiedje JM, *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yatsunenko T, Rey FE, Manary MJ, *et al.* (2012) Human gut microbiome viewed across age and geography. *Nature*, **486**, 222–227.

Zoetendal E, Akkermans A, Akkermans-van Vliet W, *et al.* (2001) The host genotype affects the bacterial community in the human gastronintestinal tract. *Microbial Ecology in Health and Disease*, **13**, 129–134.

J.F.B. conceived the research. M.L., E.A.H. and S.K. performed research. M.L., J.W., D.M. and J.F.B. analysed data. M.L., J.W. and J.F.B. wrote the manuscript.

Data accessibility

MtDNA sequences: GenBank Accessions KC139091-KC139211.
Microsatellite data: Supplementary information, Tables S2 and S3.
Bacterial 16S rRNA gene sequences: European Nucleotide Archive, www.ebi.ac.uk/ena/data/view/ERP001970.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Comparison of mucosal microbiota according to the presence/absence of caecal contents.

Fig. S1 NeighbourNet network of 121 mitochondrial D-loop sequences.

Fig. S2 Constrained analysis of principle coordinates of Bray-Curtis distance based on the presence/absence of caecal contents.

Fig. S3 (a) Chao1 index and (b) Shannon diversity measures based on genus-level composition in mucosa communities.

Fig. S4 Constrained analysis of principle coordinates of Bray-Curtis distance based on mucosa vs. content communities.

Fig. S5 Similarity-distance decay in bacterial communities in (a) mucosa and (b) content.

Table S1 SRA archive of 454 Runs, number of reads and number of reads after filtering for individual samples.

Table S2 Summary of mitochondrial D-loop and microsatellite variation.

Table S3 Microsatellite data.

Table S4 Phylum and genus level bacterial composition of both caecal mucosa and content communities.

Table S5 Analysis of bacterial genera in the current and two previously published studies of lab mice.

Table S6 Summary of Mantel tests and linear regression of similarity-distance decay for geographic distance.

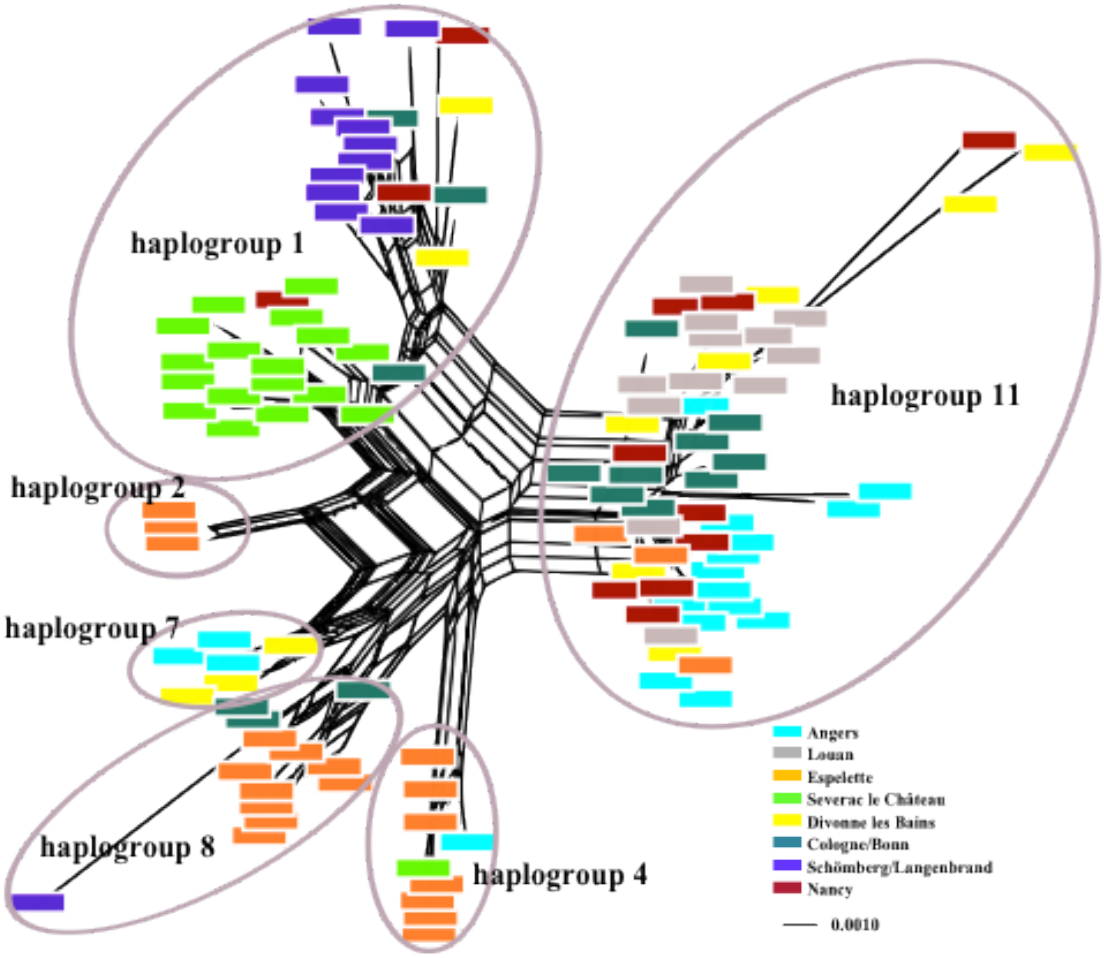


Fig. S1. NeighbourNet network of 121 mitochondrial D-loop sequences. The network was generated using "SplitsTree" under the default settings. Colors indicate the geographic sampling location. The numbering of the haplogroups (indicated by grey ovals) corresponds to that of Bonhomme *et al.* (2011).

Mucosal microbiota differ according to presence/absence of cecal contents

Differences between the cecal mucosal communities in mice lacking or containing luminal contents may reflect a lack of recent food intake and/or an influence of the luminal contents on mice *with* recent food intake. Among the major phyla we found significant differences in Bacteroidetes (mean for mice with contents 47.34%, without contents 36.99%, ANOVA $P = 0.0497$). Among the genera with average abundances >1%, we found significant differences in *Bacteroides* (with contents 34.79%, without contents 22.6%, ANOVA $P = 0.0228$), *Helicobacter* (with contents 19.32%, without contents 28.72%, ANOVA $P = 0.0489$), *Alistipes* (with contents average 1.06%, without contents 3.10%, ANOVA $P = 0.0072$), and *Dorea* (with contents 0.95%, without contents 1.95%, ANOVA $P = 0.0023$). Beta-diversity measures also display significant differences, although only a small proportion of variation is explained (*Adonis* for Bray-Curtis dissimilarities: $r^2 = 0.01326$, $P = 0.007$, Jaccard: $r^2 = 0.0111$, $P = 0.004$, weighted UniFrac: $r^2 = 0.02774$, $P = 0.024$, unweighted UniFrac: $r^2 = 0.01151$, $P = 0.239$; Fig. S2).

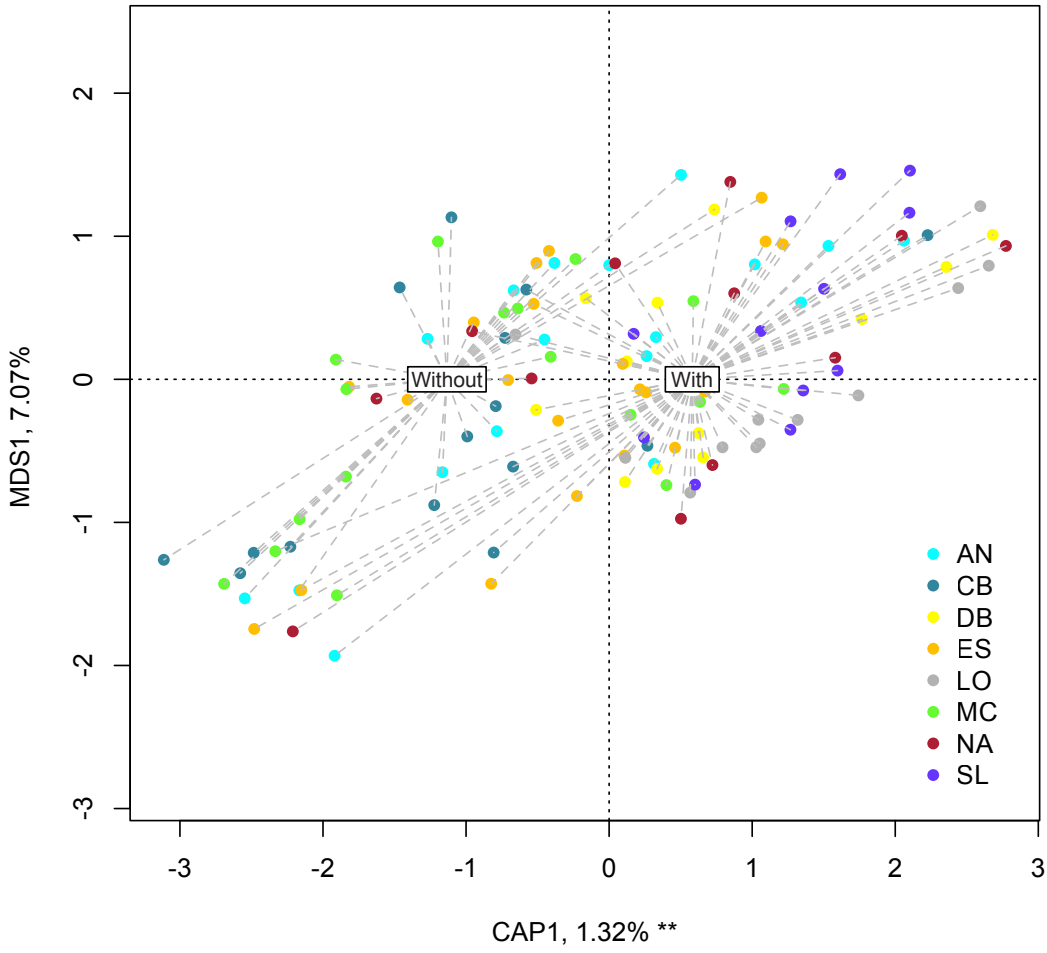


Fig S2. Constrained analysis of principle coordinates of Bray-Curtis distance based on presence (with) or absence (without) of cecal contents. CAP1 is the axis from the constrained analysis of principle coordinates ("*capscale*", see Methods) and MDS1 is the first axis of the unconstrained analysis. Each of the eight geographic sampling locations is indicated by a unique color; abbreviations for the sampling locations (*e.g.* "AN" for Angers) are given in Table 1. **Represents significance from the "anova.cca" test with respect with/without cecum content as a categorical variable with 1000 permutations (see Methods; $P < 0.01$).

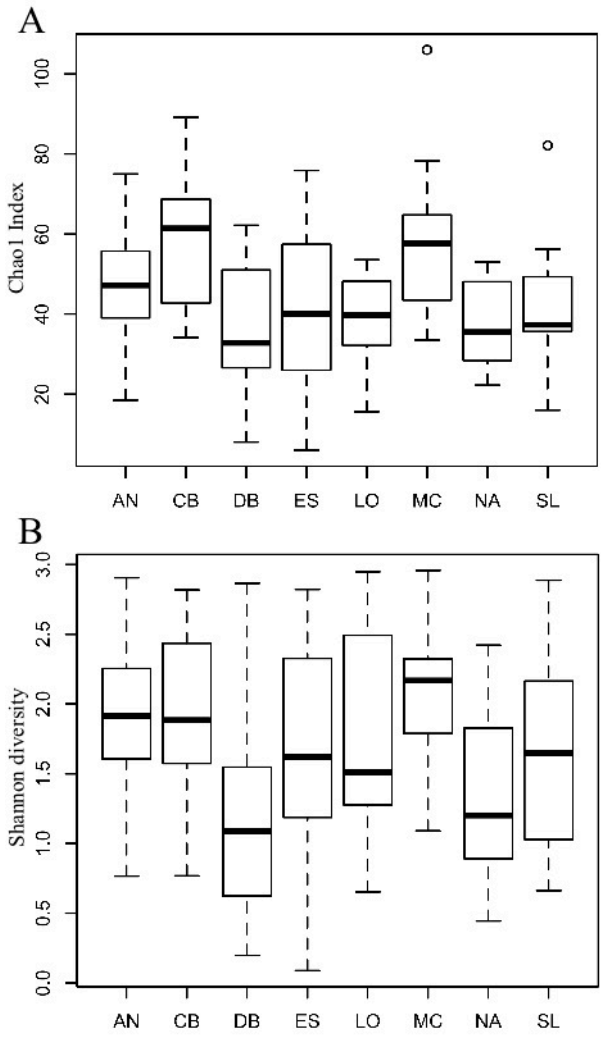


Fig S3. (a) Chao1 index and (b) Shannon diversity measures based on genus level composition in mucosa communities. Abbreviations are listed in Table 1 and geographical locations can be found in Fig 1.

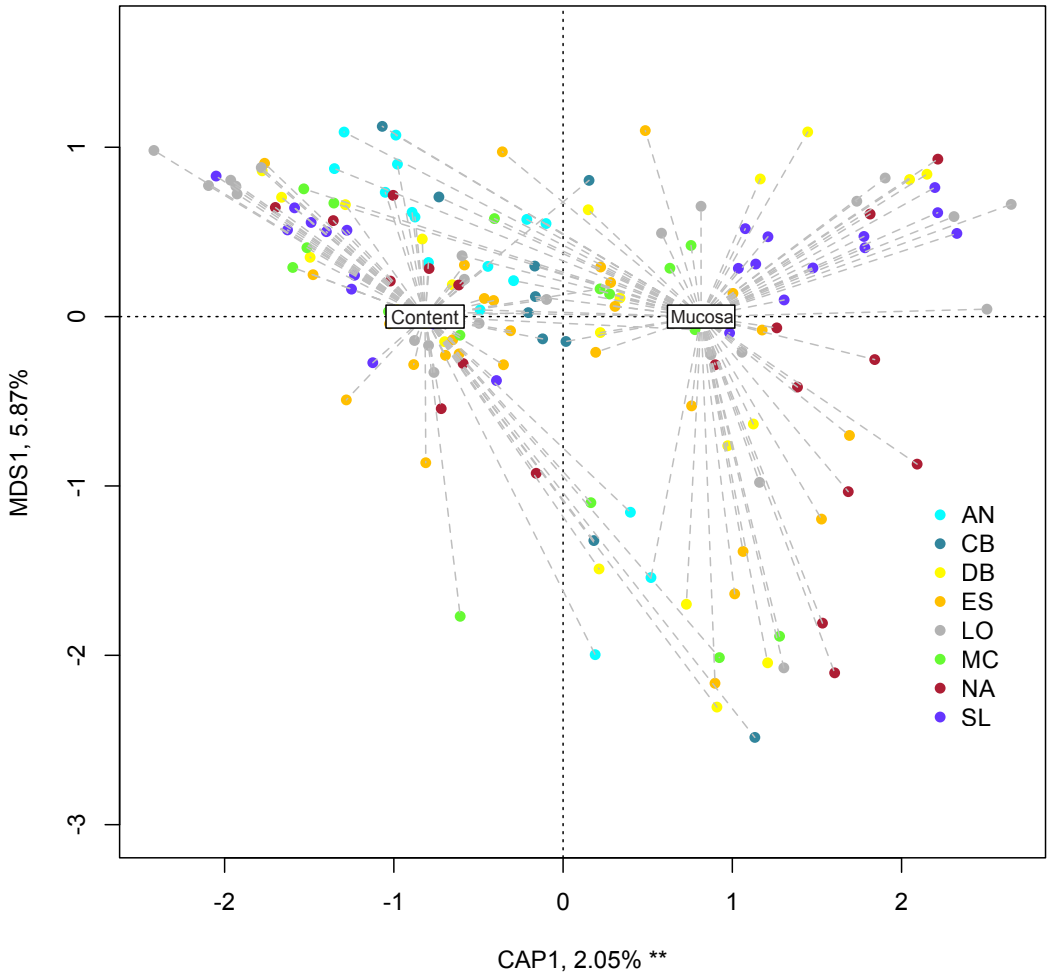
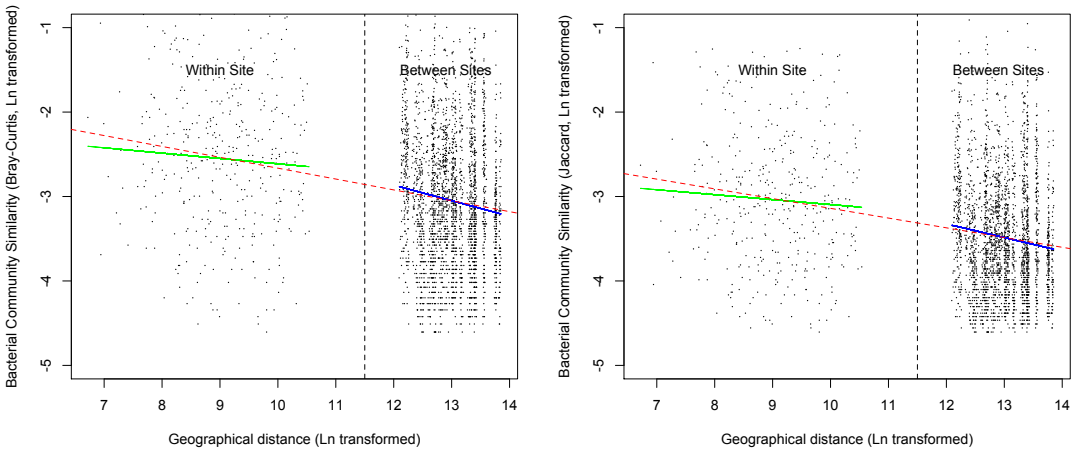


Fig S4. Constrained analysis of principle coordinates of Bray-Curtis dissimilarity based on mucosa *vs.* content communities. CAP1 is the axis from the constrained analysis of principle coordinates ("*capscale*", see Methods) and MDS1 is the first axis of the unconstrained analysis. Each of the eight geographic sampling locations is indicated by a unique color; abbreviations for the sampling locations (*e.g.* "AN" for Angers) are given in Table 1. **Represents significance from the "anova.cca" test with respect to content/mucosa as a categorical variable with 1000 permutations (see Methods; $P < 0.01$).

a.



b.

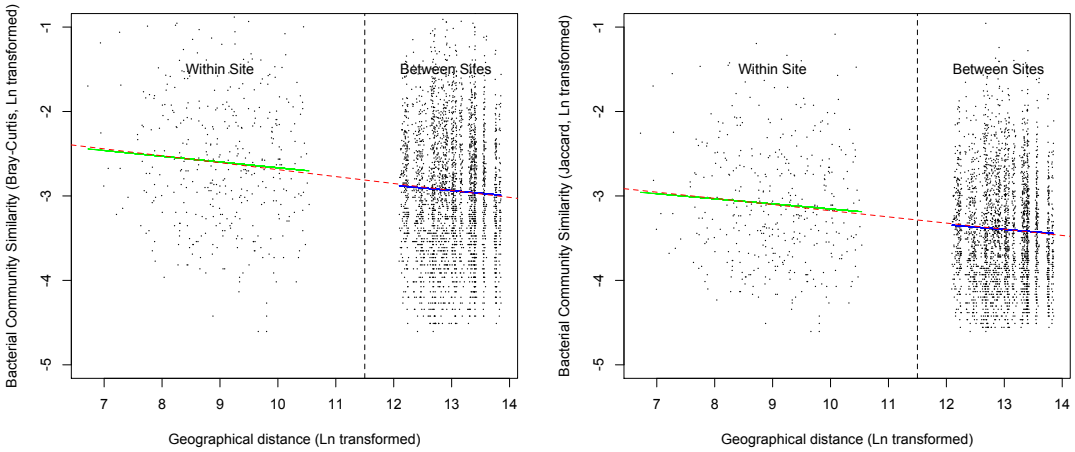


Fig S5. Similarity-distance decay in bacterial communities in a) mucosa and b) content. The red dashed lines denote regression based on the whole data set, green solid lines denote regression at a local scale (within sampling locations/regions) and blue lines denote regression at a continental scale (between sampling locations/regions). Details are provided in Table 3.

(Note: we only included supplementary tables that could fit in the thesis page: Table S1, S2, S5 and S6, which are essential for the major conclusions of the paper. Table S3 is microsatellites information for each mice and Table S4 is the genera composition of the microbial communities for all samples, and can be found online together with the publication: (www.onlinelibrary.wiley.com/doi/10.1111/mec.12206/suppinfo)

Supplementary Table 1: Original SFF files, number of reads and number of reads after filtering.

Location	Sample	ERR Accession (EBI)	Reads	Filtered reads
AN	AN13_content	ERR197719	2944	2929
	AN13_tissue	ERR197720	4436	4413
	AN15_tissue	ERR197721	4449	4432
	AN23_content	ERR197722	1639	1628
	AN23_tissue	ERR197723	3507	3500
	AN26_content	ERR197724	4362	4337
	AN26_tissue	ERR197725	3919	3912
	AN27_content	ERR197726	2630	2622
	AN27_tissue	ERR197727	3021	3011
	AN29_tissue	ERR197728	2740	2733
	AN30_content	ERR197729	2415	2402
	AN30_tissue	ERR197730	4314	4292
	AN31_tissue	ERR197731	3125	3110
	AN55_tissue	ERR197732	3171	3162
	AN60_tissue	ERR197733	2264	2257
	AN61_tissue	ERR197734	3110	3096
	AN62_content	ERR197735	2125	2121
	AN62_tissue	ERR197736	3425	3384
	AN64_content	ERR197737	11876	11789
	AN64_tissue	ERR197738	5534	5512
	AN65_content	ERR197739	3620	3601
	AN65_tissue	ERR197740	1351	1349
	AN66_content	ERR197741	3072	3042
	AN66_tissue	ERR197742	2042	2022
	AN6_tissue	ERR197743	4080	4044
	AN8_tissue	ERR197744	5653	5630
CB	CB10_tissue	ERR197745	3002	2997
	CB11_content	ERR197746	7796	7555
	CB11_tissue	ERR197747	3166	3145
	CB14_content	ERR197748	3629	3601
	CB14_tissue	ERR197749	4533	4504
	CB15_content	ERR197750	4466	4430
	CB15_tissue	ERR197751	7541	7512
	CB16_content	ERR197752	4266	4231
	CB16_tissue	ERR197753	2964	2957
	CB17_tissue	ERR197754	2940	2923
	CB19_tissue	ERR197755	6379	6330
	CB1_tissue	ERR197756	3415	3403
	CB24_content	ERR197757	3953	3922
	CB24_tissue	ERR197758	4189	4160
	CB26_tissue	ERR197759	4300	4278
	CB27_tissue	ERR197760	4987	4964
	CB28_tissue	ERR197761	5084	5047
	CB29_tissue	ERR197762	9166	9097
	CB2_tissue	ERR197763	4139	4117
	CB3_tissue	ERR197764	4377	4367
DB	DB11_content	ERR197765	4963	4943
	DB11_tissue	ERR197766	3613	3601

	DB13.1_content	ERR197767	4324	4310
	DB13.1_tissue	ERR197768	7618	7570
	DB14_content	ERR197769	3216	3204
	DB14_tissue	ERR197770	3253	3235
	DB15.1_content	ERR197771	3407	3394
	DB15.1_tissue	ERR197772	4841	4807
	DB18_content	ERR197773	4454	4443
	DB18_tissue	ERR197774	3357	3345
	DB19.1_content	ERR197775	3946	3921
	DB19.1_tissue	ERR197776	4556	4524
	DB2.2_content	ERR197777	7083	6921
	DB2.2_tissue	ERR197778	4806	4770
	DB20.1_content	ERR197779	2727	2725
	DB20.1_tissue	ERR197780	4314	4280
	DB22.1_tissue	ERR197781	4293	4270
	DB3_content	ERR197782	6621	6581
	DB3_tissue	ERR197783	2234	2223
	DB7_content	ERR197784	6816	6774
	DB7_tissue	ERR197785	2928	2912
	DB8.1_content	ERR197786	5649	5618
	DB8.1_tissue	ERR197787	2674	2663
ES	ES11_tissue	ERR197788	5754	5715
	ES12_content	ERR197789	3035	2982
	ES12_tissue	ERR197790	8060	8012
	ES13_content	ERR197791	5618	5566
	ES13_tissue	ERR197792	5640	5615
	ES14_content	ERR197793	5561	5374
	ES14_tissue	ERR197794	6490	6447
	ES15_content	ERR197795	5113	4983
	ES15_tissue	ERR197796	3351	3334
	ES18_content	ERR197797	4646	4563
	ES18_tissue	ERR197798	5736	5697
	ES19_tissue	ERR197799	7402	7381
	ES22_content	ERR197800	4521	4448
	ES22_tissue	ERR197801	4038	4014
	ES23_content	ERR197802	4341	4309
	ES23_tissue	ERR197803	4821	4797
	ES26_content	ERR197804	5320	5269
	ES26_tissue	ERR197805	7367	7297
	ES27_content	ERR197806	4789	4741
	ES27_tissue	ERR197807	8815	8744
	ES28_content	ERR197808	4092	4004
	ES28_tissue	ERR197809	5653	5620
	ES29_tissue	ERR197810	4181	4169
	ES2_tissue	ERR197811	5242	5220
	ES30_content	ERR197812	7145	7082
	ES30_tissue	ERR197813	4397	4363
	ES31_content	ERR197814	5318	4999
	ES31_tissue	ERR197815	1368	1358
	ES4_tissue	ERR197816	4349	4322
	ES5_tissue	ERR197817	3335	3323

	ES6_content	ERR197818	4992	4932
	ES6_tissue	ERR197819	2650	2632
	ES7_tissue	ERR197820	2841	2830
	ES8_content	ERR197821	4009	3958
	ES8_tissue	ERR197822	1874	1870
	ES9_tissue	ERR197823	2344	2333
LO	LO1.1_content	ERR197824	5463	5391
	LO1.1_tissue	ERR197825	5945	5921
	LO10.1_content	ERR197826	3660	3647
	LO10.1_tissue	ERR197827	3688	3661
	LO13XX_content	ERR197828	3798	3744
	LO13XX_tissue	ERR197829	3244	3228
	LO14.1_content	ERR197830	3831	3803
	LO14.1_tissue	ERR197831	2894	2881
	LO2.1_content	ERR197832	5485	5397
	LO2.1_tissue	ERR197833	4061	4033
	LO3.1_content	ERR197834	8397	8308
	LO3.1_tissue	ERR197835	4881	4852
	LO4.1_content	ERR197836	5644	5579
	LO4.1_tissue	ERR197837	4258	4229
	LO5.1_content	ERR197838	7076	7009
	LO5.1_tissue	ERR197839	4002	3984
	LO6.1_content	ERR197840	6523	6479
	LO6.1_tissue	ERR197841	3322	3313
	LO7.1_content	ERR197842	5150	5122
	LO7.1_tissue	ERR197843	5162	5130
	LO8.1_content	ERR197844	2736	2712
	LO8.1_tissue	ERR197845	4694	4677
	LO9.1_content	ERR197846	3222	3212
	LO9.1_tissue	ERR197847	4600	4577
MC	MC11_content	ERR197848	3450	3429
	MC12_content	ERR197849	4175	4151
	MC13_content	ERR197850	6015	5996
	MC15_content	ERR197851	4437	4409
	MC02_content	ERR197852	2501	2487
	MC22_content	ERR197853	2866	2856
	MC28_content	ERR197854	2925	2912
	MC06_content	ERR197855	2874	2858
	MC02_tissue	ERR197856	4672	4120
	MC03_tissue	ERR197857	3591	3152
	MC11_tissue	ERR197858	3704	3283
	MC12_tissue	ERR197859	3464	3458
	MC13_tissue	ERR197860	5815	5113
	MC15_tissue	ERR197861	5341	4770
	MC20_tissue	ERR197862	4917	4381
	MC21_tissue	ERR197863	3432	3010
	MC22_tissue	ERR197864	5954	5201
	MC23_tissue	ERR197865	5462	4758
	MC25_tissue	ERR197866	7249	7217
	MC28_tissue	ERR197867	4244	3813
	MC29_tissue	ERR197868	4929	4342

	MC30_tissue	ERR197869	5080	5063
	MC33_tissue	ERR197870	3639	3628
	MC34_tissue	ERR197871	3056	3048
	MC35_tissue	ERR197872	2152	2149
	MC06_tissue	ERR197873	3588	3581
NA	NA10.1_content	ERR197874	4896	4844
	NA10.1_tissue	ERR197875	2874	2865
	NA11.1_content	ERR197876	4938	4837
	NA11.1_tissue	ERR197877	3402	3386
	NA12.1_tissue	ERR197878	3713	3686
	NA13_tissue	ERR197879	2907	2898
	NA14.1_content	ERR197880	2886	2878
	NA14.1_tissue	ERR197881	1031	1028
	NA2.1_content	ERR197882	4587	4560
	NA2.1_tissue	ERR197883	3420	3404
	NA3.1_content	ERR197884	8642	8528
	NA3.1_tissue	ERR197885	3246	3227
	NA4.1_content	ERR197886	8171	8011
	NA4.1_tissue	ERR197887	2792	2780
	NA6.1_content	ERR197888	8508	8381
	NA6.1_tissue	ERR197889	2704	2685
	NA7.1_content	ERR197890	5069	5019
	NA7.1_tissue	ERR197891	6789	6761
	NA8.1_content	ERR197892	10361	10164
	NA8.1_tissue	ERR197893	4080	4053
	NA9.2_content	ERR197894	4600	4546
	NA9.2_tissue	ERR197895	4463	4446
SL	SL11_content	ERR197896	9740	9661
	SL11_tissue	ERR197897	2968	2956
	SL13.1_content	ERR197898	8919	8884
	SL13.1_tissue	ERR197899	5512	5479
	SL14.1_content	ERR197900	6005	5971
	SL14.1_tissue	ERR197901	5097	5082
	SL15.1_content	ERR197902	6887	6847
	SL15.1_tissue	ERR197903	6077	6020
	SL16_content	ERR197904	12890	12717
	SL16_tissue	ERR197905	4473	4448
	SL17.1_content	ERR197906	9438	9319
	SL17.1_tissue	ERR197907	4328	4308
	SL18.1_content	ERR197908	9949	9826
	SL18.1_tissue	ERR197909	4078	4050
	SL3_content	ERR197910	6712	6661
	SL3_tissue	ERR197911	4747	4723
	SL4.1_content	ERR197912	6078	6018
	SL4.1_tissue	ERR197913	5131	5100
	SL5.1_content	ERR197914	5134	5087
	SL5.1_tissue	ERR197915	4190	4178
	SL8.3_content	ERR197916	11974	11830
	SL8.3_tissue	ERR197917	3510	3499
	SL9.1_content	ERR197918	9122	9003
	SL9.1_tissue	ERR197919	3449	3439

Table S2. Summary of mitochondrial D-loop and microsatellite variation.

	D-loop sequences		18 autosomal microsatellite loci		
population	N_{ind}	N_{haplo}	N	H_{exp}	H_{obs}
CB	15	11	29.33	0.87	0.62
SL	12	6	22.56	0.81	0.54
MC	18	8	35.67	0.86	0.71
ES	22	9	42.44	0.79	0.59
AN	18	9	35.44	0.83	0.61
NA	12	9	23.78	0.84	0.67
LO	12	3	23.56	0.79	0.67
DB	12	7	24	0.9	0.6
mean	15.13	7.75	29.6	0.84	0.63

N_{ind}	Number of individuals analyzed
N_{haplo}	Number of different haplotypes
N	Number of gene copies
H_{exp}	expected heterozygosity
H_{obs}	observed heterozygosity
A_{av}	average number of alleles across all loci

Supplementary Table 5: Analysis of bacterial genera in the current and two previously published studies of lab mice.

	Current Study	Staubach et al. 2012	Ubeda et al. 2012
Number of genera +/- SE	36.9 +/- 1.4	23.5 +/- 2.6	29.3 +/- 1.9
Average abundances of shared genera between wild and lab mice			
Acetanaerobacterium	4.0625E-04	2.2519E-04	1.7797E-03
Alistipes	1.3513E-02	1.4137E-02	9.9356E-02
Allobaculum	5.0000E-05	1.1513E-03	1.1525E-03
Anaerophaga	6.7250E-03	3.2822E-05	1.1356E-03
Anaerostipes	4.7313E-03	2.4515E-02	1.4068E-02
Bacteroides	3.5662E-01	1.2438E-02	4.2525E-02
Barnesiella	5.2626E-02	8.8574E-02	1.4112E-01
Blautia	5.9126E-03	1.1625E-04	2.3051E-03
Butyricicoccus	6.4375E-04	8.1675E-03	1.9153E-03
Butyrivibrio	7.0565E-03	1.4397E-05	1.9136E-02
Clostridium	8.1375E-03	1.2579E-03	4.9153E-04
Coprobacillus	1.6319E-02	1.8498E-04	3.7288E-04
Coprococcus	6.4315E-03	2.3979E-02	2.3237E-02
Dorea	9.2438E-03	5.9184E-03	1.9237E-02
Enterococcus	1.4375E-04	1.4626E-03	1.6949E-05
Hespellia	5.9191E-03	3.6376E-03	3.7627E-02
Lactobacillus	2.2238E-02	2.0751E-02	3.6966E-02
Lawsonia	6.2565E-03	3.9583E-05	5.0847E-05
Moryella	1.0378E-03	2.1657E-02	1.2712E-03
Mucispirillum	2.5620E-02	4.0738E-03	1.3898E-03
Natronincola	1.8750E-05	8.0138E-05	1.3559E-04
Odoribacter	1.1313E-03	4.3644E-03	2.6102E-03
Oscillibacter	2.2133E-02	1.1003E-02	2.3525E-02
Oxobacter	1.3125E-04	7.2869E-06	4.7458E-04
Papillibacter	8.0814E-03	1.5902E-06	1.4576E-03
Parabacteroides	1.6363E-02	2.9289E-03	5.5932E-04
Paralactobacillus	1.3125E-04	9.6916E-06	5.0847E-05
Paraprevotella	1.7563E-03	1.0014E-02	6.2881E-03
Parasporobacterium	7.4756E-03	1.0643E-03	3.7373E-02
Parasutterella	1.2875E-03	4.6173E-03	1.1356E-03
Persicitalea	1.2500E-05	6.1951E-07	1.6949E-05
Robinsoniella	6.2501E-02	4.5235E-05	1.9198E-01
Roseburia	2.3813E-03	8.4727E-05	4.6102E-03
Ruminococcus	1.2500E-04	2.2552E-05	1.1525E-03
Soehngenia	1.2500E-04	3.3956E-05	3.7288E-04
Sporobacter	2.3876E-03	1.1655E-02	2.2525E-02
Sporobacterium	3.6001E-03	6.4006E-05	2.5271E-02
Staphylococcus	7.8125E-04	5.2411E-04	2.8814E-04
Streptococcus	2.1875E-04	1.4180E-02	8.4746E-05
Syntrophococcus	8.5003E-04	1.7139E-04	5.2373E-03

Table S6. Summary of Mantel tests and linear regression of similarity-distance decay for geographic distance.

Community location	Beta diversitymeasure	Mantel *r*	Mantel *P*	Overall slope[a]	Overall *P*[b]	Local slope[c]	Local *P*	Continental slope[d]	Contenintal *P*
Mucosa	Bray-Curtis	0.179	**0.001**[e]	-0.129	**0.0001**	-0.064	0.1179	-0.184	0.0001
Mucosa	Jaccard	0.172	**0.001**	-0.115	**0.0001**	-0.058	0.1132	-0.161	0.0001
Mucosa	Weighted UniFrac	0.016	0.304	8.96E-09	0.3942	-7.86E-07	0.4259	9.74E-09	0.6552
Mucosa	Unweighted UniFrac	0.033	0.197	6.47E-09	0.2559	-6.05E-07	0.6964	7.48E-09	0.0815
Content	Bray-Curtis	0.128	**0.001**	-0.082	**0.0001**	-0.068	0.1733	-0.063	0.0198
Content	Jaccard	0.125	**0.001**	-0.073	**0.0001**	-0.06	0.0811	-0.055	0.0207
Content	Weighted UniFrac	-0.008	0.601	-5.48E-09	0.5893	-1.41E-06	0.6273	-1.72E-08	0.9969
Content	Unweighted UniFrac	0.061	0.064	1.22E-08	0.1341	-3.08E-07	0.6916	1.76E-08	0.0006

Chapter II Examining enterotypes in wild mice

Introduction

The intestinal microbiota play many important biological roles in mammals and has been extensively studied. In humans, the gut microbiota were for a long time considered to be extremely variable and display no discernable general patterns. However, Arumugam *et al.* (2011) suggested the existence of enterotypes in humans, where 3 major clusters of microbial communities were found based on metagenomic data, for which signature bacterial taxa were found (*Bacteroides*, *Prevotella* and *Ruminococcus*, respectively, for each of the three clusters). This idea was further tested by Wu *et al.* (2011), who suggested that protein and animal fat versus carbohydrates in long-term dietary patterns influence the abundant bacteria species, and thus enterotypes. Shortly thereafter, the existence of enterotypes was reported both in chimps (Moeller *et al.,* 2012) and lab mice (Hildebrand *et al,* 2013). In contrast, by applying alternative testing approaches some scientists view the microbiome as a continuum rather than distinct clusters (Koren *et al.,* 2013),

It is questionable though, whether the conclusion from lab mice is representative of real-world scenarios, since the microbiota of lab mice are typically eliminated and replaced by an artificial collection of microbiota (e.g. the altered Schaedler flora). To analyze enterotypes in wild mice, we used wild-caught mice that were dissected on site. With extensive examination of clustering quality, we established two enterotypes in wild mice. Additional observations come from mice of wild origin, but maintained in the lab for various periods of time, where we found only one enterotype to dominate the gut microbiome. Functional metagenomic analysis indicates that these changes in the microbiome might be associated with the change of the protein/carbohydrate ratio in their diet. Thus, we further analyzed the isotope signatures of mouse tissues that were dissected on site and reconstructed their possible diet patterns. This again supports the conclusion that dietary patterns, mainly differences in protein/carbohydrates, are the leading cause of enterotypes.

This study provides insights into an issue of debate, with so far the most comprehensive evidence supporting the role of diet in shaping enterotypes. Additionally, the fast turnover with regard to enterotypes is also interesting and suggests the value of mice in general for studies of the microbiome, since Wu *et al.* (2011) did not observe a change in humans with a dietary intervention time course of ten days, similar to other recent studies that used extreme diets (David *et al.,* 2013) that are uncommon for humans. In contrast, mice fed lab chow changed their enterotype classification within one week. Further investigation in mice could explore the important biological consequences of enterotypes, for instance functionality of the microbiome as well as the consequences for host health.

Manuscript

Jun Wang, Miriam Linnenbrink, Sven Künzel, Ricardo Fernandes, Marie-Josée Nadeau, Philip Rosenstiel, John F. Baines. **Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice.**
Submitted to Proc. Natl. Acad. Sci. U. S. A..

Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice

Jun Wang [1, 2], Miriam Linnenbrink [1], Sven Künzel [1], Ricardo Fernandes[3,4], Marie-Josée Nadeau[3], Philip Rosenstiel[5] and John F. Baines[1,2,§]

[1]Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, D-24306 Plön, Germany
[2]Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Arnold-Heller-Str. 3, D-24105 Kiel, Germany
[3]Leibniz-Laboratory for Radiometric Dating and Isotope Research, Christian-Albrechts-University of Kiel, Max-Eyth-Str. 11-13, D-24118 Kiel, Germany
[4]Graduate School Human Development in Landscapes, Christian-Albrechts-University of Kiel, Leibnizstr. 3, D-24118 Kiel, Germany
[5]Institute of Clinical Molecular Biology (Team, 2011), Christian-Albrechts-University of Kiel, Schittenhelmstr. 12, D-24105 Kiel, Germany

[§]Corresponding author
E-mail: baines@evolbio.mpg.de

Abstract

Understanding the origins of gut microbial community structure is critical for the identification and interpretation of potential fitness-related traits for the host. The presence of distinct community clusters characterized by differences in the abundance of signature taxa, referred to as enterotypes, is a debated concept first reported in humans and later extended to other mammalian hosts. In this study, we provide a thorough assessment of their existence in wild house mice using a panel of evaluation criteria. We identify support for two clusters that are compositionally similar to clusters identified in humans, chimpanzees and lab mice, characterized by differences in *Bacteroides*, *Robinsoniella* and unclassified genera belonging to the family Lachnospiraceae. To further evaluate these clusters, we (i) monitored community changes associated with moving mice from the natural- to a lab environment, (ii) performed functional metagenomic sequencing and (iii) subjected wild-caught samples to stable isotope analysis to reconstruct dietary patterns. This reveals differences in the proportions of genes involved in carbohydrate *versus* protein metabolism in the functional metagenome, as well as differences in plant- *versus* meat-derived food sources between clusters. In conjunction with wild-caught mice quickly changing their enterotype classification upon transfer to a standard lab chow diet, these results provide strong evidence that dietary history contributes to the presence of enterotype-like clustering in wild mice.

