The Evolution of the Genus *Bacteroides* in the House Mouse Species Complex

Dissertation

in fulfilment of the requirements for the degree *Doctor rerum naturalium*

of the Faculty of Mathematics and Natural Sciences of Kiel University

submitted by

Hanna Fokt

Kiel, April 2021

First referee: Prof. Dr. John F. Baines

Second referee: Prof. Dr. Tal Dagan

Date of the oral examination: 29.06.2021

Approved for publication: 29.06.2021

Table of Contents

Zusammenfassung

Der Darmtrakt von Säugetieren beherbergt eine komplexe Gemeinschaft von Mikroorganismen, die eine uralte Evolutionsgeschichte teilen und wechselseitig vorteilhafte Beziehungen mit ihren Wirten eingehen. Trotz der interindividuellen Variation in der Zusammensetzung der mikrobiellen Gemeinschaft im Darm sind die wichtigsten bakteriellen Phyla über die Zeit der Säugetierevolution hinweg konserviert. Über die evolutionären Prozesse im Säugetierdarm ist jedoch wenig bekannt. *Bacteroides* sind häufig vorkommende Darmbakterien und werden mit vielen gesundheitsrelevanten Merkmalen des Wirts in Verbindung gebracht. Trotz der Bedeutung dieser Gattung sind nur wenige Arten gut untersucht. Daher gibt es immer noch einen Mangel an Informationen über die Muster der Diversität innerhalb der Spezies über verschiedene Wirtsarten hinweg, was mit einer möglichen lokalen Anpassung an unterschiedliche Wirtsumgebungen verbunden sein könnte.

Unter Verwendung des Hausmaus-Artenkomplexes als Modell habe ich zunächst versucht, potenzielle Signaturen für die Differenzierung der *Bacteroides*-Häufigkeit in Abhängigkeit von der Wirtsunterart zu identifizieren. Durch eine geographische Untersuchung der Darmmikrobiota von *Mus musculus musculus* und *M. m. domesticus* mittels 16S rRNA-Gen-Sequenzierung fand ich heraus, dass die Wirtsunterarten eine geringere Rolle für die Struktur der Darmgemeinschaft spielen als der Einfluss der Geographie. Nichtsdestotrotz identifizierte die Indikatorspeziesanalyse der Gattung *Bacteroides* konsistente Wirtsunterarten-*Bacteroides*-Assoziationen über verschiedene geographische Standorte hinweg. Als nächstes wollte ich Kandidaten-*Bacteroides*-Taxa [Amplikonsequenzvarianten (ASVs) auf Stammebene] charakterisieren und die Unterschiede in ihren Genomen identifizieren, die zur bakteriellen Anpassung an die verschiedenen Mausunterarten beitragen könnten. Dazu wurde eine Kombination aus Kultivierungs- und genomischen Analysemethoden verwendet und vollständig sequenzierte Genome von 146 *Bacteroides*-Isolaten generiert. Die taxonomische Einordnung zeigt,

dass neben *B. acidifaciens*, *B. caecimuris* und *B. sartorii*-Stämmen zwei potenziell neue *Bacteroides*-Arten isoliert wurden. Darüber hinaus wurde eine Übereinstimmung zwischen einem Kandidatenindikator *Bacteroides* ASV, der stark mit *M. m. musculus* assoziiert, und den beiden nicht klassifizierten Isolaten festgestellt, was auf die Beteiligung dieser potenziell neuen *Bacteroides*-Spezies an der faszinierenden Wirt-Mikroben-Assoziation hindeutet.

Schließlich habe ich kontaktunabhängige, antagonistische Interaktionen zwischen Darm-assoziierten *Bacteroides*-Stämmen aus den beiden Hausmaus-Subspezies untersucht. Ich fand heraus, dass einige *Bacteroides*-Isolate antagonistische Interaktionen eingehen, und die beobachteten hemmenden Interaktionen vor allem zwischen Isolaten auftreten, die zu verschiedenen *Bacteroides*-Spezies (Inter-Spezies-Antagonismus) und zu verschiedenen Mauspopulationen gehören, und nicht zwischen Stämmen, die von verschiedenen Wirtsunterarten isoliert wurden.

Zusammenfassend lässt sich sagen, dass die vorliegende Arbeit die erste Studie ist, die systematisch Darm-assoziierte *Bacteroides* zwischen zwei Hausmaus-Subspezies untersucht. Es wurden starke Wirts-*Bacteroides*-Assoziationen identifiziert, die über verschiedene geographische Standorte hinweg konsistent sind. Die Sequenzierung des gesamten Genoms der isolierten Stämme wirft ein Licht auf das *Bacteroides*-Pan-Genom in Bezug auf Proteingehalt und Funktionen. Darüber hinaus ist dies die erste Studie, die antagonistische Interaktionen zwischen mausassoziierten *B. acidifaciens*-, *B. caecimuris*- und *B. sartorii*-Stämmen aufzeigt, die aus zwei Wirtssubspezies isoliert wurden.

Abstract

The mammalian intestinal tract harbors a complex community of microorganisms that share an ancient evolutionary history and establish mutually beneficial relationships with their hosts. Despite inter-individual variation in gut microbial community composition, the major bacterial phyla remain conserved over the time of mammalian evolution. However, much less is known about evolutionary processes in the mammalian gut. *Bacteroides* are dominant intestinal bacteria and linked to many health-related traits of the host. Despite the importance of this genus, only a few species are well studied. Thus, there is still a lack of information regarding the patterns of within-species diversity across different host species, which could be linked to potential local adaption to different host environments.

Using the house mouse species complex as a model, I first aimed to identify potential signatures of differentiation in *Bacteroides* abundance according to host subspecies. By performing a geographical survey of *Mus musculus musculus* and *M. m. domesticus* gut microbiota using 16S rRNA gene sequencing, I found host subspecies to play minor role in gut community structure compared to the impact of geography. Nevertheless, indicator species analysis of the *Bacteroides* genus identified consistent host subspecies-*Bacteroides* associations across different geographic locations. Next, I aimed to characterize candidate *Bacteroides* taxa [strain level amplicon sequence variants (ASVs)] and identify the differences in their genomes that might contribute to bacterial adaptation to the different mouse subspecies. For this, a combination of culturing and genomic analysis methods was used, which yielded fully sequenced genomes of 146 *Bacteroides* isolates. Taxonomic classification indicates that two potentially new *Bacteroides* species were isolated, along with *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains. Furthermore, a perfect match between a candidate indicator *Bacteroides* ASV, which strongly associates to *M. m. musculus*, and both unclassified isolates was detected, suggesting the involvement of this potentially new *Bacteroides* species in the intriguing host-microbe association.

Finally, I aimed to identify contact-independent antagonistic interactions between gut-associated *Bacteroides* strains among these two house mouse subspecies. I found some *Bacteroides* isolates to engage in antagonistic interactions, and the observed inhibitory interactions seem to occur mostly between isolates belonging to different *Bacteroides* species (inter-species antagonism) and to different mouse populations than between strains isolated from different host subspecies.

In conclusion, the present work is the first study systematically investigating gut-associated *Bacteroides* among two house mouse subspecies. Strong host-*Bacteroides* associations were identified to be consistent across different geographic locations. Whole genome sequencing of the isolated strains sheds light on the *Bacteroides* pan genome in terms of protein content and functions. Moreover, this is the first study identifying antagonistic interactions among mouse-associated *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains isolated from two host subspecies.

General introduction

The mammalian intestinal tract is densely populated by microorganisms that share an ancient evolutionary history with the host, such that a mutually beneficial relationship has evolved (Bäckhed *et al.*, 2005). Together with microbial communities from other body sites and the host, the gut microbiota make up the mammalian metaorganism. In a healthy individual the relationship between the host and microbiota is symbiotic. The host provides the microbiota with nutrients and a stable environment, while the microbiota offers many benefits to the host through a variety of physiological functions, including modulation of the immune system (Gensollen *et al.*, 2016), protection against pathogens (Pickard *et al.*, 2017) and energy supply by digesting dietary carbohydrates (Nakata *et al.*, 2017).

The diversity of the gut microbiota is significantly lower compared to other bodily sites and it is characterized by functional redundancy (Pérez-Cobas *et al.*, 2013; Moya and Ferrer, 2016) - when different microbial taxa share similar functions. The gut of a healthy human is dominated by the bacterial phyla Bacteroidetes and Firmicutes, followed by smaller proportions of Actinobacteria, Proteobacteria and Verrucomicrobia (Reyes *et al.*, 2010). Numerous studies identified different factors contributing to inter-individual variation in gut microbial community composition in humans and other mammals such as diet (David *et al.*, 2014; Donaldson, Lee and Mazmanian, 2015), geography (Yatsunenko, Federico E. Rey, *et al.*, 2012; Rehman *et al.*, 2016), host genetics and other extrinsic factors (Ley *et al.*, 2008; Bonder *et al.*, 2016; Kurilshikov *et al.*, 2017). Despite this variation between individual hosts (for example, at the bacterial species level), the major bacterial phyla remain conserved over mammalian evolutionary timescales (Ley *et al.*, 2008).

Members of the genus *Bacteroides* belong to the order Bacteroidales, the predominant bacteria in the mammalian gut. Moreover, *Bacteroides* are linked to many health-related traits of the host. The members of this group are obligate anaerobic, Gram-negative bacteria with approximately

60 species described to date. *Bacteroides* compose approximately 25% of mammalian fecal community (Ochoa-Repáraz *et al.*, 2010) and also displays considerable within-genus diversity in wild mice (Linnenbrink *et al.*, 2013). They are metabolically versatile and offer key benefits to the host including the breakdown of dietary carbohydrates (Comstock, 2009). However, in certain contexts *Bacteroides* can shift from a member of the normal flora to pathogenic state, i.e. are so called "pathobionts" (Round and Mazmanian, 2009).

Recently, *Bacteroides* are gaining more attention and becoming an important model for understanding the dynamics of the human gut environment and the role of the microbiome in health and disease (Bencivenga-Barry *et al.*, 2020; Donaldson *et al.*, 2020). Despite the importance of this genus, only a few species have been well studied and cover mainly clinically relevant *Bacteroides* such as members of the *B. fragilis* group. Thus, there is still a lack of information regarding the patterns of within-species diversity across different host species, which could be tied to potential local adaption to different host environments, and thus contribute to overall evolution at the metaorganism level.

Notably, *Bacteroides* abundance appears to be a heritable genetic trait (Turpin *et al.*, 2016) and repeatedly displays genetic associations in mapping studies (Wang, Kalyan, Steck, Turner, Harr, Künzel, Vallier, Häsler, Franke, H. H. Oberg, *et al.*, 2015; Bubier *et al.*, 2020). Wang *et al*., (2015) evaluated whether there is divergence in the genetic basis of intestinal microbiota regulation between the *M. musculus musculus* (*mus*) and *M. m. domesticus* (*dom*) house mouse subspecies. They performed a QTL mapping on the gut microbiota of a set of *mus*/*dom* F2 laboratory hybrids. Fourteen SNPs were identified to be associated with 29 bacterial traits, including a *Bacteroides* species-level operational taxonomic unit (OTU). Another recent genetic mapping study performed on 8,956 human samples from German individuals revealed gene loci to be significantly associated with *Bacteroides* taxa (Rühlemann *et al.*, 2021).