Keywords: Gut microbiome, wild mice, diet, enterotype, metagenomics

Introduction

First large-scale sequencing surveys of the human intestinal microbiome emphasized considerable differences between individuals [1]. Numerous studies expanded upon these initial observations and identified factors such as geography, host genetics, diet and other environmental factors that contribute to inter-individual variability in humans [2,3,4] and other animals [5,6,7]. However, important questions pertaining to the nature and origins of gut microbial community structure remain to be answered. In particular, the presence of enterotypes, or distinct clusters characterized by the abundances of signature bacterial genera, is a debated concept first reported in humans [3]. Wu *et al*. [8] soon after provided evidence that the proportion of protein and animal fat versus carbohydrates in long-term dietary habits contribute to determining and individual's enterotype, whereas the intriguing existence of analogous enterotypes in chimpanzees suggests that they may reflect more ancient features of host-microbial physiology and homeostasis in the gut [9]. On the other hand, a meta-analysis of enterotypes across human body sites found their identification to be sensitive to distance metrics and clustering methods used, in addition to a majority of gradients- rather than distinct clusters being present in signature taxa [10].

Most recently, Hildebrand *et al.* [11] provide a first assessment of the possibility of enterotypes in the house mouse, a critical and widely employed model of gut microbiome research, whereby two enterotype-like clusters among the five lab strains studied are identified. Furthermore, differences in low-grade inflammation between these two groups suggest possible mechanisms, *e.g*. differences in inflammation-inducing taxa such as Enterobacteriaceae. However, which aspects of inflammation and differences in taxon abundance represent cause or consequence remains unclear, as does the possibility of other contributing factors. In addition, the natural state of bacterial communities is potentially misrepresented in laboratory settings due to practices such as inoculating lab mice with limited cocktails of bacterial strains (*e.g*. altered Schaedler flora) and feeding standard diets.

In this study, we use the criteria outlined by Koren *et al.* [10] to first provide an assessment of enterotype-like clustering in a panel of wild-caught house mice previously included in a biogeographic survey of intestinal communities [7] and

compare them to wild-caught mice housed in a lab environment. We find that two clusters similar to those identified in humans, chimpanzees and lab mice are frequently present in the wild, but are nearly lost among mice housed in the lab for one year. As a follow-up, we caught additional mice, transferred them to the lab and regularly monitored the dynamics of their fecal communities over a period of 12 weeks beginning from the time of sampling. This documents the rapid loss of one cluster abundant in the wild under lab conditions. Furthermore, deep shotgun metagenomic sequencing reveals significant differences in functional microbiomic categories, in particular pertaining to protein *versus* carbohydrate metabolism between the two groups. Finally, stable isotope analysis performed on wild-caught samples reveals significant differences in long-term diet between the two groups, consistent with the observations based on functional metagenomic data.

Results

Enterotype-like clusters in wild mice

To evaluate the presence of enterotypes in wild mice, we first analyzed the cecal contents of 80 samples collected and dissected on-site at eight geographic locations in Western Europe (on average 530 km apart, across Germany and France; [7]) following established guidelines [10,11]. To this end, optimal clustering was evaluated for five beta diversity metrics, six clustering methods and three judging criteria (see Methods; Table 1). Three of the beta diversity indices were applied to genus-level community composition (Bray-Curtis (BC), Jensen-Shannon (JS) and Jaccard (JA)) and two are phylogenetic-based (weighted and unweighted UniFrac distances (WU and UW, respectively)). Using BC, three out of six clustering methods (partitioning around medoids (pam), kmeans and hierarchical clustering using ward linkage) revealed an optimum of two clusters using all three judging criteria (Table 1). Nearly identical results were obtained for the remaining four beta diversity metrics by the same three clustering methods. To assess the robustness of clustering, we focused on prediction strength (PS) and the average silhouette score (SI), which are absolute measures of clustering quality. While SI offers at best weak support across beta diversity metrics and clustering methods, PS provides moderate (>0.8) to strong (>0.9) support for two clusters in numerous instances (Table 1), which is comparable to fecal samples in the human microbiome project [4,10]. This suggests that enterotype-like

clusters are also present in wild mice based on genus-level composition, although the same analyses applied to operational taxonomic units (OTUs) revealed less consistent results (Supplemental Table 1).

Compositional analysis of enterotype-like clusters

In order to describe and compare our results to previous studies of humans [3,8], chimpanzees [9] and lab mice [11], we further focus on the two clusters identified with genus-level composition and pam clustering, which is the clustering method used in these previous studies. Furthermore, we focus on the BC beta diversity measure as it (1) gives the highest prediction strength using pam (Table 1), (2) is highly correlated to JS (Mantel test, r=0.9632, p=0.001) and JA (Mantel test, r=0.9887, p=0.001) and (3) allows the application of SIMPER [12], a well-established method to identify taxa contributing to similarity within- and dissimilarity between groups. Based on this combination (BC and pam), we define two enterotype-like clusters, termed "E1" and "E2", which are present in all eight geographic regions (80 sites distributed across eight geographic locations on average 530 km apart [7]) (Fig. 1). Using the multivariate ANOVA implemented in analysis of dissimilarity ('*adonis'*; see Methods), we determined that these two clusters explain 39.45% of the variation in BC distances.

To identify signature taxa, we first applied the SIMPER method, which identifies *Bacteroides* as the largest contributor (31.68%) to dissimilarity between groups and similarity within E1 (77.78%). The second largest contributor (8.59%) to dissimilarity between groups is *Robinsoniella*, which contributes 14.04% to similarity within E2 and is also the most abundant genus in this group (average 13.6%). In addition, we identified an unclassified genus belonging to Lachnospiraceae (unclassified_Lachnospiraceae) that contributes 17.15% of the similarity within E2, but is second in terms of its abundance in E2 (average 12.45%). Second, we tested for significant differences in abundance among all taxa displaying ≥1% abundance in the whole data set. Ten out of these 16 genera displayed significant differences in abundance (Table 2). Finally, we also applied alpha diversity analyses focusing on species richness (Chao1), evenness (Shannon) and phylogenetic diversity, revealing significantly lower diversity in E1 by all three measures (Supplemental Fig. 1).

Wild mice moved to a lab environment

During the course of fieldwork conducted in 2009 [7], additional mice sampled in the Massif Central region of France were brought back to the breeding facility of the Max Planck Institute for Evolutionary Biology. In total, ten mice captured from seven different farms were analyzed after one year of lab housing. In order to assess the possible influence of lab-housing on the distribution of enterotype-like clusters among mice, we added these ten cecal content samples to the 80 samples dissected on-site and repeated the evaluation of clustering performed above. This reveals nearly identical results as described above, whereby most consistent support is provided for two clusters (Supplemental Table 2). Intriguingly, however, when using BC and pam clustering as described above, all ten lab-housed samples belong to E2. This is in stark contrast to the mice sampled from the same Massif Central region but dissected on-site, which display a proportion of 5:3 of E1:E2, respectively (Table 3), and is highly unlikely to be observed by chance (Fisher's exact test, p=0.001508). However, we lack knowledge of the status of these ten mice at the time of capture. Thus, these results offer only circumstantial evidence that common environmental conditions experienced in the lab influence enterotype-like clustering.

In order to directly document whether individuals shift their enterotype classification upon moving from the wild to a lab environment, additional sampling of wild mice was performed in the Cologne/Bonn region of Germany in 2012, which previously displayed an E1:E2 ratio of 1:4 in 2010 (Table 2) [7]. In total 14 mice were sampled ("week 0"), of which eight were transferred to the lab breeding facility to be monitored at regular intervals (1, 2, 3, 4, 6, 8, 12 weeks) by sampling feces (see Material and Methods). One mouse was not sampled at week 2 due to treatment for a skin injury, while another was sampled only until week 4, after which it gave birth to a litter and was removed from the experiment. Thus, in total, 66 feces samples from eight different time points were analyzed using the same distance measures, clustering methods and judging criteria as described above.

To remain consistent with our previous analysis we used the panel of 80 wild-caught samples as a reference for classification and applied the suite of beta diversity metrics, clustering methods and judging criteria together with the 66 additional feces samples (Supplemental Table 3). Again, this yielded the most consistent support for two clusters, for which *adonis* analysis explains a large portion of variation in all five beta diversity measures (BC: 32.11%; JS: 45.42%; JA: 21.96%; WU: 41.02%; UW: 20.12%). These effects are two- to three-fold greater than the differences that can be

explained between fecal *vs*. cecal content sampling (BC: 11.78%; JS: 17.92%; JA: 8.24%; WU: 6.82%; UW: 3.33%), which is unavoidable for the purpose of monitoring changes over time. Based on BC and pam clustering, eight E1 and six E2 classifications are present among the 14 mice at the initial point of sampling (Figure 2). Subsequently, eight mice (three E1 and five E2) were sampled at regular intervals in the lab up to 12 weeks post-capture. Among the five individuals initially classified as E2, all remained E2, with the exception of two individuals that displayed a short-term (*i.e*. a single sampling timepoint) transition to E1. In contrast, all three individuals classified as E1 at the onset of sampling quickly (*i.e*. after one week) shifted to E2 and remained so throughout the remaining sampling period (Figure 2). These changes are recapitulated by observing the differences in abundance of the signature taxa *Bacteroides* and *Robinsoniella*, where a drastic initial drop in abundance is observed for *Bacteroides* among the mice initially classified as E1 (Figure 3a). Accordingly, temporary spikes in *Bacteroides* abundance are also associated with fluctuations between E1 and E2 in two mice initially classified as E2 (Figure 3c).

Functional metagenomic analysis of enterotype-like clusters

To determine what functional genomic changes may be associated with different enterotype-like clusters, we sequenced the 14 mice captured in the Cologne/Bonn region as well as the week 1 and 12 time points of the eight mice brought back to the lab using an Illumina Hiseq platform (Supplemental Table 4; Methods). Raw sequence data was subsequently submitted and analyzed using the MG-RAST pipeline [13]. This pipeline provides four levels of functional categories, from most general to most specific, based on the results of mapping to KEGG pathway maps [14]. First, we evaluated community-level differences in functional genomic content after standardizing the abundances of each category to the total number of proteins predicted for each metagenomic sample (*i.e*. in a manner analogous to taxon abundance relative to the number of 16S rRNA gene sequences in a sample). Based on the application of *adonis* to each of the three applicable beta diversity measures (BC, JS, JA), significant differences are apparent for both abundance-based measures (BC and JS) at all four levels of functional categories, for which enterotype status explains between 15% and 30% of the variation. Comparatively less variation was

explained (~10%) by a presence/absence measure (JA), although significant differences are present at three category levels (Table 4).

Next, we identified the major functional differences in the microbiome between the two enterotype classes using the SIMPER method in a manner analogous to that used to identify signature genera among 16S rRNA sequence data, but instead using the standardized functional categories from each level produced by MG-RAST. At the highest level of functional categories, genes involved in protein metabolism are the largest contributor to differences in BC between E1 and E2 (10.87%), whereby E1 displays a higher proportion (9.13% compared to 8.62% in E2) (Figure 4a, Table 5). The next largest contributors are "mobility and chemotaxis" (explaining 9.33% of the differences between E1/E2) and "carbohydrates" (explaining 8.29% of the differences between E1/E2), which are on average lower in E1 (1.04% and 14.94% compared to 1.89% and 15.51% in E2, respectively) (Figure 4b). At the second level of functional categories, "protein biosynthesis" and "di/oligo-saccharides" contribute most to the differences between enterotypes (4.63% and 3.58%, respectively), where again genes involved in protein metabolism are higher in E1 (6.57% compared to 6.12% in E2), but genes involved in carbohydrate metabolism are lower (2.03% in E1 compared to 2.76% in E2). Similar patterns are observed at the remaining two levels of functional categories, although the contribution of each function becomes less prominent (Table S5), likely due to the increase in number of pathways.

Stable isotope analysis of wild mice

The results of the functional metagenomic analysis suggests that differences in diet, *e.g*. at the level of protein *vs*. carbohydrate intake, might contribute to the occurrence of enterotype-like clusters in wild mice. To test this hypothesis, we further analyzed stable isotopes ($\delta^{15}N$ and $\delta^{13}C$) in muscle and liver tissues of original wild-caught mouse samples from Linnenbrink *et al*. [7], then quantified likely diets using the Bayesian mixing model FRUITS [15], which incorporates information from representative potential food sources (see Materials and Methods). We found that on average, food sources similar to standard lab chow (chow) are a major component of the long-term diet (39.95±1.31%) of wild mice, followed by grain (36.42±4.06%), meat (16.37±2.67%) and insects (8.14±2.90%). We also found significant differences in the major components of plant-derived food sources (chow) between the two enterotypes (mean for E1 38.77% and E2 39.35%, Wilcoxon test p=0.027), and the

opposite pattern for animal-derived food sources (meat, mean for E1 17.02% and E2 16.80%) although the difference is not significant (Wilcoxon test p=0.187). No significant differences were found for grain or insects between enterotypes (Figure 5). Additionally, the application of linear models reveals *Bacteroides* abundance to be positively correlated with %meat and negatively correlated with %chow in the diet, while the opposite pattern is present for *Robinsoniella* abundance, although neither is significant (all p>0.05). We do, however, observe a significant positive correlation between unclassified_Lachnospiraceae and %chow (r^2=0.1089, p= 0.002), as well as a negative correlation between unclassified_Porphyromonadaceae and %meat (r^2= 0.0964, p=0.005), both consistent with their differences in abundance between enterotypes (Table 2).

Discussion

Our study provides several points of biological insight into the origin and functionality of enterotype-like community clusters. In particular, we describe the state of enterotypes in natural populations of house mice, and therefore identify the taxa and structuring of bacterial communities with which the mouse lineage most likely coevolved. Furthermore, we provide detailed information on the circumstances in which communities change their enterotype classification as well as the functional genomic consequences of such changes. Finally, the results of stable isotope analysis are consistent with underlying functional genomic changes (*i.e*. protein *vs*. carbohydrate metabolism) providing additional evidence that differences in recent dietary history/behavior may contribute to the existence of different enterotype states.

Overall, our study adds to a growing picture of similarity in signature taxa across host species. Following the nomenclature in humans [3] and chimps [9], the major contributing genus to E1 in our study, *Bacteroides*, is also a significant contributor to "enterotype 1" in humans and chimps, as is *Parabacteroides* (Table 2). Similarly, unclassified Lachnospiraceae are significantly more abundant in E2 in our study and follow the same pattern in the human/chimp "enterotype 2". Finally, the human/chimp "enterotype 3" is characterized in part by *Ruminococcus*. Although we identify no equivalent of a third enterotype in mice, *Robinsoniella*, a signature taxon of E2, is a close relative of *Ruminococcus*.

The study of Hildebrand *et al.* [11] identifies somewhat different signature taxa, which can possibly be attributed to investigating inbred lab mice. Nevertheless, we identify the same number of enterotypes and similarities in the signature taxa between the two mouse studies. In particular, although a different nomenclature is used, their "ET1" displays increased Lachnospiraceae as does our E2. Interestingly, however, the other major contributor to our E2, *Robinsoniella*, appears absent in the Hildebrand *et al.* [11] study. We previously identified this genus as the most consistent and abundant member of the Firmicutes in wild mice [7], suggesting an important biological role that may be either absent or replaced by functional equivalents in the communities of lab mice. In addition, we identify one enterotype with reduced alpha diversity, E1, for which a similar phenomenon is present in "ET2" from Hildebrand et al. [11] A possible reason for this correspondence may be the increase in *Bacteroides* and Bacteroidaceae in the current study (E1) and Hildebrand *et al.* [11] (ET2), respectively, at the expense of other community members.

Soon after the first report of enterotypes in humans [3], Wu *et al.* [8] reported that long-term diets were strongly associated with enterotype status, where individuals with more animal fat and protein intake were more likely to present the *Bacteroides*-dominated enterotype compared to those with more carbohydrate intake being associated with a *Prevotella*-dominated enterotype. The most compelling evidence for dietary influences on enterotypes given by our study is that the *Bacteroides*-dominated enterotype in wild mice, E1, quickly shifts to the alternative state of E2 after being transferred to the lab. Presumably, the standard chow fed to the mice in our breeding facility is in contrast to the greater variety of food sources available in the wild, which includes animal material such as insects, worms, *etc.* Thus, a lab chow-restricted diet would be expected to deliver a more consistent source of plant-derived nutrients. Indeed, when comparing the major functional genomic categories that differentiate mouse enterotypes by shotgun metagenomic sequencing, the *Robinsoniella*-dominated E2 displays a greater proportion of categories attributed to carbohydrate metabolism. Finally, by reconstructing dietary patterns via stable isotope analysis of liver and muscle tissue together with surrogates of food sources similar to the range available in the wild, we revealed a pattern of diet composition that is consistent with a more meat-like food intake in E1 and more chow-like food intake in E2. The fact that the wild mice included in our study change their enterotype status within one week after being transferred to the lab is in contrast to the ten-day dietary

intervention conducted in humans by Wu *et al.* [8], where no change was observed. This may in part be explained by the higher metabolic rate of small mammals such as the mouse.

The clear presence of a single enterotype among the wild mice transferred to our lab facility raises the question as to why two enterotypes are observed among inbred lab mice that are also fed a standard diet. In their study, Hildebrand *et al*. [11] identify "low-grade" inflammation as a possible contributing factor, and others have indeed suggested that lab mice may display compromised immune functioning due to their limited housing/environmental conditions [16,17]. Another possibility is that alternatively, related enterotypes may emerge over time among the altered species assemblages present in lab mice. Future studies incorporating additional lab facilities, time series- and gnotobiotic experiments may help shed light on these questions.

Conclusions

In summary, we provide evidence for the existence of two functionally different enterotype-like clusters present in wild house mice. Remarkably, these display several characteristics in common with those of the distantly related human and chimpanzee hosts, suggesting the existence of ancient shared traits among the bacterial communities that assemble within mammalian hosts. Finally, we provide additional evidence that dietary habits may be the most important contributing factor changes in enterotype status, which warrants more intensive future research to understand the impact of diet-microbiome interactions on human health and disease.

Methods

Animal material and sampling

The cecal contents of 80 wild-caught mice described by Linnenbrink et al. [7] and an additional ten mice transferred from the Massif Central region of France were separated from cecal tissue and preserved in 4 mL RNALater according to manufacturer's instructions until further processing, as previously described [7]. Of the ten mice from the Massif Central that were transferred to the breeding facility of the Max Planck Institute for Evolutionary Biology, six were housed in individual cages and four were maintained as two breeding pairs to generate offspring to be

included in an outbred colony derived from this region [18]. The four mice involved in breeding were maintained as pairs (*i.e*. co-housed) for one to two months, but were housed separately for > six months prior to sacrifice.

To directly monitor changes in bacterial communities associated with transfer to the lab, 14 mice were newly sampled in 2012 from the same Cologne/Bonn region of Germany sampled in 2010 [7]. Feces were collected from all mice within several hours of retrieving the traps. Six mice failed to survive transportation to the lab and were thus only included as "week 0" samples. The remaining eight surviving mice were brought back to the lab breeding facility and monitored at regular intervals from the point of capture (1, 2, 3, 4, 6, 8, 12 weeks) by sampling feces. One mouse was not sampled at week 2 due to treatment for a skin injury, while another was sampled only until week 4, after which it gave birth to a litter and was removed from the experiment. Thus, in total, 66 feces samples from seven different time points were analyzed.

DNA Extraction

Bacterial DNA from cecal contents and feces was extracted using the QIAmp DNA stool mini kit (QIAGEN). Approximately 200 mg of material was transferred to 2-mL screw-cap tubes containing 50 mg each of 0.1 mm, 0.5 mm and 1 mm glass beads (BioSpec Products). The tubes/beads were treated with UV exposure for 2 h prior to performing the extraction. After adding 1.4 mL ASL lysis buffer, bead beating was performed using the Precellys (Peqlab) for 3 x 15 s at 4723 g. Samples were then heated to 95 °C for 10 min, after which the manufacturer's protocol was followed.

Bacterial 16S rRNA gene sequencing and processing

The 27F-338R primer pair spanning the hypervariable regions V1 and V2 was used for PCR and barcoded pyrosequencing on the 454 GS-FLX platform with Titanium sequencing chemistry as previously described [2]. Sequences were filtered using MOTHUR version 1.22.2 [19] with the inclusion criteria of mean quality score >20 and length ≥250 bp. Sequences were assigned to samples by exact matches of 10 bp barcodes. For each sample, random subsets of 1000 sequences were extracted to normalize coverage and taxonomical classification was performed using RDP Classifier [20]. USEARCH/UCHIME verion 5.2.32 [21,22] was used to identify chimeric sequences and perform sequencing clustering into operational taxonomic

units using default parameters (OTU-based analyses are presented in the Supplemental Material).

Bacterial community analysis

Bacterial community analyses including beta diversity metrics (Bray-Curtis, Jaccard) and analysis of dissimilarity (‘*adonis’*, which performs a multidimensional analysis of variance on distance matrices and was applied to beta diversity metrics) were carried out using the ‘VEGAN’ R package [23]. The Jensen-Shannon distance was calculated using the “flexmix” R package [24]. Weighted and unweighted UniFrac distance matrices were calculated from a maximum-likelihood tree constructed by FastTree [25] as implemented in MOTHUR.

To evaluate the presence of enterotype-like clustering, we used the original criteria of Arumugam *et al.* [3], which uses the Calinski-Harabasz index as an indicator for best clustering, in addition to the silhouette score and prediction strength suggested by Hildebrand *et al*. [11] and Koren *et al*. [10], respectively. The following clustering algorithms implemented in R packages were used: partitioning around medoids (pam; R package "cluster"), kmeans (R package “flexclust”) and hierarchical clustering using “single”, “complete”, “average” and “ward” linkage (R function “hclust”). To identify taxa contributing to groups based on Bray-Curtis distance, we applied the SIMPER method [12], which identifies taxa contributing to similarity within- and dissimilarity between groups and ranks their contribution.

Shotgun metagenomic sequencing and analysis

Bacterial DNA from fecal samples was subject to full metagenomic sequencing on the Illumina HiSeq 2000 platform with paired-end 100 bp reads. Each sample was sequenced on a ¼-lane, resulting in an average 8.54E9 bp of sequence per sample. After stringent quality filtering using the Fastq Tool Kit (each sequence required to have >99% of its nucleotides with a quality score ≥ 30), sequences were assembled using Meta-Velvet [26] under default parameters, resulting in an average of 1.36E8 bp of assembled reads per sample and then submitted to MG-RAST (http://metagenomics.anl.gov/) [13] for further analysis. Assembly prior to submission to was necessary due to the Fastq data exceeding the limits of MG-RAST. The automated pipeline provided by MG-RAST was used to obtain taxonomic classification and functional annotation of genes predicted by FragGeneScan using the

BLAT program referencing the M5NR database. For each sample, taxonomical and functional information was extracted and subjected to the same statistical analyses applied to 16S rRNA gene data using the VEGAN package in R [27].

Stable isotope analysis and diet reconstruction

Approximately 200 mg each of liver and muscle tissue were recovered from enthanol preserved carcasses subjected to extraction with a methanol/dichloromethane mixture to remove lipids [28] and dried at 70 ºC for 48 hours. The stable isotopes and concentrations of nitrogen and carbon were analyzed simultaneously with a THERMO/Finnigan MAT V isotope ratio mass spectrometer and THERMO Flash EA 1112 elemental analyzer at Braford University, UK. The ratios of stable isotopes are given using the conventional delta notion ($\delta^{15}N$ and $\delta^{13}C$), with respect to atmospheric nitrogen and Vienna PeeDee Belemnite standards. Additionally we included four main foods groups that represent the range of possible dietary sources of wild mice. These included wheat grains from a rat food mixture and standard mouse lab chow (chow) as plant-derived food sources, as well as dried meal worms (insect) and meat-based dog food (meat). To re-construct the most probable diet structure with respect to the relative proportion of each food standard listed above, we combined the isotope signatures from both muscle and liver and obtained dietary estimates using the Bayesian mixing model FRUITS [15] (http://sourceforge.net/projects/fruits/).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) grants BA 2863/2-2 to J.F.B., Excellence Cluster “Inflammation at Interfaces” to P.R. and J.F.B. and the Max Planck Society.

References

1, Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. *Science* 308(5728):1635-1638.

2, Rausch P, Rehman A, Kunzel S, Hasler R, Ott SJ, Schreiber S, Rosenstiel P, Franke A, Baines JF. 2011. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proc. Natl. Acad. Sci. U S A* 108(47):19030-19035.

3, Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM et al. 2011. Enterotypes of the human gut microbiome. *Nature* 473(7346):174-180.

4, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C et al. 2009. The NIH Human Microbiome Project. *Genome Res*. 19(12):2317-2323.

5, Ochman H, Worobey M, Kuo CH, Ndjango JBN, Peeters M, Hahn BH, Hugenholtz P. 2010. Evolutionary Relationships of Wild Hominids Recapitulated by Gut Microbial Communities. *PLoS Biol.* 8(11):e1000546.

6, Phillips CD, Phelan G, Dowd SE, McDonough MM, Ferguson AW, Hanson JD, Siles L, Ordonez-Garza N, San Francisco M, Baker RJ. 2012. Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Mol. Ecol.* 21(11):2617-2627.

7, Linnenbrink M, Wang J, Hardouin EA, Kunzel S, Metzler D, Baines JF. 2013. The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol.* 22(7):1904-1916.

8, Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R et al. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* 333(6052):105-108.

9, Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. 2012. Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat. Commun*. 3(1179).

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

11, Hildebrand F, Nguyen TLA, Brinkman B, Yunta RG, Cauwe B, Vandenabeele P, Liston A, Raes J. 2013. Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* 14(1).

12, Clarke KR. 2003. Nonparametric Multivariate Analyses of Changes in Community Structure. *Aust. J. Ecol.* 18(1):117-143.

13, Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9.

14, Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27(1):29-34.

15, Fernandes R, Millard AR, Brabec M, Nadeau MJ, Grootes P. Food Reconstruction Using Isotopic Transferred Signals (FRUITS): a Bayesian model for diet reconstruction. *PloS One,* in press.

16, Martin B, Ji S, Maudsley S, Mattson MP. 2010. "Control" laboratory rodents are metabolically morbid: why it matters. *Proc. Natl. Acad. Sci. U S A* 107(14):6127-6133.

17, Abolins SR, Pocock MJO, Hafalla JCR, Riley EM, Viney ME. 2011. Measures of immune function of wild mice, *Mus musculus*. *Mol. Eco*. 20(5):881-892.

18, Ihle S, Ravaoarimanana I, Thomas M, Tautz D. 2006. An analysis of signatures of selective sweeps in natural populations of the house mouse. *Mol. Biol. Evol.* 23(4):790-797.

19, Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ et al. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microb.* 75(23):7537-7541.

20, Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microb.* 73(16):5261-5267.

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

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

23, Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14(6):927-930.

24, Grun B, Leisch F. 2008. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *J. Stat. Softw.* 28(4):1-35.

25, Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* 26(7):1641-1650.

26, Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40(20).

27, R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

28, Cequier-Sanchez E, Rodriguez C, Ravelo AG, Zarate R. 2008. Dichloromethane as a solvent for lipid extraction and assessment of lipid classes and fatty acids from samples of different natures. *J. Agr. Food. Chem.* 56(12):4297-4303.

Table 1. Summary statistics of enterotype analysis.

Beta-diversity indices	Clustering method	Optimal cluster for CH[1]	CH	Optimal cluster for SC	SC	Optimal cluster for PS	PS[2]
Bray-Curtis	pam	**2**	50.4910	**2**	0.3590	**2**	**0.8365**
	kmeans	**2**	42.0325	**2**	0.3492	**2**	0.6972
	ward linkage HC	**2**	44.0426	**2**	0.3260	**2**	0.7290
	average linkage HC	5	20.2763	5	0.3683	2	0.5524
	single linkage HC	**2**	2.9384	**2**	0.3281	**2**	0.4020
	complete linkage HC	3	27.9506	4	0.3520	2	0.5728
Jensen-Shannon	pam	**2**	78.9419	**2**	0.4488	**2**	0.7902
	kmeans	**2**	75.3964	**2**	0.4462	**2**	0.6997
	ward linkage HC	**2**	70.1722	**2**	0.4280	**2**	0.7785
	average linkage HC	5	37.1731	2	0.5097	2	0.5663
	single linkage HC	4	5.3837	2	0.4605	2	0.3335
	complete linkage HC	3	49.3278	5	0.4619	2	0.5778
Jaccard	pam	**2**	32.3442	**2**	0.2673	**2**	**0.8262**
	kmeans	2	31.9530	2	0.2634	5	0.8520
	ward linkage HC	**2**	30.2973	**2**	0.2500	**2**	0.7146
	average linkage HC	5	11.4750	5	0.2707	2	0.5943
	single linkage HC	4	2.2775	2	0.2299	2	0.2873
	complete linkage HC	3	17.4573	4	0.2632	2	0.6376
Weighted Unifrac	pam	**2**	20.1507	**2**	0.3110	**2**	0.7351
	kmeans	**2**	38.6987	**2**	0.3126	**2**	**0.9274**
	ward linkage HC	**2**	38.6076	**2**	0.3049	**2**	0.7462
	average linkage HC	6	15.4513	2	0.3620	2	0.6313
	single linkage HC	3	4.4317	2	0.3084	2	0.3413
	complete linkage HC	5	19.7068	2	0.3620	2	0.6327
Unweighted Unifrac	pam	**2**	5.9405	**2**	0.0701	**2**	0.7930
	kmeans	**2**	6.0814	**2**	0.0767	**2**	**0.9801**
	ward linkage HC	**2**	4.8617	**2**	0.0752	**2**	**0.9193**
	average linkage HC	**2**	4.8668	**2**	0.0764	**2**	**0.9177**
	single linkage HC	5	1.1163	2	0.0264	2	0.2788
	complete linkage HC	**2**	5.8697	**2**	0.0543	**2**	**0.8883**

[1]Bold indicates beta diversities and clustering methods which consistently produce two optimal clustering based on all three judging criteria.
[2]Bold indicates moderate to strong support for two clusters (PS>0.8).
Abbrevations: CH=Calinski-Harabasz index, SC=Silhouette score, PS=prediction strength, HC=hierarchical clustering.

Table 2. Average abundances of major genera (overall mean abundance >1%) between E1 and E2 enterotypes.

Major genera	E1_Bacteroides	E2_Robinsoniella	ANOVA p-value (FDR corrected)[1]
Bacteroides	0.6026	0.1013	**3.20E-15**
unclassified_Lachnospiraceae	0.0297	0.1246	**1.40E-07**
Robinsoniella	0.0098	0.1356	**5.52E-05**
Barnesiella	0.0325	0.0944	**9.55E-04**
Helicobacter	0.0480	0.0364	6.13E-01
Lactobacillus	0.0513	0.0103	1.53E-01
Oscillibacter	0.0139	0.0423	**5.52E-05**
Coprobacillus	0.0273	0.0145	2.88E-01
unclassified_Ruminococcaceae	0.0157	0.0272	**1.28E-02**
unclassified_Porphyromonadaceae	0.0122	0.0300	**1.39E-02**
Parabacteroides	0.0257	0.0095	8.38E-02
unclassified_Rikenellaceae	0.0032	0.0317	**9.55E-04**
Alistipes	0.0053	0.0281	**2.19E-03**
Mucispirillum	0.0123	0.0204	2.77E-01
unclassified_Prevotellaceae	0.0082	0.0138	3.61E-01

[1]Bold numbers denotes significant difference revealed by ANOVA (FDR corrected).

Table 3. Distribution of enterotypes with respect to sampling location and time.

Location	E1_Bacteroides	E2_Robinsoniella
AN[1]	5	3
CB	1	4
DB	7	4
ES	4	10
LO	8	5
MC	5	3
NA	5	5
SL	8	4
MC (after >1 year in lab)	0	10
CB (newly captured in 2012)	8	6

[1]Sampling locations and time are those given in Linnenbrink et al. 2013, unless otherwise noted.

Table 4. Differences in hierarchical functional categories (KEGG pathway levels) between E1 and E2 enterotypes.

KEGG pathway level	Dissimilarity Index	Adonis p-value between E1 and E2	Variances explained by enterotype
Level 1	Bray-Curtis	0.011	0.13065
	Jensen-Shannon	0.003	0.34091
	Jaccard	NA[1]	NA
Level 2	Bray-Curtis	0.007	0.14713
	Jensen-Shannon	0.003	0.34037
	Jaccard	0.001	0.1011
Level 3	Bray-Curtis	0.003	0.1488
	Jensen-Shannon	0.002	0.32386
	Jaccard	0.002	0.10017
Level 4	Bray-Curtis	0.001	0.14434
	Jensen-Shannon	0.002	0.2893
	Jaccard	0.004	0.08484

[1]Jaccard index not applicable at this level due to no differences in presence /absence.

Table 5. Level 1 and 2 KEGG pathways differing between E1 and E2 enterotypes.

Level	Pathways	Mean value in E1 (%)	Mean value in E2 (%)	Contribution to variance(%)[1]	Wilcox test p[2]
Level 1	Protein Metabolism	9.13	8.64	10.78	0.1629
	Motility and Chemotaxis	1.04	1.84	9.22	**0.0062**
	Carbohydrates	14.94	15.54	8.48	**0.0178**
	Phages Prophages Transposable elements/Plasmids	4.23	3.92	6.78	0.0966
Level 2	Protein biosynthesis	6.57	6.14	4.60	0.1416
	Di and oligosaccharides	2.03	2.77	3.64	**0.0009**
	Flagellar motility in Prokaryota	0.85	1.46	3.41	**0.0083**
	Phages Prophages	2.81	2.92	2.94	0.2968
	Transposable elements	1.18	0.75	2.86	**0.0349**
	RNA processing and modification	5.09	4.73	2.22	0.0572
	Fermentation	0.98	1.36	1.93	**0.0021**
	Capsular and extracellular polysacchrides	1.84	2.17	1.86	0.0021
	Ton and Tol transport systems	0.71	0.36	1.74	**0.0072**
	Central carbohydrate metabolism	3.48	3.31	1.74	**0.0013**
	Plant/Prokaryote	6.44	6.2	1.73	0.1132
	Monosaccharides	2.42	2.36	1.46	0.4470

[1]The contribution of each pathway to the total variance is calculated via SIMPER.
[2]Significant differences ($p<0.05$) are shown in bold.

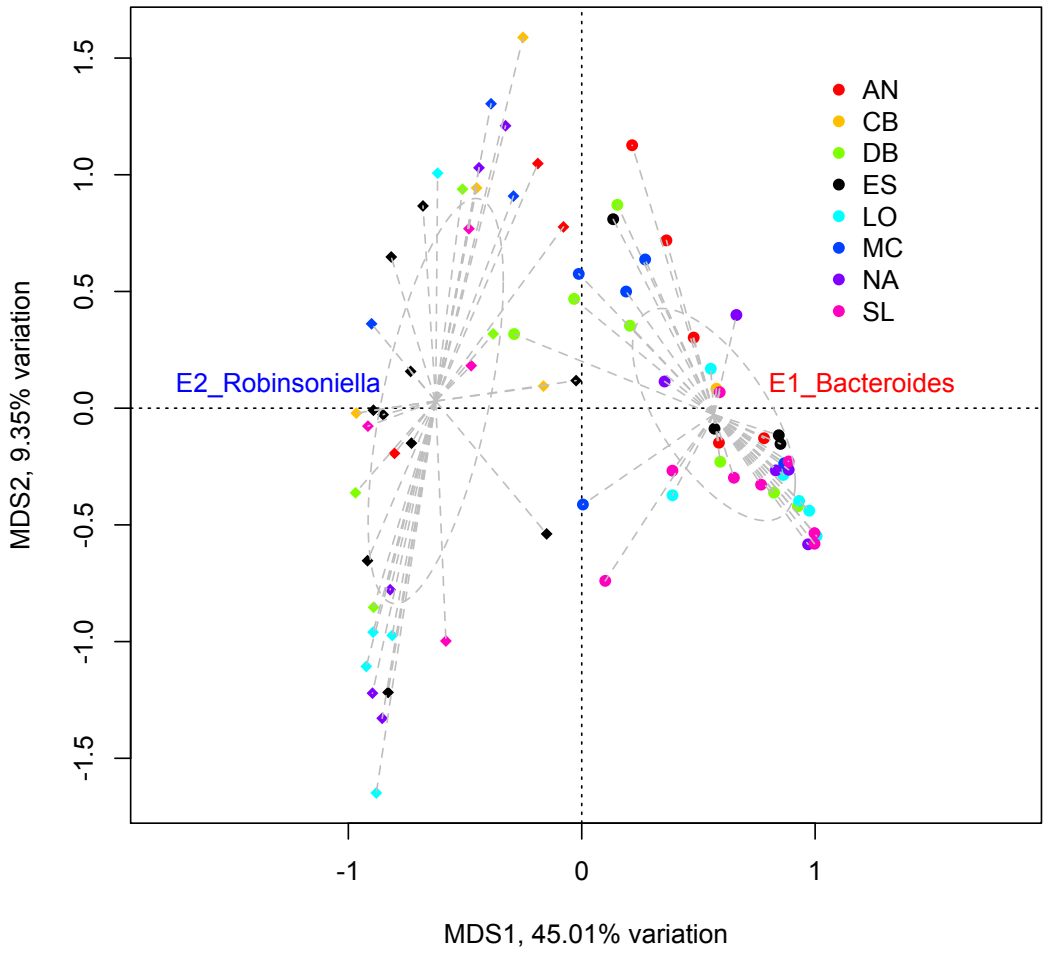


Figure 1. Principal coordinate analysis (PCoA) of wild captured mice based on Bray-Curtis dissimilarity.

Colors denote each of eight unique sampling locations across Germany and France that are on average 530 km apart (Linnenbrink *et al.,* 2013; n = 80). Dots and diamonds represent the E1 enterotype (*Bacteroides*-dominant) and E2 enterotype (*Robinsoniella*-dominant) classification, respectively.

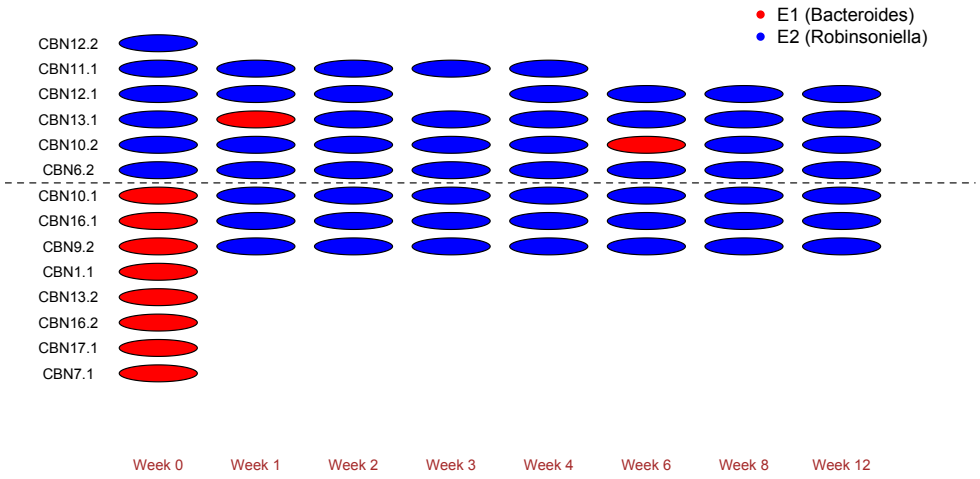


Figure 2. Enterotype classification of mice transferred to the lab.

Timepoints range from time of capture in the Cologne-Bonn (Week 0; n=14) through a 12-week period of regular sampling in the lab (n=8). Each oval denotes one sample and colors correspond to enterotype classification (red for E1 and blue for E2).

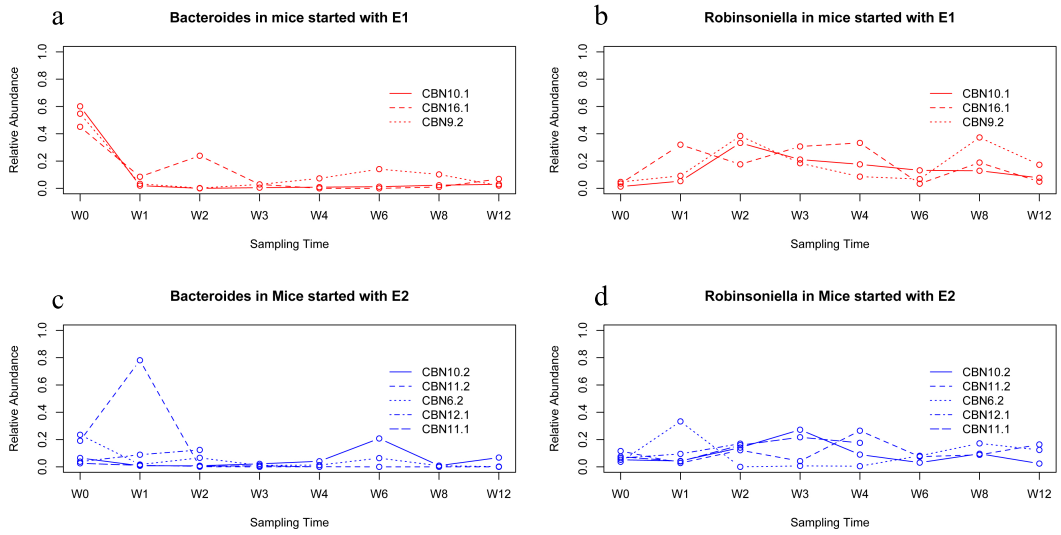


Figure 3. Abundances of signature genera *Bacteroides* and *Robinsoniella.*
Panels a) and b) display *Bacteroides* and *Robinsoniella* abundance over time among mice classified as E1 at the time of capture (n=3; red). Panels c) and d) display *Bacteroides* and *Robinsoniella* abundance over time among mice classified as E2 at the time of capture (n=5; blue).

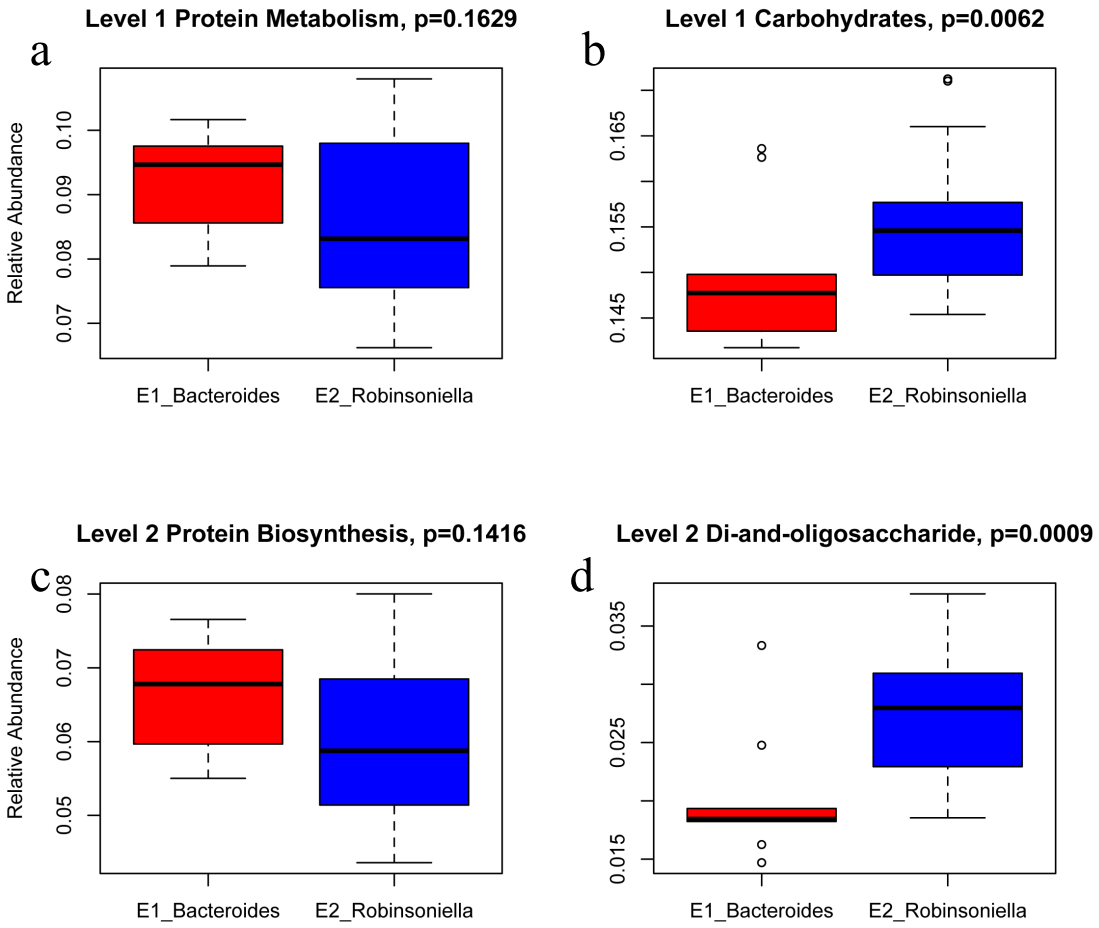


Figure 4. Differences in hierarchical KEGG pathways between enterotypes.

a) Relative abundance of genes involved in level 1 protein metabolism. b) Relative abundance of genes involved in level 1 carbohydrates. c) Relative abundance of genes involved in level 2 protein biosynthesis. d) Relative abundances of genes involved in level 2 di-and-oligo saccharide metabolism.

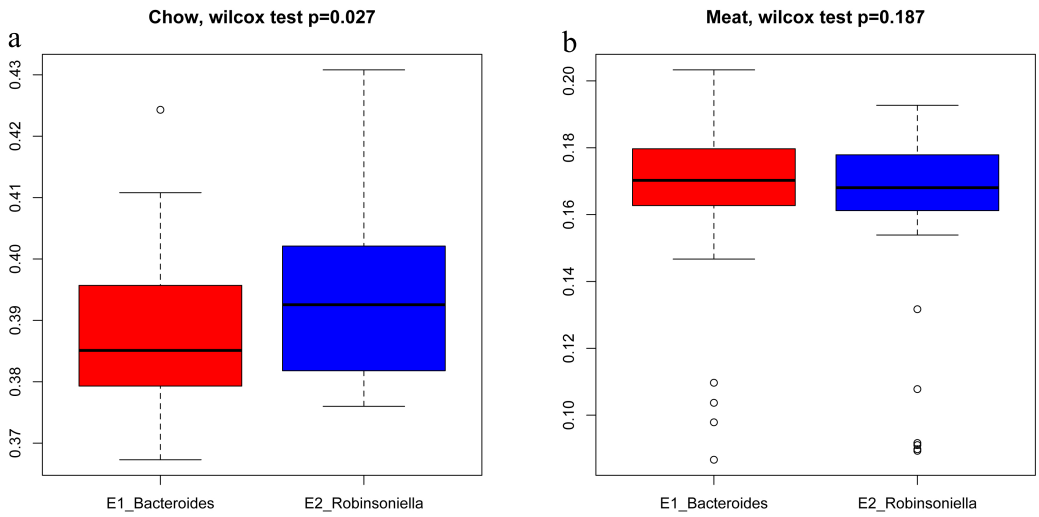


Figure 5. Abundance of major dietary components reconstructed for the two enterotypes.

a) Relative abundances of "chow" and b) "meat" in the diet (see Materials and Methods for details on food standards).

Supplementary Table 1. Summary statistics of enterotype analyses based on OTUs (97% and 95% similarity, as suggested by Koren *et al.*, 2013).

Beta-diversity indices	Clustering method	Optimal cluster for CH[1]	CH index	Optimal cluster for SC	SC	Optimal cluster for PS	PS[2]
Bray-Curtis of 97% OTUs	pam	3	6.5041	5	0.0901	2	0.7266
	kmeans	2	8.3578	5	0.0992	2	**0.9981**
	ward linkage HC	2	7.5696	5	0.0949	2	0.6603
	average linkage HC	**2**	1.5635	**2**	0.0877	**2**	0.5787
	single linkage HC	**2**	1.3388	**2**	0.0809	**2**	0.1085
	complete linkage HC	5	4.6775	9	0.0918	2	0.6717
Bray-Curtis of 95% OTUs	pam	3	7.4625	6	0.0936	2	0.7454
	kmeans	2	9.4648	8	0.1115	2	**0.9839**
	ward linkage HC	2	8.9416	5	0.1070	2	0.6682
	average linkage HC	6	3.1479	2	0.1034	2	0.5867
	single linkage HC	5	1.3573	2	0.0785	2	0.1590
	complete linkage HC	2	8.1797	6	0.1019	2	0.6742

[1]Bold indicates beta diversities and clustering methods which consistently produce two optimal clustering based on all three judging criteria.
[2]Bold indicates moderate to strong support for two clusters (PS>0.8).
Abbrevations: CH=Calinski-Harabasz index, SC=Silhouette score, PS=prediction strength, HC=hiearchical clustering.

Supplementary Table 2. Summary statistics of enterotype analyses with addition of MC mice from the lab.

Beta-diversity indices	Clustering method	Optimal cluster for CH[1]	CH index	Optimal cluster for SC	SC	Optimal cluster for PS	PS[2]
Bray-Curtis	pam	**2**	61.1704	2	0.3694	**2**	**0.8485**
	kmeans	**2**	64.8375	**2**	0.3803	**2**	0.7617
	ward linkage HC	**2**	50.6673	**2**	0.3357	**2**	0.7009
	average linkage HC	5	24.3138	5	0.3821	2	0.6011
	single linkage HC	4	2.9623	2	0.3180	2	0.3339
	complete linkage HC	3	38.6640	4	0.3906	2	0.5879
Jensen-Shannon	pam	**2**	102.4611	**2**	0.4726	**2**	**0.8022**
	kmeans	**2**	104.8259	**2**	0.4768	**2**	**0.8360**
	ward linkage HC	**2**	95.0824	**2**	0.4601	**2**	0.6932
	average linkage HC	4	45.4335	2	0.4925	2	0.5264
	single linkage HC	3	5.0954	2	0.4447	2	0.2239
	complete linkage HC	3	62.5293	5	0.4732	2	0.5439
Jaccard	pam	**2**	39.4323	**2**	0.3012	**2**	**0.8482**
	kmeans	2	38.8670	2	0.2442	5	0.8736
	ward linkage HC	**2**	39.1975	**2**	0.2794	**2**	0.6910
	average linkage HC	2	34.3224	2	0.2564	2	0.5829
	single linkage HC	4	0.2814	4	0.2228	2	0.3457
	complete linkage HC	3	22.5214	2	0.1912	2	0.6273
Weighted Unifrac	pam	**2**	23.8045	**2**	0.1990	**2**	0.7869
	kmeans	2	22.9956	2	0.2116	3	0.7472
	ward linkage HC	**2**	24.0998	**2**	0.2212	**2**	0.7421
	average linkage HC	5	10.0403	2	0.2652	2	0.5771
	single linkage HC	**2**	2.7165	**2**	0.2652	**2**	0.2952
	complete linkage HC	**2**	2.6442	**2**	0.2477	**2**	0.6452
Unweighted Unifrac	pam	**2**	6.7574	**2**	0.0669	**2**	0.6397
	kmeans	**2**	7.1394	**2**	0.0630	**2**	0.8769
	ward linkage HC	**2**	6.6720	**2**		**2**	0.6343
	average linkage HC	**2**	6.4030	**2**	0.0702	**2**	0.8120
	single linkage HC	3	1.1295	2	0.0184	2	0.1013
	complete linkage HC	**2**	5.8171	**2**	0.0513	**2**	0.6973

[1]Bold indicates beta diversities and clustering methods which consistently produce two optimal clustering based on all three judging criteria.
[2]Bold indicates moderate to strong support for two clusters (PS>0.8).
Abbrevations: CH=Calinski-Harabasz index, SC=Silhouette score, PS=prediction strength, HC=hiearchical clustering.

Supplementary Table 3. Summary statistics of enterotype analyses with addition of newly captured CB mice.

Beta-diversity indices	Clustering method	Optimal cluster for CH[1]	CH	Optimal cluster for SC	SC	Optimal cluster for PS	PS[2]
Bray-Curtis	pam	**2**	43.5121	**2**	0.2959	**2**	0.8501
	kmeans	**2**	46.9388	**2**	0.3046	**2**	**0.6930**
	ward linkage HC	**2**	39.8734	**2**	0.2602	**2**	**0.8202**
	average linkage HC	5	18.7827	2	0.3162	2	0.4519
	single linkage HC	4	3.0702	2	0.3162	2	0.2463
	complete linkage HC	3	25.9587	2	0.3402	2	0.5756
Jensen-Shannon	pam	**2**	67.3913	**2**	0.3726	**2**	**0.8209**
	kmeans	2	72.0155	2	0.3909	3	0.5345
	ward linkage HC	**2**	67.3732	**2**	0.3697	**2**	0.7608
	average linkage HC	5	30.0716	2	0.4827	2	0.4797
	single linkage HC	4	5.1218	2	0.4508	2	0.2259
	complete linkage HC	**2**	65.0644	**2**	0.3667	**2**	0.4953
Jaccard	pam	**2**	27.7191	**2**	0.2147	**2**	**0.8290**
	kmeans	**2**	30.0080	**2**	0.2166	**2**	**0.9886**
	ward linkage HC	**2**	28.0477	**2**	0.1968	**2**	0.7686
	average linkage HC	4	11.3550	5	0.2166	2	0.4992
	single linkage HC	4	2.1646	2	0.2154	2	0.2862
	complete linkage HC	3	15.6887	2	0.2298	2	0.5669
Weighted Unifrac	pam	**2**	42.5299	**2**	0.2730	**2**	0.6224
	kmeans	**2**	42.6002	**2**	0.2744	**2**	**0.8631**
	ward linkage HC	**2**	40.1802	**2**	0.2580	**2**	0.7760
	average linkage HC	4	12.1693	2	0.2884	2	0.5606
	single linkage HC	2	4.2344	2	0.2884	2	0.3094
	complete linkage HC	3	27.2311	2	0.2752	2	0.5893
Unweighted Unifrac	pam	2	7.0446	2	0.0702	3	0.7004
	kmeans	**2**	7.2124	**2**	0.0694	**2**	**0.9835**
	ward linkage HC	**2**	6.6615	**2**	0.0540	**2**	**0.8974**
	average linkage HC	**2**	5.2940	**2**	0.0811	**2**	**0.9200**
	single linkage HC	4	1.0946	2	0.0215	2	0.2688
	complete linkage HC	**2**	6.7316	**2**	0.0567	**2**	**0.9232**

[1]Bold indicates beta diversities and clustering methods which consistently produce two optimal clustering based on all three judging criteria.
[2]Bold indicates moderate to strong support for two clusters (PS>0.8).
Abbrevations: CH=Calinski-Harabasz index, SC=Silhouette score, PS=prediction strength, HC=hiearchical clustering.

Supplementary Table 4. Summary statistics for functional metagenomics performed in this study.

MG-RAST ID	MiceID	SAMPLING TIME	Enterotype	Total Reads	Assemblies	Assembled Nucleotides
4528728.3	CBN10.2	W0	E2_Robinsoniella	86,288,058	161,004	129,756,632
4528729.3	CBN11.2	W0	E2_Robinsoniella	96,148,052	168,808	119,956,533
4528730.3	CBN12.1	W0	E2_Robinsoniella	100,758,326	574,218	188,500,012
4528731.3	CBN11.1	W0	E2_Robinsoniella	85,766,392	1,117,827	242,605,462
4528732.3	CBN10.1	W0	E1_Bacteroides	79,685,072	346,730	131,175,401
4528733.3	CBN16.1	W0	E1_Bacteroides	86,298,056	184,715	118,927,889
4528740.3	CBN13.1	W0	E2_Robinsoniella	96,788,344	164,361	141,893,546
4528741.3	CBN15.1	W0	E2_Robinsoniella	81,535,318	282,510	178,586,899
4528742.3	CBN1.1	W0	E1_Bacteroides	87,261,610	127,821	90,347,526
4528743.3	CBN13.2	W0	E1_Bacteroides	90,544,464	292,813	135,516,320
4528744.3	CBN16.1	W0	E1_Bacteroides	95,274,100	16,689	45,346,046
4528745.3	CBN17.1	W0	E1_Bacteroides	84,011,316	99,197	71,656,879
4529796.3	CBN9.2	W0	E1_Bacteroides	68,228,848	204,999	106,126,963
4529797.3	CBN6.2	W0	E2_Robinsoniella	76,805,110	179,628	104,883,011
4529799.3	CBN7.1	W0	E1_Bacteroides	74,029,640	41,788	51,325,238
4528746.3	CBN6.2	W1	E2_Robinsoniella	82,214,742	172,270	144,419,325
4528747.3	CBN10.2	W1	E2_Robinsoniella	78,924,366	169,215	126,436,911
4528748.3	CBN11.2	W1	E1_Bacteroides	91,561,660	296,881	134,632,504
4528749.3	CBN12.1	W1	E2_Robinsoniella	87,677,022	165,482	138,049,815
4528750.3	CBN11.1	W1	E2_Robinsoniella	98,531,562	182,522	172,818,770
4528751.3	CBN10.1	W1	E2_Robinsoniella	84,361,184	203,838	152,877,234
4529800.3	CBN16.1	W1	E2_Robinsoniella	83,251,762	162,422	94,600,966
4529801.3	CBN9.2	W1	E2_Robinsoniella	80,599,392	113,237	89,693,422
4528734.3	CBN6.2	W12	E2_Robinsoniella	95,010,460	155,824	119,932,702
4528735.3	CBN10.2	W12	E2_Robinsoniella	88,215,294	203,543	147,954,096
4528736.3	CBN11.2	W12	E2_Robinsoniella	86,318,332	179,629	131,012,946
4528737.3	CBN12.1	W12	E2_Robinsoniella	89,284,256	205,592	116,036,273
4528739.3	CBN10.1	W12	E2_Robinsoniella	86,433,192	223,013	165,017,679
4529798.3	CBN9.2	W12	E2_Robinsoniella	72,326,248	143,302	81,903,896
4528738.3	CBN11.1	W4[1]	E2_Robinsoniella	77,047,990	153,818	116,014,755

[1] Mice CBN11.1 was removed from the experiment after week four, thus the 4th week sampling is the last sampling of this mice.

Supplementary Table 5. Level 3 and 4 KEGG pathways between E1 and E2 enterotypes.

Level	Pathways	Mean value in E1 (%)	Mean value in E2 (%)	Contribution to variance (%)[1]	Wilcox test p[2]
Level 3	Conjugative transposon Bacteroidales	1.05	0.56	2.22	**0.0463**
	Phage integration and excision	1.5	1.26	1.37	0.1599
	Ton and Tol transport systems	0.71	0.36	1.25	**0.0071**
	Ribosome LSU bacterial	1.3	1.06	1.17	0.0539
	Flagellum	0.42	0.7	1.14	**0.0241**
	Flagellar motility	0.32	0.57	0.99	**0.0170**
	Serine glyoxylate cycle	1.43	1.22	0.99	**0.0382**
	Iron acquisition in Vibrio	0.69	0.4	0.98	**0.0049**
	r1t like streptococcal phages	0.38	0.51	0.86	0.0698
	Restriction Modification System	0.82	0.85	0.77	0.3661
	Bacterial Chemotaxis	0.19	0.37	0.75	**0.0199**
	Ribosome SSU bacterial	0.93	0.75	0.73	0.0382
	FructooligosaccharidesFOS and Raffinose Utilization	0.24	0.44	0.72	**0.0091**
	Type I Restriction Modification	0.71	0.72	0.72	0.4028
	Sugar utilization in Thermotogales	1.71	1.82	0.68	0.0580
	Bacterial Cytoskeleton	0.8	0.83	0.63	0.3022
	Biotin biosynthesis	0.41	0.23	0.57	**0.0032**
	Group II intron associated genes	0.31	0.4	0.56	0.0898
	Macromolecular synthesis operon	0.51	0.38	0.51	**0.0241**
	Beta Glucoside Metabolism	0.2	0.33	0.5	**0.0170**
	High affinity phosphate transporter and control of PHO regulon	0.19	0.33	0.5	**0.0296**
	Sporulation gene orphans	0.11	0.25	0.49	**0.0071**
	Coenzyme B12 biosynthesis	0.26	0.37	0.47	**0.0272**
	YgfZ	1.2	1.08	0.47	**0.0250**
	Universal GTPases	0.89	0.92	0.46	0.3933
	BlaR1 Family Regulatory Sensor transducer Disambiguation	0.24	0.35	0.46	**0.0382**
	COG3533	0.48	0.56	0.45	**0.0382**
	Maltose and Maltodextrin Utilization	0.46	0.59	0.45	**0.0071**
	Cellulosome	0.23	0.11	0.45	**0.0170**
	Phage capsid proteins	0.18	0.14	0.43	0.2067
	Lactose and Galactose Uptake and Utilization	0.52	0.62	0.43	**0.0349**
	Phosphate metabolism	0.42	0.52	0.4	0.0539
	HtrA and Sec secretion	0.44	0.33	0.4	**0.0241**
	tRNA modification Bacteria	1.06	0.98	0.39	0.0698
	Phage packaging machinery	0.11	0.23	0.38	**0.0032**
	C jejuni colonization of chick caeca	0.21	0.3	0.38	**0.0250**
	DNA replication	1.08	0.99	0.38	**0.0349**
	Two cell division clusters relating to chromosome partitioning	0.38	0.44	0.38	0.1302
	tRNA processing	0.32	0.2	0.37	**0.0032**
	RNA polymerase bacterial	0.41	0.36	0.36	0.1231
	RNA modification and chromosome partitioning cluster	0.39	0.4	0.36	0.3661
	Alginate metabolism	0.18	0.3	0.36	**0.0003**
	P uptake cyanobacteria	0.23	0.32	0.36	0.0623
	Translation elongation factors bacterial	0.52	0.49	0.35	0.3290

	Respiratory Complex I	0.2	0.1	0.35	**0.0071**
	Flagellum in Campylobacter	0.11	0.19	0.34	**0.0241**
	Plasmid replication	0.24	0.26	0.34	0.4470
	Butanol Biosynthesis	0.15	0.25	0.34	**0.0199**
	Acetone Butanol Ethanol Synthesis	0.18	0.27	0.33	**0.0159**
	LSU ribosomal proteins cluster	0.35	0.28	0.33	0.0669
	Galactosylceramide and Sulfatide metabolism	0.38	0.4	0.33	0.2448
Level 4	TonB-dependent receptor	1.11	0.51	1.18	0.1984
	Integrase	1.03	0.68	0.78	**0.0305**
	Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	0.82	0.42	0.69	**0.0019**
	DNA primase (EC 2.7.7.-)	0.68	0.38	0.65	**0.0179**
	Cysteine desulfurase (EC 2.8.1.7)	0.63	0.96	0.63	0.0082
	Butyryl-CoA dehydrogenase (EC 1.3.99.2)	0.16	0.43	0.58	0.0576
	Beta-galactosidase (EC 3.2.1.23)	0.74	0.91	0.55	0.0883
	Chromosome (plasmid) partitioning protein ParB	0.59	0.53	0.50	0.3012
	Alpha-galactosidase (EC 3.2.1.22)	0.47	0.67	0.45	**0.0158**
	Chromosome (plasmid) partitioning protein ParA	0.38	0.49	0.40	0.1354
	DNA topoisomerase III (EC 5.99.1.2)	0.30	0.43	0.40	0.1166
	Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)	0.55	0.59	0.38	0.4719
	Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)	0.44	0.52	0.36	0.1275
	Beta-lactamase repressor BlaI	0.12	0.29	0.33	**0.0118**
	UDP-glucose 4-epimerase (EC 5.1.3.2)	0.73	0.75	0.33	0.4505
	Site-specific recombinase	0.29	0.38	0.32	0.1715
	Retron-type reverse transcriptase	0.31	0.40	0.32	0.1107
	Malate dehydrogenase (EC 1.1.1.37)	0.30	0.17	0.32	0.0823
	Serine hydroxymethyltransferase (EC 2.1.2.1)	0.48	0.58	0.31	0.1241
	Beta-hexosaminidase (EC 3.2.1.52)	0.32	0.19	0.31	**0.0237**
	Phage major capsid protein	0.21	0.24	0.30	0.3127
	Aspartate aminotransferase (EC 2.6.1.1)	0.57	0.66	0.30	0.1715
	Arylsulfatase (EC 3.1.6.1)	0.24	0.08	0.30	**0.0059**
	Translation elongation factor G	0.48	0.49	0.29	0.4210
	tRNA-i(6)A37 methylthiotransferase	0.40	0.26	0.28	**0.0082**
	Flagellin protein FlaA	0.10	0.22	0.28	0.0211
	tRNA delta(2)-isopentenylpyrophosphate transferase (EC 2.5.1.8)	0.29	0.14	0.28	**0.0047**
	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)	0.43	0.33	0.28	0.0823
	Beta-glucosidase (EC 3.2.1.21)	0.15	0.30	0.28	**0.0059**
	Phosphate transport ATP-binding protein PstB (TC 3.A.1.7.1)	0.10	0.23	0.28	0.0509
	Acetyl-CoA acetyltransferase (EC 2.3.1.9)	0.11	0.23	0.26	0.0509
	Glycosyltransferase	0.36	0.44	0.24	0.0777
	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	0.31	0.36	0.24	0.3715
	Putative mobilization protein BF0133	0.18	0.10	0.23	0.0777
	ClpB protein	0.41	0.34	0.23	0.1107
	Phage capsid and scaffold	0.13	0.04	0.23	0.3350

Methylmalonyl-CoA mutase (EC 5.4.99.2)	0.22	0.14	0.22	0.1275
Ferrous iron transport protein B	0.34	0.41	0.22	0.1166
UDP-N-acetylglucosamine 4,6-dehydratase (EC 4.2.1.-)	0.13	0.19	0.22	0.0777
RecA protein	0.25	0.37	0.22	0.0082
Chemotaxis regulator - transmits chemoreceptor signals to flagelllar motor components CheY	0.08	0.19	0.22	0.0130
DNA gyrase subunit A (EC 5.99.1.3)	0.18	0.29	0.21	0.0158
Ferric siderophore transport system, periplasmic binding protein TonB	0.17	0.05	0.21	**0.0047**
UDP-N-acetylglucosamine 2-epimerase (EC 5.1.3.14)	0.29	0.29	0.21	0.4861
GTP cyclohydrolase I (EC 3.5.4.16) type1	0.29	0.34	0.21	0.1553
Translation initiation factor 2	0.42	0.41	0.21	0.3832
Phosphoenolpyruvate carboxykinase [ATP] (EC 4.1.1.49)	0.27	0.18	0.21	0.0576
Endo-1,4-beta-xylanase A precursor (EC 3.2.1.8)	0.26	0.26	0.21	0.4210
Succinate dehydrogenase iron-sulfur protein (EC 1.3.99.1)	0.20	0.12	0.20	0.0576
Translation elongation factor Ts	0.22	0.11	0.20	**0.0091**
Multi antimicrobial extrusion protein (Na(+)/drug antiporter), MATE family of MDR efflux pumps	0.15	0.26	0.20	**0.0077**
RNA polymerase sigma-54 factor RpoN	0.25	0.24	0.20	0.3127
tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA	0.39	0.39	0.20	0.4505
Pyruvate,phosphate dikinase (EC 2.7.9.1)	0.23	0.28	0.20	0.1275
tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase (EC 2.1.1.61)	0.30	0.27	0.20	0.3243
8-amino-7-oxononanoate synthase (EC 2.3.1.47)	0.17	0.06	0.20	**0.0082**
Ferric siderophore transport system, biopolymer transport protein ExbB	0.15	0.05	0.20	0.0129
Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8)	0.31	0.32	0.20	0.4719
Ribonuclease BN (EC 3.1.-.-)	0.18	0.07	0.19	**0.0042**
Phosphate regulon transcriptional regulatory protein PhoB (SphR)	0.07	0.16	0.19	**0.0144**
D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)	0.22	0.25	0.19	0.2141
2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (EC 2.7.6.3)	0.15	0.05	0.19	**0.0058**
Glutamate synthase [NADPH] large chain (EC 1.4.1.13)	0.10	0.21	0.19	**0.0034**
Conjugative transposon protein TraG	0.16	0.10	0.19	0.0958
Iojap protein	0.15	0.04	0.19	**0.0037**
Acetaldehyde dehydrogenase (EC 1.2.1.10)	0.13	0.20	0.19	0.0777
Pyruvate-flavodoxin oxidoreductase (EC 1.2.7.-)	0.28	0.24	0.19	0.1885
Branched-chain amino acid aminotransferase (EC 2.6.1.42)	0.36	0.30	0.19	0.1275
Alpha-1,2-mannosidase	0.16	0.06	0.18	**0.0082**

Pyruvate formate-lyase (EC 2.3.1.54)	0.24	0.30	0.18	0.1659
Copper-translocating P-type ATPase (EC 3.6.3.4)	0.16	0.24	0.18	0.0622
Alpha-L-fucosidase (EC 3.2.1.51)	0.22	0.14	0.18	**0.0305**
ATP-dependent DNA helicase RecQ	0.23	0.15	0.18	0.0756
Flagellin	0.09	0.16	0.18	0.0560
3-ketoacyl-CoA thiolase (EC 2.3.1.16)	0.06	0.15	0.18	**0.0082**
Signal transduction histidine kinase CheA (EC 2.7.3.-)	0.07	0.15	0.18	**0.0187**
Glutamine synthetase type I (EC 6.3.1.2)	0.19	0.27	0.17	**0.0470**
Phosphoglycerate kinase (EC 2.7.2.3)	0.35	0.30	0.17	0.1553
tRNA (Guanine37-N1) -methyltransferase (EC 2.1.1.31)	0.22	0.22	0.17	0.4650
decarboxylase	0.46	0.42	0.17	0.1715
UDP-glucose dehydrogenase (EC 1.1.1.22)	0.20	0.28	0.17	**0.0187**
Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1)	0.19	0.15	0.17	0.1808
Acetate kinase (EC 2.7.2.1)	0.36	0.37	0.17	0.4210
Cysteine desulfurase (EC 2.8.1.7), SufS subfamily	0.15	0.10	0.17	0.1275
Probable poly(beta-D-mannuronate) O-acetylase (EC 2.3.1.-)	0.07	0.16	0.17	**0.0034**
GTPase and tRNA-U34 5-formylation enzyme TrmE	0.34	0.30	0.16	0.3243
Translation elongation factor Tu	0.25	0.25	0.16	0.4719
NAD-specific glutamate dehydrogenase (EC 1.4.1.2)	0.16	0.09	0.16	**0.0187**
Ferric uptake regulation protein FUR	0.14	0.05	0.16	**0.0034**
Signal peptidase I (EC 3.4.21.89)	0.20	0.11	0.16	**0.0019**
Dihydrolipoamide dehydrogenase (EC 1.8.1.4)	0.19	0.10	0.16	**0.0042**
Translation elongation factor LepA	0.22	0.28	0.16	**0.0254**
Phage terminase	0.09	0.17	0.16	**0.0044**
NADP-specific glutamate dehydrogenase (EC 1.4.1.4)	0.23	0.25	0.16	0.3012
S-adenosylmethionine:tRNA ribosyltransferase-isomerase (EC 5.-.-.-)	0.24	0.19	0.15	0.0883
3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)	0.21	0.27	0.15	0.0576
Acetolactate synthase large subunit (EC 2.2.1.6)	0.12	0.19	0.15	**0.0470**
Enolase (EC 4.2.1.11)	0.25	0.29	0.15	0.1241
Outer membrane protein H precursor	0.12	0.04	0.15	**0.0140**
Alcohol dehydrogenase (EC 1.1.1.1)	0.19	0.21	0.15	0.1885
tRNA pseudouridine synthase B (EC 4.2.1.70)	0.19	0.11	0.15	**0.0104**
Cytoplasmic axial filament protein CafA and Ribonuclease G (EC 3.1.4.-)	0.18	0.10	0.15	**0.0042**
Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4)	0.15	0.19	0.15	0.1715
Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	0.27	0.21	0.15	0.0777
Uracil phosphoribosyltransferase (EC 2.4.2.9)	0.15	0.08	0.15	**0.0254**
Phosphate regulon sensor protein PhoR (SphS) (EC 2.7.13.3)	0.11	0.19	0.15	**0.0091**
Purine nucleoside phosphorylase (EC	0.26	0.31	0.15	0.1354

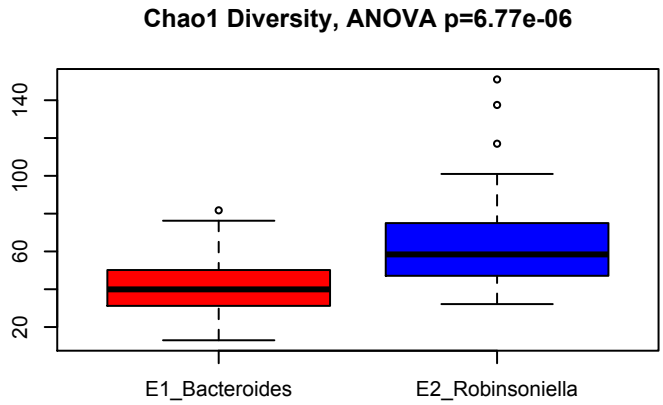
2.4.2.1)				
Peptide chain release factor 1	0.24	0.25	0.14	0.4210
RecD-like DNA helicase YrrC	0.12	0.16	0.14	0.1038
Cysteinyl-tRNA synthetase (EC 6.1.1.16)	0.22	0.19	0.14	0.3012
Phosphate transport system permease protein PstC (TC 3.A.1.7.1)	0.06	0.14	0.14	**0.0284**
Two-component system response regulator	0.16	0.12	0.14	0.1808
RNA polymerase sigma factor	0.17	0.24	0.14	**0.0067**
YgjD/Kae1/Qri7 family, required for N6-threonylcarbamoyl adenosine t(6)A37 modification in tRNA	0.23	0.27	0.14	0.1166
Cell division transporter, ATP-binding protein FtsE (TC 3.A.5.1.1)	0.09	0.16	0.14	**0.0041**
Pyruvate kinase (EC 2.7.1.40)	0.30	0.31	0.14	0.5000
Signal recognition particle receptor protein FtsY (=alpha subunit) (TC 3.A.5.1.1)	0.08	0.16	0.14	**0.0019**
DNA gyrase subunit B (EC 5.99.1.3)	0.32	0.30	0.14	0.3832
Multiple sugar ABC transporter, ATP-binding protein	0.04	0.11	0.14	**0.0042**
DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6)	0.20	0.14	0.14	**0.0470**
3-oxoacyl-[acyl-carrier-protein] synthase, KASIII (EC 2.3.1.41)	0.16	0.08	0.14	**0.0044**
Maltodextrin phosphorylase (EC 2.4.1.1)	0.05	0.11	0.13	**0.0187**
SSU ribosomal protein S1p	0.21	0.15	0.13	0.0576
6-phosphofructokinase (EC 2.7.1.11)	0.30	0.29	0.13	0.3982
Glucokinase (EC 2.7.1.2)	0.19	0.14	0.13	0.0622
Phosphopentomutase (EC 5.4.2.7)	0.05	0.11	0.13	**0.0254**
3-dehydroquinate synthase (EC 4.2.3.4)	0.13	0.06	0.13	**0.0082**
GMP synthase [glutamine-hydrolyzing] (EC 6.3.5.2)	0.19	0.22	0.13	0.1885
Regulatory sensor-transducer, BlaR1/MecR1 family	0.09	0.15	0.13	0.0823
Glycogen phosphorylase (EC 2.4.1.1)	0.18	0.13	0.13	**0.0305**

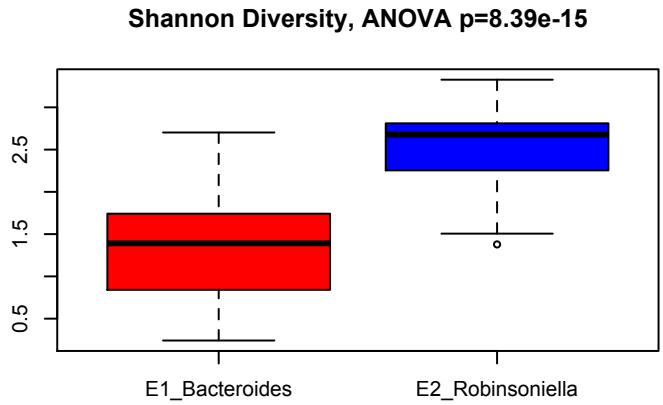
[1]The contribution of each pathway to the total variance is calculated via SIMPER.
[2]Significant differences ($p<0.05$) are shown in bold.

Supplementary Table 6. Stable isotope measures (left rows) and reconstructed diets using FRUITS program.