The house mouse is a widely used model in biomedicine research (Guénet and Bonhomme, 2003) and was also used in the present thesis to study gut-associated *Bacteroides* across closely

related host subspecies. In addition to the knowledge gained from laboratory strains, wild mouse populations are increasingly used in the studies of the gut microbiome (Linnenbrink *et al.*, 2013; Wang *et al.*, 2014), due to their natural genetic and environmental diversity (Guénet and Bonhomme, 2003). The evolutionary history of the *M. musculus* species complex is well defined (Guénet and Bonhomme, 2003). It originated on the Indian subcontinent, and the *dom* and *mus* subspecies split approximately 0.5 Myr ago (Guénet and Bonhomme, 2003; Neme and Tautz, 2016). The *mus* subspecies colonized much of Asia and Eastern Europe, while *dom* populated first the Near East and from there Western Europe (Cucchi, Vigne and Auffray, 2005).

The competition mediated by contact-dependent and secreted antimicrobial toxins play an important role of the gut microbiota composition and stability. A closer study of the antagonistic interactions among bacteria in the metaorganism is crucial to better understanding of the symbiosis. Despite the fact that Bacteroidales, specifically members of *Bacteroides* genus, are one of the most abundant bacteria in the mammalian gut, the antimicrobial interactions between these bacteria are poorly studied. In contrast to some members of the Firmicutes phylum, where antimicrobial compounds have been studied for decades, antimicrobial toxins produced by Bacteroidetes started to gain more attention only in last decade (Mattick, Hirsch and Berridge, 1947; Gardner, 1950). The first antimicrobials produced by *Bacteroides* were identified in the last years and mainly concern the strains of several human gut-associated species: *B. fragilis*, *B. ovatus*, *B. vulagtus*, *B. thetaiotaomicron*, *B. ovatus*, *B. dorei*, *B. cellulosilyticus*, and *B. stercoris* (Chatzidaki-Livanis, Coyne and Comstock, 2014a; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Coyne *et al.*, 2019).

Bacteroides are physically in contact with each other in the mammalian gut, and it was already shown that gut colonization with more than one strain from the same species is common in the human gut (Bjerke *et al.*, 2011; Zitomersky, Coyne and Comstock, 2011). Additionally, these bacteria evolved mechanisms to antagonize each other (Wexler and Goodman, 2017). Bacteroidales were shown to engage in two different types of antagonistic interactions: contact-dependent type VI secretion

systems (T6SSs) (Russell *et al.*, 2014; Chatzidaki-Livanis *et al.*, 2016) and secreted diffusible antimicrobial toxins (Chatzidaki-Livanis, Coyne and Comstock, 2014b; Coyne *et al.*, 2019). The previous study by Coyne, Roelofs and Comstock (2016) revealed most of human gut *B. fragilis* strains to carry genetic loci encoding T6SSs. Moreover, some of these systems have been shown to antagonize nearly all gut Bacteroidales species tested (Chatzidaki-Livanis *et al.*, 2016). Similarly to T6SSs, a family of diffusible peptide toxins called bacteroidetocins produced by some *Bacteroide*s species were identified to have broad spectrum activity and inhibit not only across genera, but also across families (Coyne *et al.*, 2019). Furthermore, Bacteroidales secreted antimicrobial proteins (BSAPs) were revealed to contain membrane attack/perforin (MACPF) domains, and contrary to T6SS systems and bacteroidetocins, these proteins target a subset of closely related strains (Chatzidaki-Livanis, Coyne and Comstock, 2014b; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Shumaker *et al.*, 2019).

In this context, the present PhD thesis aims to detect signatures of differentiation in gut-associated *Bacteroides* strains according to host subspecies, which may help to identify potential coadaptive processes. Furthermore, an investigation of the antagonism between *dom* and *mus* gut-associated *Bacteroides* strains not only helps to understand whether such interactions could mediate host subspecies-specific differences in *Bacteroides* composition, but also contributes to the better understanding of the basic principles of symbiosis in the context of the mouse metaorganism. First, I performed a bacterial 16S rRNA gene survey and indicator species analysis applied to the *Bacteroides* genus, which yielded several interesting candidates differentially abundant in *mus.* Next, using a combination of culturing and genome analyses I obtained fully sequenced genomes of 146 *Bacteroides* isolates, including two potentially new *Bacteroides* species, and studied the *Bacteroides* pan genome in terms of protein content and functions. Finally, I identified antagonistic interactions between different *Bacteroides* isolates, which seem to be mediated by a so far uncharacterized toxin.

Chapter I:

Geographic screen of the gut microbiome in the house mouse species complex

Introduction

The mammalian intestinal tract harbors a vast community of microorganisms that share an ancient evolutionary history with the host, such that a mutually beneficial relationship has evolved (Bäckhed *et al.*, 2005). Together with microbial communities from other body sites and the host, the gut microbiota compose a mammalian metaorganism. The microbiota provides many benefits to the host through a variety of physiological functions, including modulation of the immune system (Gensollen *et al.*, 2016), protection against pathogens (Pickard *et al.*, 2017) and energy supply by digesting dietary carbohydrates (Nakata *et al.*, 2017).

The gut of a healthy human is dominated by bacterial phyla Bacteroidetes and Firmicutes, followed by smaller proportions of Actinobacteria, Proteobacteria and Verrucomicrobia (Reyes *et al.*, 2010). Different studies identified factors contributing to inter-individual variation in gut microbial community composition in humans and other mammals such as diet (David *et al.*, 2014; Donaldson, Lee and Mazmanian, 2015), geography (Yatsunenko, Federico E. Rey, *et al.*, 2012; Rehman *et al.*, 2016), host genetics and other extrinsic factors (Ley *et al.*, 2008; Bonder *et al.*, 2016; Kurilshikov *et al.*, 2017). Despite this variation within individual hosts (for example, at the bacterial species level), the major bacterial phyla remain conserved over the time of mammalian evolution (Ley *et al.*, 2008).

The house mouse (*Mus musculus*) is a widely used model in biomedicine (Guénet and Bonhomme, 2003). In addition to the knowledge gained from laboratory strains, wild mouse populations are increasingly used in the studies of the gut microbiome (Linnenbrink *et al.*, 2013; Wang *et al.*, 2014), due to their natural genetic and environmental diversity (Guénet and Bonhomme, 2003). The evolutionary history of *M. musculus* species complex is well known (Guénet and Bonhomme, 2003). It originated on the Indian subcontinent, and the subspecies *M. m. domesticus* (*dom*) and *M. m. musculus* (*mus*) split approximately 0.5 Myr ago (Guénet and Bonhomme, 2003; Neme and Tautz,

2016). *M. m. musculus* colonized much of Asia and Eastern Europe, while *M. m. domesticus* populated first the Near East and from there Western Europe (Cucchi, Vigne and Auffray, 2005).

Bacteroides is an important genus in the mammalian intestine related to many health-related traits of the host. The members of this group are obligate anaerobic, Gram-negative bacteria with approximately 60 species described to date. *Bacteroides* compose approximately 25% of mammalian fecal community (Ochoa-Repáraz *et al.*, 2010) and also has considerable within-genus diversity in wild mice (Linnenbrink *et al.*, 2013). They are metabolically versatile and offer key benefits to the host including the breakdown of dietary carbohydrates (Comstock, 2009). However, in certain contexta *Bacteroides* can shift from a member of the normal flora to pathogenic state, i.e. are so called "pathobionts" (Round and Mazmanian, 2009).

Recently, *Bacteroides* are gaining more attention and becoming an important model for understanding the dynamics of the human gut environment and the role of the microbiome in health and disease (Bencivenga-Barry *et al.*, 2020; Donaldson *et al.*, 2020). Despite the importance of this genus, only a few species have been well studied and cover mainly clinically relevant *Bacteroides* such as members of *B. fragilis* group. Thus, there is still a lack of information regarding the patterns of within-species diversity across different host species, which could be tied to potential local adaption to different host environments.

Notably, *Bacteroides* abundance appears to be a heritable genetic trait (Turpin *et al.*, 2016) and repeatedly displays genetic associations in mapping studies (Wang, Kalyan, Steck, Turner, Harr, Künzel, Vallier, Häsler, Franke, H. H. Oberg, *et al.*, 2015; Bubier *et al.*, 2020). Wang *et al*., (2015) evaluated whether there is divergence in the genetic basis of intestinal microbiota regulation between *mus* and *dom* house mouse subspecies. They performed a QTL mapping on the gut microbiota of a set of *mus*/*dom* F2 laboratory hybrids. Fourteen SNPs were identified to be associated with 29 bacterial traits, including a *Bacteroides* species-level OTU. Another genetic mapping study performed on 8,956

human samples from German individuals revealed gene loci to be significantly associated with *Bacteroides* taxa (Rühlemann *et al.*, 2021).

In this chapter, I aimed to identify potential signatures of differentiation in microbial taxon abundance according to host subspecies, focusing mainly on *Bacteroides* as a candidate genus to identify potential coevolutionary processes. To achieve this, the house mouse species complex was used as a model, comprising multiple wild-derived outbred mouse colonies maintained at the MPI in Plön. These colonies originate from five locations across the geographic range of the subspecies *dom* (Germany, France, and Iran) and *mus* (Austria and Kazakhstan). Despite an impact of geography on the inter-individual variation in gut microbiota composition, a subset of important/reliable host species-specific differences would be expected to be common to all geographic locations. A survey of the 16S rRNA gene and indicator species analysis applied to the *Bacteroides* genus yielded several interesting bacterial candidates, which are further characterized at the strain-level with respect to host subspecies in Chapter II.

Results

I. Gut microbiota of the *M. m. musculus* and *M. m. domesticus* subspecies

I.1 Alpha and beta diversity measures in house mice

In order to obtain insight into the composition and structure of bacterial communities inhabiting the gut of the house mice, different mouse colonies originating from five locations across the geographic range of the subspecies *domesticus* and *musculus* were sampled (Figure 1). Cecum contents were collected from 120 adult males as described in Methods. To assess the composition of the cecal microbiota, V1-V2 region of the 16S rRNA gene was sequenced at the DNA level.