Mice	Isotope measurements				Reconstructed diets			
	Muscle d15N	Muscle d13C	Liver d15N	Liver d13C	Grains	Chow	Insect	Meat
AN14	6.44	-25.58	6.73	-25.28	0.4040	0.4021	0.1022	0.0917
AN26	7.39	-25.80	7.11	-25.59	0.4112	0.3865	0.1045	0.0979
AN27	6.00	-26.00	4.84	-25.07	0.4140	0.3759	0.0434	0.1667
AN30	9.60	-23.70	12.01	-18.18	0.3281	0.3849	0.1097	0.1773
AN62	7.82	-26.15	7.44	-26.56	0.3928	0.4092	0.0943	0.1037
AN64	8.26	-25.25	7.79	-25.80	0.3375	0.3708	0.1017	0.1900
AN65	6.86	-25.67	6.57	-25.96	0.4118	0.3760	0.0462	0.1660
AN66	10.62	-23.12	8.08	-24.69	0.3246	0.4213	0.0959	0.1582
CB11	6.22	-25.94	5.42	-27.02	0.4006	0.3927	0.0459	0.1608
CB14	5.99	-22.86	5.82	-22.81	0.4165	0.3804	0.0388	0.1643
CB15	5.90	-26.05	5.60	-26.39	0.4093	0.3808	0.0347	0.1751
CB16	4.43	-25.93	5.48	-26.07	0.4135	0.3810	0.0368	0.1687
CB24	6.81	-22.39	6.49	-22.71	0.3319	0.3826	0.0958	0.1896
DB02	6.80	-21.33	7.30	-16.15	0.3284	0.3937	0.1101	0.1678
DB03	8.35	-12.58	7.74	-13.43	0.3273	0.3865	0.0937	0.1924
DB07	7.44	-25.34	5.98	-24.69	0.3975	0.4071	0.1088	0.0867
DB08	8.01	-21.30	7.95	-20.55	0.3315	0.4070	0.1003	0.1612
DB11	6.12	-24.92	5.23	-25.46	0.3315	0.4070	0.1003	0.1612
DB13	6.93	-24.91	6.26	-24.77	0.4151	0.3673	0.0405	0.1770
DB14	7.40	-22.71	7.45	-19.74	0.3263	0.4064	0.1077	0.1597
DB15	6.77	-25.99	6.98	-25.92	0.3993	0.3976	0.0934	0.1097
DB18	10.82	-15.31	9.25	-14.28	0.3254	0.4065	0.0952	0.1729
DB19	7.66	-22.81	7.73	-22.67	0.3349	0.3843	0.0943	0.1865
DB20	6.78	-24.90	7.00	-24.66	0.4033	0.3994	0.1080	0.0893
ES06	6.18	-16.52	5.25	-16.48	0.3263	0.3907	0.1034	0.1797
ES08	5.90	-18.11	5.72	-19.40	0.4125	0.3763	0.0428	0.1684
ES12	7.47	-21.71	8.21	-21.70	0.3234	0.3919	0.1082	0.1765
ES13	5.18	-23.81	5.70	-23.62	0.4118	0.3818	0.0422	0.1642
ES14	8.49	-14.03	8.47	-14.80	0.3255	0.4108	0.1004	0.1634
ES15	9.37	-14.04	9.26	-15.35	0.3235	0.3954	0.0977	0.1834
ES18	8.60	-17.43	10.02	-16.17	0.3256	0.4001	0.0937	0.1805
ES22	8.48	-12.74	7.45	-11.93	0.3250	0.3916	0.1055	0.1779
ES23	9.07	-13.35	6.90	-14.42	0.3298	0.3914	0.1017	0.1771
ES26	9.30	-16.41	8.58	-14.62	0.3259	0.4092	0.1026	0.1623
ES27	8.22	-16.59	8.13	-18.23	0.3266	0.3924	0.1011	0.1799
ES28	5.31	-22.78	5.63	-22.26	0.4126	0.3760	0.0469	0.1645
ES30	7.52	-23.58	6.16	-24.45	0.3288	0.3968	0.1046	0.1698
ES31	10.53	-13.68	9.61	-15.97	0.3241	0.3996	0.1085	0.1677
LO01	5.55	-25.83	4.97	-25.87	0.4043	0.3831	0.0343	0.1784
LO02	5.39	-25.80	5.13	-25.48	0.4238	0.3712	0.0433	0.1617
LO03	4.89	-26.04	5.13	-25.96	0.4015	0.3798	0.0383	0.1803
LO04	6.00	-26.10	6.51	-25.63	0.4070	0.3800	0.0465	0.1665
LO05	4.98	-26.69	6.39	-26.40	0.4076	0.3766	0.0455	0.1703
LO06	5.73	-25.70	5.76	-25.31	0.4150	0.3793	0.0393	0.1665
LO07	5.93	-25.97	5.85	-26.25	0.4123	0.3853	0.0434	0.1589
LO08	9.19	-18.91	8.64	-20.06	0.3284	0.4308	0.1091	0.1317
LO09	5.87	-25.32	6.18	-25.27	0.4093	0.3828	0.0395	0.1684
LO10	7.25	-21.36	7.01	-21.53	0.3273	0.3984	0.1130	0.1613
LO13	5.84	-25.65	6.04	-25.62	0.4124	0.3869	0.0467	0.1539
LO14	5.58	-23.07	7.38	-19.02	0.3317	0.4011	0.1036	0.1635
MC02	8.95	-25.80	9.04	-26.48	0.3260	0.3817	0.0890	0.2033
MC06	6.19	-25.90	6.22	-25.76	0.4048	0.3797	0.0445	0.1710
MC11	7.03	-25.24	5.65	-24.95	0.4132	0.3766	0.0464	0.1637
MC12	6.55	-25.19	5.57	-25.18	0.4046	0.3861	0.0350	0.1743
MC13	8.32	-25.52	8.82	-25.10	0.3349	0.3803	0.1018	0.1830

MC15	6.81	-25.71	6.34	-25.20	0.4012	0.3840	0.0455	0.1693
MC22	5.78	-24.96	5.24	-24.69	0.4171	0.3724	0.0420	0.1685
MC28	10.75	-24.46	10.23	-24.00	0.3350	0.3885	0.1003	0.1762
NA02	5.39	-16.37	5.76	-18.57	0.4104	0.3751	0.0344	0.1801
NA03	10.26	-23.82	10.80	-23.73	0.3246	0.4038	0.0923	0.1793
NA04	7.70	-25.35	7.66	-25.10	0.3271	0.3937	0.1032	0.1760
NA06	6.90	-20.28	6.34	-15.85	0.3322	0.3870	0.0985	0.1823
NA07	8.66	-25.16	8.68	-24.90	0.3298	0.3784	0.1165	0.1752
NA08	6.21	-25.14	6.46	-25.12	0.4210	0.3738	0.0425	0.1627
NA09	7.54	-25.61	6.77	-25.26	0.4047	0.4059	0.0995	0.0898
NA10	7.27	-26.11	7.68	-26.09	0.3959	0.4052	0.0911	0.1078
NA11	7.52	-23.89	7.09	-24.87	0.3288	0.3997	0.0922	0.1794
NA14	11.79	-18.78	11.48	-20.03	0.3236	0.4034	0.0910	0.1820
SL03	6.97	-21.29	8.05	-18.74	0.3218	0.3947	0.1120	0.1715
SL04	8.58	-25.22	8.53	-24.57	0.3243	0.4000	0.0974	0.1783
SL05	6.33	-26.27	6.53	-26.47	0.4069	0.3874	0.0463	0.1594
SL08	7.33	-23.76	7.66	-22.28	0.3257	0.3820	0.0992	0.1932
SL09	7.54	-20.62	7.16	-19.65	0.3238	0.3875	0.0960	0.1927
SL11	7.82	-22.91	7.76	-22.71	0.3362	0.3767	0.0951	0.1921
SL13	8.96	-26.15	8.78	-26.50	0.3287	0.3957	0.1054	0.1702
SL14	7.20	-25.38	6.68	-24.89	0.4035	0.4041	0.1013	0.0911
SL15	7.58	-24.11	7.19	-24.17	0.3360	0.3921	0.1152	0.1568
SL16	10.72	-25.90	10.34	-25.80	0.3247	0.3734	0.1173	0.1846
SL17	7.59	-23.81	8.16	-24.05	0.3278	0.3848	0.0964	0.1909
SL18	8.28	-23.18	7.91	-22.60	0.3238	0.4243	0.1053	0.1467





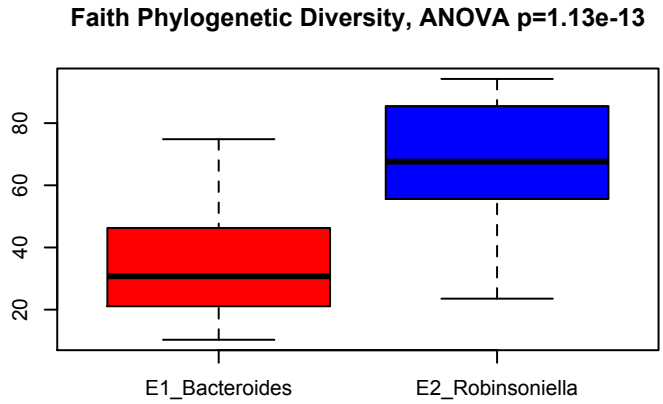


Figure S1. Alpha diversity measures between the two enterotypes. All display on average lower values in Enterotype 1 compared to Enterotype 2. Upper panel: Chao1 diversity based on genus-level composition; middle panel: Shannon diversity based on genus-level composition; lower panel: phylogenetic diversity based on ML-tree produced by FastTree.

Chapter III Gut microbiota in hybrid mice

Introduction

Research on the intestinal microbiota in mice is mainly carried out using laboratory strains (mainly of *Mus musculus domestics* origin), which have provided deep insights into different aspects of host-microbiota interaction. At the same time, limitations are also reported due to the artificial construction of microbiota in laboratory mice and their limited genetic diversity. In Chapter I of my thesis, I show that wild mice, on the other hand, have higher microbial diversity than laboratory mice and that a dominant role of geography surpasses the smaller effect of host genetics. Our working hypothesis, however is that the *M. m. domesticus* populations studied in the first Chapter, due to their close genetic relatedness, do not fully reflect the true level of genetic control / co-evolution between mouse hosts and their microbiota. A study using mice with greater genetic divergence, and at the same time negligible geographical differences, would therefore provide a better assessment of theses effect.

Thus, Chapter III focuses on the natural hybrid zone between *M. m. domesticus* and *M. m. musculus*, as well as a complementary lab cross. The two house mouse subspecies originated from central Asia, diverged approximately one million years ago and migrated to Europe by different routes, whereby their secondary contact in central Europe formed a narrow hybrid zone. Due to the special characteristics of asymmetric reproductive isolation and a stable species boundary (Good *et al.,* 2008), the hybrid zone became a widely studied model for population genetics, hybrid fitness and speciation in a natural environment. Other contributing factors for hybrid fitness besides asymmetric sterility are reported, for instance parasite loads (Sage *et al.,* 1986), and studies report that a compromised immune system is the cause for higher parasite loads in hybrid mice (Moulia *et al.,* 1991).

In this chapter, investigation of the hybrid mouse microbiota demonstrates several interesting findings. In both wild and lab mice, I discovered a distinct microbiota present in hybrid mice compared to the parental subspecies, characterized by lower species richness and an abnormal abundance of the major bacterial genera compared to the pure subspecies. I then applied a quantitative trait loci (QTL) approach using an

expanded lab cross and mapped the genomic regions contributing to the altered microbiota. In addition, I investigated the possible influence of the microbiota on hybrid fitness by studying the immune function using a mouse-specific immune gene expression panel. This reveals an over-all immune deficiency of hybrid mice; while histology revealed a higher frequency of pathology in hybrid mice. This presents new insight into host genetic regulation and host-microbiota co-evolution, whereby purifying selection maintains a relatively similar microbiome in diverged hosts, and conflicts arising in hybrid genomes may contribute to reduced hybrid fitness via their influence on the microbiome.

Manuscript

Jun Wang, Shirin Kalyan, Natalie Steck, Leslie M. Turner, Bettina Harr, Sven Künzel, Robert Häsler, Andre Franke, Guntram A. Grassl, Dieter Kabelitz and John F. Baines. **Analysis of the intestinal microbiota of hybrid house mice reveals evolutionary divergence in a vertebrate hologenome**.
To submit to Nat. Genet.

Analysis of the intestinal microbiota of hybrid house mice reveals evolutionary divergence in a vertebrate hologenome

Jun Wang [1, 2], Shirin Kalyan[3], Natalie Steck[2], Leslie M. Turner[1], Bettina Harr[1], Sven Künzel [1], Robert Häsler[4], Andre Franke[4], Guntram A. Grassl[2], Dieter Kabelitz and John F. Baines[1,2]

[1] Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, D-24306 Plön, Germany

[2] Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Arnold-Heller-Str. 3, D-24105 Kiel, Germany

[3]Institute of Immunology, Christian-Albrechts University Kiel, Arnold-Heller-Str. 3, Haus 17, 24105, Kiel, Germany

[4]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Schittenhelmstr. 12, D-24105 Kiel, Germany

Corresponding author:

Prof. Dr. John F. Baines

Institute for Experimental Medicine

Christian-Albrechts-University of Kiel

Arnold-Heller-Str. 3, D-24105 Kiel, Germany

Email: j.baines@iem.uni-kiel.de

Keywords: gut microbiota, wild mouse, diet, enterotype, metagenomics

Running title: Wild mouse gut microbiome moved to the lab environment

Abstract

Recent evidence suggests that natural selection operating on host genomes to maintain an appropriate, co-adapted microbiome may ultimately contribute to the emergence of new species, *i.e*. the “hologenomic basis of speciation”. In this study, we analyze the gut microbiota of house mice across a hybrid zone between the eastern and western subspecies *Mus musculus musculus* and *M. m. domesticus*, which is a long-standing model system for studying the origins of reproductive isolation, in addition to a F2 intercross between inbred strain representatives performed under controlled lab conditions. This reveals the gut microbiota of hybrid mice to display widespread transgressive phenotypes in a variety of measures of bacterial community structure, consistent with purifying selection being a predominating force operating on the intestinal microbiome between species. Further, QTL-, immune gene expression- and histopathological analyses indicate that genetic incompatibilities and/or immuno-deficiencies are the most likely cause of the altered community structure among hybrids. These results provide unique insight into the consequences of evolutionary divergence in a vertebrate "hologenome", which may be an important aspect to consider in the origin of new host species in this taxonomic group.

Introduction

It is widely assumed that animal hosts and their assemblages of associated microbes, particularly those inhabiting the intestine, represent the outcome of millions of years of coevolution. Support for this assumption lies in the distinctness of host-associated versus free-living bacterial communities [1], the reasoning that eukaryotic life forms never existed without bacteria in their environment [2] and examples of phylogenies of bacterial communities mirroring those of their hosts [3,4]. However, host-associated communities do not follow a strict pattern of vertical transmission, as external factors such as diet and geography also play a role in determining their composition [5-7]. Thus, although undoubtedly present, the extent to which coevolutionary processes influence the genomes of hosts and associated microbes, *i.e.* the "hologenome" [8], remains largely unknown for the majority of individual host and microbial taxa.

In a recent landmark study, Brucker and Bordenstein [4] demonstrate a gut microbial basis for the lethality of hybrids between two species of parasitoid jewel wasps (*Nasonia*). As a byproduct of evolutionary divergence, the structure of the original host species-specific ("phylosymbiotic") communities breaks down due to incompatibilities between loci in mosaic host genomes. Moreover, the involvement of innate immune gene expression and the fact that hybrid lethality is also observed when non-native bacteria (*Escherischia coli*) are introduced indicates a general problem with regulating the microbiome. These results carry significant implications for a role of the microbiome in host speciation; thus, understanding the extent to which related phenomena occur in other taxa is of critical importance. In particular, given the potential for a "large immune effect" in hybrid incompatibilities [9] and the hypothesis that the vertebrate adaptive immune system itself may have evolved as a means to manage more complex microbial communities [10], the vertebrates are an important group to investigate.

The house mouse is a key and widely applied model for many areas of biology and medicine. In addition to the wealth of information gained from inbred laboratory strains, natural mouse populations are gaining attention for use in areas such as evolutionary genetics [11], immunology [12,13] and studies of the gut microbiome [6] due to their wide range of "real-world" genetic and environmental diversity [14]. In particular, a naturally occurring hybrid zone between the eastern and western house mouse subspecies (*Mus musculus musculus* and *M. m. domesticus*, respectively) located in

central Europe has long served as a model for evolutionary biologists studying the evolution and maintenance of species boundaries. These two subspecies share a common ancestor approximately 0.5 Myr ago [15,16] and hybrid mice display reduced fitness, most notably in the form of reduced fertility [17-20] as well as higher parasite loads in the gut [21], although the latter was recently called into question [22]. These observations can be generally understood within the framework of Bateson-Dobzhansky-Muller incompatibilities (BDMIs) [23-25], which describe deleterious interaction between genetic loci in hybrid genomes that can emerge as a byproduct of evolutionary divergence between isolated (allopatric) parental populations, and are likely widespread between the *M. m. musculus – M. m. domesticus* species pair [26].

In this study, we employ several complementary approaches to determine whether the evolutionary divergence experienced by the *M. m. musculus – M. m. domesticus* species pair also extends to the regulation of the intestinal microbiome. We first collected fresh material from a well-studied transect of their hybrid zone located in Bavaria, Germany and then complemented these observations with an F2 intercross between inbred strains of *M. m. musculus* and *M. m. domesticus* under controlled laboratory conditions. For both groups of mice we performed profiling of bacterial community structure, histology and immune gene expression analysis of intestinal tissue, revealing consistent, systematic differences in hybrid- compared to parental mice. Finally, performing quantitative trait locus (QTL) mapping in a larger cohort of F2 mice provides insight into the genetic architecture of transgressive phenotypes of the microbiome we observe in hybrid mice.

Results

Intestinal microbiome in pure species versus hybrids

To determine whether divergence in the genetic basis of regulating the intestinal microbiome exists between the *M. m. musculus* - *M. m. domesticus* species pair, we first used hybridization as a means to expose possible evidence of incompatibilities. To this end, we included two complementary groups: one including 69 wild-caught mice sampled across a Bavarian transect of the *mus* – *dom* hybrid zone and dissected on site (termed "WILD" mice), and a second comparable group including 55 *mus* and *dom* inbred strains representatives (WSB/EiJ and PWD/PhJ, respectively, n=7 each) and their second-generation (F2) hybrids (n=41) housed under common laboratory conditions (termed "LAB" mice). Primary analysis of cecal mucosa-associated communities based on pyrosequencing of the bacterial 16S rRNA gene (Supplementary Table 1) indicates significant differences between WILD and LAB mice ("analysis of dissimilarity" (*adonis*) applied to four beta diversity measures; see Table 1, Methods). This includes significant differences in numerous taxa including phyla level Bacteroidetes, Firmicutes and Proteobacteria, and genus level *Bacteroides*, *Barnesiella*, *Helicobacter* and *Mucispirillum* (Supplementary Table 2). Thus, the following analyses were performed separately for these groups.

To identify hybrid versus "pure species" individuals among WILD mice, we used 46 diagnostic SNPs [27] and defined three groups based on the percentage of *M. m. domesticus*: *mus* (<12%, n=13), *hybrid* (12-85%, n=37) and *dom* (>85%, n=19; see Supplementary Figure 1). This revealed overall significant differences between groups for three out of four beta diversity measures (Table 1). Intriguingly, however, a linear model applied to the coordinates of a constrained principle coordinate analysis ("*capscale*" [28], see Methods) reveals the *hybrid* group to be significantly different from either *dom* or *mus* ($p<0.05$), whereas the major axis separating *dom* and *mus* themselves is not significant (Fig. 1a).

Similarly, among LAB mice we compared the parental strains WSB/EiJ and PWD/PhJ to their F2 hybrids, which based on the same 46 diagnostic SNPs display a range of % *M. m. domesticus* from 40 to 80%. These three groups display significant differences in community structure based on all four beta diversity measures (Table 1; Fig. 1b), in addition to a linear model applied to *capscale* coordinates revealing significant differences in all three possible comparisons (Table 1). Interestingly,

however, these comparisons are equidistant to each other in terms of Bray-Curtis distance (PWD/PhJ-WSB/EiJ = 0.84, PWD/PhJ-F2 hybrid = 0.85, F2 hybrid-WSB/EiJ = 0.82), which is not expected under an additive genetic model (see QTL analysis and Discussion).

Having identified differences between parental species and their hybrids in both the wild and laboratory settings using beta diversity analyses, we next focused on identifying underlying contributing aspects of the intestinal microbiome by applying alpha diversity (compositional richness) analyses. WILD mice display significantly higher richness than LAB mice at both the genus (Chao1: 52.9 vs. 41.7, p=3.42E-4) and species level (Chao1: 766.3 vs. 427.6, p=4.48E-7). Although categorical analysis of pure species and hybrids within each group revealed no significant differences, a polynomial linear model revealed a trend of decreasing richness among hybrid mice of both categories. Thus, we further analyzed this relationship by incorporating interspecific heterozygosity as a measure of the degree of "hybridness" in generalized linear models. Because age is a potential factor influencing diversity, we also incorporated body length and weight as proxies into each model. After controlling for body length (coefficient=-0.08, z=-23.10, p<2e-16) and weight (coefficient =0.0004, z=-43.80, p<2e-16), a negative relationship between interspecific heterozygosity and species-level Chao1 is apparent among WILD mice (coefficient =-0.144, z=-22.744, p<2e-16; Fig. 2). A negative relationship between interspecific heterozygosity and Chao1 is confirmed among F2 hybrid LAB mice (coefficient =- 0.974, z=- 12.028, p< 2e-16; Fig. 2), for which body length and weight played no significant role, likely due to a uniform age range (12-14 weeks; see Methods). Similar trends were observed for genus-level Chao1 but linear models did not show significance. Thus, in summary, individuals with a greater degree of hybridness display less diverse intestinal bacterial communities.

To identify potential contributors to lower alpha diversity among hybrids, we applied generalized linear models to individual bacterial genus abundances. Interestingly, despite overall different communities, this reveals several contributors common to both WILD and LAB hybrids: *Bacteroides* and *Helicobacter* are higher among animals with lower Chao1 diversity (p<0.05 after FDR correction), while other taxa including *Alistipes*, *Parasporobacterium* and unclassified_Lachnospiraceae, are lower among animals with reduced diversity (p<0.05 after FDR correction). Additional individual contributors include *Mucispirillum* (negatively correlated to

Chao1 in LAB mice), *Blautia* and unclassified_Ruminococcaceae (positively correlated to Chao1 in WILD mice). This indicates that an increase in abundance of relatively few well-known intestinal community members, at the expense of other taxa, contributes to lower diversity among hybrids.

Genetic architecture of hybrid mouse microbiomes

To further understand the genetic factors contributing to hybrid microbiomes, we performed a quantitative trait locus (QTL) mapping of the intestinal microbiota in an expanded set of 334 F2 LAB hybrids (WSB/EiJ x PWD/PhJ) using 234 informative SNPs. The log relative abundance of bacterial taxa belonging to a defined "Core-Measurable-Microbiota" (CMM [29]) in addition to Chao1 diversity were used as traits in single- and two-locus models (see Methods). We first determined family (cohort) and cage effects on all traits measured using a generalized linear model and then proceeded with mapping using the residuals from this model. To correct for multiple testing, 1000 global permutations were performed for all traits combined and the 95% quantile from all log of odds ratio (LOD) distributions was used to determine the threshold for significant QTLs (an LOD of 3.74 corresponds to the top 5%, while 3.12 corresponds to the top 10% and was used to define "suggestive" QTLs [30]).

Despite high environmental (cohort/cage) effects, a considerable amount of the variation in bacterial abundances and diversity is explained by a modest number of genomic loci, whereby 15 loci (unique SNPs) influence a total of 29 different traits, (Fig. 3; Supplementary Table 3). Among the 15 loci detected by QTL analysis, five display pleiotropic effects, as they are correlated to the abundances of several different bacterial taxa (Supplementary Table 3). Furthermore, the effects of 13 marker-trait pairs (*i.e*. roughly half of all such pairs) were found to be *transgressive*, whereby the mean phenotypic value of heterozygotes at a given locus is significantly outside the range of both homozygotes (examples in Fig. 4; a full list is given in Supplementary Table 3; see Methods). These effects influence several major community members including Bacteroidetes and Proteobacteria at the phylum level, *Bacteroides*, *Helicobacter* and *Barnesiella* at the genus level as well as Chao1 diversity. Thus, a large portion of the microbiome is influenced by distinct regions of the mouse genome, and the prevailing transgressive effects are in line with the distinct composition of the microbiome observed in hybrids.

For Chao1 diversity we identified particularly complex genetic control. First, a suggestive locus on chromosome 13 (rs6411888) displays a transgressive effect (Fig. 4b, Supplementary Table 3). Second, multi-locus mapping detects evidence for epistasis between two loci (rs30213049 on chromosome 14 and rs31610566 on chromosome 5) contributing to lower Chao1 values among hybrids (r^2=0.06322, p=0.0003). Furthermore, this interaction follows the expectations of a BDMI, as the double heterozygotes and one combination of homozygotes display the lowest mean values of all nine possible genotype combinations (Fig. 5).

Differential immune gene expression among hybrids

Given the possibility of large immune effects [9] and intervals of the current (Supplementary Table 3) and previous microbiota QTL studies [29,30] containing numerous immune-related genes, we next tested for an overall immune effect among hybrids by performing expression profiling of cecal tissue from 69 WILD- and the subset of 55 LAB mice using TaqMan Mouse Immune Arrays, which contain 90 genes involved in immune response in addition to six house-keeping genes. Applying linear models reveals the expression levels of 14 and 27 genes to be significantly negatively correlated to interspecific heterozygosity in WILD and LAB mice, respectively (Supplementary Table 4). Among these, two (Ccr7 and Tgfb1) are shared between WILD and LAB mice (Fig. 6).

To determine whether differences in immune gene expression contribute to the overall patterns of intestinal bacterial communities structure, we first applied the "envfit" function [31] (see Methods) to identify genes that correlate to the major axes of bacterial community ordinations. This reveals five and nine genes correlated to beta diversity among WILD and LAB mice, respectively (Supplementary Fig. 2), explaining from 8 to 19% of the variation in Bray-Curtis distance, although the effects are not additive. Of these, one (Ctla4) and three (Il18, Socs1 and Socs2) are among those negatively correlated to interspecific heterozygosity in WILD and LAB mice, respectively. Further, numerous significant associations (three bacteria genera to total 14 genes; p<0.05 after FDR correction) between individual immune genes and bacterial genera abundances support these results (Supplementary Table 5).

Inflammation and immunophenotypes in pure species versus hybrids

Given the aberrant expression of numerous immune-related genes in cecal tissue and potential for members of the intestinal microbiota to act as opportunistic pathogens, we next performed histopathological analysis of cecal tissue samples to reveal possible differences in immune cell influx and/or edema of the intestinal mucosa. Samples were evaluated based on both the presence/absence of moderate to severe pathology (e.g. ulcerations of the epithelium, accumulation of inflammatory cells, appearance of organized lymphoid structures; see Methods for full details). This revealed hybrids to display a higher proportion of animals with moderate/severe pathology for both the WILD and LAB groups (Fig. 7).

Next, we evaluated whether the individuals displaying moderate/severe pathology display differences in their immune gene expression and/or intestinal microbiome. First, among WILD mice we identified 52 genes with on average higher expression values in normal tissue samples compared to those with moderate/severe pathology, whereby two interleukins, Il18 and Il1b, display statistical significance (Wilcoxon test; p=0.007 and 0.029, respectively). Among LAB mice, only nine genes display higher expression in normal tissue, with one gene, interleukin Il2ra, displaying statistical significance (Wilcoxon test; p=0.025).

Second, we determined whether differences in abundance of the main bacterial phyla and genera (those with >1% overall abundance; n=5 and n=6, respectively) are present in normal tissue compared to that with moderate/severe pathology (Supplementary Table 6). Among WILD mice, tissues with pathology display higher Bacteroidetes and Deferribacteres and lower Firmicutes and Proteobacteria at the phylum level, whereas at the genus level *Barnesiella* and unclassified_Porphyromonadaceae are significantly lower. Among LAB mice, tissues with pathology display higher Tenericutes and lower Bacteroidetes at the phylum level, whereas at the genus level *Alistipes*, *Parasporobacterium* and unclassified_Lachnospiraceae are significantly lower. The general lack of correspondence between WILD and LAB mice in the taxa differing in the context of inflammation likely reflects the overall differences between these mice observed at the community level (see above). However, a small number of taxa differing among WILD mice do display similar trends in LAB mice, whereby Deferribacteres is higher- and Proteobacteria, *Barnesiella,* Lachnospiraceae and Porphyromonadaceae are lower in tissue displaying pathology (Supplementary Table 6).

Discussion

In our study, we make use of the *M. m. musclus – M. m. domesticus* hybrid system as a unique resource offering insight into the consequences of evolutionary divergence in a complex vertebrate holobiont. This is in particular made possible by the shared human-associated ecology (*i.e*. commensalism) and intermediate level of divergence between these hosts, whereby fitness-related traits such as fertility or parasite resistance are often reduced, yet reproductive isolation is incomplete, making them amenable to genetic study. Our previous work revealed considerable influences of geography on the intestinal microbiota of wild mice [6], thus, the close physical proximity of the hybrid zone (spanning approx. 40 km) provides a valuable opportunity to examine the influence of substantial genetic variability with minimal geographic influence. Furthermore, crosses between inbred strain representatives of *M. m. musclus* and *M. m. domesticus* enabled the results of a natural system to be compared to a controlled laboratory setting where environmental influences are further minimized.

Arguably, the most salient feature of hybrid intestinal microbiomes compared to those of pure *M. m. musclus* and *M. m. domesticus* is their overall transgressive pattern of community structure, which is similar in both the natural- and lab setting. Indeed, analysis of hybrid zone mice reveals significant differences in beta diversity to be limited to comparisons involving hybrids, whereby *M. m. musclus* and *M. m. domesticus* are more similar to each other than either is to their hybrids. This provides strong evidence that purifying selection is a predominating force operating on intestinal bacterial community composition and structure, whereby different genetic architectures for maintaining microbial communities that evolved over time are revealed only in the context of genomic mosaicism.

In the laboratory setting, differences between the PWD/PhJ (*musculus*) and WSB/EiJ (*domesticus*) inbred strains are also present. This may in part be due to the overall reduced diversity in lab mice combined with well-known "drift" among communities housed separately over time [32]. We note, however, that these effects, ultimately attributable to maternal transmission in the lab, are controlled for in the design of our F2 intercross (see Methods), whereby the intermixing between PWD/PhJ and WSB/EiJ communities is made possible *via* corprophagy. However, despite the opportunity for intermixing at the level of both the microbiome and host

genome, the PWD/PhJ-WSB/EiJ hybrid mice do not display an intermediate (additive) phenotype, but rather communities equidistant to either parental strain. This again supports the notion that the interaction between portions of divergent host genomes may lead to an altogether different community.

Several scenarios that could lead to divergent mechanisms for maintaining the same overall community "phenotype" are possible. For example, multiple host-pathobiont coevolutionary processes may accumulate over time, as evidenced by numerous taxonomic groups displaying transgressive phenotypes in our QTL analysis, some of which are known pathobionts and/or inhabit the intimately-associated mucosal layer (*e.g*. *Bacteroides*, *Helicobacter* and Deferribacteres). Alternatively or in addition, the evolution of resistance to other intestinal pathogens may lead to "collateral damage" in the context of maintaining commensal/symbiotic communities, which would then in turn select for other, compensatory genetic modifiers. For example, loss-of-function mutations at the *FUT2* (*Secretor*) locus in humans are known to provide resistance against Norovirus infection [33], but in turn alter bacterial community composition and structure to the potential detriment of the host [34]. Further, although the members of the house mouse species complex share a similar ecology of omnivory and commensalism, hosts that adapt to shifts in diet may do so in part *via* their gut microbiome, as in the case of the giant panda [35], which could further drive divergence of the host genome.

It is not yet possible to identify individual genes controlling variation in the microbiome with our current mapping intervals. However, as previous studies revealed a major role of immune system in shaping bacterial communities [4,9,36], we investigated the role of immune system functioning in the hybrid gut. Interestingly, a "transgressive" phenotype is also observed for a majority of immune genes, for which lower gene expression levels were observed in hybrids compared to the pure subspecies. Furthermore, they are tightly associated with either community-level variation or individual bacterial taxa. Combined with this immune deficiency, elevated abundances of Proteobacteria and Deferribacteres--who are transgressive and on average higher in heterozygotes in the F2 intercross--are also correlated to pathology in hybrid mice. Though a causal relationship has yet to be established between host and bacteria, this more frequent pathology is an indication of lower fitness in hybrid mice and may represent a phenomenon similar to that observed in *Nasonia* [4].

The additional discovery of epistasis between loci following a BDM-model is intriguing, though admittedly our mapping population does not have enough power to detect more other interacting loci. Arguably, the alpha diversity (Chao1 in our study) correlated to this pair reflects essential characteristics of the microbial community, such as community stability and anti-infectious abilities that can be crucial to the host's health [37]; and the finding of a gene interaction following a classical model of genetic incompatibility further reflects the existence of selective pressure for maintaining a stable bacterial community between diverging subspecies. In their "Microbial-assisted BDM" model, Brucker & Bordenstein [9] extend such incompatibilities to include microbial components, whereby numerous classes of incompatibilities within and between mosaic host genomes and microbiomes are possible. Due to the lack of both microbial genomic data and insufficient time for co-diversification of bacteria since the common ancestor of *M. m. musculus and M. m. domesticus* at the 16S rRNA species level threshold [38], our taxonomic-based approach is limited in its ability to detect incompatibilities beyond those involving the host genome. The identification of characterization of such "microbial assistance" indeed represents a major future experimental challenge that would be expected to significantly escalate with increasing complexity of the bacterial community involved.

In summary, our study provides novel microbial insight into the house mouse hybrid zone as an important, long-studied evolutionary model system. We demonstrate that the aberrant genetic architecture and/or immune functioning of hybrids are the most likely causes of their divergent gut microbiome. This provides support for "hologenome" concept in complex vertebrates, in which the host genome and microbiome are involved in crucial interactions that may eventually shape the evolution of host species.

Methods

Wild mouse sampling

House mice were captured in farms and horse barns in the Bavarian hybrid zone of *Mus musculus musculus* and *Mus musculus domesticus* during May and June 2011 [6] along a transect from Augsburg to Landshut. The sampling transect spans the whole hybrid zone as well as part of the pure subspecies distributions (Figure S1). In total, 69 mice were captured from 34 unique locations, with a maximum of three mice per location. All mice were dissected on site according to the procedure described by Linnenbrink et al [6]. Cecal tissues were preserved in RNALater to obtain both bacterial DNA and host RNA, and cecal lymphoid tissues were fixed in formalin for histological analysis.

Lab mouse husbandry and crossing

WSB/EiJ (The Jackson Laboratory, *M.m.domesticus*) and PWD/PhJ (The Jackson Laboratory, *M.m.musculus*) mice strains are kept under conventional conditions at Max-Planck-Institute for Evolutionary Biology, Ploen, Germany. F1 mice (WXP) were generated from crosses (n=4) between WSB/EiJ females and PWD/PhJ males (the reciprocal cross produced sterile males), and F2 mice (WP) are litters of sibling mate pairs (n=8) of F1 mice. All mice were weaned at 21 days and transferred to cages according to gender. Food and water were given *ad libitum* until the mice reached 12 +-1 weeks of age. We dissected 41 F2 mice as well as seven each of WSB/EiJ and PWD/PhJ parental mice using the identical procedure performed for the wild house mice. Additionally, a total of 293 F2 mice were used for quantitative trait loci (QTL) analysis, for which only cecal tissue was used to extract bacterial DNA and ears for mouse genomic DNA.

Pyrosequencing and bacterial community analysis

Bacterial DNA from cecal tissues was extracted using the QIAmp DNA stool mini kit (QIAGEN), using the modified protocol described by Linnenbrink et al [6]. We focused primarily on cecal tissue due to the finding that host genetics has a greater influence on this mucosal site than on the luminal [6]. The V1-V2 region of 16S rRNA gene was amplified using the 27F-338R primer pair according to the conditions described by Linnenbrink et al [6], and sequencing was performed on a 454 GS-FLX with Titanium

sequencing chemistry. Raw sequences were trimmed and filtered using Mothur version 1.22.2 [39] with the requirements of average quality score >25 and minimum length of 250 bp. Sequences were assigned to each sample by exact matches of MID (multiplex identifier, 10 nt) sequences. Chimeras were removed using the Uchime (Usearch) program [40] using the recommended database as a reference, and species-level OTUs (operational taxonomical unit, 97% sequence similarity threshold) were clustered using Uclust (Usearch) [41]. To normalize the reads per sample we picked a random subset of 1000 sequences per sample for all mice, and taxonomical classification of sequences was performed using RDP classifier with a bootstrap value of 80% for all taxonomical levels [42]. Alpha- (Chao1 index) and beta-diversity measures (Bray-Curtis and Jaccard distances) were calculated in R using Vegan package [31]. Phylogenetic-based alpha-diversity (Phylogenetic diversity [43]) and beta-diversity (weighted and unweighted UniFrac [44]) were obtained using Mothur with a phylogenic tree produced by FastTree [45]. The Vegan package in R [46] was used for analysis of dissimilarity ("*adonis*", a multidimensional analysis of significance based on the variance in distance matrices) and constrained analysis of principal coordinates ("*capscale*"), a hypothesis-driven ordination that limits the separation of the communities based on the variable tested [28].

Mouse genotyping

Genomic DNA was extracted from ears using DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's protocol. For all the 69 wild mice (WILD mice), as well as subsets of 41 F2 (WP) mice from WSB/PWD cross (LAB mice), the relative proportion of the *M.m.domesticus* and *M.m.muculus* subspecies was determined by genotyping a diagnostic set of 46 SNPs roughly equally distributed along the chromosomes using the SNPstream platform [27]. For the larger set of F2 mice generated for QTL analysis, genotyping of 370 SNP markers was performed with the Sequenom iPLEX MassARRAY system. These SNPs were designed according to the NCBI mouse SNP database and are evenly distributed along mouse chromosomes (average interval 10M bp). Genomic DNA from WSB/EiJ and PWD/PhJ mice were used for genotyping controls.

QTL analysis

Vigorous quality filtering of the genotyping data was performed according to the procedure described by White et al [47], whereby SNPs failing the criteria were removed. Mice missing >20% of the SNP marker information and/or containing less than 900 bacterial 16S rRNA gene reads were also removed. This resulted in a final data set of 334 mice genotyped at 234 SNP markers, which contains 96.2% of the genotype data for this subset. The genetic and physical maps were calculated using the "est.map" function in the R/qtl package [48] assuming zero genotyping error and SNP marker positions corresponding to the position in the NCBI SNP database. We included the log-transformed relative abundances of a total of 77 Core-Measurable-Microbiota traits (CMM [29], defined as taxa with more than of 30 sequences per sample on average in our dataset) at different taxonomical levels: phylum, class, order, family, genus as well as 97% similarity OTUs (species-level OTUs). In addition, alpha-diversity (Chao1) was included as a trait. Family and cage effects were controlled for using a general linear model in the "lmer" R package, which treats family as a random factor and cohort/cage as a random factor nested within family. The residuals from this model were then used for QTL analysis. To obtain the significance threshold, we first performed 1000 permutations on all traits combined using the "scanone" function in R/qtl, and for each permutation we extracted the LOD value ($\log_{10}$ likelihood ratio) and used the top 5% of the overall LOD distributions as the significance threshold for QTLs and the top 10% for suggestive QTLs. A standard interval mapping for each trait was then performed using the "scanone″ function with default parameters. The variance in a trait explained by a given locus was calculated using a linear model in R. The detection of epistatic interactions between two loci was carried out using the "scantwo" function in R/qtl, which performs a step-wise, two-dimensional mapping, for which the significance threshold was determined using 10000 permutations.

Immune gene expression analysis

RNA was extracted from RNALater-preserved cecal tissues (WILD and LAB mice) using the QIAgen RNA miniprep kit according to the manufacturer's protocol. After RNA extraction, cDNA was synthesized using the High-Capacity cDNA synthesis kit (Invitrogen) according to the manufacturer's protocol. The TaqMan® Array Mouse Immune Panel microfluidic card chambers were loaded with 100 uL of mixture containing 35 uL cDNA (corresponding to 35 ng of total RNA), 15 uL of nuclease-

free water and 50 uL of TaqMan® Universal PCR Master Mix. This panel contains 90 immune- and six house-keeping genes. The results were analyzed using the SLqPCR package [49] in R. This is conducted by first transforming the raw data to relative expression values and then normalizing based on geometric averaging of all six house-keeping genes, as proposed by Vandesompele et al [49]. The normalized expression data were analyzed using a generalized linear model with interspecific heterozygosity as continuous variable.

Histological analysis

Cecal tissues preserved in formalin were processed and embedded in paraffin. Tissue sections (5 μm) were deparaffinized and stained with hematoxylin and eosin (H&E). Tissue pathology was determined according to the presence of cell debris in the intestinal lumen, infiltrating lymphocytes and polymorphnuclear cells to the submucosa and mucosa, epithelial desquamation, crypt abscesses and ulcerations. The presence of "moderate-severe" inflammation is defined by the presence of three or more of the above-listed pathological traits.

References

1. Ley, R.E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647-1651 (2008).
2. McFall-Ngai, M. *et al.* Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 3229-36 (2013).
3. Ochman, H. *et al.* Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology* **8**, - (2010).
4. Brucker, R.M. & Bordenstein, S.R. The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus *Nasonia*. *Science* **341**, 667-9 (2013).
5. Ley, R.E., Lozupone, C.A., Hamady, M., Knight, R. & Gordon, J.I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6**, 776-788 (2008).
6. Linnenbrink, M. *et al.* The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Molecular Ecology* **22**, 1904-1916 (2013).
7. Staubach, F., Baines, J.F., Kunzel, S., Bik, E.M. & Petrov, D.A. Host species and environmental effects on bacterial communities associated with Drosophila in the laboratory and in the natural environment. *PLoS One* **8**, e70749 (2013).
8. Zilber-Rosenberg, I. & Rosenberg, E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *Fems Microbiology Reviews* **32**, 723-35 (2008).
9. Brucker, R.M. & Bordenstein, S.R. Speciation by symbiosis. *Trends in Ecology and Evolution* **27,** 443-451 (2012).
10. McFall-Ngai, M. Adaptive immunity: care for the community. *Nature* **445**, 153 (2007).
11. Staubach, F. *et al.* Expression of the blood-group-related glycosyltransferase *B4galnt2* influences the intestinal microbiota in mice. *ISME Journal* **6**, 1345-1355 (2012).
12. Pedersen, A.B. & Babayan, S.A. Wild immunology. *Molecular Ecology* **20**, 872-80 (2011).
13. Abolins, S.R., Pocock, M.J.O., Hafalla, J.C.R., Riley, E.M. & Viney, M.E. Measures of immune function of wild mice, *Mus musculus*. *Molecular Ecology* **20**, 881-892 (2011).
14. Guenet, J.L. & Bonhomme, F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in Genetics* **19**, 24-31 (2003).
15. Boursot, P. & Belkhir, K. Mouse SNPs for evolutionary biology: beware of ascertainment biases. *Genome Research* **16**, 1191-2 (2006).
16. Geraldes, A. *et al.* Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Molecular Ecology* **17**, 5349-63 (2008).
17. Forejt, J. & Ivanyi, P. Genetic studies on male sterility of hybrids between laboratory and wild mice (*Mus musculus* L.). *Genetic Research* **24**, 189-206 (1974).
18. Britton-Davidian, J., Fel-Clair, F., Lopez, J., Alibert, P. & Boursot, P. Postzygotic isolation between the two European subspecies of the house

mouse: estimates from fertility patterns in wild and laboratory-bred hybrids. *Biological Journal of the Linnean Society* **84**, 379-393 (2005).
19. Good, J.M., Handel, M.A. & Nachman, M.W. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* **62**, 50-65 (2008).
20. Turner, L.M., Schwahn, D.J. & Harr, B. Reduced male fertility is common but highly variable in form and severity in a natural house mouse hybrid zone. *Evolution* **66**, 443-458 (2012).
21. Sage, R.D., Heyneman, D., Lim, K.C. & Wilson, A.C. Wormy mice in a hybrid zone. *Nature* **324**, 60-63 (1986).
22. Baird, S.J. *et al.* Where are the wormy mice? A reexamination of hybrid parasitism in the European house mouse hybrid zone. *Evolution* **66**, 2757-72 (2012).
23. Dobzhansky, T. Studies on Hybrid Sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* **21**, 113-35 (1936).
24. Bateson, W. Heredity and variation in modern lights. *Darwin and Modern Science:* **85**(1909).
25. Muller, H. Bearing of the *Drosophila* work on systematics. *The new systematics.*, 185–268. (1940).
26. Janousek, V. *et al.* Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. *Molecular Ecology* **21**, 3032-47 (2012).
27. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Research* **16**, 730-737 (2006).
28. Anderson, M.J. & Willis, T.J. Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* **84**, 511-525 (2003).
29. Benson, A.K. *et al.* 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* **107**, 18933-18938 (2010).
30. Srinivas, G. *et al.* Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nature Communications* **4**, 2462 (2013).
31. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**, 927-930 (2003).
32. Ubeda, C. *et al.* Familial transmission rather than defective innate immunity shapes the distinct intestinal microbiota of TLR-deficient mice. *Journal of Experimental Medicine* **209**, 1445-1456 (2012).
33. Lindesmith, L. *et al.* Human susceptibility and resistance to Norwalk virus infection. *Nature Medicine* **9**, 548-553 (2003).
34. Rausch, P. *et al.* Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and *FUT2* (Secretor) genotype. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 19030-5 (2011).
35. Zhu, L.F., Wu, Q., Dai, J.Y., Zhang, S.N. & Wei, F.W. Evidence of cellulose metabolism by the giant panda gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17714-17719 (2011).

36. Franzenburg, S. *et al.* MyD88-deficient Hydra reveal an ancient function of TLR signaling in sensing bacterial colonizers. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19374-19379 (2012).
37. Stecher, B. *et al.* Like Will to Like: Abundances of Closely Related Species Can Predict Susceptibility to Intestinal Colonization by Pathogenic and Commensal Bacteria. *PLoS Pathogens* **6**, - (2010).
38. Ochman, H., Elwyn, S. & Moran, N.A. Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 12638-12643 (1999).
39. Schloss, P.D. *et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* **75**, 7537-7541 (2009).
40. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200 (2011).
41. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-1 (2010).
42. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-5267 (2007).
43. Faith, D.P. Conservation Evaluation and Phylogenetic Diversity. *Biological Conservation* **61**, 1-10 (1992).
44. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228-8235 (2005).
45. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**, 1641-1650 (2009).
46. R Core Development Team. R: A Language and Environment for Statistical Computing. (2011).
47. White, M.A., Steffy, B., Wiltshire, T. & Payseur, B.A. Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics* **189**, 289-U988 (2011).
48. Broman, K.W., Wu, H., Sen, S. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890 (2003).
49. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology* **3**(2002).

Table 1. Summary statistics of beta-diversity comparisons among different groups. The variances (r^2) and p-values are calculated using analysis of dissimilarity (*adonis*, see methods), and all four beta-diversity measures as well as all comparisons in the results are presented. Bold p-values are significant comparisons (<0.05).

	Bray-Curtis		Jaccard		Weighted Unifrac		Unweighted Unifrac	
	r^2	p	r^2	p	r^2	p	r^2	p
Between WILD and LAB	0.0654	**0.0010**	0.0434	**0.0010**	0.1500	**0.0010**	0.0640	**0.0010**
Hybrid/Pure subspecies in WILD	0.0570	**0.0110**	0.0540	**0.0010**	0.0530	**0.0300**	0.0600	0.1030
Between subspecies in WILD	0.0330	0.3810	0.0326	0.4010	0.0342	0.3560	0.0378	0.0960
Hybrid/Pure subspecies in LAB	0.3062	**0.0010**	0.2340	**0.0010**	0.1430	**0.0130**	0.1110	**0.0010**
Between subspecies in LAB	0.1618	**0.0490**	0.1380	**0.0390**	0.1384	0.0790	0.0762	0.2240

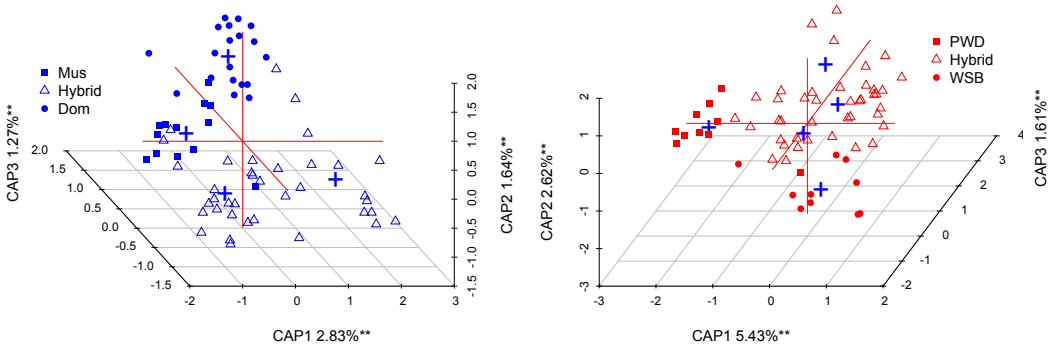


Figure 1. Constrained principal coordinate analysis ("capscale") of the gut microbiome in WILD (left panel) and LAB (right panel) mice. Mice belonging to either of the subspecies are denoted with filled circles/diamonds, and hollow triangles are hybrid mice. Only the first three axes and explained variances are displayed. **Represents significance from the "anova.cca" test with respect to geographic location as a categorical variable with 1000 permutations (see Methods; $P < 0.01$).

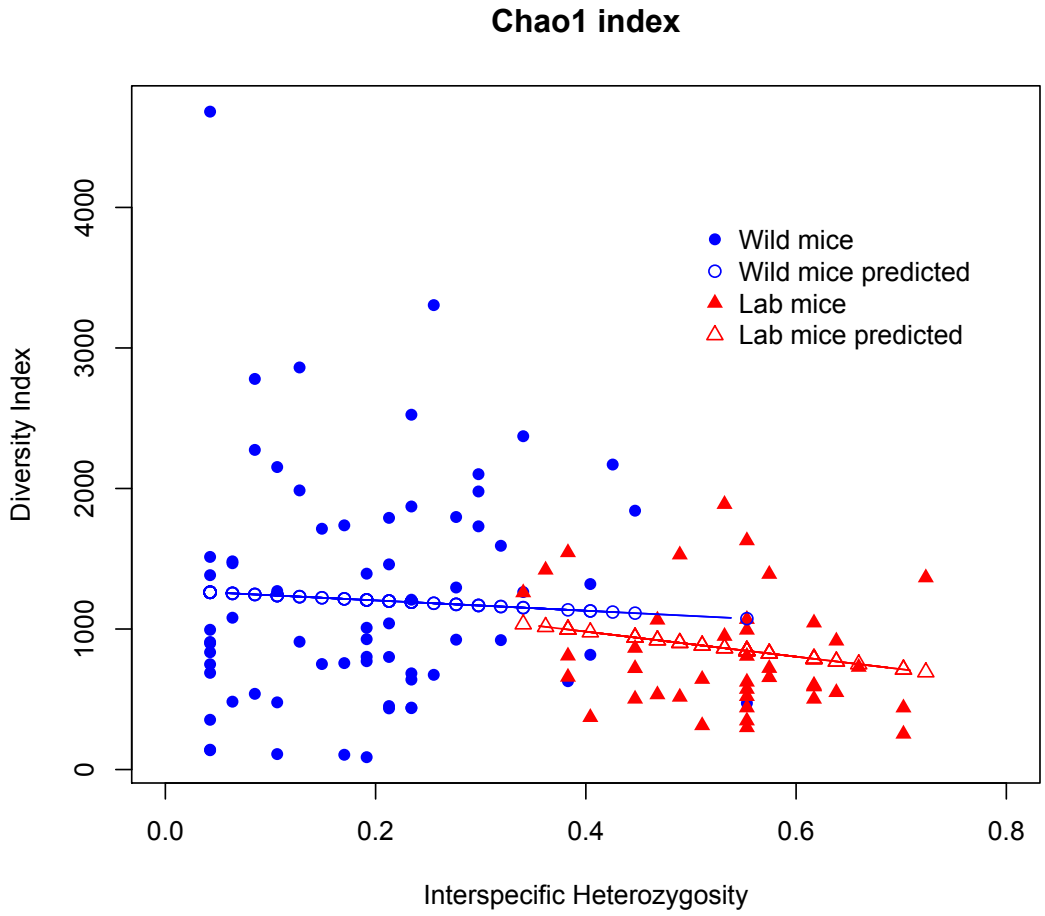


Figure 2. Chao1 diversity decreases with increasing interspecific heterozygosity. Chao1 diversity is calculated based on species-level OTUs for WILD (blue filled circles) and LAB (red filled triangles) mice, and hollow circles/triangles show the decrease of Chao1 diversity based on a general linear regression (see results).

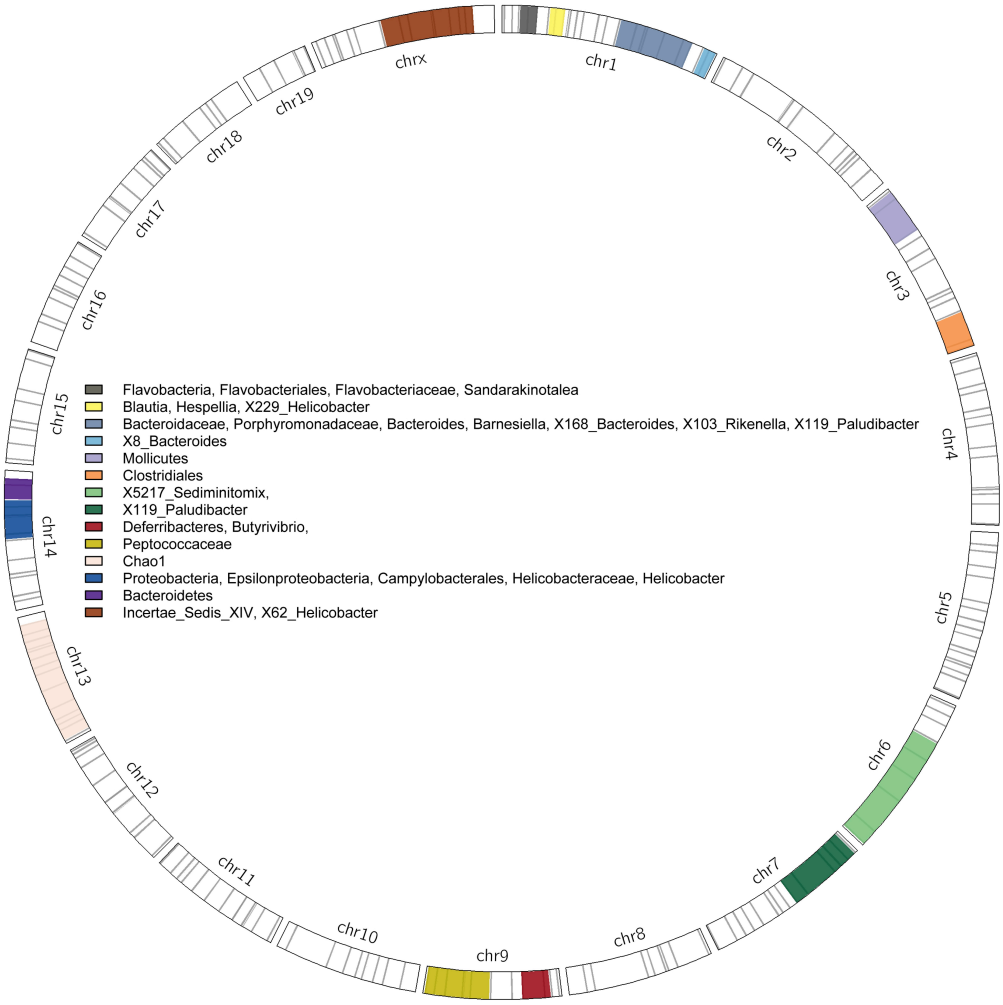


Figure 3. QTL mapping of the gut microbiota in F2 hybrids of WSB X PWD. Black lines on the chromosomes are SNPs used in the QTL mapping, and each colored region denotes the confidence interval of a QTL. Traits significantly correlated to given QTLs are listed in the legend.

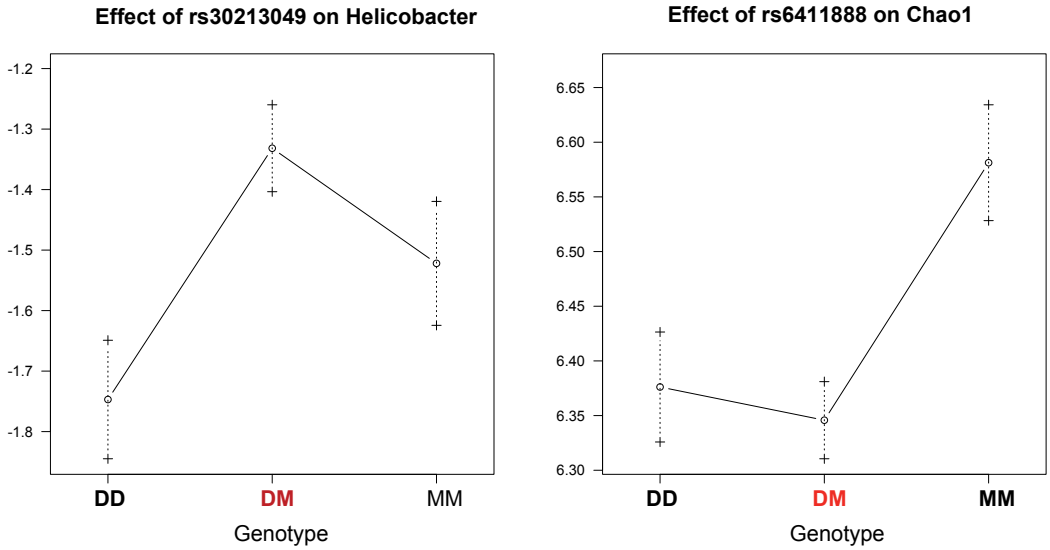


Figure 4. Examples of transgressive phenotypes identified by QTL analysis. Left panel: SNP rs30213049 on chromosome 14 is significantly correlated to the abundance of *Helicobacter,* whereby the heterozygotes at this SNP locus have significantly higher *Helicobacter* than the homozygotes. Right panel: SNP rs6411888 on chromosome 13 is significantly correlated to the Chao1 diversity of the microbial communities, and heterozygotes at this SNP locus have significantly lower Chao1 than the homozygotes (Supplementary Table 3). For both panels, M denotes *M. m. musculus* alleles and D denotes *M. m. domesticus* alleles.

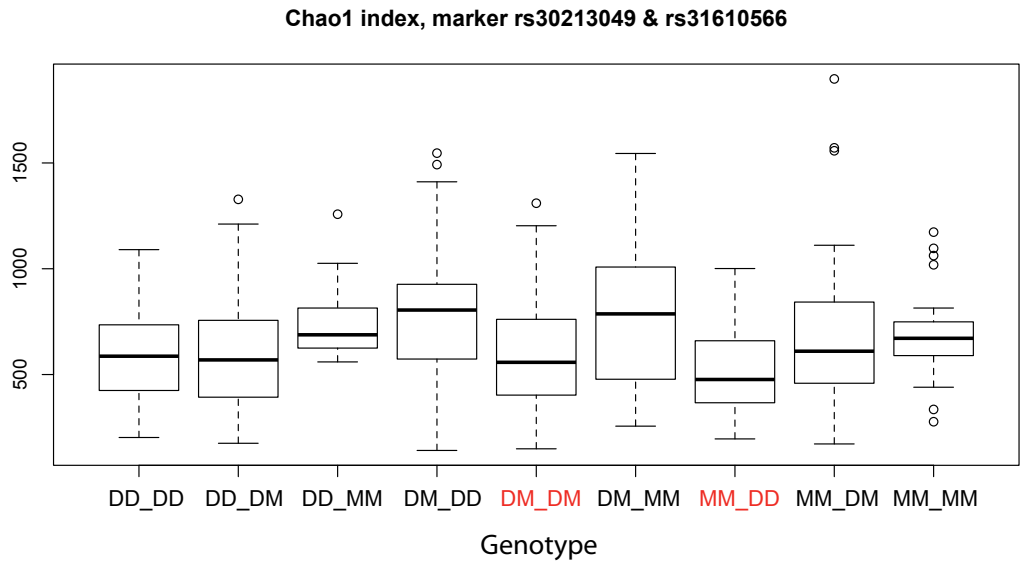


Figure 5. Boxplot of two-locus epistasis between rs30213049 (Chr 14, 84.8M) and rs31610566 (Chr 5, 81.9M). The epistasis of the two loci follows a Bateson-Dobzhansky-Muller model, whereby the genotypes marked in red have the two lowest Chao1 measures (double heterozygotes DM_DM and another genotype MM_DD), and the differences among genotypes are significant (ANOVA p=0.015). For all genotypes, M denotes *M. m. musculus* alleles and D denotes *M. m. domesticus* alleles.

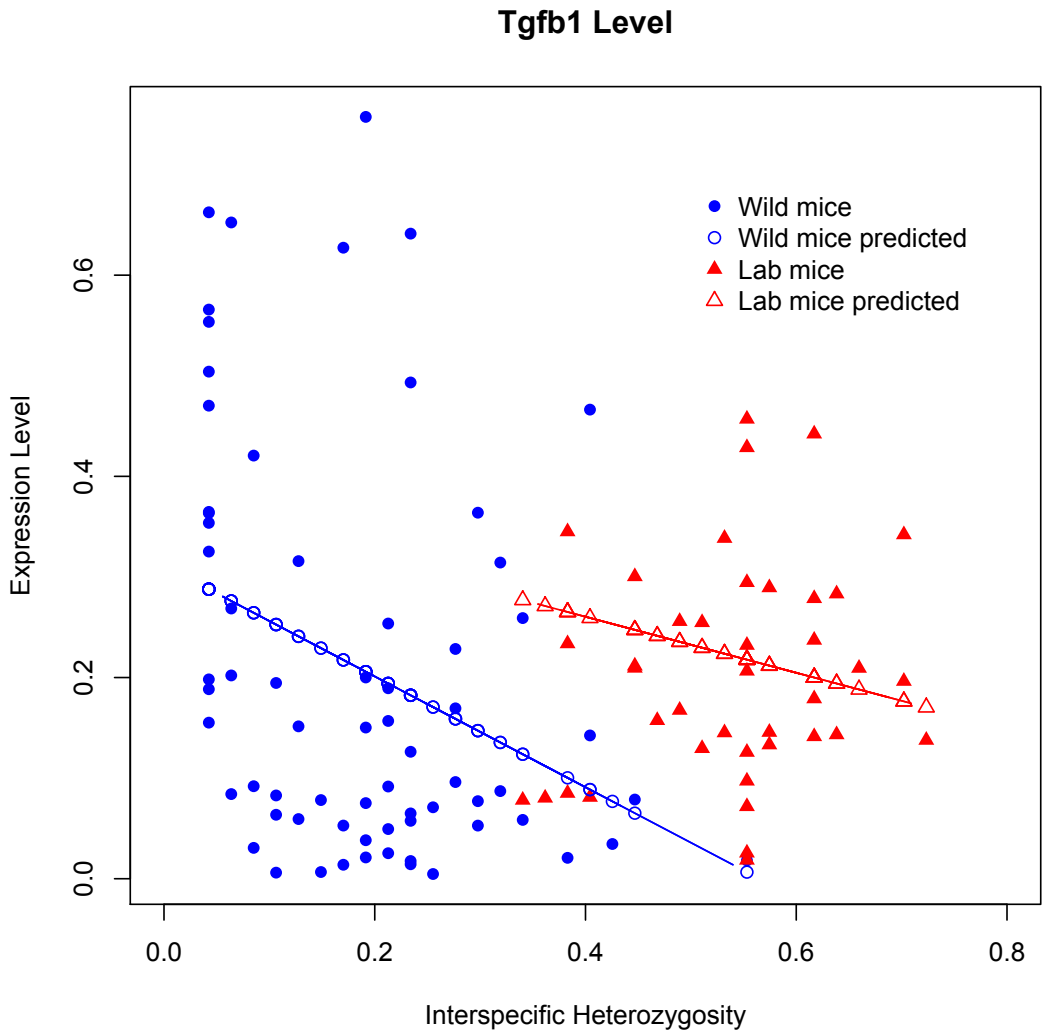


Figure 6. Expression level of *Tgfb1* decreases with increasing interspecific heterozygosity. Expression levels were determined by the Mouse Immune Panel microfluidic card (see methods) for WILD (blue filled circles) and LAB (red filled triangles) mice, and hollow circles/triangles show the decrease of expression level based on a general linear regression (see results).

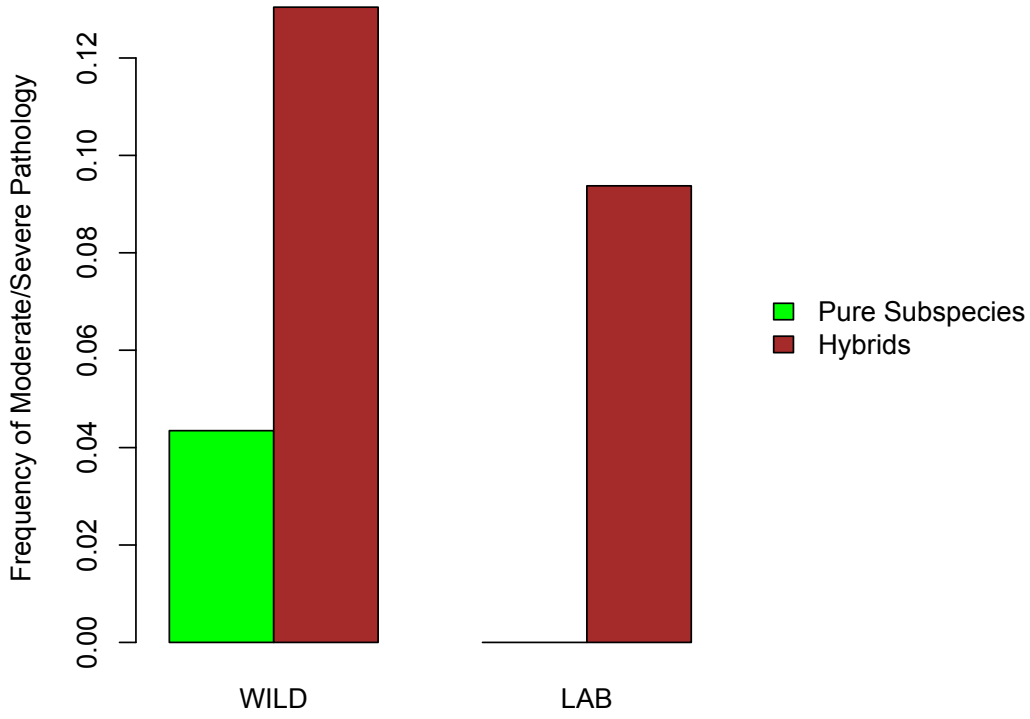
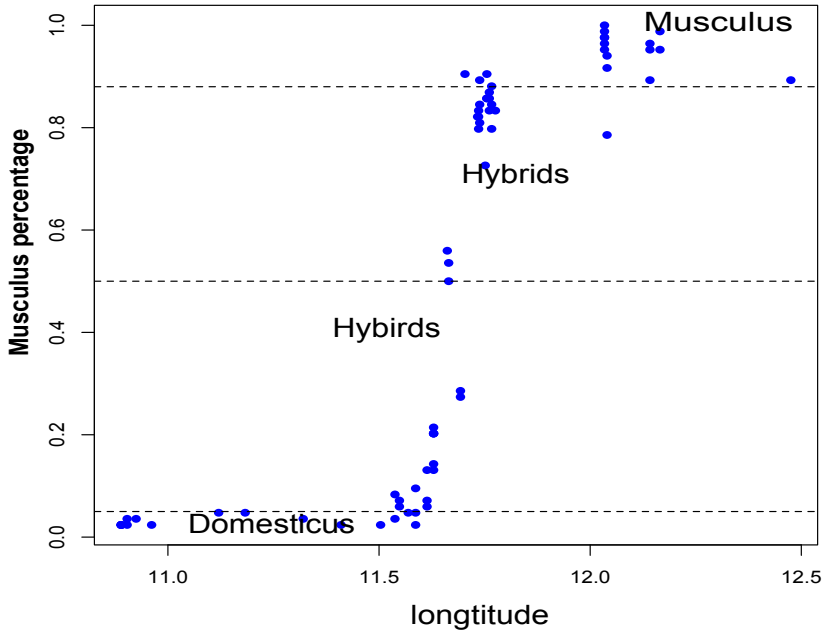
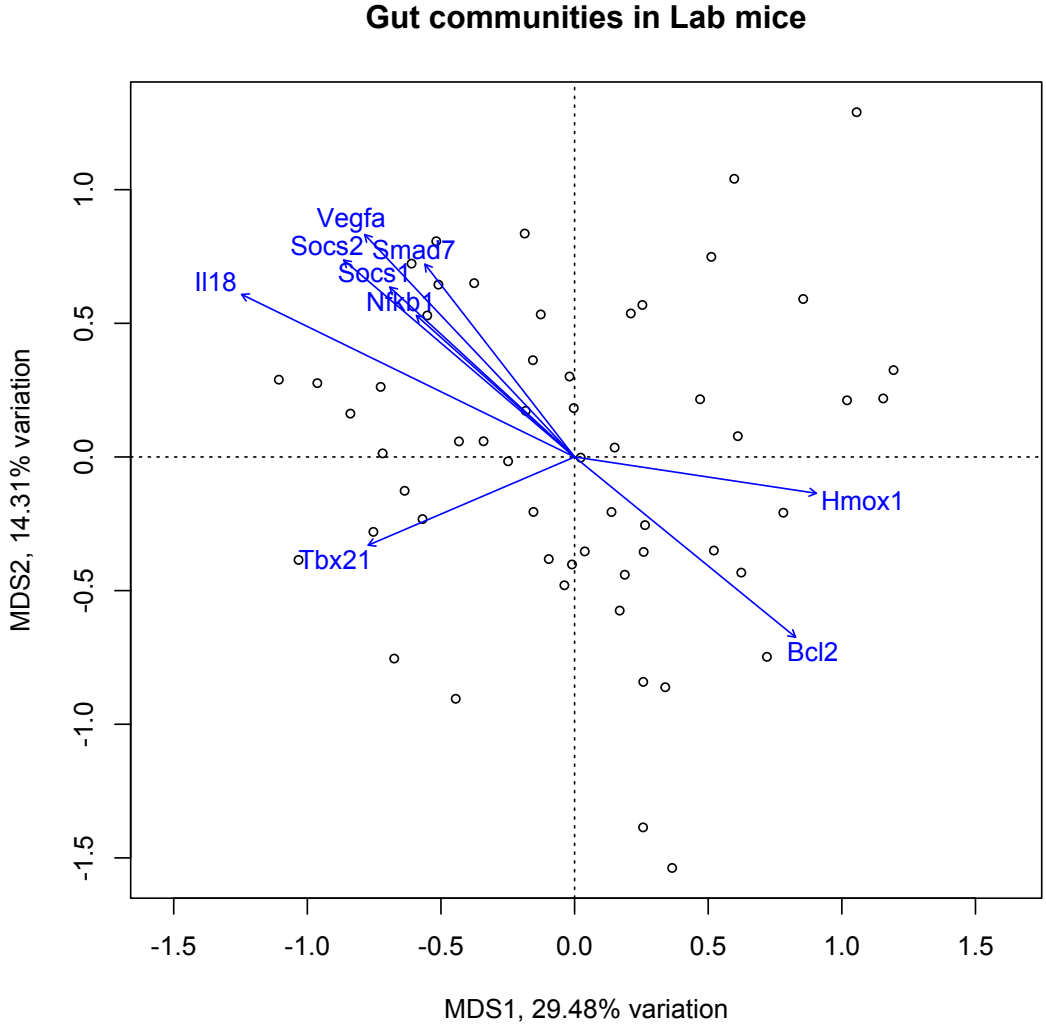


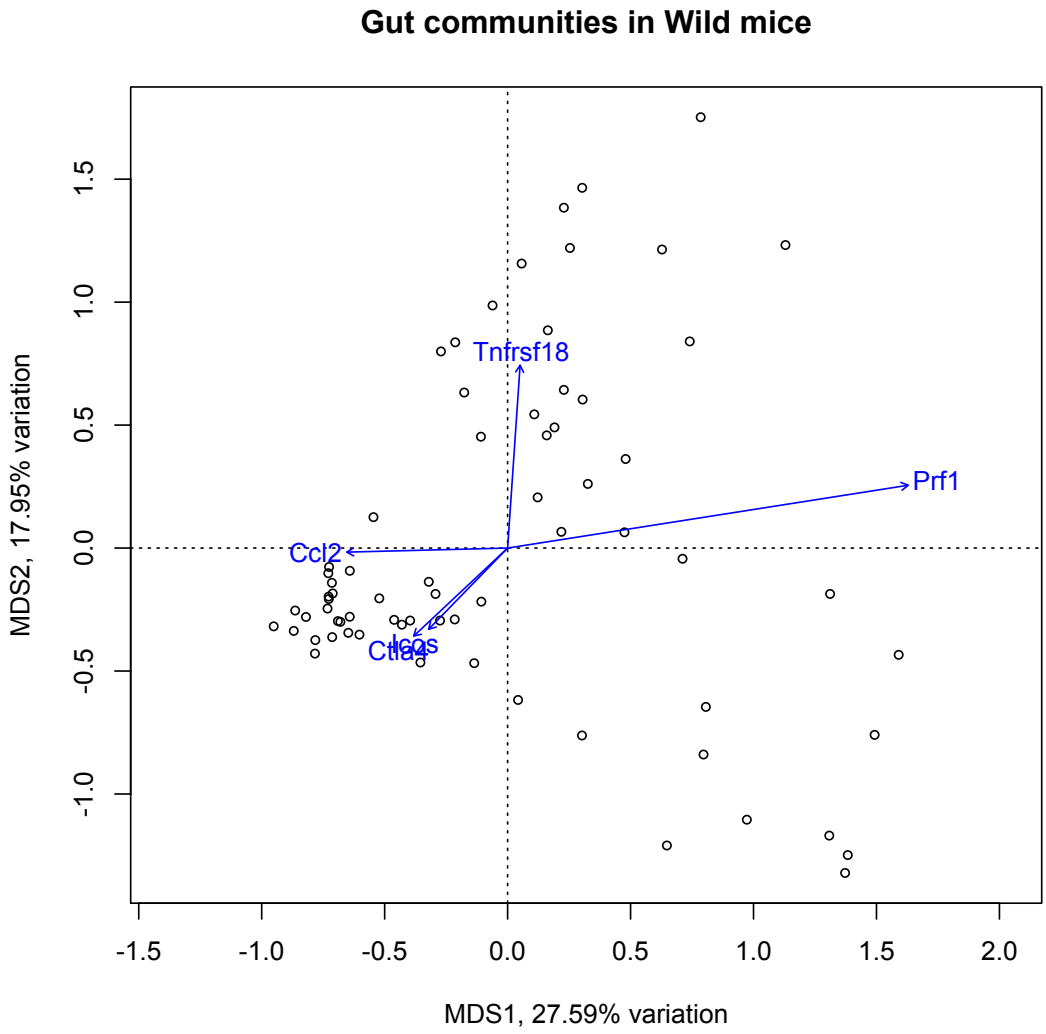
Figure 7. Frequencies of moderate-to-severe pathology in cecal tissue. For both WILD and LAB mice, the frequencies of pathology are presented for pure subspecies (green) and hybrid mice (red).



Supplementary Figure 1. Genomic composition of mice across the hybrid zone. Percentage of *M. m. musculus* is calculated from the number of SNPs unique to this subspecies in the whole SNP array (Harr *et al.,* 2009). The X-axis denotes the transition of *M .m. domesticus* to *M. m. musuclus* from west (Augsburg) to east (Landshut).


Gut communities in Lab mice
Vegfa
Socs2
Smad7
Socs1
Nfkb1
Il18
Hmox1
Tbx21
Bcl2
MDS2, 14.31% variation
MDS1, 29.48% variation
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

a)



b)

Supplementary Figure 2. Correlation of immune genes in shaping the microbiome. The upper panel shows the correlation of nine genes in shaping the lab mouse gut microbiome, the lower panel shows the correlation of five genes in wild mice. Both plots are un-constrained PCoA plots of the Bray-Curtis index, and the direction of immune genes is determined by "envfit" function, which applies a linear-regression based on the PCoA plot. All genes here have significant correlations ($p<0.05$ in envfit with FDR corrections).

Supplementary Table 1. Overview of WILD and LAB mice mentioned in this study. The number of 16S rRNA reads after quality filtering is presented.

Mice	Origin	Specification	Reads number
FS01	WILD	Hybrid	3226
FS02	WILD	Hybrid	3343
FS03	WILD	Hybrid	2358
FS04	WILD	Hybrid	2437
FS05	WILD	Hybrid	2117
FS06	WILD	Hybrid	1825
FS07	WILD	Hybrid	4117
FS08	WILD	*M.m.musculus*	2389
FS09	WILD	*M.m.musculus*	2085
FS10	WILD	Hybrid	2554
FS11	WILD	Hybrid	2061
FS12	WILD	Hybrid	3736
FS13	WILD	Hybrid	3295
FS14	WILD	Hybrid	2850
FS15	WILD	Hybrid	3152
FS16	WILD	Hybrid	3556
FS17	WILD	*M.m.musculus*	2589
FS18	WILD	Hybrid	1934
FS19	WILD	*M.m.musculus*	6133
FS20	WILD	*M.m.musculus*	3320
FS21	WILD	*M.m.musculus*	4949
FS22	WILD	*M.m.musculus*	2568
FS23	WILD	*M.m.musculus*	2141
FS24	WILD	*M.m.musculus*	3065
FS25	WILD	Hybrid	4752
FS26	WILD	Hybrid	3790
FS27	WILD	*M.m.domesticus*	4855
FS28	WILD	*M.m.domesticus*	3963
FS29	WILD	Hybrid	3268
FS30	WILD	*M.m.domesticus*	3426
FS31	WILD	*M.m.domesticus*	5249
FS32	WILD	Hybrid	3288
FS33	WILD	*M.m.domesticus*	4622
FS34	WILD	Hybrid	3749
FS35	WILD	*M.m.domesticus*	6631
FS36	WILD	*M.m.domesticus*	3911
FS37	WILD	Hybrid	3347
FS38	WILD	Hybrid	6972
FS39	WILD	Hybrid	6092
FS40	WILD	Hybrid	1166
FS41	WILD	Hybrid	1337
FS42	WILD	Hybrid	3898
FS43	WILD	Hybrid	3771
FS44	WILD	Hybrid	1281
FS45	WILD	Hybrid	2456
FS46	WILD	Hybrid	1373
FS47	WILD	Hybrid	2407
FS48	WILD	Hybrid	3317

FS49	WILD	Hybrid	2996
FS50	WILD	Hybrid	1561
FS51	WILD	Hybrid	1994
FS52	WILD	*M.m.musculus*	1209
FS53	WILD	*M.m.musculus*	2127
FS54	WILD	*M.m.musculus*	4111
FS55	WILD	Hybrid	3898
FS56	WILD	Hybrid	5382
FS57	WILD	*M.m.domesticus*	4251
FS58	WILD	*M.m.domesticus*	2904
FS59	WILD	*M.m.domesticus*	6084
FS60	WILD	*M.m.domesticus*	3565
FS61	WILD	*M.m.domesticus*	3403
FS62	WILD	*M.m.domesticus*	2454
FS63	WILD	*M.m.domesticus*	2435
FS64	WILD	*M.m.domesticus*	1940
FS65	WILD	*M.m.domesticus*	1446
FS66	WILD	*M.m.domesticus*	6590
FS67	WILD	*M.m.domesticus*	3399
FS68	WILD	*M.m.domesticus*	3959
FS69	WILD	*M.m.musculus*	3430
WP101.C.1F	LAB	F2_hybrids	3882
WP101.C.1M	LAB	F2_hybrids	2345
WP101.C.2F	LAB	F2_hybrids	4084
WP101.C.2M	LAB	F2_hybrids	2436
WP101.C.3M	LAB	F2_hybrids	2053
WP101.C.4M	LAB	F2_hybrids	1808
WP101.C.5M	LAB	F2_hybrids	3846
WP106.F2C.1F	LAB	F2_hybrids	2601
WP106.F2C.1M	LAB	F2_hybrids	3746
WP106.F2C.2F	LAB	F2_hybrids	2178
WP106.F2C.3F	LAB	F2_hybrids	3111
WP106.F2C.4F	LAB	F2_hybrids	1750
WP106.F2C.5F	LAB	F2_hybrids	1724
WP107.F2C.1F	LAB	F2_hybrids	3373
WP107.F2C.1M	LAB	F2_hybrids	1543
WP107.F2C.2F	LAB	F2_hybrids	3460
WP107.F2C.2M	LAB	F2_hybrids	868
WP107.F2C.3F	LAB	F2_hybrids	1382
WP107.F2C.3M	LAB	F2_hybrids	6759
WP107.F2C.4M	LAB	F2_hybrids	4830
WP107.F2C.5M	LAB	F2_hybrids	4103
WP108.F2C.1F	LAB	F2_hybrids	2489
WP108.F2C.1M	LAB	F2_hybrids	1892
WP108.F2C.2F	LAB	F2_hybrids	1738
WP108.F2C.2M	LAB	F2_hybrids	2736
WP108.F2C.3M	LAB	F2_hybrids	1635
WP108.F2C.4F	LAB	F2_hybrids	1243
WP109.F2C.1F	LAB	F2_hybrids	2261
WP109.F2C.1M	LAB	F2_hybrids	3018
WP109.F2C.2F	LAB	F2_hybrids	1963
WP109.F2C.2M	LAB	F2_hybrids	2694
WP109.F2C.3F	LAB	F2_hybrids	1120

WP110.F2C.1F	LAB	F2_hybrids	5274
WP110.F2C.1M	LAB	F2_hybrids	1802
WP110.F2C.2F	LAB	F2_hybrids	2837
WP110.F2C.2M	LAB	F2_hybrids	1955
WP110.F2C.3M	LAB	F2_hybrids	4301
WP110.F2C.4M	LAB	F2_hybrids	3733
WP111.F2C.1F	LAB	F2_hybrids	4580
WP111.F2C.1M	LAB	F2_hybrids	1021
WP111.F2C.2M	LAB	F2_hybrids	1743
PWD131A1F	LAB	*M.m.musculus*	1146
PWD131A1M	LAB	*M.m.musculus*	1201
PWD131A2F	LAB	*M.m.musculus*	2981
PWD131A2M	LAB	*M.m.musculus*	1236
PWD131A3F	LAB	*M.m.musculus*	1459
PWD131A4F	LAB	*M.m.musculus*	3134
PWD131A5F	LAB	*M.m.musculus*	4154
PWD132A1M	LAB	*M.m.musculus*	1784
WSB112A1F	LAB	*M.m.domesticus*	4001
WSB112A2F	LAB	*M.m.domesticus*	1412
WSB113A1F	LAB	*M.m.domesticus*	1404
WSB113A1M	LAB	*M.m.domesticus*	2007
WSB113A2F	LAB	*M.m.domesticus*	4246
WSB114A1M	LAB	*M.m.domesticus*	4689
WSB114A2M	LAB	*M.m.domesticus*	975

Supplementary Table 2. Summary of relative abundance of major bacterial phyla/genera in different groups of mice. Relative abundances in each group are presented as mean values, and p-values from ANOVA analysis between WILD and LAB mice, as well as among different groups in WILD/LAB mice are presented.