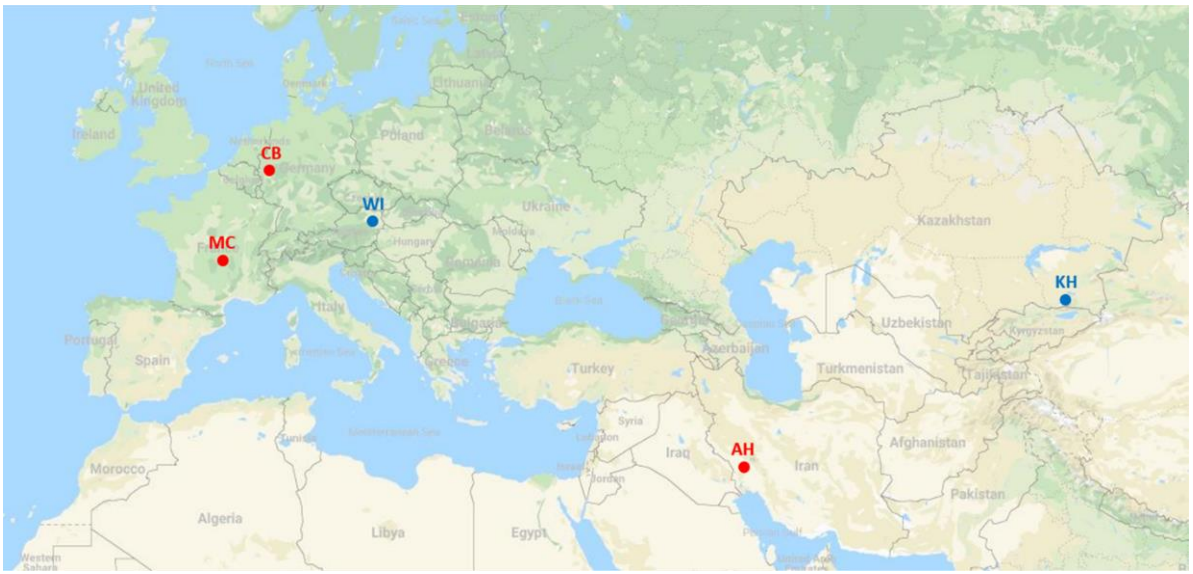


Figure 1. Geographic location of sampled house mice populations. The circles indicate the geographic origins of the mice and are color coded by mouse subspecies: red circles – *M. m. domesticus*, blue circles – *M. m. musculus*. The letters correspond to the geographic location name. Abbreviations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

The patterns of diversity within and between house mice were assessed by calculating alpha (Observed, Chao1 and Shannon) and beta diversity (Bray-Curtis and Jaccard) indices at the level of host subspecies and mouse populations. Alpha diversity measures (within sample diversity) show no significant differences between host subspecies (Figure 2A). However, pairwise comparisons of alpha diversity measures among different mice populations showed that bacterial richness in Vienna (WI)

mice was slightly, but significantly lower compared to Ahvaz (AH) mice (Wilcoxon test, Richness: $p = 0.029$ and 0.018 for Observed and Chao1, respectively) (Figure 2B). The same trend was observed in Cologne-Bonn (CB) compared to AH mice (Wilcoxon test, Chao1: $p = 0.047$). No significant differences were observed in taxa evenness among different mouse populations (Wilcoxon test, Shannon index: $p = 1.000$). Taken together, the estimates of alpha diversity in *mus* and *dom* mice suggest that overall, the two host subspecies do not differ in alpha diversity of their cecal communities. However, local geography does appear to influence the species richness of individual populations within host subspecies.

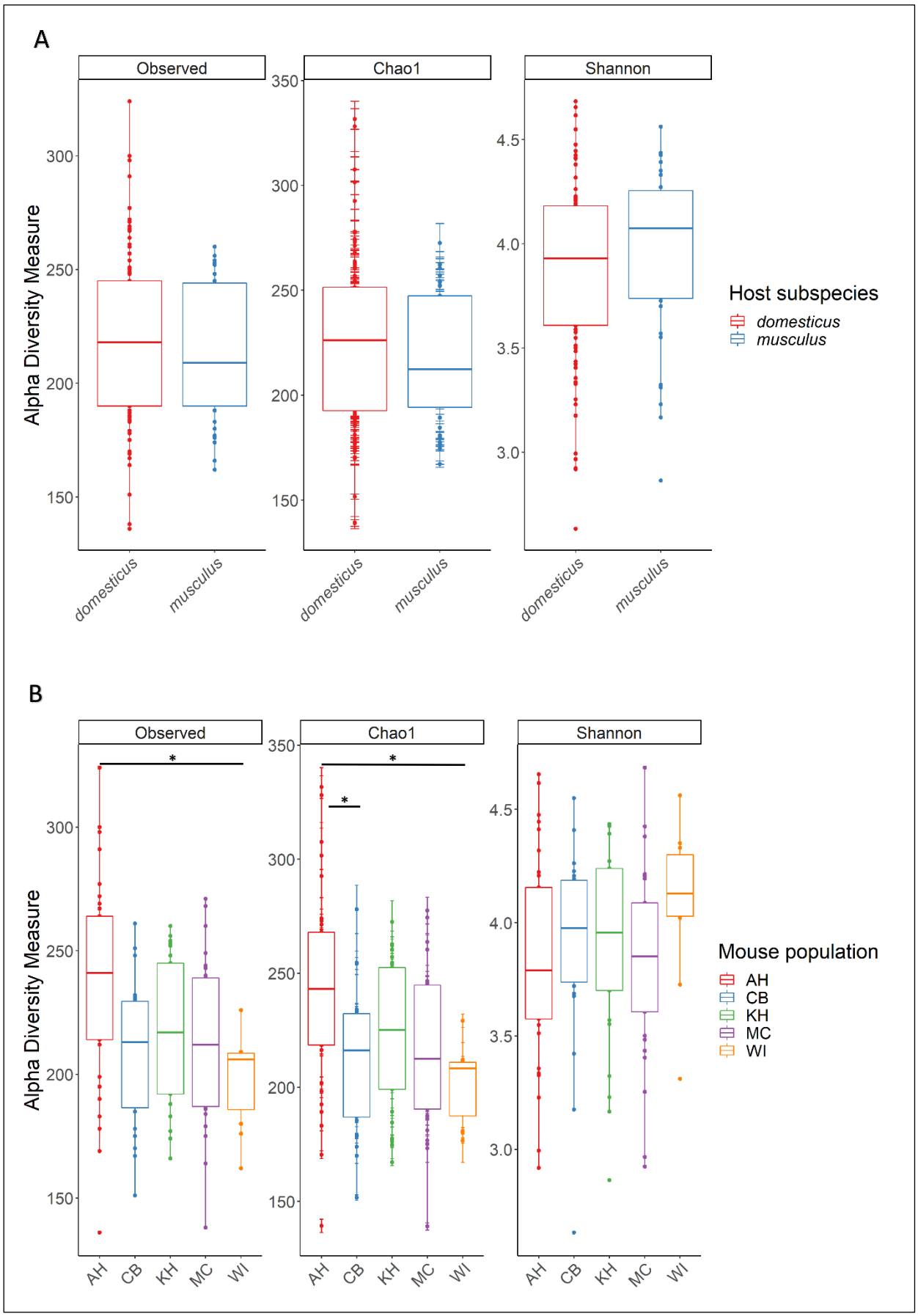

A
Observed
Chao1
Shannon
Alpha Diversity Measure
domesticus
musculus
Host subspecies
domesticus
musculus
B
Observed
Chao1
Shannon
Alpha Diversity Measure
AH
CB
KH
MC
WI
Mouse population
AH
CB
KH
MC
WI

Figure 2. Alpha diversity of the cecal microbiota of *dom* and *mus* mouse subspecies. Alpha diversity of bacterial taxa in **A -** mouse subspecies and **B -** mouse populations from different geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria. The analysis was performed with three alpha diversity measures: Observed and Chao1 (richness) and Shannon index (richness and evenness). The calculation of pairwise comparisons was performed using Pairwise Wilcoxon. Significance levels are denoted by stars: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

To assess the differences in gut microbial composition and to investigate to what extent these differences are associated to host subspecies- or the geographical origin of the mice, beta diversity was evaluated (Suppl. table 1). First, principal coordinate analysis (PCoA) was applied to weighted (quantitative) Unifrac index, which incorporates the abundance and phylogenetic relatedness of the observed taxa. Comparisons of beta diversity between *mus* and *dom* mice showed that gut bacterial communities cluster independently of the host subspecies (Figure 3A). However, the non-parametric multivariate analysis of variance (Adonis) applied to weighted Unifrac distances revealed a small influence of mouse subspecies, which explains 2.6% of total variance ($r^2 = 0.0261$, $p = 9.999 \times 10^{-5}$, 10000 permutations). Also, Adonis reveals a significant influence of mouse population (Figure 3B), which explains 12.8% of the total variance ($r^2 = 0.1284$, $p = 9.999 \times 10^{-5}$, 10000 permutations).

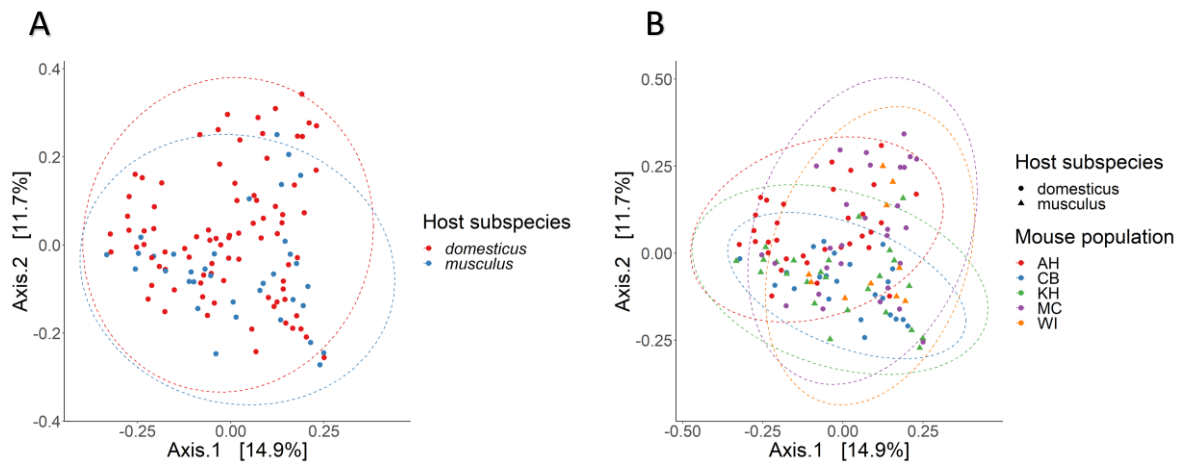


Figure 3. Principal coordinate analysis (PCoA) of weighted Unifrac index. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0261$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.1284$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

Next, PCoA was performed on unweighted Unifrac index, which is based on presence/absence and phylogenetic relatedness of the observed microbial taxa. The clustering according to host subspecies was clearer in comparison to weighted Unifrac, but Adonis analysis of variance revealed a similar influence of host subspecies (2.9%) and slightly higher influence of host population (14.5%) on the gut microbiota composition (Figure 4A and B).