	LAB mice			WILD mice			Anova p between	Anova p among	Anova among
Main Phyla	**WSB**	**PWD**	**WP**	**Dom**	**Mus**	**Hybrid**	**LAB/WILD**	**LAB**	**WILD**
Firmicutes	0.265	0.230	0.377	0.441	0.462	0.427	**1.960E-02**	**2.480E-02**	9.210E-01
Bacteroidetes	0.194	0.263	0.185	0.291	0.372	0.371	**8.590E-06**	1.720E-01	4.220E-01
Proteobacteria	0.389	0.446	0.294	0.247	0.159	0.179	**2.790E-04**	**4.780E-02**	4.490E-01
Deferribacteres	0.042	0.133	0.119	0.014	0.002	0.016	**2.650E-13**	8.910E-02	1.540E-01
Tenericutes	0.015	0.018	0.020	0.004	0.001	0.001	**1.640E-04**	9.440E-01	3.220E-01
Main Genera									
Helicobacter	0.381	0.436	0.148	0.184	0.140	0.287	**8.770E-06**	5.060E-02	7.980E-01
Robinsoniella	0.113	0.080	0.066	0.162	0.085	0.182	**5.880E-03**	6.230E-02	1.890E-02
Lachnospiraceae	0.068	0.045	0.104	0.108	0.117	0.092	1.100E-01	2.050E-01	9.250E-01
Bacteroides	0.020	0.015	0.156	0.144	0.145	0.014	**7.090E-06**	8.350E-01	9.770E-01
Barnesiella	0.012	0.056	0.091	0.062	0.071	0.030	**3.210E-03**	2.180E-01	6.390E-01
Mucispirillum	0.133	0.042	0.016	0.014	0.002	0.118	**2.820E-13**	8.820E-02	1.540E-01
Alistipes	0.107	0.074	0.067	0.016	0.038	0.026	**1.090E-08**	1.920E-01	1.580E-01
Oscillibacter	0.004	0.006	0.039	0.036	0.035	0.011	**2.610E-08**	1.110E-01	8.200E-01
Porphyromonadaceae	0.013	0.009	0.027	0.013	0.024	0.004	**5.600E-06**	9.630E-02	9.310E-01
Ruminococcaceae	0.004	0.005	0.021	0.017	0.019	0.009	**9.310E-07**	6.150E-02	1.030E-01
Sandarakinotalea	0.014	0.032	0.000	0.001	0.000	0.034	**1.030E-10**	2.470E-01	5.710E-01
Parasporobacterium	0.014	0.014	0.011	0.012	0.022	0.014	8.120E-01	4.090E-01	2.850E-01
unclassified_ Rikenellaceae	0.000	0.010	0.018	0.014	0.034	0.006	**3.660E-02**	6.050E-01	7.850E-02
Butyrivibrio	0.007	0.003	0.022	0.009	0.016	0.006	1.020E-01	7.000E-01	4.740E-01
Prevotellaceae	0.000	0.017	0.016	0.014	0.024	0.000	**3.090E-03**	**2.610E-07**	1.190E-01
Blautia	0.004	0.003	0.017	0.006	0.017	0.007	**6.160E-03**	3.730E-01	7.320E-01
Streptococcus	0.000	0.000	0.024	0.013	0.015	0.000	**1.420E-05**	0.000E+00	3.970E-01
Sporobacterium	0.004	0.004	0.013	0.004	0.011	0.008	4.080E-01	7.810E-01	3.970E-01

Supplementary Table 3. Summary of quantitative trait loci (QTLs) identified in this study. The confidence interval for a single QTL is defined by a 1.5 drop in the LOD score in flanking SNPs. DD/DM/MM denote the average value of a specific trait in each of the three possible genotypes. If the DM genotype (heterozygotes) displays values that are significantly outside the range of both homozygotes (DD and MM), the trait is defined as a transgressive phenotype.

Trait	Chrom	qtl peak	location	lower CI	higher CI	variations	LOD	MM	DM	MM	Note
Bacteroidetes	14	rs6359032	11.33	6.92	11.86	0.039	4.35	-0.770	-0.722	-0.614	
Deferribacteres	9	rs6210093	1.61	0.98	3.59	0.037	6.57	-1.441	-0.487	-1.727	Transgressive
Proteobacteria	14	rs30213049	8.48	6.35	11.33	0.029	3.47	-0.643	-0.556	-0.724	Transgressive
Flavobacteria	1	rs30718061	2.45	1.63	3.26	0.022	56.30	-1.974	-5.111	-2.203	Transgressive
Campylobacterales	14	rs30213049	8.48	6.35	11.33	0.030	2.52	-0.664	-0.574	-0.755	Transgressive
Clostridiales	3	rs13477506	15.45	12.73	15.96	0.009	3.19	-0.481	-0.555	-0.483	Transgressive
Flavobacteriales	1	rs30718061	2.45	1.63	3.26	0.022	77.70	-1.974	-5.111	-2.203	Transgressive
Bacteroidaceae	1	rs6309584	13.30	10.68	17.41	0.096	7.72	-4.919	-3.498	-2.436	
Flavobacteriaceae	1	rs30718061	2.45	1.63	3.26	0.016	73.10	-2.048	-5.111	-2.220	Transgressive
Helicobacteraceae	14	rs30213049	8.48	6.35	11.33	0.030	3.63	-0.654	-0.413	-0.628	Transgressive
Incertae_Sedis_XIV	X	rs6174454	8.98	8.19	14.77	0.030	7.11	-3.193	-2.691	-8.000	
Peptococcaceae	9	rs6237640	9.01	6.47	12.31	0.045	3.22	-3.167	-4.294	-4.677	
Porphyromonadaceae	1	rs6309584	13.30	0.32	19.65	0.036	3.15	-2.124	-1.764	-1.510	
Bacteroides	1	rs6309584	13.30	10.68	17.41	0.096	7.72	-4.919	-3.498	-2.436	
Barnesiella	1	rs6309584	13.26	7.45	16.63	0.098	7.78	-3.411	-2.296	-1.739	
Blautia	1	rs32363391	4.93	0.35	15.96	0.015	7.97	-3.095	-8.000	-3.435	Transgressive
Butyrivibrio	9	rs6210093	1.61	0.98	3.59	0.017	23.90	-3.152	-8.000	-2.895	Transgressive
Helicobacter	14	rs30213049	8.48	6.35	11.33	0.030	3.74	-0.665	-0.574	-0.757	Transgressive
Hespellia	1	rs32363391	4.93	4.31	5.70	0.033	114.00	-2.691	-8.000	-2.536	Transgressive
Sandarakinotalea	1	rs30718061	2.45	1.63	3.26	0.014	13.40	-2.139	-5.111	-2.439	Transgressive
X168	1	rs6309584	13.30	12.25	19.65	0.093	7.33	-6.224	-4.649	-3.658	
X8	1	rs31672460	19.10	18.49	9.65	0.086	6.79	-6.699	-5.194	-4.335	
X229	1	rs32363391	4.93	1.63	5.70	0.028	12.75	-1.829	-8.000	-2.100	Transgressive
X62	X	rs6297442	8.19	6.39	8.98	0.085	20.96	-1.697	-1.584	-8.000	
X119	1	rs8256196	13.04	10.68	19.65	0.065	5.19	-7.073	-5.604	-4.924	
X119	7	rs32418253	1.23	0.36	7.15	0.035	3.20	-6.727	-5.524	-5.303	
X103	1	rs6309584	13.30	10.68	19.65	0.046	3.76	-5.355	-4.365	-3.602	
X5217	6	rs6248135	9.97	5.63	14.83	0.048	4.01	-7.078	-6.800	-5.550	
Chao1	13	rs6411888	5.18	3.43	9.13	0.038	3.16	784.000	626.000	628.000	

Supplementary Table 4. List of genes that are significantly reduced in expression among hybrid mice. Tests were performed with linear models between gene expression values and interspecific heterozygosity. Genes with p-values lower than 0.05 after FDR correction are presented. Genes marked in red denote those that also significantly correlate to overall microbial community structure (as measured by the Bray-Curtis index, Figure S2).

Mice	Genes	Linear model p-value
WILD	Bax	2.83E-02
	Bcl2	6.60E-03
	Ccr7	4.36E-02
	Cd28	6.60E-03
	Cd3e	6.60E-03
	Cd4	4.21E-02
	Cd68	7.82E-05
	Ctla4	4.21E-02
	Fn1	6.78E-03
	Hprt1	6.60E-03
	Stat4	6.78E-03
	Tbx21	4.21E-02
	Tgfb1	4.21E-02
	Vcam1	4.14E-02
LAB	Ccl19	2.54E-02
	Ccl2	3.85E-02
	Ccr2	1.04E-02
	Ccr4	1.57E-02
	Ccr7	1.07E-02
	Cd40	2.75E-02
	Cd68	2.49E-02
	Cd86	2.08E-02
	Gapdh	4.19E-02
	H2.Ea	3.02E-04
	Icos	6.09E-03
	Il15	6.85E-05
	Il18	2.81E-02
	Il2ra	2.89E-02
	Il7	1.73E-02
	Nfkb1	8.70E-04
	Pgk1	2.78E-03
	Ptprc	3.68E-02
	Smad7	4.18E-05
	Socs1	6.09E-04
	Socs2	6.83E-04
	Stat1	1.93E-04
	Stat3	1.85E-02
	Tgfb1	5.39E-03
	Tnfrsf18	4.49E-02
	Vcam1	2.28E-02
	Vegfa	1.70E-05

Supplementary Table 5. Summary of bacterial genera with significant correlations to immune gene expression levels. Only genes with significant p-values (<0.05 after FDR correction) are presented.

	Bacteria genus	Gene	Linear model p-value
LAB	*Barnesiella*	Cxcl10	2.86E-04
		Cxcl11	2.86E-04
		Nfkb2	1.95E-02
		Tnf	5.76E-07
	Bacteroides	Cd28	3.36E-02
		Cd86	3.60E-02
		Cxcl10	1.12E-06
		Cxcl11	1.77E-05
		Stat4	3.39E-02
		Tnf	1.97E-02
WILD	*Helicobacter*	Cd38	2.93E-02
		Gusb	2.93E-02
		Pgk1	4.38E-02
		Tfrc	2.93E-02

Supplementary Table 6. Summary of abundance of major bacterial phyla/genera with regard to pathological states of the cecum. Phyla and genera that showed similar trend in WILD and LAB mice are marked in bold.

		WILD			LAB		
		Normal tissues	**Pathological tissues**	**Wilcox p-value**	**Normal tissues**	**Pathological tissues**	**Wilcox p-value**
Phyla	Bacteroidetes	3.685E-01	4.253E-01	3.4650E-01	2.087E-01	5.833E-02	9.1860E-03
	Deferribacteres	**1.464E-02**	**2.350E-02**	**2.8420E-01**	**1.095E-01**	**2.090E-01**	**1.9280E-01**
	Firmicutes	4.051E-01	3.698E-01	4.4770E-01	3.211E-01	3.490E-01	3.9480E-01
	Proteobacteria	**2.075E-01**	**1.530E-01**	**4.4570E-01**	**3.399E-01**	**3.283E-01**	**4.6450E-01**
	Tenericutes	8.571E-04	1.525E-02	4.8090E-01	1.619E-02	5.500E-02	1.7830E-02
Genera	*Alistipes*	1.971E-02	6.450E-02	2.4110E-01	7.728E-02	1.967E-02	2.9310E-02
	Bacteroides	1.550E-01	2.795E-01	3.9240E-01	1.663E-02	2.667E-03	1.1570E-01
	Barnesiella	**8.812E-02**	**1.725E-02**	**3.7980E-02**	**3.614E-02**	**7.333E-03**	**8.0120E-02**
	Helicobacter	1.835E-01	1.433E-01	4.8440E-01	3.315E-01	3.193E-01	4.4690E-01
	Parasporobacterium	1.036E-02	1.850E-02	1.3600E-01	1.465E-02	2.000E-03	2.3670E-02
	Mucisprillum	**1.464E-02**	**2.325E-02**	**2.8430E-01**	**1.093E-01**	**2.087E-01**	**1.9280E-01**
	Sandarakinotalea	4.524E-04	0.000E+00	4.0850E-01	3.009E-02	1.800E-02	3.0420E-01
	Robinsoniella	1.050E-01	6.125E-02	2.7920E-01	1.433E-01	2.830E-01	8.7340E-02
	Lachnospiraceae	**1.047E-01**	**7.375E-02**	**3.9250E-01**	**7.670E-02**	**2.633E-02**	**4.1340E-02**
	Porphyromona-daceae	**2.440E-02**	**4.000E-03**	**8.6650E-03**	**6.256E-03**	**2.333E-03**	**1.3590E-01**

Chapter IV Bacterial genome sequencing of *Staphylococcus epidermidis*

Introduction

Since the first full genome sequencing of *Haemophilus influenzae* (Fleischmann *et al.,* 1995), the number of available genomes has increased exponentially in the public database. Similar to other studies in microbiology, disease-related bacterial genomes predominate the organisms with sequenced genomes. Consequently, comparative genomics lead largely to understandings of molecular mechanisms of bacterial infection and diseases, as well as the development of important traits such as antibiotic resistance. With the development of next-generation-sequencing techniques, the number of genomes from bacteria with probiotic- (e.g. *Barnesiella spp*.) (Ubeda *et al.,* 2013) or immunomodulatory characteristics (e.g. SFB bacteria and Th17 cells, Prakash *et al.,* 2011) has also increased.

In a recent study carried out jointly between our group and researchers at Luebeck University, we re-mapped the genetic segments involved in the development of an auto-immune skin disease, epidermolysis bullosa acquisita (EBA), with an inter-cross of four mice strains (provided as supplementary material in this chapter, I contributed to both microbiome analysis and QTL mapping, Srinivas *et al.,* 2013). Different from previous mapping however, we also profiled the skin microbiome in all individuals in the mapping population and used community members as co-variates in the mapping. One bacterial genus, *Staphylococcus*, is shown to influence the gene-disease relationship and potentially probiotic/anti-inflammatory activity, as higher abundance of *Staphylococcus* could prevent disease development despite the disease alleles being present.

To understand the genomic basis of probiotic function of *Staphylococcus* as well as possible host-adaption / co-evolution of this genus and its members, I cultivated skin swabs from five different species of rodent hosts (three from *Mus musculus*, of which two are in the EBA cross; *Mus speciligus* and *Apodemus uralensis* as an outgroup) on Tryptocase Soy Agar (TSA). After 16S rRNA sequencing of the selected colonies, we

found the majority of the *Staphylococcus* inhabiting mice belong to *Staphylococcus epidermidis*. We then performed genome sequencing of representative strains and report the first overview of genome sequencing, assembly and annotation to serve as a basis for future in-depth studies.

Publication

Jun Wang, Sven Kuenzel, John F. Baines, 2014**, Draft genomes of 11 *Staphylococcus epidermidis* strains isolated from wild mouse species,** Genome Announcement, 2(1).

genomeA
Journals.ASM.org

Draft Genome Sequences of 11 *Staphylococcus epidermidis* Strains Isolated from Wild Mouse Species

Jun Wang,[a,b] Sven Kuenzel,[a] John F. Baines[a,b]

Max Planck Institute for Evolutionary Biology, Plön, Germany[a]; Institute for Experimental Medicine, Christian-Albrechts-University (CAU) of Kiel, Kiel, Germany[b]

We report here the draft genome sequences of 11 strains of *Staphylococcus epidermidis*, a common bacterium inhabiting the skin of humans and other animals. These isolates, obtained from five mouse species, provide valuable information on the native *Staphylococcus* spp. of this important model organism and form a basis for studying host-bacterial interactions in their natural environment.

Received 2 December 2013 **Accepted** 10 December 2013 **Published** 16 January 2014

Citation Wang J, Kuenzel S, Baines JF. 2014. Draft genome sequences of 11 *Staphylococcus epidermidis* strains isolated from wild mouse species. Genome Announc. 2(1): e01148-13. doi:10.1128/genomeA.01148-13.

Address correspondence to Jun Wang, wang@evolbio.mpg.de.

The genus *Staphylococcus* contains important members of the human skin microbiome (1). The major members are commensal under normal circumstances but can also be pathogenic. *Staphylococcus aureus* is so far the main species of interest, as it is a major source of nosocomial infections and can afflict numerous organs (2). Other examples include *Staphylococcus haemolyticus* (causing infective endocarditis [3]) and *Staphylococcus saprophyticus* (causing urinary tract infections [4]). *Staphylococcus epidermidis*, on the other hand, is of critical importance, as it is the most common source of medical device-associated infections (5), but at the same time, it is capable of inhibiting *S. aureus* colonization in human nasal cavities (6). Given this important clinical relevance, infection models are established in mice but are so far limited to human *S. aureus* isolates (7). Understanding the interaction and coevolutionary history between mice and their native bacterial species has attracted recent attention (8), although native *Staphylococcus* strains remain unexplored. Furthermore, we recently discovered *Staphylococcus* to contain important members of the native mouse skin microbiota influencing susceptibility to autoimmune skin blistering (9).

In order to provide insight into the native species of *Staphylococcus* inhabiting mice, we isolated *Staphylococcus* spp. from 11 wild mice representing five species and subspecies (*Mus musculus musculus, M. musculus domesticus, M. musculus castaneus, M. musculus spicilegus*, *and Apodemus uralensis*), which were captured from the wild and maintained in conventional animal facilities at the Max Planck Institute for Evolutionary Biology, Ploen, Germany. The majority of isolates belonged to *S. epidermidis*, and we subsequently selected 11 strains for genome sequencing. The sequencing libraries were prepared using the Illumina Nextera XT kit and run on the MiSeq platform with paired-end reads of 250 bp, with a minimum coverage of 31× and a maximum of 61×. The reads were assembled *de novo* using Velvet (10) with parameters optimized by VelvetOptimiser (http://www.vicbioinformatics.com/software.velvetoptimiser.shtml). The contigs were annotated by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) version 2.0 (11).

For the 11 strains, we obtained a minimum of 91 and maximum of 277 contigs, and the total number of assembled nucleotides ranged from 2,458,755 to 2,762,809 per strain. The average G+C contents ranged from 31.7% to 32.0%, which is close to those of the available reference strains (32.1% for *S. epidermidis* ATCC 12228 and 32.2% for *S. epidermidis* RP62A). A range of 2,259 to 2,541 proteins were predicted and annotated using the NCBI PGAP, with 83.6% to 92.2% of the proteins having homologs in the *S. epidermidis* ATCC 12228 (NCBI accession no. NC_004461) (12) and/or RP62A genomes (NCBI accession no. NC_002976) (13) (BLASTp [14] with an *E* value of 1E-20 and similarity threshold of 0.8). Thirteen to 70 tRNA genes and 3 to 24 rRNA genes are predicted for each strain. Further analyses of the genomic content may reveal important aspects of the interaction and coevolution of *S. epidermidis* and mouse hosts.

Nucleotide sequence accession numbers. The draft genome sequences are deposited in GenBank under accession no. ATCU00000000, ATCV00000000, ATCW00000000, ATCX00000000, ATCY00000000, ATDA00000000, ATDC00000000, ATDE00000000, ATDF00000000, ATDG00000000, and ATDH00000000. The second versions are described in this paper: ATCU02000000, ATCV02000000, ATCW02000000, ATCX02000000, ATCY02000000, ATDA02000000, ATDC02000000, ATDE02000000, ATDF02000000, ATDG02000000, and ATDH02000000.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Excellence Cluster 306 "Inflammation at Interfaces" and the Max Planck Society.

We thank the Department of Evolutionary Genetics of the Max Planck Institute for Evolutionary Biology for mouse stock resources and Miriam Linnenbrink for assistance with sampling.

REFERENCES

1. **Otto M.** 2010. *Staphylococcus* colonization of the skin and antimicrobial peptides. Expert Rev. Dermatol. 5:183–195. http://dx.doi.org/10.1586/edm.10.6.

2. **Lowy FD.** 1998. *Staphylococcus aureus* infections. N. Engl. J. Med. **339:** 520–532. http://dx.doi.org/10.1056/NEJM199808203390806.
3. **Falcone M, Campanile F, Giannella M, Borbone S, Stefani S, Venditti M.** 2007. *Staphylococcus haemolyticus* endocarditis: clinical and microbiologic analysis of 4 cases. Diagn. Microbiol. Infect. Dis. **57:**325–331. http://dx.doi.org/10.1016/j.diagmicrobio.2006.08.019.
4. **Ishihara S, Yokoi S, Ito M, Kobayashi S, Deguchi T.** 2001. Pathologic significance of *Staphylococcus saprophyticus* in complicated urinary tract infections. Urology **57:**17–20. http://dx.doi.org/10.1016/S0090-4295(00)00883-9.
5. **Kozitskaya S, Cho SH, Dietrich K, Marre R, Naber K, Ziebuhr W.** 2004. The bacterial insertion sequence element IS256 occurs preferentially in nosocomial *Staphylococcus epidermidis* isolates: association with biofilm formation and resistance to aminoglycosides. Infect. Immun. **72:** 1210–1215. http://dx.doi.org/10.1128/IAI.72.2.1210-1215.2004.
6. **Iwase T, Uehara Y, Shinji H, Tajima A, Seo H, Takada K, Agata T, Mizunoe Y.** 2010. *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. Nature **465:**346–349. http://dx.doi.org/10.1038/nature09074.
7. **Prabhakara R, Harro JM, Leid JG, Harris M, Shirtliff ME.** 2011. Murine immune response to a chronic *Staphylococcus aureus* biofilm infection. Infect. Immun. **79:**1789–1796. http://dx.doi.org/10.1128/IAI.01386-10.
8. **Colton L, Kosoy M.** 2013. Experimental infection of laboratory mice with two *Bartonella tribocorum* strains from wild *Mus* species: a homologous host-bacteria model system at the genus level. Parasitology **140:**61–68. http://dx.doi.org/10.1017/S0031182012001333.
9. **Srinivas G, Möller S, Wang J, Künzel S, Zillikens D, Baines JF, Ibrahim SM.** 2013. Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. Nat. Commun. **4:**2462. http://dx.doi.org/10.1038/ncomms3462.
10. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
11. **Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O.** 2008. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. Omics **12:** 137–141. http://dx.doi.org/10.1089/omi.2008.0017.
12. **Zhang YQ, Ren SX, Li HL, Wang YX, Fu G, Yang J, Qin ZQ, Miao YG, Wang WY, Chen RS, Shen Y, Chen Z, Yuan ZH, Zhao GP, Qu D, Danchin A, Wen YM.** 2003. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). Mol. Microbiol. **49:**1577–1593. http://dx.doi.org/10.1046/j.1365-2958.2003.03671.x.
13. **Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance IR, Nelson KE, Fraser CM.** 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J. Bacteriol. **187:**2426–2438. http://dx.doi.org/10.1128/JB.187.7.2426-2438.2005.
14. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Supplementary Materials

nature COMMUNICATIONS

ARTICLE

Received 15 May 2013 | Accepted 19 Aug 2013 | Published 17 Sep 2013

DOI: 10.1038/ncomms3462 OPEN

Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering

Girish Srinivas[1,2], Steffen Möller[2], Jun Wang[1,3], Sven Künzel[1], Detlef Zillikens[2], John F. Baines[1,3,*] & Saleh M. Ibrahim[2,*]

Susceptibility to chronic inflammatory diseases is determined by immunogenetic and environmental risk factors. Resident microbial communities often differ between healthy and diseased states, but whether these differences are of primary aetiological importance or secondary to the altered inflammatory environment remains largely unknown. Here we provide evidence for host gene–microbiota interactions contributing to disease risk in a mouse model of epidermolysis bullosa acquisita, an autoantibody-induced inflammatory skin disease. Using an advanced intercross, we identify genetic loci contributing to skin microbiota variability, susceptibility to skin blistering and their overlap. Furthermore, by treating bacterial species abundances as covariates with disease we reveal a novel disease locus. The majority of the identified covariate taxa are characterized by reduced abundance being associated with increased disease risk, providing evidence of a primary role in protection from disease. Further characterization of these putative probiotic species or species assemblages offers promising potential for preventative and therapeutic treatment development.

[1] Max Planck Institute for Evolutionary Biology, August-Thienemann-Street 2, D-24306 Plön, Germany. [2] Department of Dermatology, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany. [3] Institute for Experimental Medicine, Christian-Albrechts-University of Kiel, Arnold-Heller-Street 3, D-24105 Kiel, Germany. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.B. (email: baines@evolbio.mpg.de).

Diverse communities of symbiotic bacteria inhabit nearly every surface of the body exposed to the environment. The skin in particular is in constant contact with the environment and serves a critical barrier function, yet provides a range of niches to inhabiting microbial communities. A multitude of interactions between the skin microbiota, host and environment contribute to community structure over space and time and its potential contribution to changes in health status[1,2]. Recent landmark studies of the mouse gut microbiota using a quantitative trait locus (QTL) mapping approach unequivocally demonstrate the role of host genetics in shaping diversity between individuals[3,4]. Likewise, inflammatory disorders afflicting the skin such as psoriasis and atopic dermatitis harbour clear immunogenetic components, but whether these associations may be mediated by alterations in microbial community structure is unknown[1,5].

Epidermolysis bullosa acquisita (EBA) is a chronic skin blistering disease of autoimmune origin characterized by antibodies to type VII collagen (COL7)[6]. We previously demonstrated the contribution of MHC haplotype and other non-MHC genes to EBA susceptibility in an immunization-induced model of EBA in mice[7]. Intriguingly, autoimmunity against COL7 is also a common observation among patients with Crohn disease[8], one of two major forms of inflammatory bowel disease with clear host genetic and microbial components[9]. In this study, we investigate the contribution of host genetic control of the skin microbiota in mice from a four-way autoimmune-prone advanced intercross, enabling loci influencing microbial community structure and disease (EBA) susceptibility to be simultaneously analysed. In addition to identifying host genetic loci that contribute to variability in bacterial taxon abundances in the skin, we find that individual genotype-dependent microbial risk factors modify susceptibility to EBA and increase the power to detect disease-associated loci.

Results

Composition and diversity of skin microbiota. To measure the contribution of host genetics to variation in the mouse skin microbiota, we first analysed 261 individuals from the fourth generation of an advanced intercross lines (183 immunized individuals, of which 64 developed EBA, plus 78 non-immunized controls, Methods) using pyrosequencing of the 16S rRNA gene. At the phyla level, the Firmicutes are most abundant (54%), followed by Proteobacteria (21%), Actinobacteria (12%) and Bacteroidetes (6%) (Supplementary Fig. S1), revealing communities similar to those observed in previous studies of the skin[2,10–12]. At the genus level, *Staphylococcus* (36%), *Corynebacterium* (9%) and *Ralstonia* (8%) are most abundant (Supplementary Fig. S2). To characterize the level and pattern of diversity within individuals, we applied different measures of alpha diversity, which focus on species richness, evenness and abundance. The Chao1 species richness index is higher in the healthy individuals compared with those afflicted with EBA (Fig. 1) (Wilcoxon rank-sum test, $W=170$, $P=0.005$). The same pattern is also observed for Faith's[13] phylogenetic diversity index (PD whole tree) (Wilcoxon rank-sum test, $W=176$, $P=0.008$), the observed number of species (Wilcoxon rank-sum test, $W=212$, $P=0.05$) and the Shannon evenness measure (Wilcoxon signed-rank Test, $Z=-4.3726$, $P<0.001$) (Supplementary Fig. S3). To analyse bacterial community composition and structure *between* individuals (that is, beta diversity,) we first used the weighted and unweighted Unifrac metric[14,15], which is a phylogenetic-based measure weighted by taxon abundance and based on presence–absence information, respectively. Constrained analysis of principal coordinates (CAP) using EBA status as an explanatory variable and the weighted Unifrac metric as a response variable also reveals a small, but significant effect ($P=0.015$), with the first principal component axis (or CAP1 axis) explaining 1.8% of the variation between individuals (Supplementary Fig. S4a), while the CAP1 axis for unweighted Unifrac explains 2% of the variation ($P=0.005$; Supplementary Fig. S4b). Analysis of beta diversity using OTU-based approaches yields very similar results (Bray–Curtis index, with CAP1 explaining 1.8% of the variation ($P=0.015$) (Supplementary Fig. S4c) and Jaccard index, with CAP1 explaining 1% of the variation ($P=0.001$) (Fig. 2)). For further analysis a 'core measurable microbiota' (CMM) of 131 OTUs was defined[3], which

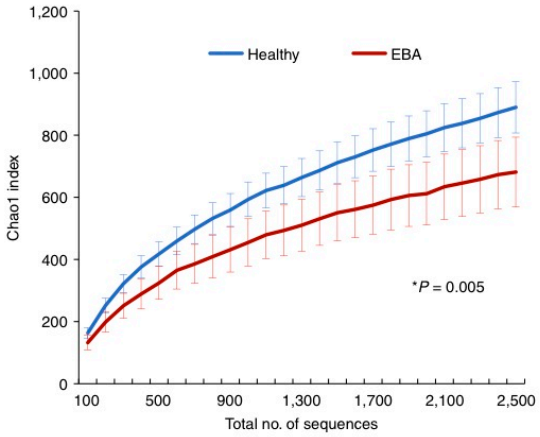


Figure 1 | Chao1 species richness. The Chao1 index based on species-level OTUs was estimated for immunized healthy ($n=119$) and EBA ($n=64$) samples. Error bars represent the 95% confidence interval. *Significance was determined by the Wilcoxon rank-sum test.

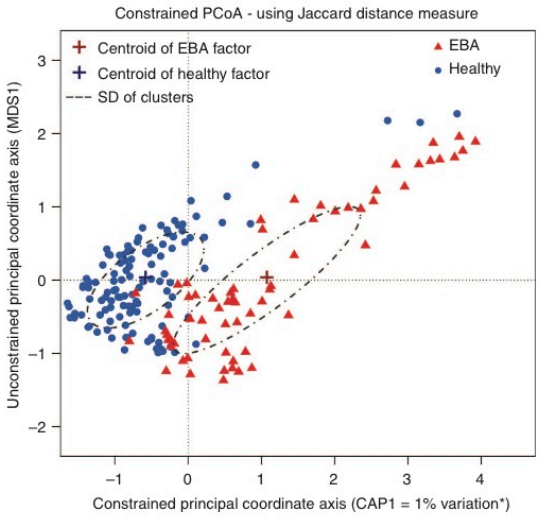


Figure 2 | Constrained analysis of principal coordinates of the Jaccard index. The Jaccard index was calculated for immunized healthy ($n=119$) and EBA ($n=64$) samples. The disease status was taken as the constrained factor and differed significantly by permutation test ($P=0.005$, see Methods).

contain nearly 80% of the total sequences in the data set (Supplementary Figs S5 and S6) (Methods).

QTL analysis of skin microbiota. To account for both environmental and genetic contributions to variation in CMM OTU abundances, a linear mixed model analysis was performed including cage, family and sex as factors (see Methods). This reveals significant cage and family effects accounting for 28% and 3% of the variation in CMM species abundances, respectively. No significant influence of sex is observed for any of the CMM taxa and was thus removed from the model. To measure the genetic contribution, CMM abundances were tested for co-segregation against 1,199 informative single-nucleotide polymorphism (SNP) markers after accounting for cage and family effects. This reveals host genetics to have significant control over members of the skin microbiota, which can be seen in Fig. 3. Nine out of 131 CMM OTUs are associated with three significant (*E*-value cutoff <0.05) and six suggestive (*E*-value cutoff <0.1) (see Methods) species-level OTU QTLs, hereafter referred to as 'spQTLs' (Supplementary Data 1). Mapping at higher taxonomic levels including phylum, class, order, family and genus reveals a total of six QTLs, including three with one or more significant associations and three with suggestive associations, hereafter referred to as 'gpQTLs' (Supplementary Data 2). Two out of the nine spQTLs are contained within gpQTLs, thus, in total thirteen unique QTLs are identified. To gain further insight we compared our results with previously published QTL studies of the gut microbiota[3,4] and reveal evidence of overlap greater than expected by chance (Fig. 3; Methods). Interestingly, the confidence intervals of our spQTLs and gpQTLs contain five and four genes related to innate immunity, respectively (see Discussion, Supplementary Data 3).

Effect of immunization and disease status on QTL mapping. As the model of EBA used in this study is immunization-based, we also included 78 non-immunized mice to control for the effect of immunization in the QTL mapping. Accordingly, we analysed a subset where both EBA-afflicted and non-immunized individuals are removed (that is, only the 119 healthy, autoimmunized samples). Despite decreasing the sample size from 261 to 119, two out of nine spQTLs and two out of six gpQTLs are still detected (Supplementary Data 1 and 2). Next, we analysed a subset where the EBA-afflicted mice are removed (that is, including 119 healthy autoimmunized and 78 non-immunized samples). One out of nine spQTLs and two out of six gpQTLs are still detectable despite lowering the sample size from 261 to 197 (Supplementary Data 1 and 2). Thus, the presence of QTLs among subsamples not influenced by differences in disease/autoimmunization status supports the presence of true genetic effects.

Gene–microbe covariation in disease susceptibility. To investigate the potential role of host genetic variation for skin bacterial

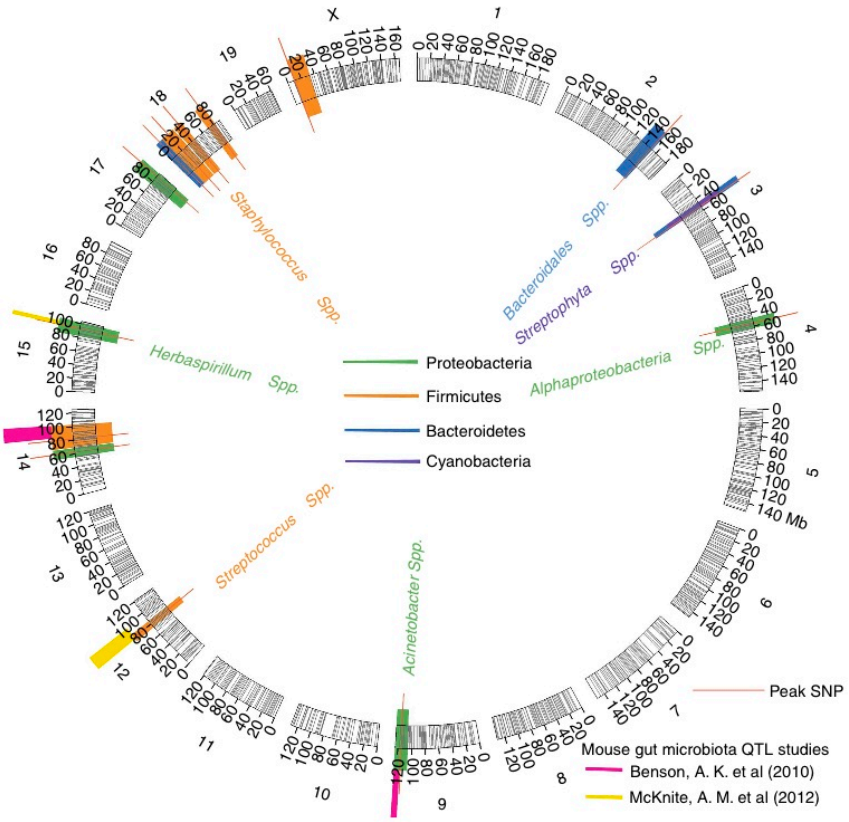


Figure 3 | QTL mapping of skin microbiota. Nineteen mouse autosomes and the X chromosome are depicted with 1,199 SNPs indicated by black lines. QTLs for species level OTUs (spQTLs) are colour coded according to their phylum along with their highest reliable taxonomic classification in adjacent text. QTLs from the genus to phylum level (gpQTLs) are colour coded according to their phylum classification (see Supplementary Data 1 and 2).

abundances in disease, we first re-analysed the subset of 183 immunized mice common to this and our previous study on EBA[7]. This reveals no significant QTL for EBA presence/absence at an *E*-value[16] cutoff <0.1 (see Methods), likely due to the reduced number of animals (Fig. 4a). However, because bacterial taxon abundance does display a clear genetic component, we next sought to evaluate potential interactions between bacterial species and disease susceptibility by applying a covariate analysis between each of the 131 CMM species abundances and EBA disease susceptibility. This reveals significant covariation (*E*-value <0.1) involving 10 out of 131 taxa, which, intriguingly, increases the power of detecting EBA QTLs, as a novel locus (covariate QTL) (Chr.19, CI 53–60, peak at 57 Mb) is detected (Fig. 4b, Supplementary Data 4). Two OTUs belonging to the genus *Staphylococcus* clearly display a gene–bacterial interaction (*E*-value <0.05) (Fig. 4c).

To further characterize the nature of the identified covariate QTL, we arbitrarily divided individuals into 'high' (top 50%) and 'low' (bottom 50%) groups with respect to their individual OTU abundances and analysed the proportion of individuals developing EBA with respect to host genotype. This reveals that for most cases the proportion of animals developing EBA is higher among the low OTU abundance category ($n=10$; one of which is also significant by Fisher's exact test between these defined abundance categories (Fig. 4d); we note, however, that all 10 taxa display significant covariation), suggesting a probiotic role (Supplementary Data 5). Although community-level alterations of the skin microbiota in the context of EBA are present (for example, Supplementary Figs S3 and S4), we note that the putative probiotic covariate taxa identified do not vary in abundance simply according to disease status. Namely, both healthy and diseased individuals are found among the low abundance categories, thus, low abundance of, for example, *Staphylococcus spp.* is not a simple byproduct of disease, but increases the probability of developing symptoms.

The large number of covariate bacterial taxa interacting with a single host locus suggests that individual bacterial taxa may not be acting independently. Thus, to identify potential interactions among covariate taxa we performed a pairwise correlation analysis (Fig. 5). Indeed, this reveals significant positive correlations (Pearson's correlation; *P*-value ≤ 0.05 after correction for multiple testing (Benjamini–Hochberg[17])) between many taxa, suggesting interactions between the host locus and bacterial species assemblages or individual driver species.

Discussion

Our results provide strong evidence that host genetic variation contributes to differences in the bacterial communities observed in the skin. In a previous QTL analysis of the mouse faecal community, Benson *et al.*[3] reported 13 significant and 5 additional suggestive QTLs for 26 out of 64 taxonomic groups

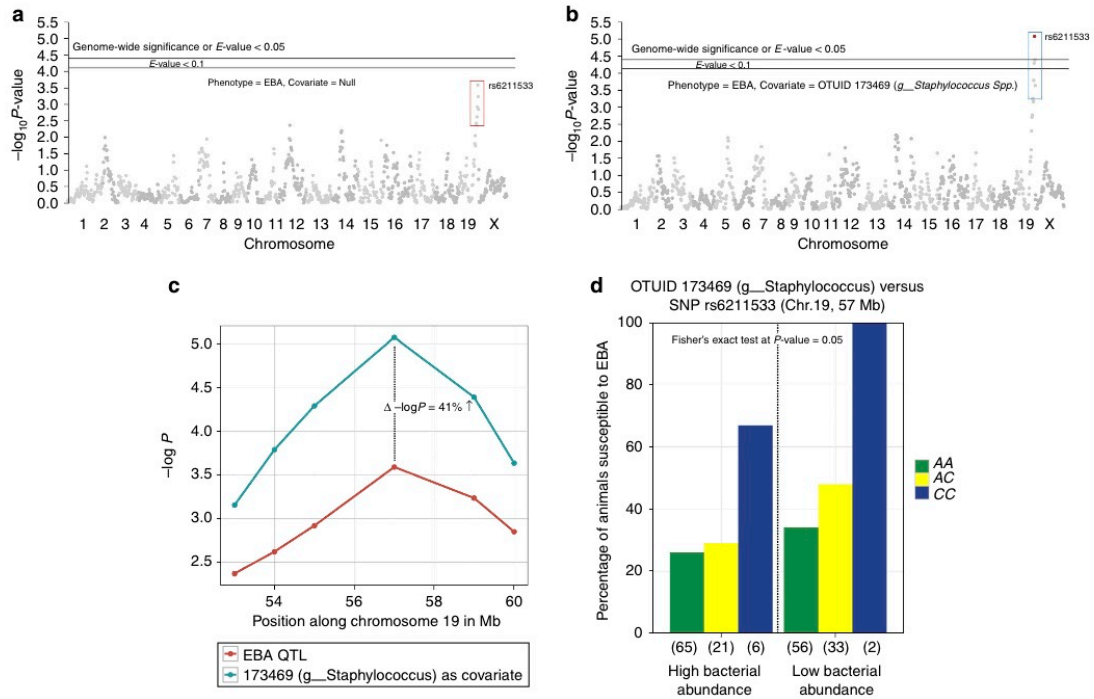


Figure 4 | Gene–microbe interaction in EBA susceptibility. (**a**) Manhattan plot of $-\log_{10}$ *P*-values for each SNP (1,199 SNPs in *x* axis according to their position on each chromosome) tested against EBA disease phenotype (presence/absence). (**b**) Manhattan plot showing $-\log_{10}$ *P*-values for each SNP tested against EBA including *Staphylococcus spp.* (OTUID 173469) abundance as a covariate. SNPs falling below an *E*-value of 0.05 are shown in red. (**c**) Portion of chromosome 19 containing the covariate QTL with a peak at SNP rs6211533. (**d**) Percentage of animals developing EBA among high (top 50%) and low (bottom 50%) *Staphylococcus spp.* (OTUID 173469) abundance categories with respect to host genotype at rs6211533 (represented by green (*AA*), yellow (*AC*) and blue (*CC*)). Numbers in parentheses indicate the sample size within each genotype category.

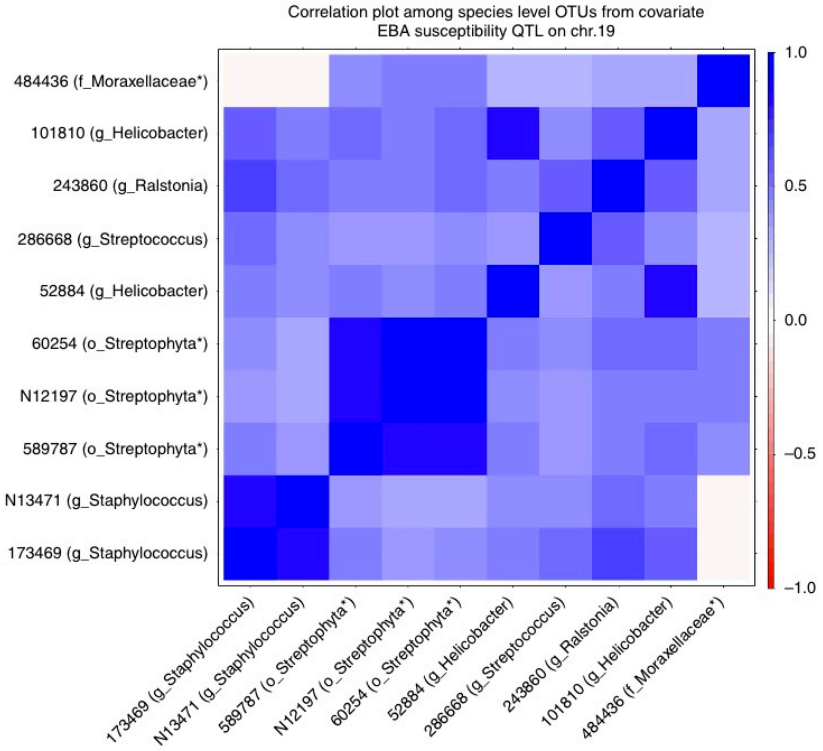


Figure 5 | Correlation matrix of covariate OTUs. Pearson's correlation values between OTUs that significantly co-vary with an EBA susceptibility locus on chromosome 19 are displayed. For OTUs not classified at the genus level, the next highest taxonomic level for which classification was possible is displayed by an asterisk. The taxonomic level of classification is indicated by k, p, c, o, f and g for kingdom, phylum, class, order, family and genus, respectively. Only values significantly differing from zero after correction for multiple testing[17] are shown by either blue (positive correlation) or red (negative correlation) squares.

tested. However, their analysis was not extended beyond the level of bacterial genera. Despite our more inclusive set of phenotypic traits extending to the bacterial species (OTU) level, we detected a smaller number of loci, with nine significant QTLs for 9 out of 131 species-level traits. Differences in sample size, sequencing coverage, mouse strains used and the obvious distinctions between the gut and skin environments may all contribute to these differences in QTL detection. Interestingly, we nonetheless find evidence of overlap between the two studies. The confidence intervals of two out of 18 QTLs controlling bacterial abundance in murine faeces contain the peak SNP of a skin QTL, which overlaps more than expected by chance (P-value <0.05; Fig. 3; Methods). One of these QTLs is consistent at the phyla level (Firmicutes, Chromosome 14), while the other is at the order level (*Pseudomonadales*, Chromosome 9). Similarly, an additional two of our skin spQTLs (OTU N31208 belonging to *Streptococcus* on Chromosome 12 and OTU 130241 belonging to *Herbaspirillum* on Chromosome 15) overlap with faecal QTLs from another more recently published study[4], although the taxonomic assignments do not agree at even the phylum level.

The confidence intervals of our skin microbiota QTLs contain nine genes known to be involved in the functioning of the innate immune system (Supplementary Data 3). *Interleukin-1 receptor-associated kinase (IRAK)-4* is an interesting candidate found within the confidence interval of spQTL6, which modulates an OTU (ID 130241) belonging to the genus *Herbaspirillum*. Deficiencies of this gene in humans lead to increased susceptibility to pyogenic bacterial infections including *Staphylococcus aureus*[18], and its interaction with the MYD88 adapter protein is used by several Toll-like receptor pathways in host defence[19], as well as being involved in controlling commensal bacteria[20]. Another gene coding for CD14 antigen is found within spQTL8 on chromosome 18, which modulates an OTU (ID N10459) belonging to the genus *Staphylococcus*. Increasing CD14 expression enhances Toll-like receptor 2 activation in skin in the presence of vitamin D_3—1,25-dihydroxyvitamin D_3 (1,25D3)[21], which in turn influences the skins sensitivity to microbial challenge. Furthermore, several studies have shown that components of *S. aureus* (LTA and peptidoglycan) interact with the CD14 molecule[22–24]. Finally, by treating bacterial abundances as covariates with the presence/absence of EBA, we identified an additional significant EBA QTL on chromosome 19 (Fig. 4). One potential candidate gene lying within this chromosomal interval (53–60 Mb) is *caspase-7* (*casp7*), a member of the cytosolic cysteine protease family known to be involved with inflammatory disorders[25,26] and defence against pathogens[27].

Similar to previous studies of chronic inflammatory skin diseases, our findings support a role of resident microbial communities in disease pathogenesis. The differences in community composition and structure between mice with and without EBA symptoms are akin to shifts in the skin microbiota associated with atopic dermatitis disease flares and treatment[28] or between psoriatic lesions and both unaffected skin in patients and healthy controls[29]. Although other examples such as

endemic pemphigus foliaceus (fogo selvagum), where exposure to haematophagous insects is implicated[30], suggest direct environmental/microbial triggers of disease, the roughly three-fold increase over the last 30 years of atopic dermatitis in industrialized countries suggests more complex environmental influences, possibly mediated by changes in microbial communities. By investigating disease provocation in a large mouse mapping population under controlled environmental conditions, we were able to identify individual, genotype-dependent microbial risk factors among a core set of taxa inhabiting the skin of both healthy and diseased mice, more closely resembling a disease-modifying effect. Although further validation and characterization of these interactions awaits more intensive experimental interrogation in, for example, gnotobiotic animals, our investigation of another EBA-susceptible mouse strain (SJL/J; not included in this study) before and after autoimmunization reveals that other aspects of the skin community, in particular alpha diversity (species richness and evenness), are predictive of disease outcome (unpublished results). Thus, the further identification and functional analysis of host genetic and probiotic bacterial factors represent promising avenues for research in preventative and therapeutic treatment development.

Methods

Generation of a four-way advanced intercross line. Parental mouse strains (MRL/MpJ, NZM2410/J, BXD2/TyJ, Cast) for generating a heterogeneous intercross line[31] were purchased from the Jackson laboratory (Maine, USA). Briefly, strains were intercrossed at an equal strain and sex distribution. First generation (G1) offspring mice were then mated considering their parental origin to maintain an equal distribution of parental alleles for successive generations by maintaining at least 50 breeding pairs per generation. Male and female offspring used in the study were transferred to separate cages according to sex after weaning. Depending on the number of animals per cage, females from multiple families were also housed together. Animals were held under specific pathogen-free conditions at a 12-h light/dark cycle with food and water *ad libitum*. All 261 animals (135 males and 126 females) in this study were taken from the fourth generation of this advanced intercross line at 6 months of age. All animal experiments were approved by the state of Mecklenburg-Vorpommern, Germany.

Induction of experimental EBA. EBA was induced by immunization with an immunodominant peptide within the murine NC1 domain of type VII collagen (GST-mCOL7C)[7]. In brief, 60 μg GST-mCOL7C emulsified in 60 μl adjuvant (TiterMax, Alexix, Lörrach, Germany) was injected subcutaneously into the foot pad and tail base. After immunization mice were screened for skin inflammation every 4th week for a period of 12 weeks, after which the ears were taken for analysis at 6 months of age. Ear skin samples were fixed in 4% buffered formalin and snap frozen at −80 °C. A total of 183 immunized- and 78 non-immunized mice were included in skin microbiota QTL mapping.

DNA extraction and 16S rRNA gene pyrosequencing. Bacterial DNA from mouse ears was extracted using the PowerSoil Kit (MoBio, Carlsbad, CA). During the killing of the mice, both ears were taken for either bacterial DNA extraction or other analyses (that is, histology, RNA), and the ear chosen for microbial analysis was chosen at random. Approximately one third of an ear was transferred to the Power Bead tubes containing 60 μl of C1 solution and 20 μl of 20 mg ml^{-1} Proteinase K. Samples were incubated at 50 °C for 2 h at 850 r.p.m. and the remaining steps were performed according to the manufacturer. Amplification of the hypervariable V1 and V2 (27F–338R) region of the 16S rRNA gene was performed using composite forward (5′-*CTATGCGCCTTGCCAGCCCGC*TCAGTCAGA GTTTGATCCTGGCTCAG-3′) and reverse (5′-*CGTATCGCCTCCCTCGCGCCA* TCAGXXXXXXXXXXCATGCTGCCTCCCGTAGGAGT-3′) primers. These primers include the 454 Life Sciences adaptor A (reverse) and B (forward), denoted by italics. The underlined sequences represent the broadly conserved bacterial primers 27F and 338R. Ten base pair multiplex identifiers (MIDs; designated as "XXXXXXXXXX") were added to reverse primers to uniquely tag PCR products. Duplicate 25 μl PCR reactions, each containing 100 ng of template DNA, were performed using Phusion Hot Start DNA Polymerase (Finnzymes, Espoo, Finland) with the following cycling conditions: initial denaturation for 30 s at 98 °C; 35 cycles of 9 s at 98 °C, 30 s at 55 °C and 30 s at 72 °C; final extension for 10 min at 72 °C. Duplicate reactions were combined after PCR and products were extracted with the MiniElute Gel Extraction Kit (Qiagen, Hilden, Germany). Quantification was performed with the Quant-iT dsDNA BR Assay Kit on a Qubit fluorometer (Invitrogen, Darmstadt, Germany). Purified PCR products were pooled in equimolar amounts and further purified using Agencourt Ampure Beads (Beckman Coulter, Krefeld, Germany). Aliquots of each library were run on an Agilent Bioanalyser before emulsion PCR and sequencing according to the manufacturer's instructions on a Roche 454 GS-FLX using Titanium sequencing chemistry.

454 Pyrosequencing data analysis. A Perl script using the Smith–Waterman algorithm[32] was written to match the forward primer and barcode allowing no insertions or deletions. Sequences were required to have a length between 290 to 370 nucleotides, an average quality score ≥20 and contain no ambiguous bases. ChimeraSlayer (http://microbiomeutil.sourceforge.net/) was used to remove chimeras. An average of 5,732 reads per sample was obtained for 261 animals.

Taxonomic classification. RDP classifier[33] (RDP Multi-Classifier version 1.0) was applied to assign taxonomy to the genus level using 0.80 as minimum confidence. The 'fixrank' option was used to assign taxonomy from kingdom to genus level. QIIME[34] scripts were used to pick OTUs at a 97% similarity threshold. The most abundant sequence from each OTU bin was chosen as the representative sequence. Species-level taxonomy was obtained by NCBI BLAST against the greengenes reference database (http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/Caporaso_Reference_OTUs/gg_otus_4feb2011.tgz) with an *E*-value cutoff of 0.001. Sequences unclassifiable at a given taxonomic rank were classified at the next possible higher rank.

Data preparation for QTL analysis. The proportion values (number of reads for a given taxon/total sequence reads for a given animal) for each taxonomic level were normalized as suggested by Benson *et al.*[3] The taxon bins were then $\log_{10}$ transformed. This resulted in 34,624 species-level OTUs, 863 genera, 376 families, 218 orders, 93 classes and 51 phyla. Out of 34,624 species-level OTUs, 19,443 cover nearly 99% of the total sequences and 762 species OTUs represents nearly 90% of the total sequences (Supplementary Fig. S6).

Core measurable microbiota. A CMM was determined by a manner similar to that of Benson *et al.*[3] using two technical repeats from five different samples. Sequences were processed, classified into taxonomic bins and the $\log_{10}$ transformed values for each bin were plotted for all pairwise combinations of the two repeats (Supplementary Fig. S5). A threshold of >20 reads per bin leads to a correlation >0.97. Thus, the CMM taxa were defined as bins containing more than 20 reads in at least 20 animals. The resulting 131 species-level OTUs represent nearly 80% of the total sequences (Supplementary Fig. S6).

Genotyping and QTL analysis. The Illumina murine HD array was used to genotype 1,449 SNPs (of which 1,199 are informative) from 261 G4 animals. The Happy package[35] was used for QTL and covariate QTL analysis, expecting additive contributions from inherited parental haplotypes at each locus. Each $\log_{10}$ transformed taxonomic bin was treated as an individual phenotypic trait. The hfit function in Happy was used to correct for cage and family effects in a manner similar to Johnsen *et al.*[36] Covariates were included in the model by specifying an additional design matrix in the input file. The genome-wide significance, or *E*-value, for each phenotype was estimated by a permutation test based on 1,000 shuffled reassignments of the phenotypes[16]. To illustrate the procedure used to obtain the *E*-value for a given trait, we produced a 'QQ-plot' (Supplementary Fig. S7), which shows the distribution of maximum $-\log_{10}$ *P*-scores recorded across 1,199 SNPs for 1,000 random permutations of the phenotype scores (for OTUID: N26684). The threshold for significant QTLs was set at 5% and that for suggestive QTLs[3] at 10%. This corresponds to analysis of variance (ANOVA) $-\log_{10}$ *P*-values ≥4.39 and ≥4.10, respectively, as described by Valdar *et al.*[37] Confidence intervals were determined manually by a drop of 1.5 in ANOVA $-\log_{10}$ *P*-score[16]. The probability of overlap for QTLs was calculated as described by Graham *et al.*[38] with a size of 2,500 Mb used as the size of the mouse genome. Chromosome visualization was performed with circus[39]. All mouse genotype, phenotype and Happy input files are provided in Supplementary Data 6.

Alpha and beta diversity analysis. Alpha and beta diversity analyses were largely performed using QIIME[34]. Alpha diversity[34,40] analysis was performed on the whole data set with a normalized sequence number of 2,500 per individual (the lowest read number in the data set). An OTU table was generated using a 97% sequence similarity threshold and singletons were removed. The Chao1 and Shannon indices, both common measures of diversity within a sample (that is, alpha diversity), were used to describe species richness and evenness, respectively. Between-sample diversity (that is, beta diversity) was analysed using CAP implemented in the Vegan package in R[41], with disease status as a constraint. This analysis is very similar to a redundancy analysis, but additionally it allows non-Euclidian dissimilarity indices (we used Bray–Curtis[42], Jaccard and UniFrac[14,15] distances). Statistical significance for CAP was determined by an ANOVA-like permutation test function with 1,000 permutations (anova.cca) in Vegan. We used R version 2.15.2 on Linux (*R* Development Core Team (2012)).

References

1. Rosenthal, M., Goldberg, D., Aiello, A., Larson, E. & Foxman, B. Skin microbiota: microbial community structure and its potential association with health and disease. *Infect. Genet. Evol.* **11,** 839–848 (2011).
2. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
3. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl Acad. Sci. USA* **107,** 18933–18938 (2010).
4. McKnite, A. M. *et al.* Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* **7,** e39191 (2012).
5. Bowcock, A. M. & Cookson, W. O. C. M. The genetics of psoriasis, psoriatic arthritis and atopic dermatitis. *Hum. Mol. Genet.* **13 Spec No 1,** R43–R55 (2004).
6. Chen, M., Kim, G. H., Prakash, L. & Woodley, D. T. Epidermolysis bullosa acquisita: autoimmunity to anchoring fibril collagen. *Autoimmunity* **45,** 91–101 (2012).
7. Ludwig, R. J. *et al.* Identification of quantitative trait Loci in experimental epidermolysis bullosa acquisita. *J. Invest. Dermatol.* **132,** 1409–1415 (2012).
8. Hundorfean, G., Neurath, M. F. & Sitaru, C. Autoimmunity against type VII collagen in inflammatory bowel disease. *J. Cell Mol. Med.* **14,** 2393–2403 (2010).
9. Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. & Krawczak, M. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat. Rev. Genet.* **6,** 376–388 (2005).
10. Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome Res.* **18,** 1043–1050 (2008).
11. Gao, Z., Tseng, C., Pei, Z. & Blaser, M. J. Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl Acad. Sci. USA* **104,** 2927–2932 (2007).
12. Fierer, N., Hamady, M., Lauber, C. L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl Acad. Sci. USA* **105,** 17994–17999 (2008).
13. Faith, D. P. & Baker, A. M. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol. Bioinform.* **2,** 121–128 (2006).
14. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
15. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4,** 17–27 (2010).
16. Dupuis, J. & Siegmund, D. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151,** 373–386 (1999).
17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R Statis. Soc.* **Ser. B 57,** 289–289 (1995).
18. Picard, C. *et al.* Pyogenic bacterial infections in humans with IRAK-4 deficiency. *Science* **299,** 2076–2079 (2003).
19. Akira, S., Uematsu, S. & Takeuchi, O. Pathogen recognition and innate immunity. *Cell* **124,** 783–801 (2006).
20. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by Toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–241 (2004).
21. Schauber, J. *et al.* Histone acetylation in keratinocytes enables control of the expression of cathelicidin and CD14 by 1,25-dihydroxyvitamin D3. *J. Invest. Dermatol.* **128,** 816–824 (2007).
22. Han, S. H., Kim, J. H., Martin, M., Michalek, S. M. & Nahm, M. H. Pneumococcal lipoteichoic acid (LTA) is not as potent as staphylococcal LTA in stimulating Toll-like receptor 2. *Infect. Immun.* **71,** 5541–5548 (2003).
23. Kusunoki, T., Hailman, E., Juan, T. S., Lichenstein, H. S. & Wright, S. D. Molecules from Staphylococcus aureus that bind CD14 and stimulate innate immune responses. *J. Exp. Med.* **182,** 1673–1682 (1995).
24. Schröder, N. W. J. *et al.* Lipoteichoic acid (LTA) of Streptococcus pneumoniae and Staphylococcus aureus activates immune cells via Toll-like receptor (TLR)-2, lipopolysaccharide-binding protein (LBP), and CD14, whereas TLR-4 and MD-2 are not involved. *J. Biol. Chem.* **278,** 15587–15594 (2003).
25. García-Lozano, J. R. *et al.* Caspase 7 influences susceptibility to rheumatoid arthritis. *Rheumatology* **46,** 1243–1247 (2007).
26. Teixeira, V. H. *et al.* Genetic and expression analysis of CASP7 gene in a European Caucasian population with rheumatoid arthritis. *J. Rheumatol.* **35,** 1912–1918 (2008).
27. Akhter, A. *et al.* Caspase-7 activation by the Nlrc4/Ipaf inflammasome restricts Legionella pneumophila infection. *PLoS Pathog.* **5,** e1000361 (2009).
28. Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* **22,** 850–859 (2012).
29. Fry, L., Baker, B. S., Powles, A. V., Fahlen, A. & Engstrand, L. Is chronic plaque psoriasis triggered by microbiota in the skin? *Br. J. Dermatol.* **169,** 47–52 (2013.
30. Aoki, V. *et al.* Environmental risk factors in endemic pemphigus foliaceus (fogo selvagem). *J. Investig. Dermatol. Symp. Proc.* **9,** 34–40 (2004).
31. Asghari, F. *et al.* Identification of quantitative trait loci for murine autoimmune pancreatitis. *J. Med. Genet.* **48,** 557–562 (2011).
32. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
33. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).
34. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
35. Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C. & Flint, J. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl Acad. Sci. USA* **97,** 12649–12654 (2000).
36. Johnsen, A. K. *et al.* Genome-wide and species-wide dissection of the genetics of arthritis severity in heterogeneous stock mice. *Arthritis. Rheum.* **63,** 2630–2640 (2011).
37. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38,** 879–887 (2006).
38. Graham, A. M. *et al.* Support for the reproductive ground plan hypothesis of social evolution and major QTL for ovary traits of Africanized worker honey bees (Apis mellifera L.). *BMC Evol. Biol.* **11,** 95 (2011).
39. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
40. Tuomisto, H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* 2–22 (2010).
41. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veget. Sci.* **14,** 927–930 (2003).
42. Bray, J. & Curtis, J. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 325–349 (1957).

Acknowledgements

We thank Katja Cloppenborg-Schmidt, Ilona Klamfuß, Silke Carstensen and Heinke Buhtz for providing technical assistance; Susen Müller and Andreia Marques for their help in the mouse house, Ralf Ludwig, Andreas Recke, Yask Gupta and Philipp Rausch for valuable discussion and Derk Wachsmuth and Werner Wegner for IT assistance. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Excellence Cluster 306 'Inflammation at Interfaces', Research Training Groups 'Modulation of Autoimmunity' (GRK1727/1) and 'Genes, Environment and Inflammation' (GRK1743/1) and the Max Planck Society.

Author contributions

G.S. and S.K. performed experiments, G.S., S.M. and J.W. performed analysis; G.S., J.F.B. and S.M.I. wrote the paper; J.F.B., S.M.I., and D.Z. edited the paper and assisted in interpretation; J.F.B. and S.M.I. designed the experiments.

Additional information

Accession code: European Nucleotide Archive (ENA). ERP002614. PRJEB1934 (http://www.ebi.ac.uk/ena/data/view/PRJEB1934)

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Srinivas, G. *et al.* Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun.* 4:2462 doi: 10.1038/ncomms3462 (2013).

Conclusion and Perspectives

The major chapters of this study demonstrate the power of wild house mice for gut microbiome research. Under a more natural setup than lab mice, we were able to investigate various essential factors that determine the variation of gut microbial communities in this important model organism: geography, diet and genetics. Additionally, we have further explored the host-microbiome relationship in the context of evolution, providing additional insight into the role of the microbiome in reducing hybrid fitness, and thus consequently to speciation. As a primer for future perspectives in understanding host-microbe interactions, we followed up another study on the skin microbiome and sequenced the genomes of one potential probiotic bacterium. The findings provide a better fundamental understanding of the mouse-, and by extrapolation human microbiome, thus benefiting biological, evolutionary and possibly even medical researchers.

Biogeography of the wild mouse gut microbiome

With the first profiling of gut microbiome in wild house mice, we revealed a microbiome composed mainly of *Bacteroides*, *Robinsoniella* and *Helicobacter*. Lab mice rarely contain *Robinsoniella,* as it is not included in the original Schaedler flora and thus not inoculated to majority of the lab mice. *Helicobacter,* on the other hand, is generally a genus under strict control in the lab due to the pathogenic potential of several species within this genus. A sharp contrast in the microbiome of wild- and lab mice may indicate many further discrepancies in mouse biology in addition to the already reported deficiencies in immunology (Abolins *et al.,* 2011) and higher morbidity (Martin *et al.,* 2010) of lab mice.

The other fundamental ecological finding of this study is the dominant role of biogeography in wild mice gut microbiome, especially the isolation by distance at different geographic scales. With an overall decay of similarity among communities in the gut, the global pattern of the gut microbiome mirrors that of the environmental microbiome (*e.g.* soil, water, *etc*), suggesting similar processes to shape the ecology of free-living and host-associated communities. More interestingly, the significant decay is observed at a continental scale (>520km) of isolation, while within one geographical location (<20km distance) other factors or stochastic processes dominate

variation. The exact nature of the biogeographical influence on the mouse gut microbiome remains to be examined by experiments, as neutral dispersal and/or local selection could lead to the observed pattern. Furthermore, the sampling covered the majority of the original distribution of western house mice, but "new world" populations remain to be explored, as they may reveal the consequences of living in more distant locations and diverse environments (for instance, additional climate zones).

The relatively smaller effect of genetics compared to geography is intriguing, and inspired our study in the hybrid zone where geographical effects can be neglected and the genetic contrast is stronger. This finding may well have important indications for human studies. A consensus of biogeography in the human gut microbiome has yet to be reached, and other analyses such as genome-wide association studies (GWAS) should take the potentially larger role of geography into account when samples from different geographic regions are compared. Additional perspective could be won by studying biogeography at the functional level, as currently the conclusions are achieved with regard to community structure, while functional metagenomic analysis might uncover finer-scale variation of community functioning.

Enterotypes in wild mice

Our unique dataset of wild mouse gut/fecal microbiomes, as well as the experiments and analysis carried out in this study, provided strong evidence to solve the debate since the introduction of the "enterotype" concept. Available studies have not yet reached a conclusion regarding whether gut microbiomes are composed of distinct clusters or rather form a continuum, and if clusters (enterotypes) exist, what the underlying reason(s) is. Our analysis first used the evaluation criteria from the original enterotype study in humans, as well as additional approaches, which unanimously support enterotypes in wild mice. The striking similarity of enterotypes between wild- and lab mice, and to the two enterotypes in humans and chimpanzee, may indicate either that microbial communities have a similar strategy in adapting to different diets, and/or the existence of enterotypes is an ancient trait of gut microbiomes.

Next, by transferring wild mice to the lab environment and monitoring changes in gut microbial composition/function, as well as isotope analysis and diet reconstruction of previously capture mice, we provide strong support for a role of diet contributing to enterotypes. We observed a rapid change of enterotype status in mice compared to the initial experiment in humans, without using an extreme and unusual diet as reported in a recent study (David *et al*., 2013). We propose the faster metabolism of mice compared to humans as a possible explanation. This special feature may help with human research by reducing unnecessary effort, but also require more consideration in applying results found in mice to humans.

Several questions remain to be addressed in future studies. The prevalence of enterotypes should be examined in more mammalian groups, or even lower groups of vertebrates/invertebrates, although answering the question of when enterotypes appeared in evolutionary history may be difficult. Fitness consequences of enterotypes are also unclear in mice, while in humans different enterotypes appear to have difference in diabetes incidence as well as rheumatoid arthritis (unpublished data). Finally, the signature genera of enterotypes in mice as well as other animals need detailed analysis, with potentially important outcomes for understanding their function as well as evolution. Thus far, *Bacteroides* is shared among all reported enterotypes, and co-evolution with different host is an interesting perspective.

Gut microbiome in hybrid house mice

To complement the first chapter, we utilized another wild mouse system to study the influence of host genetics on the gut microbiome: hybrid mice. The narrow, long and relatively stable stretch in central Europe, where the partially reproductively isolated *M. m. musculus* and *M. m. domesticus* have breed for a few hundred to two thousand years, has always been an interesting system for evolutionary biologists. For our purposes, we can gain insight into genetic effects of the host while controlling geography, since the two subspecies that split 1.5 million years ago co-exist and mix within less than tens of kilometers. We further complemented the wild hybrids with lab-derived hybrids, with enabled more control over their environments and other additional analyses. The first striking pattern of distinct- and less diverse gut microbiomes in both wild and lab hybrids already suggested that the maintenance of a

rather consistent microbiome might be critical for the fitness of hosts even in long-diverged subspecies.

By dissecting the genetic architecture of the genetic control on the microbiome via a quantitative trait loci approach in lab hybrids, we gained additional evidence for our hypothesis. The over-representation of transgressive effects of QTLs (half of those identified) influencing bacterial traits indicates conflicts between genomes with sufficient divergence. The identification of an epistasic pair following the Bateson-Dobzhansky-Muller incompatibility model further shows the detrimental effect of hybridizing genomes and suggests that purifying selection acts to maintain the microbiome over evolutionary time scales.

We found evidence of immune-deficiency in hybrids from the wild and lab, and similar to the major "immune system effect" of other studies, a considerable part of the variation in the microbiome is correlated with the expression level of immune genes. Lastly, with the interaction of immune deficiency and distortion of the microbiome, higher frequencies of pathology are found in hybrid mice, providing direct support for lower fitness related to the microbiome.

It remains difficult to directly measure the fitness of mice, regardless of in the wild or lab environment, and we thus can not establish a causal relationship between the microbiome and fitness from our study. Furthermore, the effect of the gut microbiome in hybrid mice is not as striking as in the case of *Nasonia* species, in which a single gut bacteria is lethal for hybrids. However, it should be pointed out that these insects have less diverse microbiomes and only an innate immune system. Further, the *Nasonia* hybrids studied by Brucker and Bordenstein (2013) were haploid F1 males, for which incompatibilities should be more widely exposed if partially recessive.

Future work should first improve the resolution of the QTL study using more advanced generations of the cross and a larger mapping population to identify individual genes. Understanding the complex interaction between the immune system and microbiome using experimental approaches would improve the understanding in other aspects of host-microbiome interaction and co-evolution. Lastly, a few bacteria come to our attention in mice with pathology, seemingly probiotic or anti-

inflammatory *Barnesiella* and the potentially pathogenic or pro-inflammatory *Mucisprillum* should be further investigated.

Bacterial genomics

The study we carried out in the context of a mouse disease model was an interesting study, where we first mapped genomic regions responsible for the disease and/or microbiome. Then, by incorporating the microbiome as co-variates in the disease mapping, we not only increased the power of the mapping but also discovered additional gene-microbiome-disease interactions. Members of the *Staphylococcus* genus seem to be probiotic and prevent disease development. Cultivating *Staphylococcus* is relatively simple, as it requires only a normal atmosphere and bacterial media. Accordingly, we found the universal presence of *Staphylococcus epidermidis* as a main member of the skin microbiota of several rodent species/subspecies. We first published the genomes as a baseline for more detailed comparative genomic analysis, particularly emphasizing signatures for host adaptation. Some primary analysis revealed a specific gene, surfactin synthase, to be under strong positive selection (highest dN/dS value) making it a candidate for contributing to probiotic function, as surfactin is an important molecule with anti-fungal and anti-viral activities, which could be protective in the case of autoimmune disease (unpublished data).

Similarly, we are in the process of cultivating other gut bacteria of interest, and attempt to understand the interaction and co-evolution with host from genomic perspective. Cultivating gut bacteria however, is more difficult due to the more complex nutritional requirements of most gut microbes, their requirement of a special anaerobic atmosphere and longer incubation times. Nevertheless, we have already successfully isolated strains belonging to *Bacteroides*, *Barnesiella*, *Mucisprillum* and members of Lachnospriaceae. Our recently acquired genome sequences of *Barnesiella* isolates, cultivated from mice with different glycol-transferase and thus different mucosal sugar, already showed strong signs of different sugar preference in metabolic pathways, and can thus be viewed as a promising example for the power of bacterial genomics (unpublished data).

To summarize, the studies presented in my thesis demonstrate the fundamental effects of major factors (geography, diet and genetics) in shaping the microbiome and improve our knowledge of the natural ecology and evolution of house mice. We have identified a list of bacterial genera to prioritize, which are important in the biology and physiology of wild mice and thus deserve further attention. Based on our observations, we presented new hypotheses, for which proof-of-principle studies should be carried out in the future. Last but not least, incorporating the upcoming knowledge of the microbiome in wild- and lab mice, as well as from human and other animals, should help us to assemble a more complete picture of the biology of "super-" or "meta-organisms" (Bosch & McFall-Ngai, 2011), and thus fill one of the important gaps of the modern synthesis (McFall-Ngai *et al.,* 2013).

References

Abolins, S.R., M.J.O. Pocock, J.C.R. Hafalla, E.M. Riley & M.E. Viney, (2011) Measures of immune function of wild mice, *Mus musculus*. *Mol. Ecol.* 20: 881-892.

Anderson, M.T. & H.S. Seifert, (2011) Opportunity and means: horizontal gene transfer from the human host to a bacterial pathogen. *MBio* 2: e00005-11.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H.B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J.J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E.G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W.M. de Vos, S. Brunak, J. Dore, J. Weissenbach, S.D. Ehrlich, P. Bork & M. Consortium, (2011) Enterotypes of the human gut microbiome. *Nature* 473: 174-180.

Azad, M.B., T. Konya, H. Maughan, D.S. Guttman, C.J. Field, R.S. Chari, M.R. Sears, A.B. Becker, J.A. Scott, A.L. Kozyrskyj & C.S. Investigators, (2013) Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. *Can. Med. Assoc.* 185: 385-394.

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

Baird, S.J.E., A. Ribas, M. Macholan, T. Albrecht, J. Pialek & J.G. de Bellocq, (2012) Where are the wormy mice? A re-examination of hybrid parasitism in the European house mouse hybrid zone. *Evolution* 66: 2757-2772.

Bardoel, B.W. & J.A.G. van Strijp, (2011) Molecular battle between host and bacterium: recognition in innate immunity. *J. Mol. Recognit.* 24: 1077-1086.

Barton, N.H. & G.M. Hewitt, (1985) Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16: 113-148.

Benson, A.K., S.A. Kelly, R. Legge, F.R. Ma, S.J. Low, J. Kim, M. Zhang, P.L. Oh, D. Nehrenberg, K.J. Hua, S.D. Kachman, E.N. Moriyama, J. Walter, D.A. Peterson & D. Pomp, (2010) Individuality in gut microbiota composition is a

complex polygenic trait shaped by multiple environmental and host genetic factors. *P. Natl. Acad. Sci. USA* 107: 18933-18938.

Boerner, B.P. & N.E. Sarvetnick, (2011) Type 1 diabetes: role of intestinal microbiome in humans and mice. *Ann Ny Acad Sci* 1243: 103-118.

Boettcher, K.J. & E.G. Ruby, (1990) Depressed light-emission by symbiotic *Vibrio fischeri* of the sepiolid squid *Euprymna scolopes*. *J Bacteriol* 172: 3701-3706.

Bos, K.I., V.J. Schuenemann, G.B. Golding, H.A. Burbano, N. Waglechner, B.K. Coombes, J.B. McPhee, S.N. DeWitte, M. Meyer, S. Schmedes, J. Wood, D.J.D. Earn, D.A. Herring, P. Bauer, H.N. Poinar & J. Krause, (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478: 506-510.

Bosch, T.C.G. & M.J. McFall-Ngai, (2011) Metaorganisms as the new frontier. *Zoology* 114: 185-190.

Bottomley, M.J., E. Muraglia, R. Bazzo & A. Carfi, (2007) Molecular insights into quorum sensing in the human pathogen *Pseudomonas aeruginosa* from the structure of the virulence regulator LasR bound to its autoinducer. *J. Biol. Chem.* 282: 13592-13600.

Brucker, R.M. & S.R. Bordenstein, (2013) The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus *Nasonia*. *Science* 341: 667-669.

Brzuszkiewicz, E., A. Thurmer, J. Schuldes, A. Leimbach, H. Liesegang, F.D. Meyer, J. Boelter, H. Petersen, G. Gottschalk & R. Daniel, (2011) Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-aggregative-haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol*. 193: 883-891.

Champion, M.D., (2011) Host-pathogen O-methyltransferase similarity and its specific presence in highly virulent strains of *Francisella tularensis* suggests molecular mimicry. *PLoS One* 6: e20295.

Chandler, J.A., J.M. Lang, S. Bhatnagar, J.A. Eisen & A. Kopp, (2011) Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genet.* 7: e1002272.

Chen, J., E.E. Lewis, J.R. Carey, H. Caswell & E.P. Caswell-Chen, (2006) The ecology and biodemography of *Caenorhabditis elegans. Exp. Gerontol.* 41: 1059-1065.

Chu, K.H., J. Qi, Z.G. Yu & V. Anh, (2004) Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.* 21: 200-206.

Chu, S., S. Elliott & M.E. Maltrud, (2003) Global eddy permitting simulations of surface ocean nitrogen, iron, sulfur cycling. *Chemosphere* 50: 223-235.

Chu, W., F. Lu, W. Zhu & C. Kang, (2011) Isolation and characterization of new potential probiotic bacteria based on quorum-sensing system. *J. Appl. Microbiol.* 110: 202-208.

Coates, M.E., (1975) Gnotobiotic animals in research: their uses and limitations. *Lab Anim.* 9: 275-282.

Colton, L. & M. Kosoy, (2013) Experimental infection of laboratory mice with two *Bartonella tribocorum* strains from wild *Mus* species: a homologous host-bacteria model system at the genus level. *Parasitology* 140: 61-68.

Conlan, S., L.A. Mijares, J. Becker, R.W. Blakesley, G.G. Bouffard, S. Brooks, H. Coleman, J. Gupta, N. Gurson, M. Park, B. Schmidt, P.J. Thomas, M. Otto, H.H. Kong, P.R. Murray, J.A. Segre & N.C. Sequencing, (2012) *Staphylococcus epidermidis* pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biol.* 13: 64.

Cook, S.F., (1973) The significance of disease in the extinction of the New England Indians. *Hum. Biol.* 45: 485-508.

Coors, A., J. Vanoverbeke, T. De Bie & L. De Meester, (2009) Land use, genetic diversity and toxicant tolerance in natural populations of *Daphnia magna*. *Aquat. Toxicol.* 95: 71-79.

Costello, E.K., J.I. Gordon, S.M. Secor & R. Knight, (2010) Postprandial remodeling of the gut microbiota in Burmese pythons. *ISME J.* 4: 1375-1385.

D'Arienzo, R., F. Maurano, G. Mazzarella, D. Luongo, R. Stefanile, E. Ricca & M. Rossi, (2006) *Bacillus subtilis* spores reduce susceptibility to *Citrobacter rodentium*-mediated enteropathy in a mouse model. *Res. Microbiol*. 157: 891-897.

David, L.A., C.F. Maurice, R.N. Carmody, D.B. Gootenberg, J.E. Button, B.E. Wolfe, A.V. Ling, A.S. Devlin, Y. Varma, M.A. Fischbach, S.B. Biddinger, R.J. Dutton & P.J. Turnbaugh, (2013) Diet rapidly and reproducibly alters the human gut microbiome. *Nature*. 505: 559-563.

De Paepe, M., V. Gaboriau-Routhiau, D. Rainteau, S. Rakotobe, F. Taddei & N. Cerf-Bensussan, (2011) Trade-off between bile resistance and nutritional competence drives *Escherichia coli* diversification in the mouse gut. *PLoS Genet*. 7: e1002107.

de Vos, W.M., O.P. Kuipers, J.R. van der Meer & R.J. Siezen, (1995) Maturation pathway of nisin and other lantibiotics: post-translationally modified antimicrobial peptides exported by gram-positive bacteria. *Mol. Microbiol*. 17: 427-437.