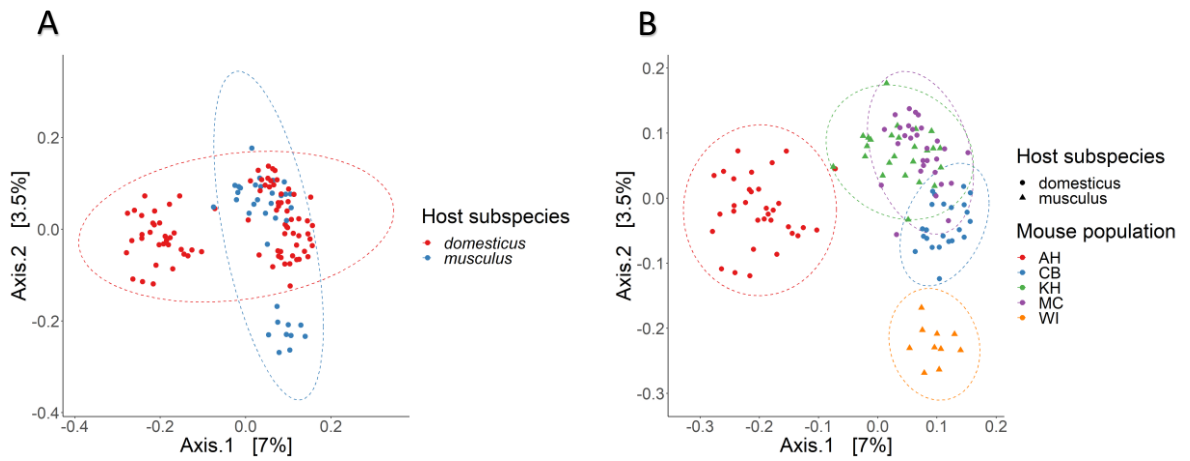


Figure 4. Principal coordinate analysis (PCoA) of unweighted Unifrac index. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0292$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.1446$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

Moreover, significant differences in overall community structure were also reflected by performing PCoA on Bray-Curtis dissimilarity (reflects differences in microbial abundances of the taxa) and on Jaccard distances (metrics based on and presence/absence of microbial taxa). Adonis analysis applied to Bray-Curtis indices revealed a similar influence of mouse subspecies and population (2.7% and 14.3%, respectively) on the microbiota structure (Suppl. figure 1, Suppl. table 1), while Jaccard distances showed a slightly lower influence of mouse subspecies and population, explaining 1.9% and 10.0% of total variance (Suppl. figure 2, Suppl. table 1).

In summary, it was detected that both host subspecies and mouse population were determinants of the gut community composition. However, the mouse population had a higher influence.

I.2 *Bacteroides* patterns of diversity among *mus* and *dom* mice

Because of the main focus of the present study, the patterns of diversity in *Bacteroides* amplicon sequence variants (ASVs) within and between groups of mice were also evaluated, by calculating alpha and beta diversity indices at the level of host subspecies and mouse populations.

Within sample diversity measures show no significant differences in *Bacteroides* diversity with respect to host subspecies (Figure 5A). Pairwise comparisons of Observed number of species and the Chao1 index among different mice populations showed that *Bacteroides* richness in CB, MC and WI mice was significantly lower compared to AH and KH mice (Figure 5B, Table 1). Similar differences were observed in taxon evenness among the same mouse populations (Shannon index: Figure 5B, Table 1). Taken together, the estimates of alpha diversity in *mus* and *dom* mice suggest that overall, host subspecies do not significantly differ in *Bacteroides* diversity in the cecal microbiota. However, geography/demography might influence species richness and evenness in populations belonging to the same or different host subspecies.

Table 1. Alpha diversity measures applied to *Bacteroides* genus among *mus* and *dom* subspecies and mouse populations. Significant p-values are shown in bold.

	Adjusted p-value		
	Observed	Chao1	Shannon
Host subspecies	0.1100	0.1200	0.1400
AH and CB populations	**1×10^{-6}**	**8.3×10^{-7}**	**1.9×10^{-7}**
AH and MC populations	**0.0041**	**0.0031**	**0.0009**
AH and KH populations	0.8060	0.7574	0.6067
AH and WI populations	**0.0020**	**0.0017**	**0.0022**
CB and MC populations	**0.017**	**0.0170**	**0.0013**
CB and KH populations	**4.2×10^{-6}**	**4.2×10^{-6}**	**1.9×10^{-6}**
CB and WI populations	0.1248	0.1248	**0.0250**
KH and MC populations	**0.0119**	**0.0119**	**0.0271**
KH and WI populations	**0.0030**	**0.0030**	**0.0100**

To assess the differences in *Bacteroides* composition and evaluate to what extent these differences are explained by host subspecies or geographic origin, PCoA was applied to weighted and unweighted Unifrac distances, Bray-Curtis and Jaccard indices (Figures 6-7, Suppl. figures 3-4, Suppl. table 1). Comparisons of beta diversity on *Bacteroides* between *mus* and *dom* mice and between different mouse populations mirror the results obtained for the entire gut microbial community (Figure 6 and 7). Host subspecies showed a low influence *Bacteroides* community structure, where the minimum total variance explained was 5.3% (Jaccard; Adonis: $r^2 = 0.0526$, $p = 9.999 \times 10^{-5}$, 10000 permutations) (Suppl. figure 4A) and maximum 7.0% (Weighted Unifrac; Adonis: $r^2 = 0.0703$, $p = 9.999 \times 10^{-5}$, 10000 permutations) (Figure 6A). Further, Adonis revealed a higher influence of mouse population (minimum 24.1% and maximum 33.0% of the total variance) (Suppl. figure 4B and 6B).

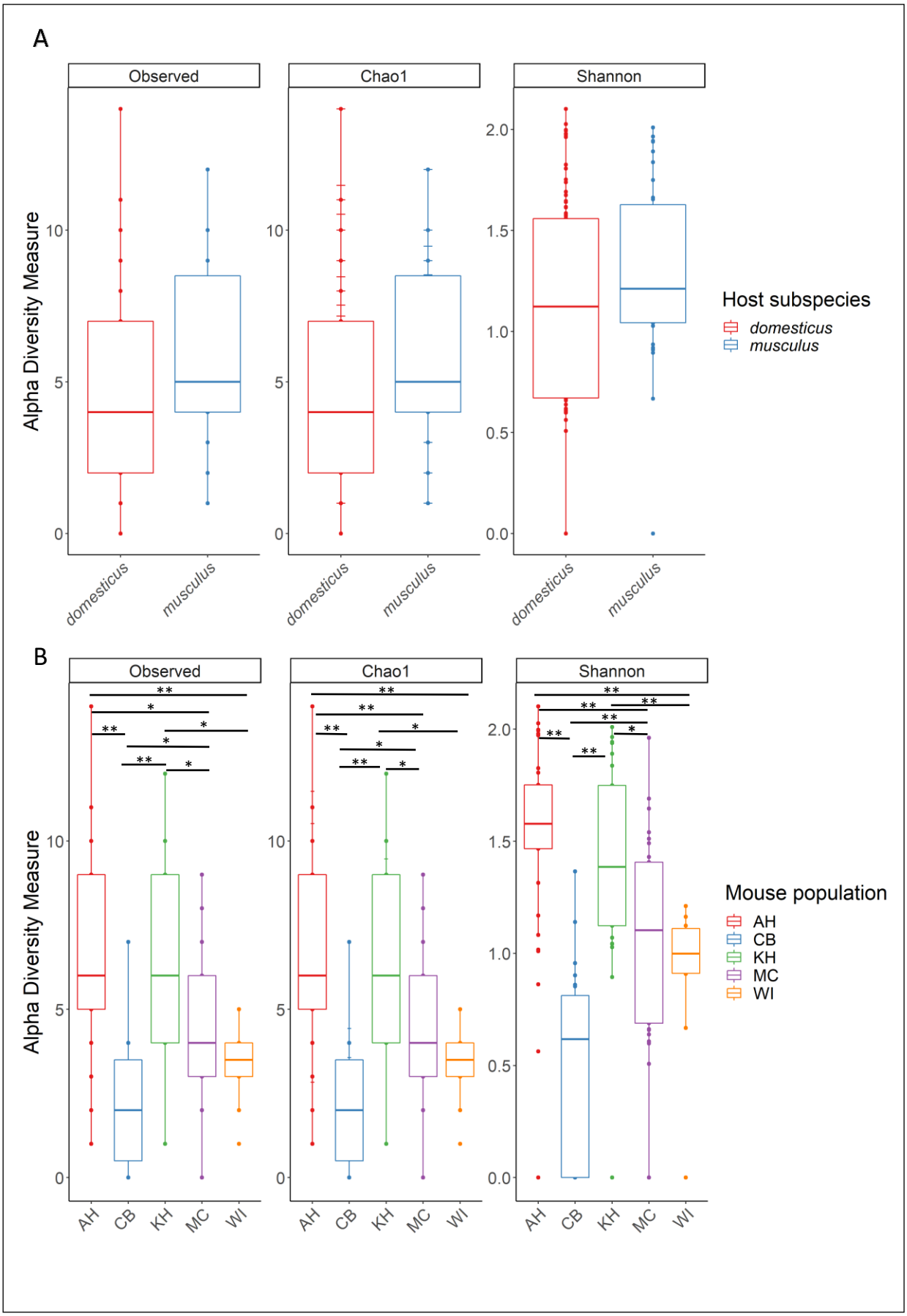

A
Observed
Chao1
Shannon
Alpha Diversity Measure
domesticus
musculus
Host subspecies
domesticus
musculus
B
Observed
Chao1
Shannon
Alpha Diversity Measure
AH
CB
KH
MC
WI
Mouse population
AH
CB
KH
MC
WI

Figure 5. Alpha diversity of *Bacteroides* taxa of *dom* and *mus* mouse subspecies. Alpha diversity of *Bacteroides* in **A -** mouse subspecies and **B -** mouse populations from different geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria. The analysis was performed with three alpha diversity measures: Observed and Chao1 (richness) and Shannon index (richness and evenness). The calculation of pairwise comparisons was performed using Pairwise Wilcoxon. Significance levels are denoted by stars: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

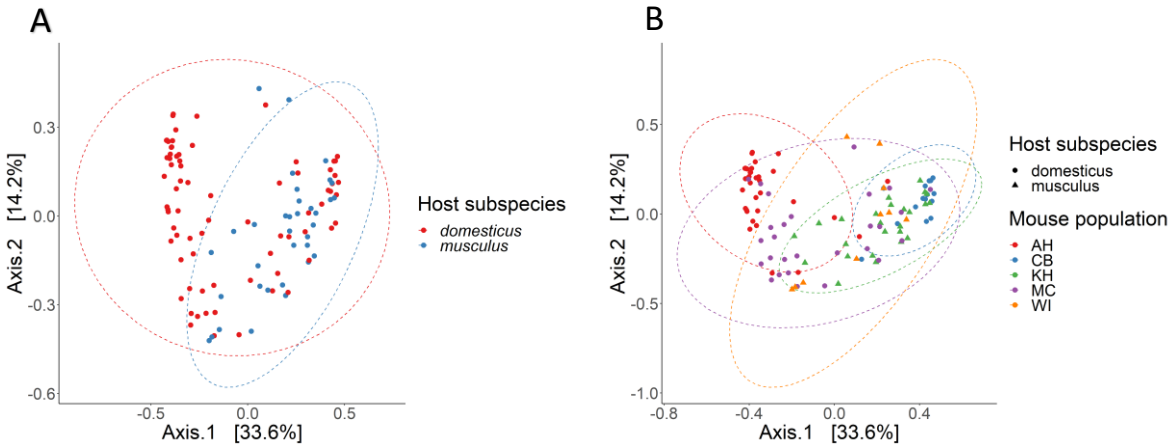


Figure 6. Principal coordinate analysis (PCoA) of weighted Unifrac index applied to *Bacteroides* taxa. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0703$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.3296$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