Deurenberg, R.H., C. Vink, S. Kalenic, A.W. Friedrich, C.A. Bruggeman & E.E. Stobberingh, (2007) The molecular evolution of methicillin-resistant *Staphylococcus aureus*. *Clin. Microbiol. Infec*. 13: 222-235.

Diamond, J., (1997) Guns, germs, and steel: the fates of human societies. W.W. Norton & Company, New York.

Dubilier, N., C. Bergin & C. Lott, (2008) Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat. Rev. Microbiol.* 6: 725-740.

Dubilier, N., C. Mulders, T. Ferdelman, D. de Beer, A. Pernthaler, M. Klein, M. Wagner, C. Erseus, F. Thiermann, J. Krieger, O. Giere & R. Amann, (2001) Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* 411: 298-302.

Dyhrman, S.T., P.D. Chappell, S.T. Haley, J.W. Moffett, E.D. Orchard, J.B. Waterbury & E.A. Webb, (2006) Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*. *Nature* 439: 68-71.

Dzur-Gejdosova, M., P. Simecek, S. Gregorova, T. Bhattacharyya & J. Forejt, (2012) Dissecting the genetic architecture of F1 hybrid sterility in house mice. *Evolution* 66: 3321-3335.

Eaton, K.A., A. Honkala, T.A. Auchtung & R.A. Britton, (2011) Probiotic *Lactobacillus reuteri* ameliorates disease due to enterohemorrhagic *Escherichia coli* in germfree mice. *Infect. Immun*. 79: 185-191.

Egert, M., A.A. de Graaf, H. Smidt, W.M. de Vos & K. Venema, (2006) Beyond diversity: functional microbiomics of the human colon. *Trends Microbiol*. 14: 86-91.

Enright, M.C., D.A. Robinson, G. Randle, E.J. Feil, H. Grundmann & B.G. Spratt, (2002) The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl. Acad. Sci. USA* 99: 7687-7692.

Erb-Downward, J.R., D.L. Thompson, M.K. Han, C.M. Freeman, L. McCloskey, L.A. Schmidt, V.B. Young, G.B. Toews, J.L. Curtis, B. Sundaram, F.J. Martinez & G.B. Huffnagle, (2011) Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS One* 6: e16384.

Erickson, A.R., B.L. Cantarel, R. Lamendella, Y. Darzi, E.F. Mongodin, C.L. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N.C. Verberkmoes, C.M. Fraser, R.L. Hettich & J.K. Jansson, (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 7: e49138.

Every, A.L., G.Z. Ng, C.D. Skene, S.N. Harbour, A.K. Walduck, M.A. McGuckin & P. Sutton, (2011) Localized suppression of inflammation at sites of *Helicobacter pylori* colonization. *Infect. Immun*. 79: 4186-4192.

Falush, D., T. Wirth, B. Linz, J.K. Pritchard, M. Stephens, M. Kidd, M.J. Blaser, D.Y. Graham, S. Vacher, G.I. Perez-Perez, Y. Yamaoka, F. Megraud, K. Otto, U. Reichard, E. Katzowitsch, X.Y. Wang, M. Achtman & S. Suerbaum, (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* 299: 1582-1585.

Fife, D.J., D.F. Bruhn, K.S. Miller & D.L. Stoner, (2000) Evaluation of a fluorescent lectin-based staining teichnique for some acidophilic mining bacteria. *Appl. Environ. Microbiol*. 66: 2208-2210.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick & et al., (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Franzenburg, S., S. Fraune, P.M. Altrock, S. Kunzel, J.F. Baines, A. Traulsen & T.C.G. Bosch, (2013) Bacterial colonization of Hydra hatchlings follows a robust temporal pattern. *ISME J* 7: 781-790.

Fukatsu, T. & T. Hosokawa, (2002) Capsule-transmitted gut symbiotic bacterium of the Japanese common plataspid stinkbug, *Megacopta punctatissima*. *Appl. Environ. Microbiol*. 68: 389-396.

Gevers, D., R. Knight, J.F. Petrosino, K. Huang, A.L. McGuire, B.W. Birren, K.E. Nelson, O. White, B.A. Methe & C. Huttenhower, (2012) The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol*. 10: e1001377.

Gilmore, M.S. & J.J. Ferretti, (2003) Microbiology: The thin line between gut commensal and pathogen. *Science* 299: 1999.

Good, J.M., M.A. Handel & M.W. Nachman, (2008) Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* 62: 50-65.

Gouaux, E., M. Hobaugh & L.Z. Song, (1997) Alpha-hemolysin, gamma-hemolysin, and leukocidin from *Staphylococcus aureus*: distant in sequence but similar in structure. *Protein Sci.* 6: 2631-2635.

Graham, D.Y. & Y. Yamaoka, (2006) Trace of human migration in *Helicobacter pylori* populations: Tracing human migration using *Helicobacter pylori* within modern populations. *Helicobacter* 11: 9-10.

Grice, E.A., H.H. Kong, S. Conlan, C.B. Deming, J. Davis, A.C. Young, G.G. Bouffard, R.W. Blakesley, P.R. Murray, E.D. Green, M.L. Turner, J.A. Segre & N.C.S. Progra, (2009) Topographical and temporal diversity of the human skin microbiome. *Science* 324: 1190-1192.

Grivennikov, S.I., K.P. Wang, D. Mucida, C.A. Stewart, B. Schnabl, D. Jauch, K. Taniguchi, G.Y. Yu, C.H. Osterreicher, K.E. Hung, C. Datz, Y. Feng, E.R. Fearon, M. Oukka, L. Tessarollo, V. Coppola, F. Yarovinsky, H. Cheroutre, L. Eckmann, G. Trinchieri & M. Karin, (2012) Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature* 491: 254.

Han, Y.P.W., R.W. Redline, M. Li, L.H. Yin, G.B. Hill & T.S. McCormick, (2004) *Fusobacterium nucleatum* induces premature and term stillbirths in pregnant mice: implication of oral bacteria in preterm birth. *Infect. Immun*. 72: 2272-2279.

Heijtza, R.D., S.G. Wang, F. Anuar, Y. Qian, B. Bjorkholm, A. Samuelsson, M.L. Hibberd, H. Forssberg & S. Pettersson, (2011) Normal gut microbiota modulates brain development and behavior. *Proc. Natl. Acad. Sci. USA* 108: 3047-3052.

Hildebrand, F., T. Ebersbach, H.B. Nielsen, X.P. Li, S.B. Sonne, M. Bertalan, P. Dimitrov, L. Madsen, J.J. Qin, J. Wang, J. Raes, K. Kristiansen & T.R. Licht, (2012) A comparative analysis of the intestinal metagenomes present in guinea pigs (*Cavia porcellus*) and humans (*Homo sapiens*). *BMC Genomics* 13: 514.

Hildebrand, F., T.L.A. Nguyen, B. Brinkman, R.G. Yunta, B. Cauwe, P. Vandenabeele, A. Liston & J. Raes, (2013) Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. *Genome Biol.* 14.

Hill, D.A. & D. Artis, (2010) Intestinal bacteria and the regulation of immune cell homeostasis. *Annu. Rev. Immunol*. 28: 623-667.

Hirsch, P.R. & T.H. Mauchline, (2012) Who's who in the plant root microbiome? *Nat. Biotechnol.* 30: 961-962.

Hogan, M.C.E., (2010) Bacteria. In: Encyclopedia of earth. S. Draggan & C.J. Cleveland (eds). National Council for Science and the Environment, Washington DC.

IJssennagger, N., M. Derrien, G.M. van Doorn, A. Rijnierse, B. van den Bogert, M. Muller, J. Dekker, M. Kleerebezem & R. van der Meer, (2012) Dietary heme alters microbiota and mucosa of mouse colon without functional changes in host-microbe cross-talk. *PLoS One* 7: e33306.

Iwase, T., Y. Uehara, H. Shinji, A. Tajima, H. Seo, K. Takada, T. Agata & Y. Mizunoe, (2010) *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature* 465: 346-349.

Jami, E. & I. Mizrahi, (2012) Composition and similarity of bovine rumen microbiota across individual animals. *PLoS One* 7.

Kashyap, P.C., A. Marcobal, L.K. Ursell, S.A. Smits, E.D. Sonnenburg, E.K. Costello, S.K. Higginbottom, S.E. Domino, S.P. Holmes, D.A. Relman, R. Knight, J.I. Gordon & J.L. Sonnenburg, (2013) Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc. Natl. Acad. Sci. USA*. 110:17059-17064.

Keane, T.M., L. Goodstadt, P. Danecek, M.A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N.A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T.G. Belgard, P.L. Oliver, R.E. McIntyre, A. Bhomra, J. Nicod, X.C. Gan, W. Yuan, L. van der Weyden, C.A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I.J. Jackson, A. Czechanski, J.A. Guerra-Assuncao, L.R. Donahue, L.G. Reinholdt, B.A. Payseur, C.P. Ponting, E. Birney, J. Flint & D.J. Adams,

(2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477: 289-294.

Kim, T. & E. Mundt, (2011) Metagenomic analysis of intestinal microbiomes in chickens. *Methods Mol. Biol*. 733: 185-194.

Klein, E., M. Ofek, J. Katan, D. Minz & A. Gamliel, (2013) Soil suppressiveness to fusarium disease: shifts in root microbiome associated with reduction of pathogen root colonization. *Phytopathology* 103: 23-33.

Koenen, M.E., R. van der Hulst, M. Leering, S.H. Jeurissen & W.J. Boersma, (2004) Development and validation of a new in vitro assay for selection of probiotic bacteria that express immune-stimulating properties in chickens *in vivo*. *FEMS Immunol. Med. Microbiol.* 40: 119-127.

Kong, H.D.H., J. Oh, C. Deming, S. Conlan, E.A. Grice, M.A. Beatson, E. Nomicos, E.C. Polley, H.D. Komarow, P.R. Murray, M.L. Turner, J.A. Segre & N.C.S. Program, (2012) Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* 22: 850-859.

Koren, O., J.K. Goodrich, T.C. Cullender, A. Spor, K. Laitinen, H.K. Backhed, A. Gonzalez, J.J. Werner, L.T. Angenent, R. Knight, F. Backhed, E. Isolauri, S. Salminen & R.E. Ley, (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* 150: 470-480.

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

Kovacs, M., A. Halfmann, I. Fedtke, M. Heintz, A. Peschel, W. Vollmer, R. Hakenbeck & R. Bruckner, (2006) A functional *dlt*-operon, encoding proteins required for incorporation of d-alanine in teichoic acids in gram-positive bacteria, confers resistance to cationic antimicrobial peptides in Streptococcus pneumoniae. *J. Bacteriol.* 188: 5797-5805.

Kroegel, C., (1994) *Helicobacter pylori*, chronic active gastritis, peptic-ulcer and Carcinoma of stomach. *Med. Welt*. 45: 281-287.

Lankau, E.W., P.Y. Hong & R.I. Mackie, (2012) Ecological drift and local exposures drive enteric bacterial community differences within species of Galapagos iguanas. *Mol. Ecol*. 21: 1779-1788.

Larue, R., Z.T. Yu, V.A. Parisi, A.R. Egan & M. Morrison, (2005) Novel microbial diversity adherent to plant biomass in the herbivore gastrointestinal tract, as revealed by ribosomal intergenic spacer analysis and *rrs* gene sequencing. *Environ. Microbiol*. 7: 530-543.

Lederberg, J., (2000) Infectious history. *Science* 288: 287-293.

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

Ley, R.E., C.A. Lozupone, M. Hamady, R. Knight & J.I. Gordon, (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* 6: 776-788.

Linnenbrink, M., J.M. Johnsen, I. Montero, C.R. Brzezinski, B. Harr & J.F. Baines, (2011) Long-term balancing selection at the blood group-related gene B4galnt2 in the genus *Mus* (Rodentia; Muridae). *Mol. Biol. Evol*. 28: 2999-3003.

Linnenbrink, M., J. Wang, E.A. Hardouin, S. Kunzel, D. Metzler & J.F. Baines, (2013) The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol*. 22: 1904-1916.

Lopez-Velasco, G., P.A. Carder, G.E. Welbaum & M.A. Ponder, (2013) Diversity of the spinach (*Spinacia oleracea*) spermosphere and phyllosphere bacterial communities. *FEMS Microbiol. Lett*. 346: 146-154.

Lozupone, C.A., J. Stombaugh, A. Gonzalez, G. Ackermann, D. Wendel, Y. Vazquez-Baeza, J.K. Jansson, J.I. Gordon & R. Knight, (2013) Meta-analyses of studies of the human microbiota. *Genome Res.* 23: 1704-1714.

Lundberg, D.S., S.L. Lebeis, S.H. Paredes, S. Yourstone, J. Gehring, S. Malfatti, J. Tremblay, A. Engelbrektson, V. Kunin, T.G. del Rio, R.C. Edgar, T. Eickhorst, R.E. Ley, P. Hugenholtz, S.G. Tringe & J.L. Dangl, (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488: 86.

Lupp, C., M.L. Robertson, M.E. Wickham, I. Sekirov, O.L. Champion, E.C. Gaynor & B.B. Finlay, (2007) Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host Microbe* 2: 119-129.

Lykke, M., A.L. Hother, C.F. Hansen, H. Friis, C. Molgaard, K.F. Michaelsen, A. Briend, T. Larsen, P.T. Sangild & T. Thymann, (2013) Malnutrition induces

gut atrophy and increases hepatic fat infiltration: studies in a pig model of childhood malnutrition. *Am. J. Transl. Res*. 5: 543-554.

Mackos, A.R., T.D. Eubank, N.M.A. Parry & M.T. Bailey, (2013) Probiotic *Lactobacillus reuteri* attenuates the stressor-enhanced severity of *Citrobacter rodentium* infection. *Infect. Immun*. 81: 3253-3263.

Mah, T.F., B. Pitts, B. Pellock, G.C. Walker, P.S. Stewart & G.A. O'Toole, (2003) A genetic basis for *Pseudomonas aeruginosa* biofilm antibiotic resistance. *Nature* 426: 306-310.

Mandrell, R.E. & M.A. Apicella, (1993) Lipo-oligosaccharides (*los*) of mucosal pathogens - molecular mimicry and host-modification of *los*. *Immunobiology* 187: 382-402.

Manichanh, C., J. Reeder, P. Gibert, E. Varela, M. Llopis, M. Antolin, R. Guigo, R. Knight & F. Guarner, (2010) Reshaping the gut microbiome with bacterial transplantation and antibiotic intake. *Genome Res*. 20: 1411-1419.

Martin, B., S. Ji, S. Maudsley & M.P. Mattson, (2010) "Control" laboratory rodents are metabolically morbid: why it matters. *Proc. Natl. Acad. Sci. USA* 107: 6127-6133.

Martinson, V.G., B.N. Danforth, R.L. Minckley, O. Rueppell, S. Tingek & N.A. Moran, (2011) A simple and distinctive microbiota associated with honey bees and bumble bees. *Mol. Ecol.* 20: 619-628.

Maslowski, K.M. & C.R. Mackay, (2011) Diet, gut microbiota and immune responses. *Nat. Immunol.* 12: 5-9.

Mathur, R., H. Oh, D.K. Zhang, S.G. Park, J. Seo, A. Koblansky, M.S. Hayden & S. Ghosh, (2012) A mouse model of *Salmonella typhi* infection. *Cell* 151: 590-602.

McFall-Ngai, M., M.G. Hadfield, T.C. Bosch, H.V. Carey, T. Domazet-Loso, A.E. Douglas, N. Dubilier, G. Eberl, T. Fukami, S.F. Gilbert, U. Hentschel, N. King, S. Kjelleberg, A.H. Knoll, N. Kremer, S.K. Mazmanian, J.L. Metcalf, K. Nealson, N.E. Pierce, J.F. Rawls, A. Reid, E.G. Ruby, M. Rumpho, J.G. Sanders, D. Tautz & J.J. Wernegreen, (2013) Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. USA* 110: 3229-3236.

McInerney, J.O., J.A. Cotton & D. Pisani, (2008) The prokaryotic tree of life: past, present... and future? *Trends Ecol. Evol*. 23: 276-281.

McKnite, A.M., M.E. Perez-Munoz, L. Lu, E.G. Williams, S. Brewer, P.A. Andreux, J.W.M. Bastiaansen, X.S. Wang, S.D. Kachman, J. Auwerx, R.W. Williams, A.K. Benson, D.A. Peterson & D.C. Ciobanu, (2012) Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PLoS One* 7: e39191.

Mellmann, A., D. Harmsen, C.A. Cummings, E.B. Zentz, S.R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S.F. McLaughlin, J.K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P.M. Brzoska, R.L. Moore, S. Guenther, J.M. Rothberg & H. Karch, (2011) Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6: e22751.

Mikuls, T.R., G.M. Thiele, K.D. Deane, J.B. Payne, J.R. O'Dell, F. Yu, H. Sayles, M.H. Weisman, P.K. Gregersen, J.H. Buckner, R.M. Keating, L.A. Derber, W.H. Robinson, V.M. Holers & J.M. Norris, (2012) *Porphyromonas gingivalis* and disease-related autoantibodies in individuals at increased risk of Rheumatoid Arthritis. *Arthritis Rheum*. 64: 3522-3530.

Moeller, A.H., P.H. Degnan, A.E. Pusey, M.L. Wilson, B.H. Hahn & H. Ochman, (2012) Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nat. Commun*. 3: 1179.

Morris, A., J.M. Beck, P.D. Schloss, T.B. Campbell, K. Crothers, J.L. Curtis, S.C. Flores, A.P. Fontenot, E. Ghedin, L. Huang, K. Jablonski, E. Kleerup, S.V. Lynch, E. Sodergren, H. Twigg, V.B. Young, C.M. Bassis, A. Venkataraman, T.M. Schmidt, G.M. Weinstock & L.H.M. Project, (2013) Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am. J. Resp. Crit. Care* 187: 1067-1075.

Moulia, C., J.P. Aussel, F. Bonhomme, P. Boursot, J.T. Nielsen & F. Renaud, (1991) Wormy mice in a hybrid zone - a genetic-control of susceptibility to parasite infection. *J. Evolution. Biol.* 4: 679-687.

Nesse, R.M. & G.C. Williams, (1995) Why We Get Sick. Times Books., New York.

Nestel, D., N.T. Papadopoulos & M.A.M. Chueca, (2008) Current advances in the study of the ecology of fruit flies from Europe, Africa and the Middle East. *J. Appl. Entomol.* 132: 681-681.

Ochman, H., M. Worobey, C.H. Kuo, J.B.N. Ndjango, M. Peeters, B.H. Hahn & P. Hugenholtz, (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol.* 8: e1000546.

Ohkuma, M., T. Sato, S. Noda, S. Ui, T. Kudo & Y. Hongoh, (2007) The candidate phylum 'Termite Group 1' of bacteria: phylogenetic diversity, distribution, and endosymbiont members of various gut flagellated protists. *FEMS Microbiol. Ecol*. 60: 467-476.

Ottesen, A.R., A.G. Pena, J.R. White, J.B. Pettengill, C. Li, S. Allard, S. Rideout, M. Allard, T. Hill, P. Evans, E. Strain, S. Musser, R. Knight & E. Brown, (2013) Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum lycopersicum* (tomato). *BMC Microbiol*. 13: 114.

Pancer, Z. & M.D. Cooper, (2006) The evolution of adaptive immunity. *Annu. Rev. Immunol.* 24: 497-518.

Paulsen, I.T., L. Banerjei, G.S.A. Myers, K.E. Nelson, R. Seshadri, T.D. Read, D.E. Fouts, J.A. Eisen, S.R. Gill, J.F. Heidelberg, H. Tettelin, R.J. Dodson, L. Umayam, L. Brinkac, M. Beanan, S. Daugherty, R.T. DeBoy, S. Durkin, J. Kolonay, R. Madupu, W. Nelson, J. Vamathevan, B. Tran, J. Upton, T. Hansen, J. Shetty, H. Khouri, T. Utterback, D. Radune, K.A. Ketchum, B.A. Dougherty & C.M. Fraser, (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* 299: 2071-2074.

Phillips, C.D., G. Phelan, S.E. Dowd, M.M. McDonough, A.W. Ferguson, J.D. Hanson, L. Siles, N. Ordonez-Garza, M. San Francisco & R.J. Baker, (2012) Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Mol. Ecol.* 21: 2617-2627.

Pomeroy, L.R., P.J.I. Williams, F. Azam & J.E. Hobbie, (2007) The Microbial Loop. *Oceanography* 20: 28-33.

Portal-Celhay, C. & M.J. Blaser, (2012) Competition and resilience between founder and introduced bacteria in the *Caenorhabditis elegans* gut. *Infect. Immun*. 80: 1288-1299.

Prager, E.M., C. Orrego & R.D. Sage, (1998) Genetic variation and phylogeography of central Asian and other house mice, including a major new mitochondrial lineage in Yemen. *Genetics* 150: 835-861.

Prakash, T., K. Oshima, H. Morita, S. Fukuda, A. Imaoka, N. Kumar, V.K. Sharma, S.W. Kim, M. Takahashi, N. Saitou, T.D. Taylor, H. Ohno, Y. Umesaki & M.

Hattori, (2011) Complete genome sequences of rat and mouse segmented filamentous bacteria, a potent inducer of Th17 cell differentiation. *Cell Host Microbe* 10: 273-284.

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

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

Ravel, J., P. Gajer, Z. Abdo, G.M. Schneider, S.S.K. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta & L.J. Forney, (2011) Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. USA* 108: 4680-4687.

Reeves, A.E., M.J. Koenigsknecht, I.L. Bergin & V.B. Young, (2012) Suppression of *Clostridium difficile* in the gastrointestinal tracts of germ-free mice inoculated with a murine isolate from the family Lachnospiraceae. *Infect. Immun*. 80: 3786-3794.

Rehman, A., C. Sina, O. Gavrilova, R. Hasler, S. Ott, J.F. Baines, S. Schreiber & P. Rosenstiel, (2011) Nod2 is essential for temporal development of intestinal microbial communities. *Gut* 60: 1354-1362.

Rinke, C., P. Schwientek, A. Sczyrba, N.N. Ivanova, I.J. Anderson, J.F. Cheng, A. Darling, S. Malfatti, B.K. Swan, E.A. Gies, J.A. Dodsworth, B.P. Hedlund, G. Tsiamis, S.M. Sievert, W.T. Liu, J.A. Eisen, S.J. Hallam, N.C. Kyrpides, R. Stepanauskas, E.M. Rubin, P. Hugenholtz & T. Woyke, (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431-437.

Ruff-Roberts, A.L., J.G. Kuenen & D.M. Ward, (1994) Distribution of cultivated and uncultivated cyanobacteria and *Chloroflexus*-like bacteria in hot spring microbial mats. *Appl. Environ. Microbiol*. 60: 697-704.

Sage, R.D., D. Heyneman, K.C. Lim & A.C. Wilson, (1986) Wormy mice in a hybrid zone. *Nature* 324: 60-63.

Sara, M., G. Bavestrello, R. Cattaneo-Vietti & C. Cerrano, (1998) Endosymbiosis in sponges: relevance for epigenesis and evolution. *Symbiosis* 25: 57-70.

Schloissnig, S., M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D.R. Mende, J.R. Kultima, J. Martin, K. Kota, S.R. Sunyaev, G.M. Weinstock & P. Bork, (2013) Genomic variation landscape of the human gut microbiome. *Nature* 493: 45-50.

Schmitt, S., P. Tsai, J. Bell, J. Fromont, M. Ilan, N. Lindquist, T. Perez, A. Rodrigo, P.J. Schupp, J. Vacelet, N. Webster, U. Hentschel & M.W. Taylor, (2012) Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J.* 6: 564-576.

Sczesnak, A., N. Segata, X. Qin, D. Gevers, J.F. Petrosino, C. Huttenhower, D.R. Littman & I.I. Ivanov, (2011) The genome of Th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. *Cell Host Microbe* 10: 260-272.

Sharon, G., D. Segal, J.M. Ringo, A. Hefetz, I. Zilber-Rosenberg & E. Rosenberg, (2010) Commensal bacteria play a role in mating preference of Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* 107: 20051-20056.

Silver, L., (1995) Mouse genetics: concepts and applications. Oxford University Press, New York.

Sinkins, S.P., (2004) Wolbachia and cytoplasmic incompatibility in mosquitoes. *Insect Biochem. Molec*. 34: 723-729.

Smith, K., K.D. McCoy & A.J. Macpherson, (2007) Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. *Semin. Immunol*. 19: 59-69.

Spor, A., O. Koren & R. Ley, (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol*. 9: 279-290.

Srinivas, G., S. Moller, J. Wang, S. Kunzel, D. Zillikens, J.F. Baines & S.M. Ibrahim, (2013) Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. *Nat. Commun*. 4: 2462.

Staubach, F., J.F. Baines, S. Kunzel, E.M. Bik & D.A. Petrov, (2013) Host species and environmental effects on bacterial communities associated with *Drosophila* in the laboratory and in the natural environment. *PLoS One* 8: e70749.

Stecher, B., S. Chaffron, R. Kappeli, S. Hapfelmeier, S. Freedrich, T.C. Weber, J. Kirundi, M. Suar, K.D. McCoy, C. von Mering, A.J. Macpherson & W.D. Hardt, (2010) Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathog.* 6: e1000711.

Stecher, B., R. Denzler, L. Maier, F. Bernet, M.J. Sanders, D.J. Pickard, M. Barthel, A.M. Westendorf, K.A. Krogfelt, A.W. Walker, M. Ackermann, U. Dobrindt, N.R. Thomson & W.D. Hardt, (2012) Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl. Acad. Sci. USA* 109: 1269-1274.

Su, W.J., P. Bourlioux, M. Bournaud, M.O. Besnier & J. Fourniat, (1986) Transfer of cecal flora from the hamster to the axenic C3h mouse - use of this model to study anti-*Clostridium-difficile* barrier flora. *Can. J. Microbiol.* 32: 132-136.

Sudakaran, S., H. Salem, C. Kost & M. Kaltenpoth, (2012) Geographical and ecological stability of the symbiotic mid-gut microbiota in European firebugs, *Pyrrhocoris apterus* (Hemiptera, Pyrrhocoridae). *Mol. Ecol*. 21: 6134-6151.

Sullam, K.E., S.D. Essinger, C.A. Lozupone, M.P. O'Connor, G.L. Rosen, R. Knight, S.S. Kilham & J.A. Russell, (2012) Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis. *Mol. Ecol.* 21: 3363-3378.

Teeter, K.C., L.M. Thibodeau, Z. Gompert, C.A. Buerkle, M.W. Nachman & P.K. Tucker, (2010) The variable genomic architecture of isolation between hybridizing species of House mice. *Evolution* 64: 472-485.

Thrash, J.C., A. Boyd, M.J. Huggett, J. Grote, P. Carini, R.J. Yoder, B. Robbertse, J.W. Spatafora, M.S. Rappe & S.J. Giovannoni, (2011) Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* 1: 13.

Tilg, H. & A. Kaser, (2011) Gut microbiome, obesity, and metabolic dysfunction. *J. Clin. Invest.* 121: 2126-2132.

Ting, C.S., G. Rocap, J. King & S.W. Chisholm, (2002) Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends Microbiol*. 10: 134-142.

Tu, S.I., J. Uknalis, A. Gehring & Y.P. He, (2007) Application of horse-radish peroxidase linked chemiluminescence to determine the production mechanism of Shiga-like toxins by E-coli O157 : H7. *SPIE Proceedings* Vol. 6761

Turnbaugh, P.J., F. Baeckhed, L. Fulton & J.I. Gordon, (2008) Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* 3: 213-223.

Turnbaugh, P.J., R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight & J.I. Gordon, (2007) The Human Microbiome Project. *Nature* 449: 804-810.

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

Ubeda, C., V. Bucci, S. Caballero, A. Djukovic, N.C. Toussaint, M. Equinda, L. Lipuma, L.L. Ling, A. Gobourne, D. No, Y. Taur, R.R. Jenq, M.R.M. van den Brink, J.B. Xavier & E.G. Pamer, (2013) Intestinal microbiota containing *Barnesiella* species cures vancomycin-resistant *Enterococcus faecium* colonization. *Infect. Immun*. 81: 965-973.

Ubeda, C., L. Lipuma, A. Gobourne, A. Viale, I. Leiner, M. Equinda, R. Khanin & E.G. Pamer, (2012) Familial transmission rather than defective innate immunity shapes the distinct intestinal microbiota of TLR-deficient mice. *J. Exp. Med.* 209: 1445-1456.

Vale, F.F., F. Megraud & J.M.B. Vitor, (2009) Geographic distribution of methyltransferases of *Helicobacter pylori*: evidence of human host population isolation and migration. *BMC Microbiol*. 9:193.

Vinatzer, B.A., R. Cai, C.L. Monteil & C.E. Morris, (2012) The role of horizontal gene transfer in bacterial crop pathogen emergence. *Phytopathology* 102: 160-160.

White, B.A., A.M. Gomez, M.F. Ho, M.B. Miller, S.M. Thomas, C.J. Yeoman, S. Yildirim, D.J. Creedon, T.L. Goldberg, S.R. Leigh, K.E. Nelson, R.M. Stumpf & B.A. Wilson, (2011a) Comparative analysis of the vaginal microbiome in health and disease. *Genome Biol*. 12: 3-4.

White, M.A., B. Steffy, T. Wiltshire & B.A. Payseur, (2011b) Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics* 189: 289-304.

Whitman, W.B., D.C. Coleman & W.J. Wiebe, (1998) Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA* 95: 6578-6583.

Widerstrom, M., C.A. McCullough, G.W. Coombs, T. Monsen & K.J. Christiansen, (2012) A multidrug-resistant *Staphylococcus epidermidis* clone (ST2) is an ongoing cause of hospital-acquired infection in a western Australian hospital. *J. Clin. Microbiol.* 50: 2147-2151.

Williams, I.R., (2011) Deep impact: deciphering mucosal microbiomes using next-generation sequencing approaches. *Mucosal Immunol*. 4: 586-587.

Wisplinghoff, H., B. Ewertz, S. Wisplinghoff, D. Stefanik, G. Plum, F. Perdreau-Remington & H. Seifert, (2005) Molecular evolution of methicillin-resistant *Staphylococcus aureus* in the metropolitan area of Cologne, Germany, from 1984 to 1998. *J. Clin. Microbiol*. 43: 5445-5451.

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

Yamaoka, Y., (2009) *Helicobacter pylori* typing as a tool for tracking human migration. *Clin. Microbiol. Infec*. 15: 829-834.

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

Heath, B. Warner, J. Reeder, J. Kuczynski, J.G. Caporaso, C.A. Lozupone, C. Lauber, J.C. Clemente, D. Knights, R. Knight & J.I. Gordon, (2012) Human gut microbiome viewed across age and geography. *Nature* 486: 222-227.

Zabalou, S., M. Riegler, M. Theodorakopoulou, C. Stauffer, C. Savakis & K. Bourtzis, (2004) *Wolbachia*-induced cytoplasmic incompatibility as a means for insect pest population control. *Proc. Natl. Acad. Sci. USA* 101: 15042-15045.

Zaman, K., (2010) Tuberculosis: A global health problem. *J. Health. Popul. Nutr*. 28: 111-113.

Zehr, J.P. & B.B. Ward, (2002) Nitrogen cycling in the ocean: new perspectives on processes and paradigms. *Appl. Environ. Microbiol*. 68: 1015-1024.

Zhang, C.H., S.F. Li, L. Yang, P. Huang, W.J. Li, S.Y. Wang, G.P. Zhao, M.H. Zhang, X.Y. Pang, Z. Yan, Y. Liu & L.P. Zhao, (2013) Structural modulation of gut microbiota in life-long calorie-restricted mice. *Nat. Commun* 4: 2163.

Acknowledgements

None of the work included in this thesis would have been possible without the excellent supervision of my supervisor, Prof. Dr. John F. Baines. He has been a role model for my research and gave not only advice for research but also on my scientific career. Neither would they be possible without great support from all members of our working group at the Max-Planck-Institute for Evolutionary Biology, Ploen as well as the Universitätsklinikum Schleswig-Holstein, CAU Kiel. Dr. Sven Kuenzel, Silke Carstensen and Katja Cloppenburg have generated invaluable datasets for numerous studies, as well as Sarah Lemcke and Theresa Arlt during their internships. Dr. Miriam Linnenbrink has provided a large collection of precious samples and assisted other sample collections, while Philipp Rausch helped with analysis and discussion in many cases. Marie Vallier and Meriem Belheouane provided insights into several aspects of this thesis and will continue some parts of the work. I also thank the devotion of my thesis committee members, Prof. Dr. Hinrich Schulenburg and Prof. Dr Bernhard Haubold, their insightful criticism and advice have greatly helped improve the work at different stages.

As my main location of work, the Max-Planck-Institute for Evolutionary Biology has provided both the technical support as well as a great atmosphere for scientific research. I would like to thank Prof. Dr. Diethard Tautz, Dr. Arne Nolte, Dr Frank Yingguang Chan, Dr. Bettina Harr and Dr. Leslie Turner for their continuous help and input into these studies; Christine Pfeifle, Till Sckerl, Maik Görtz-Sonnwald and other members of the mouse-team for maintaining the mouse resources; Heike Harre, Elke Blohm-Sievers, Heinke Buhtz, Cornelia Burghardt, the sequencing team and other technicians for their generous help. Gabriela Fischer, Elsbeth Hammerich, Hans-Joachim Hamann, Marita Hildebrandt, Petra Salenz and Sabine Meyer greatly helped with essential aspects of the administration, including attending conferences and acquiring equipment, while Dr. Kerstin Mehnert offered critical help within the International Max-Planck Research School and administrative matters with regard to the University. In addition I want to thank the recently retired Horst Hansen, Berit Hansen and Frank Korthals for their assistance in establishing the new microbiology lab and Derk Wachsmuth and Werner Wegner for the excellent IT-support.

A lot of the colleagues mentioned above are not only work associates but also friends, who I enjoyed a lot of activities in Ploen with. In addition, numerous current or

former post-docs and PhD students here also accompanied me for large part of my Ploen life and thus I am grateful: Dr. Zejka Pezer, Dr. Freddy Chain, Rafik Neme, Dr. Henrik Krehenwinkel, Dr. Jai Denton, Luisa Parralles, Joshka Kaufman, Dr Chaitanya Gokhale, Dr Jie Cheng, Dr Till Cypionka, Noemie Erin,Dr. Weini Huang, Dr. Benjamin Werner, Dr. Emilie Hardouin, Gönensin Ozan Bozdag, Dr. Sofie von Merten, Sofie Bodenstein, Dr. Dave Rogers, Ellen McConell, Dr. Duncan Greig, Dr. Primrose Boyton, Dr. Daniel Banesh, Dr. Benno Woelfling, Robert Bakaric, Dr. Mahesh Panchal, Dr Kenyon Mobley and Isabel-Monique Moreau, and I apologize if I have missed any.

I also need to thank the great collaborations we have that assisted the work to current form: Prof. Andre Franke, Prof. Dr. Philip Rosenstiel, Dr. Shirin Kalyan and Prof. Dieter Kabelitz from CAU Kiel; Prof. Dr. Derk Metzler and Kim Steige (currently in Uppsala) from LMU Munich. There are also people who are not involved in the presented work but are already helping with various follow-up studies, including Prof. Dr. Baerbel Stecher and Sandrine Brugiroux from Max-von-Pettenkofer Institute in LMU Munich, as well as co-authors of other work I am involved in, to whom I would like to give my thanks. The Max-Planck-Society generously provided a stipend for the first year of my PhD study, as well as financed all workshops and conferences I have been to, while the German Research Foundation (DFG) supported the rest time period via scholarships coming from Forschergruppe 1078.

Finally, I devote this work to my parents and grandmother, who have always supported my studies and work with their irreplaceable love, and their care and sacrifice have been continuously motivating me in pursuing higher education. At this moment, well-conducted research and a carefully prepared thesis would be the most appropriate contribution to them.

Declaration

Hereby I declare,

i. that apart from my supervisor's guidance the content and design of the paper is all of my own work, contributions of other authors are listed following this section;

ii. the thesis has not been submitted either partially or wholly as part of a doctoral degree to another examining body, and no other materials are published or submitted for publication than indicated in the thesis;

ii. that the thesis has been prepared subject to the Rules of Good Scientific Practice of the German Research Foundation.

Author contributions:

Chapter I: Miriam Linnenbrink collected mouse material from the field and performed the population genetic analysis. Jun Wang analyzed the microbiome data and wrote the paper together with Miriam Linnenbrink and John F. Baines. Emilie A. Hardouin helped with microsatellite data generation, Sven Künzel performed 454 sequencing and Dirk Metzler helped with statistical analysis.

Chapter II: Miriam Linnenbrink provided mouse samples and collected new mice. Jun Wang gathered feces, generated and analyzed all new data and wrote the paper together with John F. Baines. Sven Künzel performed the 454 sequencing, Ricardo Fernandes and Marie-Josée Nadeau helped with isotope analysis and diet reconstruction and Philip Rosenstiel contributed the Illumina HiSeq sequencing.

Chapter III: Jun Wang collected new mouse samples from the hybrid zone, set up mouse crosses for the QTL analysis, generated and analyzed all data and wrote the paper together with John F. Baines. Shirin Kalyan, Natalie Steck, Guntram A. Grassl and Dieter Kabelitz helped with immune analysis and histopathology, Sven Künzel performed the 454 sequencing, Andre Franke helped with genotyping, Robert Häsler ran the TaqMan microfluidic cards and Leslie M. Turner and Bettina Harr helped with experimental design.

Chapter IV: Jun Wang gathered all materials, generated and analyzed all data and wrote the paper together with John F. Baines. Sven Künzel performed the Illumina MiSeq sequencing.

Chapter IV, Supplementary Materials: Jun Wang assisted in microbiome and QTL data analysis. The experiments, data collection and writing of the manuscript were performed by the remaining co-authors.

Kiel, 03 Feb 2014

_____	_____
Mr. Jun Wang	Prof. Dr. John F. Baines

Curriculum Vitae

Name: Jun Wang

Date of birth: 17.06.1987

Place of birth: Shandong, China

Current residence: Ploen, Germany

Marital status: unmarried

Education

Sep 2004-Jun 2008, B.Sc Biotechnology, Ocean University of China, graduate with honor. Qingdao, China

Sep 2008-Jun 2010, M.Sc Erasmus Mundus master in Marine Biodiversity and Conservation,

Sep 2008-Jun 2009 University Ghent, Belgium

Aug 2009-Jan 2010 University Oviedo, Spain

Jan 2010-Jun 2010, University Bremen, Germany

Graduate with greatest distinction of Belgium (grootste onderscheiding)

Jan 2010- Jun 2010, Master thesis at Max-Planck-Institute for Marine Microbiology, Bremen, Germany

Jul 2010-Sep 2010, Bioinformatic internship, Max-Planck-Institute for Marine Microbiology, Bremen, Germany

Oct 2010-current, PhD student at the Max-Planck-Institute for Evolutionary Biology, Ploen & University Hospital Schleswig-Holstein, CAU Kiel, Germany.

Publications

Linnenbrink M*, **Wang J***, Hardouin E, Künzel S, Metzler D, Baines JF, (2013). The role of biogeography in shaping diversity of the intestinal microbiota in house mice. Mol. Ecol. 22(7):1904-16. *Equal author contribution

Srinivas G, Möller S, **Wang J,** Künzel K, Zillikens D, Baines JF, Ibrahim S. 2013. Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. Nat. Comm. doi:10.1038/ncomms3462

Franzenburg S, Walter J, Kuenzel S, **Wang J,** Baines JF, Bosch TCG, Fraune S. 2013. Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proc Natl Acad Sci U S A. doi: 10.1073/pnas.1304960110

Wang J, Kuenzel S, Baines JF. 2014. Draft genomes of 11 *Staphylococcus epidermidis* strains isolated from wild mouse species. Genome Announc. doi: 10.1128/genomeA.01148-13.

Tschurtschenthaler M, **Wang J**, Fricke C, Fritz T,Niederreiter L, Adolph T, Offner T, Kalinke U, Baines J, Tilg H, Kaser A, Type I interferon signaling in the intestinal epithelium determines Paneth cell function, microbial ecology and epithelial regeneration. Gut. In press.

Awards

Chinese Government Award for Excellent PhD Student Overseas 2013 (6000 USD.)