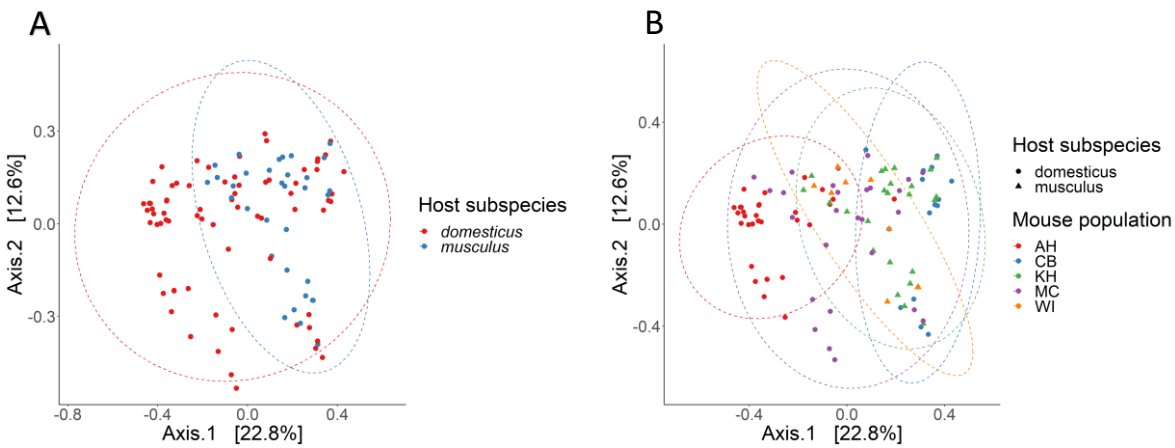


Figure 7. Principal coordinate analysis (PCoA) of unweighted Unifrac index applied to *Bacteroides* taxa. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0541$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.2694$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

I.3 Gut microbiota composition of *mus* and *dom* mice

First, the overall community composition was analyzed at phylum and genus levels, whereby the abundances of the three and six most abundant phyla and genera, respectively, were assessed between *mus* and *dom* mice (Figure 8). At the phylum level both mouse subspecies harbor similar microbial community composition (Figure 8A, Table 2, Suppl. table 2). Significant differences were observed only at the genus level and include *Eisenbergiella*, which is higher in *dom* compared to *mus* mice (Wilcoxon test, p-value=0.0211), and unclassified Bacteroidales (Wilcoxon test, p-value= 0.0035), which are significantly higher in *mus* compared to *dom* mice (Figure 8B, Table 2, Suppl. table 2).

Table 2. Pairwise comparison of relative abundances of major taxa between *mus* and *dom* mice (Wilcoxon test). Unclass: unclassified. Significant p-values (≤0.05) are indicated in bold.

Rank	Taxon	Paired Wilcoxon test (p-value)
Phylum	Bacteroidetes	0.6226
Phylum	Firmicutes	0.1020
Phylum	Proteobacteria	0.7377
Genus	*Alistipes*	0.7395
Genus	*Helicobacter*	0.9770
Genus	*Paraprevotella*	0.0995
Genus	*Eisenbergiella*	**0.0211**
Genus	Unclass. Bacteroidales	**0.0035**
Genus	Unclass. Lachnospiraceae	0.7395

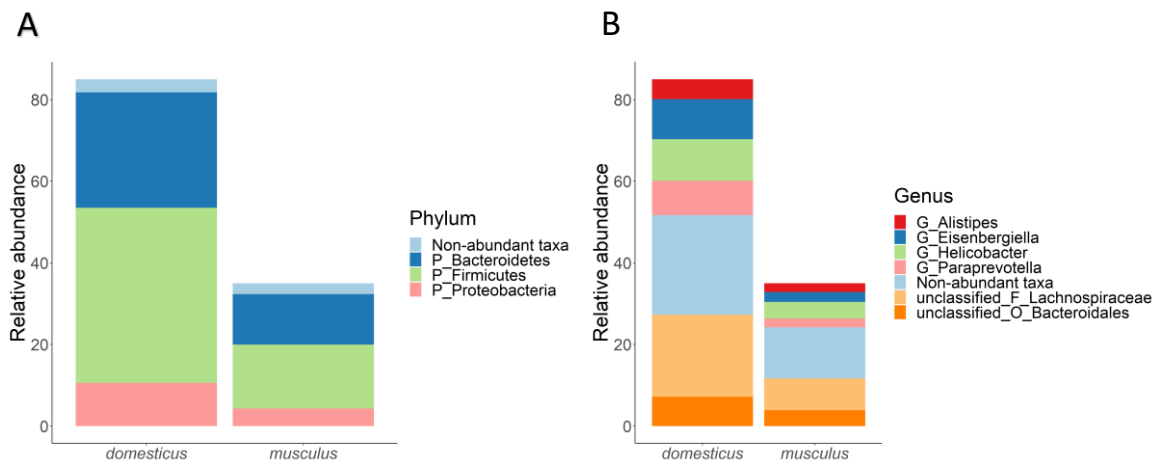


Figure 8. Composition of the gut microbiota in *dom* and *mus* mice. A - Phylum-level abundances of three most abundant phyla. **B** – genus-level relative abundances of six most abundant genera. Color represent different phyla or genera.

Notably, *Bacteroides* was not identified as one of the most abundant genera. However, it is located in the top 20 most prevalent among the whole set of mice used in this study. Data analysis revealed in total 527 *Bacteroides* ASVs. After the removal of singletons, this count however dropped to 33 ASVs. The mean relative abundance of *Bacteroides* ASVs does not differ significantly between two host subspecies (Figure 9A). Figure 9B shows relative abundances of twenty most abundant *Bacteroides* ASVs among *dom* and *mus* mice. ASV 35 is the most abundant for both *dom* and *mus*. However, the mean relative abundance is nearly three-fold higher in *mus* compared to *dom* (Suppl. table 3). Sequence variants 22 and 268 are also more abundant in *mus*, while ASV 19 and ASV 525 are more abundant in *dom* (Suppl. table 3, Figure 9B).

In summary, it was observed that the mean relative abundance of *Bacteroides* ASVs does not significantly differ between the two host subspecies. However, some of the ASVs are differentially abundant in either the *mus* or *dom* subspecies.

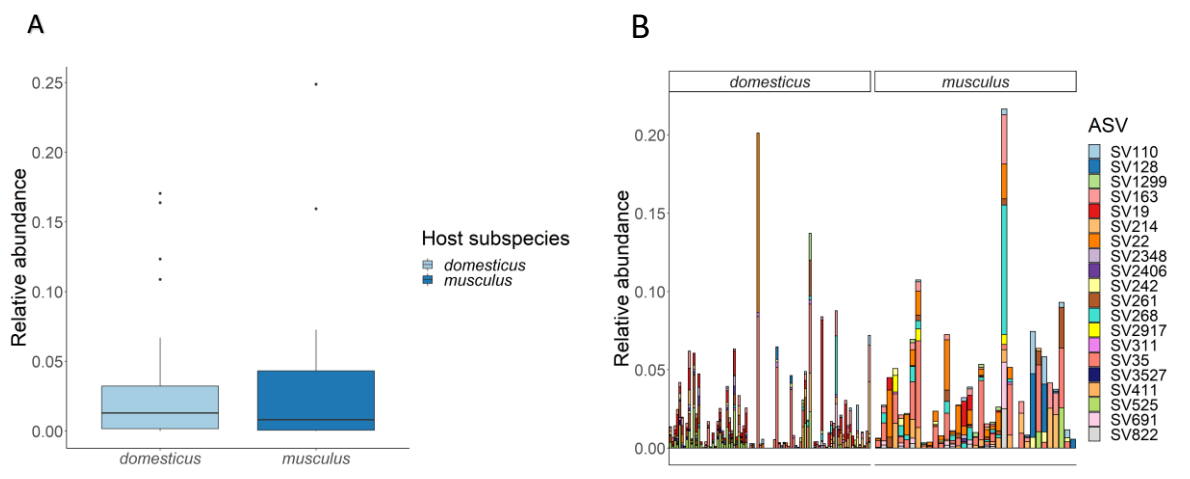


Figure 9. *Bacteroides* relative abundance across *dom* (n=85) and *mus* (n=35) mouse subspecies. A – mean relative abundance of total *Bacteroides* between *dom* and *mus* mice. The calculation of comparisons was performed using Wilcoxon (p-value = 0.8601). **B –** relative abundance of top twenty most abundant *Bacteroides* ASVs. Samples are displayed according to respective mouse subspecies.

II. Indicator genera and indicator ASVs among *mus* and *dom* mice

To identify microbial taxon abundances that differ according to host subspecies, indicator species analysis was performed at the level of genera and ASVs on the full set of 120 *musculus* and *domesticus* mice. The analysis was performed based on relative abundance and presence/absence data. Bacterial taxa specific for a given "habitat" (indicators) were found for both house mouse subspecies, comprising in total seven genera and 29 ASVs that are differentially present and/or abundant in either *mus* or *dom* mouse subspecies.

Indicator species analysis based on the presence/absence of genera, revealed *Lactobacillus, Ureaplasma, Rikenella,* unclassified Bacteroidetes, Clostridiales and Bacteroidales as indicators for *dom* mice (Table 3). The strongest association was shown by *Ureaplasma* and unclassified Clostridiales (association statistics values of 0.7559 and 0.7341, respectively; p-value=0.0074). Among genera identified for *mus* mice, *Alistipes* and *Bacteroides* were the strongest indicators (association statistics values of 0.7921 and 0.7067, respectively).

Table 3. Indicator genera analysis based on presence/absence between *musculus* (*mus*) and *domesticus* (*dom*) mice. Only genera with association statistics ≥ 0.50 are shown. Stat: association statistics; Unclass: unclassified.

Indicator genus	Group	Stat	Adjusted p-value
Lactobacillus	*dom*	0.6913	0.0074
Bacteroides	*mus*	0.7067	0.0074
Ureaplasma	*dom*	0.7559	0.0074
Alistipes	*mus*	0.7921	0.0074
Rikenella	*dom*	0.6299	0.0074
Parasutterella	*mus*	0.5853	0.0290
Helicobacter	*mus*	0.6315	0.0074
Unclass. Bacteroidetes	*dom*	0.5636	0.0433
Unclass. Clostridiales	*dom*	0.7341	0.0074
Unclass. Deltaproteobacteria	*mus*	0.5071	0.0074
Unclass. Porphyromonadaceae	*mus*	0.5316	0.0074
Unclass. Prevotellaceae	*mus*	0.5910	0.0074
Unclass. Lachnospiraceae	*mus*	0.5444	0.0433
Unclass. Bacteroidales	*dom*	0.5611	0.0338
Unclass. Bacteriodetes	*mus*	0.6758	0.0074
Unclass. Marinilabiliaceae	*mus*	0.6680	0.0074

Analysis based on relative abundance data revealed 26 ASVs, 24 of which were identified as indicators for *mus* and only two of the ASVs were indicators for *dom* mice (*Lactobacillus* ASV 4 and unclassified Ruminococcaceae ASV 976) (Table 4). For the presence/absence-based analysis, all identified indicator ASVs were associated with *mus* mice, except for ASV 452 and ASV 2292, belonging to *Ureaplasma* and unclassified Clostridiales, respectively, which were also identified as indicator genera for *dom* mice (Table 4). *Lactobacillus* ASV 93 and *Alistipes* ASV 134 showed the strongest association to *mus* mice in both abundance- and presence/absence-based analyses. Moreover, fourteen out of 17 ASV identified as differentially present in *mus* mice were also differentially abundant for the same mouse subspecies (Table 4, in bold). These include ASVs belonging to *Bacteroides* (ASV 22, 268 and 822), *Alistipes* (ASV 134), *Helicobacter* (ASV 702), *Lactobacillus* (ASV 93) and unclassified Bacteroidales, Prevotellaceae, Marinilabiliaceae, Clostridiales and Deltaproteobacteria.

Table 4. Indicator species analysis based on ASV relative abundances and presence/absence between *musculus* (*mus*) and *domesticus* (*dom*) mice. Only ASV with association statistics ≥ 0.50 are shown. Indicator ASVs common to both, the analysis based on relative abundances and presence/absence, are shown in bold. Stat: association statistics; Unclass: unclassified.

	Indicator ASV	Group	Stat	Adjusted p-value
Relative abundance	*Lactobacillus* ASV 4	*dom*	0.6828	0.0298
	***Bacteroides* ASV 22**	*mus*	0.6737	0.0141
	Bacteroides ASV 35	*mus*	0.7246	0.0298
	Unclass. Bacteroidales ASV 61	*mus*	0.7561	0.0141
	Unclass. Bacteroidales ASV 80	*mus*	0.5309	0.0141
	***Lactobacillus* ASV 93**	*mus*	0.8221	0.0141
	***Alistipes* ASV 134**	*mus*	0.7873	0.0141
	Unclass. Bacteroidales ASV 231	*mus*	0.5345	0.0141
	Alistipes ASV 250	*mus*	0.6848	0.0262
	***Bacteroides* ASV 268**	*mus*	0.6450	0.0221
	Unclass. Bacteroidales ASV 302	*mus*	0.7305	0.0141
	Unclass. Porphyromonadaceae ASV 321	*mus*	0.6288	0.0262
	Fusicatenibacter ASV 403	*mus*	0.5845	0.0467
	***Helicobacter* ASV 702**	*mus*	0.6835	0.0141
	Unclass. Prevotellaceae ASV 724	*mus*	0.6057	0.0141
	Unclass. Ruminococcaceae ASV 976	*dom*	0.7348	0.0298
	***Bacteroides* ASV 822**	*mus*	0.5373	0.0211
	Parasutterella ASV 858	*mus*	0.6277	0.0352
	Unclass. Porphyromonadaceae ASV 1117	*mus*	0.5159	0.0141
	Odoribacter ASV 1131	*mus*	0.5992	0.0298
	Unclass. Bacteroidetes ASV 1313	*mus*	0.6528	0.0141
	Unclass. Marinilabiliaceae ASV 1563	*mus*	0.6676	0.0141
	Unclass. Bacteroidales ASV 2209	*mus*	0.6325	0.0141
	Unclass. Clostridiales ASV 2569	*mus*	0.5345	0.0141
	Unclass. Deltaproteobacteria ASV 2579	*mus*	0.5071	0.0141
	Unclass. Lachnospiraceae ASV 4148	*mus*	0.5160	0.0221
Presence/absence	***Bacteroides* ASV 22**	*mus*	0.7067	0.0081
	Unclass. Bacteroidales ASV 61	*mus*	0.7413	0. 0081
	***Lactobacillus* ASV 93**	*mus*	0.7518	0. 0081
	Unclass. Bacteroidetes ASV 127	*mus*	0.6758	0. 0081
	***Alistipes* ASV 134**	*mus*	0.7921	0.0081
	Unclass. Bacteroidales ASV 231	*mus*	0.5345	0.0081
	***Bacteroides* ASV 268**	*mus*	0.6315	0. 0081
	Ureaplasma ASV 452	*dom*	0.7559	0. 0081
	***Helicobacter* ASV 702**	*mus*	0.6315	0. 0081
	Unclass. Prevotellaceae ASV 724	*mus*	0.5910	0. 0081
	***Bacteroides* ASV 822**	*mus*	0.5491	0. 0081
	Unclass. Bacteroidetes ASV 1313	*mus*	0.6458	0.0081
	Unclass. Marinilabiliaceae ASV 1563	*mus*	0.6680	0.0081

Unclass. Bacteroidales ASV 2209	*mus*	0.6325	0.0081
Unclass. Clostridiales ASV 2292	*dom*	0.7341	0. 0081
Unclass. Clostridiales ASV 2569	*mus*	0.5345	0. 0081
Unclass. Deltaproteobacteria ASV 2579	*mus*	0.5071	0. 0081

II.1 Indicator *Bacteroides* ASVs

To identify candidate *Bacteroides* ASV associated to either *mus* or *dom* mouse subspecies, indicator species analysis was performed on relative abundance and presence/absence of ASVs belonging only to the *Bacteroides* genus. This yielded a total of nine ASVs differentially present and/or abundant in either *mus* or *dom* mice (Table 5). For the analysis based on relative abundance data, all six identified *Bacteroides* ASVs were associated to *mus* mice, and the most abundant ASV 35 (Figure 10A) showed the strongest association (association statistics of 0.7246, p-value = 0.0033). Analysis based on presence/absence data revealed ASV 22 and ASV 19 to be strong indicators for *mus* and *dom* mice, respectively (Table 5). Moreover, *Bacteroides* ASV 19 was the most abundant among the ASVs associated with *dom* mice (Figure 10B). Furthermore, all six indicator ASVs identified based on relative abundance data, also showed significant associations with *mus* mice based on their presence/absence.

Interestingly, it seems that KH mouse population contributed the most to the overall relative abundances of all *Bacteroides* ASVs associated to *mus* mice, with the exception of ASV 411 and ASV 35 (Figure 11A). Similar patterns are present for ASV 525 and ASV 2348 differentially present in *dom* mice (Figure 11B).

Table 5. Indicator species analysis based on *Bacteroides* ASV abundance and presence/absence between *musculus* (*mus*) and *domesticus* (*dom*) mice. Indicator ASVs common to both, the analysis based on relative abundances and presence/absence, are shown in bold. Stat: association statistics; Unclass: unclassified.

	Indicator *Bacteroides* ASV	Group	Stat	Adjusted p-value
Relative abundance	**ASV 22**	*mus*	0.6737	0.0016
	ASV 35	*mus*	0.7246	0.0033
	ASV 268	*mus*	0.6450	0.0016
	ASV 411	*mus*	0.5417	0.0148
	ASV 691	*mus*	0.4939	0.0059
	ASV 822	*mus*	0.5373	0.0033
Presence/absence	ASV 19	*dom*	0.8165	0.0201
	ASV 22	*mus*	0.7067	0.0017
	ASV 35	*mus*	0.6315	0.0023
	ASV 268	*mus*	0.6839	0.0049
	ASV 411	*mus*	0.5437	0.0123
	ASV 525	*dom*	0.6013	0.0049
	ASV 691	*mus*	0.4953	0.0049
	ASV 822	*mus*	0.5491	0.0034
	ASV 2348	*dom*	0.4947	0.0201

In summary, the results of indicator species analysis revealed the genera *Ureaplasma* and unclassified Clostridiales together with respective ASV 452 and 2292 to be differentially present in *dom* mice. Further, *Lactobacillus* ASV 4 and unclassified Ruminococcaceae ASV 976 were differentially abundant in *dom* mice. For *mus* mice, an overlap between differentially present and differentially abundant indicator ASVs was observed, including ASVs belonging to the *Alistipes*, *Helicobacter*, *Lactobacillus* and *Bacteroides* genera. When limiting the analysis to *Bacteroides*, six ASVs were differentially present and abundant in *mus* mice, representing interesting candidates for the further characterization.

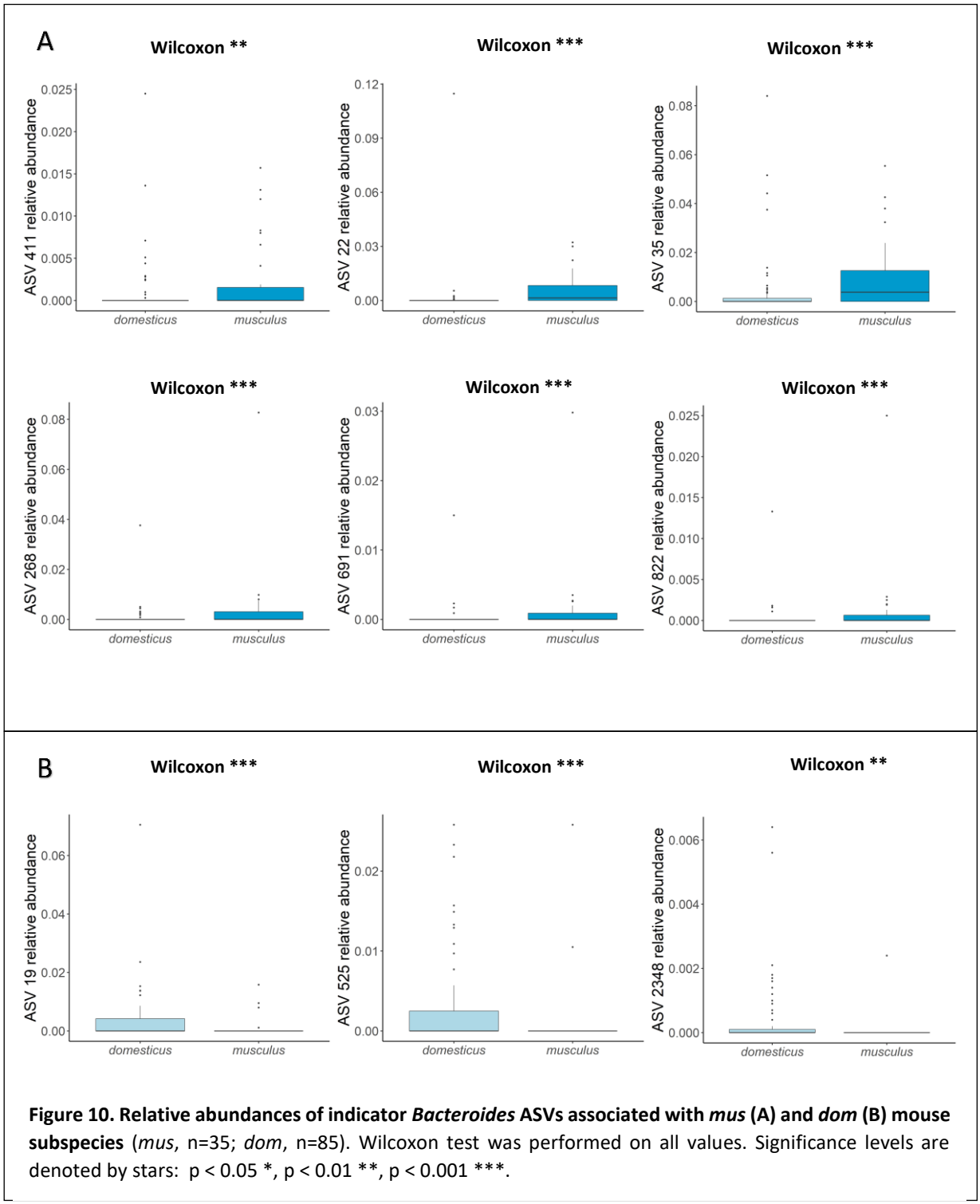


Figure 10. Relative abundances of indicator *Bacteroides* ASVs associated with *mus* (A) and *dom* (B) mouse subspecies (*mus*, n=35; *dom*, n=85). Wilcoxon test was performed on all values. Significance levels are denoted by stars: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

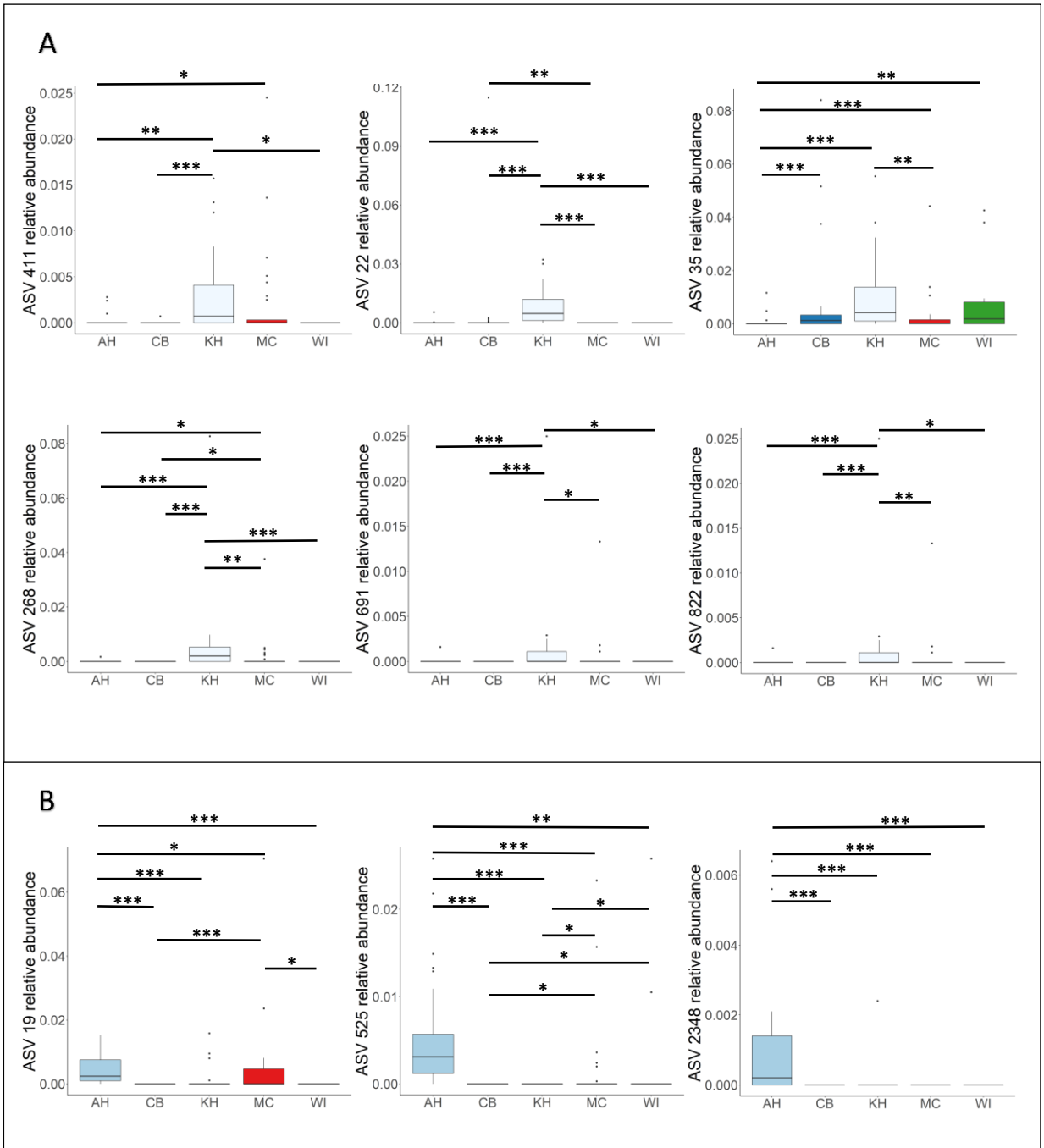


Figure 11. Relative abundances of indicator *Bacteroides* ASVs associated with *mus* (A) and *dom* (B) mouse subspecies across different mouse populations (AH, n=33; CB, n=12; MC, n=29; KH, n=25; WI, n=10). Wilcoxon test was performed on all values. Significance levels are denoted by stars: $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***. Abbreviations for the mouse populations are given in Table 8.

III. Taxonomic identification of *Bacteroides* ASVs

In this section the aim was to identify *Bacteroides* species corresponding to the candidate ASVs identified above. Given that the ASVs represent only a small portion (around 300 bp) of the full length 16S rRNA gene, it is difficult to obtain feasible taxonomic classification of indicator *Bacteroides* species. Thus, in order to gain deeper insight into the taxonomy of *Bacteroides*, a genus specific primer pair was used to amplify, clone and sequence a longer portion (approx. 700 bp) of the 16S rRNA gene (Table 11).

Ten samples were selected: two of each mouse population (AH, CB, MC, KH, and WI) representing the highest diversity in *Bacteroides* ASVs. This strategy was taken in an attempt to cover the maximum number of ASV-level taxa belonging to *Bacteroides* in the set of mice used in this study. For each of the samples, 96 clones containing a desired insert were isolated and sequenced. After quality check, 626 clone sequences were selected to be classified using the RDP classifier (see Methods).

The results obtained by comparison of the clone sequences to the 16S rRNA gene database are presented in the Table 6, for which only S_ab scores equal to or higher that 0.70 are shown. Most of the Sanger reads were classified as *B. acidifaciens* (435 sequences), showing the lowest average S_ab score of 0.86 and the highest of 0.95. *B. uniformis* was the closest match to 110 and *B. rodentium* to 37 sequences with average S_ab scores ranging from 0.73-0.82 and 0.76-0.78, respectively. Two clone sequences were identified as *B. stercorirosoris* (average S_ab score of 0.80) and 16 as *Bacteroides* sp. (average S_ab score of 0.90) (Table 6).

Table 6. Summary of clone sequences classification by RDP. Only S_ab ≥ 0.70 are shown; n – number of sequences.

Sample	Number of clones	Average S_ab score	Closest matches
AH 694 (n=55)	46	0.90	*B. acidifaciens*
	6	0.90	*Bacteroides* sp.
	3	0.82	*B. uniformis*
AH 766 (n=68)	33	0.94	*B. acidifaciens*
	34	0.79	*B. uniformis*
	1	0.80	*B. stercorirosoris*
CB 010 (n=71)	61	0.92	*B. acidifaciens*

	10	0.90	*Bacteroides* sp.
CB 150 (n=75)	41	0.90	*B. acidifaciens*
	34	0.78	*B. rodentium*
KH 365 (n=34)	13	0.87	*B. acidifaciens*
	21	0.73	*B. uniformis*
KH 051 (n=42)	31	0.86	*B. acidifaciens*
	9	0.73	*B. uniformis*
	2	0.77	*B. rodentium*
MC 945 (n=60)	60	0.87	*B. acidifaciens*
MC 362 (n=66)	66	0.95	*B. acidifaciens*
WI 270 (n=45)	33	0.91	*B. acidifaciens*
	11	0.79	*B. uniformis*
	1	0.76	*B. rodentium*
WI 296 (n=79)	51	0.92	*B. acidifaciens*
	27	0.78	*B. uniformis*
	1	0.79	*B. stercorirosoris*

Next, indicator *Bacteroides* ASV sequences were classified by performing alignments to Sanger fragments (see Methods, section V) and choosing the best match based on the nucleotide identity percentages of these alignments. Four out of six indicator ASVs align to the clone sequences classified as *B. acidifaciens*, with the highest nucleotide identity for ASV 35 (99.99%) and the lowest for ASV 691 (97.74%) (Table 7). Moreover, ASV 22 showed the closest match to *B. uniformis* and ASV 411 to *B. massiliensis*. However, the classification of the clone sequence as *B. massiliensis* showed less than 70% similarity (Table 7).

This taxonomic classification of the candidate ASVs using the alignments to 626 clone sequences enabled the preliminary identification of the *Bacteroides* species - indicators for *mus* mice. Based on these results, the isolation of *B. acidifaciens, B. massiliensis* and *B. uniformis* from the cecal contents of mice was followed and described in Chapter II.

Table 7. Classification of the ASVs belonging to *Bacteroides* genus. Indicator ASVs, respective classification and nucleotide identities are shown in bold. Nucleotide identity percentage corresponds to alignments between each ASV-clone sequence pairs.

***Bacteroides* ASV**	**Closes match**	**Nucleotide identity (%) ***
ASV 19	*B. acidifaciens*	98.85
ASV 22	***B. uniformis***	**99.67**
ASV 35	***B. acidifaciens***	**99.99**
ASV 110	*B. acidifaciens*	98.70
ASV 128	*B. acidifaciens*	99.67
ASV 163	*B. acidifaciens*	98.70
ASV 214	*B. acidifaciens*	98.70
ASV 242	*B. acidifaciens*	98.38
ASV 261	*B. uniformis*	99.67
ASV 268	***B. acidifaciens***	**98.38**
ASV 311	*B. acidifaciens*	98.70
ASV 411	***B. massiliensis***	**95.22**
ASV 525	*B. uniformis/B. faecis*	99.34
ASV 675	*B. acidifaciens*	98.70
ASV 691	***B. acidifaciens***	**97.74**
ASV 822	***B. acidifaciens***	**98.05**
ASV 1299	*B. massiliensis*	93.97
ASV 1700	*B. uniformis*	96.13
ASV 2309	*B. massiliensis*	93.65
ASV 2348	*B. uniformis*	84.40
ASV 2406	*B. massiliensis*	86.62
ASV 2917	*B. massiliensis*	95.22
ASV 3215	*B. acidifaciens*	97.39
ASV 3527	*B. uniformis*	99.02
ASV 5811	*B. massiliensis*	86.03
ASV 5872	*B. sartorii*	84.76
ASV 7409	*B. acidifaciens*	98.38
ASV 9546	*B. massiliensis*	86.62
ASV 15996	*B. acidifaciens*	98.38
ASV 17568	*B. uniformis/B. faecis*	98.69

ASV 17571	*B. acidifaciens*	98.38
ASV 32545	*B. acidifaciens*	98.05
ASV 49106	*B. rodentium*	95.80

* Nucleotide identity between *Bacteroides* ASV and matching clone sequence

Discussion

The intestinal microbiota provides a wide range of benefits to their mammalian hosts through a number of physiological functions (Nakata *et al.*, 2017; Pickard *et al.*, 2017). Members of the intestinal microbial community belonging to *Bacteroides* are of particular interest due to their important role in host health and evidence of *Bacteroides* abundance being dependent on host genetic factors. Hence, characterizing the patterns of within-bacterial species diversity across different host subspecies may provide an opportunity to understand the forces that shape *Bacteroides* variation and are potentially involved in adaptation to the host. The present chapter first aimed to describe the overall composition and diversity of gut bacterial communities, and *Bacteroides* genus members in particular across five outbred *dom* and *mus* house mouse lines. Second, the candidate *Bacteroides* ASVs differentially present and/or abundant in either *mus* or *dom* mouse subspecies were identified and classified taxonomically. This study accordingly provides a first comparison of the *Bacteroides* genus between the *mus* and *dom* house mouse subspecies.

The first finding of this study is that the overall gut bacterial diversity as well as *Bacteroides* genus diversity are similar in the *dom* and *mus* subspecies, although small differences were also detected in the composition of their microbiota. Similar results were obtained in the previous study of our group, where they detected non-significant difference in gut bacterial community composition between wild-caught *mus* and *dom* mice (Wang, Kalyan, Steck, Turner, Harr, Künzel, Vallier, Häsler, Franke, H. H. Oberg, *et al.*, 2015). Nonetheless, we found Unclassified Bacteroidales to be significantly more abundant in *mus* and *Eisenbergiella* in *dom* mice. The genus *Eisenbergiella* belongs to butyrate-producing gut bacteria, which provides the major energy source for the colonic epithelium in healthy hosts (Roediger, 1980). Also, Bao *et al*. (2018) detected an increase in *Eisenbergiella* abundance in mice infected with cystic echinococcosis. However, the candidate unclassified Bacteroidales genus belongs

to a broad taxonomic group with numerous species and diverse behavior, and it is difficult to make further conclusions, as it is only classified at the class level.

On the other hand, geography seems to influence not only the diversity of overall gut bacterial communities and that of the *Bacteroides* genus, but also its structure. These results were largely expected, since the geography is reported to considerably influence gut microbiota variability in humans (Yatsunenko, Federico E Rey, *et al.*, 2012) and in mice (Linnenbrink *et al.*, 2013). In their study, Linnenbrink *et al*. found that the gut community structure of *dom* wild mice differs significantly between the individual geographical locations where the mice were caught. In this case, geography might be viewed as a complex of different environmental factors, which does not apply to the present study. Because all mouse populations used here were wild-derived and maintained as outbred colonies in a lab facility, the environmental factors such as weather or availability of food are not expected to play a role. Also, all the mice received the same diet. However, it is expected that these mice would still maintain gut community structures compositionally close to the original wild state (Moeller *et al.*, 2018).

Another finding of the present study are the strong indicator genera for *mus* mice, *Alistipes* and *Bacteroides*. Interestingly, the overall relative abundance of *Bacteroides* does not differ significantly between *dom* and *mus* mice, which was also shown by Wang *et al.* (2015). Notably, the relative abundance of *Bacteroides* was shown to be relatively low in the present set if mice, although it was still highly prevalent. Previous studies revealed *Bacteroides* to be among the most abundant genera in mice (Linnenbrink *et al.*, 2013; Wang, Kalyan, Steck, Turner, Harr, Künzel, Vallier, Häsler, Franke, H. H. Oberg, *et al.*, 2015). On the other hand, *Bacteroides* was highly prevalent (85.1%), but low abundant (2.4%) in the study performed on a set of 101 healthy mice (Wang *et al.*, 2019). It should be however noted that differences in primer sets, etc. can lead to systematic differences between 16S rRNA gene sequencing studies (Hiergeist *et al.*, 2016). Nonetheless, the indicator species analysis performed on *Bacteroides* yielded six indicators differentially abundant in *mus* mice with ASV 35, classified as *B. acidifaciens,*

showing the strongest association. These observations suggest that these host species-specific differences might be reliable, because they are consistent across different geographic locations.

Intriguingly, the same *Bacteroides* ASV 35 was detected in an ongoing independent genetic mapping study conducted by Shauni Doms (unpublished). Employing an association mapping approach, she identified host genomic regions that influence gut microbial traits in a set of 320 F2 hybrids from the intercross between wild-derived inbred mouse strains originated from the hybrid zone of *mus* and *dom* subspecies. The results show the same *Bacteroides* ASV 35 to be more abundant in the hybrids homozygous in *mus* allele at the identified genomic locus. Despite the genetic differences between the mice populations used in both independent studies, the results yielded the same host species-specific difference in *Bacteroides* ASV 35 abundance. The overlap in the results suggests that there is evidence of a genetic basis behind this host-microbe association.

In conclusion, the findings reveal that gut bacterial- and *Bacteroides* genus diversity appear to be similar in *dom* and *mus* mouse subspecies. Host subspecies in this case seems to play a relatively minor role in gut community structure. However, despite being maintained under the same housing conditions, the geography of their origin influence the variability of the respective gut microbiota communities. The detection of strong *Bacteroides* indicators for *mus* mice suggests that these host-*Bacteroides* associations are consistent across different geographic locations and represent promising candidates for further characterization across *dom* and *mus* mice. Moreover, the host-microbe interaction involving *Bacteroides* ASV 35, which might have diverged since the common ancestor of the *dom* and *mus* subspecies, is of particular interest.

Methods

I. Sample collection

The panel of the mice used in this study comprises in total 120 male animals belonging to *M. musculus musculus* and *M. musculus domesticus* subspecies, derived from five geographic locations (Figure 1 and Table 8). All the mice colonies are wild-derived, outbred (except for WI mice) and maintained at the MPI facility in Plön.

Mice were dissected using different set of sterilized utensils for each body site (skin, peritoneal wall and cecum). The utensils were sterilized in dry glass bead sterilizer prior to each individual dissection. The caecum content samples were collected from 120 adult male mice (Table 9) and divided in two parts. The part to be used for bacterial cultivation assays (Chapter II) was stored in Brain Heart Infusion (BHI) medium (Carl Roth) with 20% glycerol at -80°C. The part to be used for DNA extraction was placed in RNA*later* solution (Thermo Fisher Scientific) for 24 hours, followed by centrifugation at 5000 rpm and 6°C for 10 min to remove the stabilizing solution. Afterwards, the cecum content samples were stored at -20°C until processing. Organ removal for scientific purposes was performed according to the German animal welfare law (Permit V 312-72241.123-34).

Table 8. Mouse populations used in this study.

Abbreviation	Subspecies	Origin	Number of mice
CB	*domesticus*	Cologne/Bonn, Germany	23
MC		Massif Central, France	29
AH		Ahvaz, Iran	33
KH	*musculus*	Almaty, Kazakhstan	25
WI		Vienna, Austria	10

Table 9. Summary of the samples taken from the mice, storage conditions and intended purpose.

Organ	Storage solution	Storage temperature	Purpose
Cecum content 1	--	- 20°C	16S rRNA gene profiling
Cecum content 2	BHI medium with 20% glycerol	- 80°C	Cultivation

II. Microbial DNA extraction for 16S rRNA gene profiling

The DNA was extracted from cecum content using the QIAamp DNA Stool Mini Kit from Qiagen. Each sample was quantified individually using a NanoDrop (Thermo Fisher Scientific).To avoid DNA degradation, the frozen cecum content samples were placed in ice. Using a spatula, bits of the frozen sample were scraped and used for the extraction procedure. Each extracted sample had around 150 mg of cecum content. Then the samples were disrupted and homogenized in 1,4 ml of ASL buffer (Qiagen), placed in a Lysing Matrix E tube (MPBio) using the Precellys 24 with run conditions of 3 x 15 s at speed 6500 rpm. To lyse the bacteria and increase the DNA yield, the suspension was heated up to 95°C during 10 min. The lysates were incubated with the InhibitEX reagent (Qiagen) to remove inhibitors, treated with proteinase K, washed with ethanol and the DNA was eluted following the manufacturer's instructions. For each extraction, the negative extraction control was included. Extracted DNA was stored at -20°C.

III. 16S rRNA gene profiling for the cecal microbiota

III.1 PCR and Next Generation Sequencing

To amplify the V1-V2 regions of the 16S rRNA gene from the DNA of cecum content samples, the primers 27F and 338R were used. The primers were barcoded to allow multiplexing. PCR reactions contained 10,25 µl H_2O, 5µL buffer, 0,50 µl dNTPs, 0,25 µl Taq polymerase (Phusion High-Fidelity DNA Polymerase - Thermo Scientific), 4 µl of 2 µM Primer and 1 µl DNA template or negative extraction control. The amplification program is presented in Table 10.

Table 10. PCR program for 16S rRNA gene sequencing

Temperature	Time	Number of cycles
98°C	30 sec	
98°C	9 sec	× 30
55°C	30 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

After the amplification, PCR products were quantified on the gel using the GelDoc XR+ (BioRad). The samples were then mixed in subpools with equal amounts of DNA. The subpools were purified by gel extraction using the MiniElute kit (Qiagen). Purified subpools were quantified with the fluorescence NanoDrop (Thermo Scientific) and mixed in a final pool so that each sample has the same final concentration. The library was sequenced on an Illumina MiSeq machine using the v2 kit with 2x300 bp reads.

III.2. Sequenced data processing

From the BCL files containing base calls obtained from sequencing machine, the demultiplexed fastq files were generated. For this, The Illumina bcl2fastq2 Conversion Software v2.20 was used, allowing 1 mismatch in the barcodes. Reads quality was checked with the FastQC tool, version 0.11.6 (Andrews, 2010).

Single-nucleotide resolution inference of sample sequences using DADA2

Further data processing including trimming and quality filtering, was performed in R (v. 3.5.0) (R Core Team, 2018) using DADA2 software package (Callahan *et al.*, 2016). The data analysis pipeline is available on the DADA2 website. The customized version of the pipeline used in this study follows below:

1. Quality control: filtering and trimming of reads.
   - `filterAndTrim`: the reads were trimmed at the first base with a quality below 5. Only trimmed reads that are 200 bp or longer were kept.

2. Dereplication: grouping the amplicon reads with same sequence into unique sequences.
   - `derepFastq`: filtered reads were dereplicated using the standard parameters.

3. Sample inference and merging the forward and reverse reads.
   - `dada`: using the core sequence-variant inference algorithm, the sequencing errors were removed from the dereplicated amplicon reads data, and the amplicon sequence variants (ASV) were inferred for each sample.
   - `mergePairs`: the denoised forward and reverse reads were merged. One mismatch between forward and reverse reads was allowed, only merged reads that are between 300 and 350 bp and that have an overlap of at least 100 bp were kept.

4. Sequence table construction
   - `makeSequenceTable`: sequence table (a table with the ASV and respective abundances for each sample) was constructed from the provided list of samples.

5. Chimera detection and removal.
   - `removeBimeraDenovo`: the chimeras were identified by consensus method across samples, meaning that each sample in the sequence table was checked for exact bimeras and a consensus decision on each ASV is made. Sequence variants identified as bimeric were removed.

6. Taxonomic classification.
   - `assignTaxonomy`: taxonomical classification was carried out using the Ribosomal Database Project (RDP) training set 14 (Cole *et al.*, 2014).

Write the data to the files

The ASV table (including unique sequence variants abundances per sample), the table with the ASV sequences and taxonomic classification were saved as TSV (tab-separated values) files for further community analysis.

III.3 Microbial community analysis

The microbial community analysis, including the alpha and Beta diversity measures, was performed in R using the Phyloseq package (Mcmurdie and Holmes, 2013). The output tables produced by DADA2 pipeline were merged to produce a phyloseq object (the pipeline is available on the DADA2 website). All the plots were generated by ggplot2 package (Wickham, 2016).

Alpha diversity and beta diversity indices were calculated based on the ASVs distribution. Alpha diversity measures were calculated using the function `estimate_richness` and plotted using the function `plot_richness`. Comparison of alpha diversity indices between mouse subspecies and between mouse populations, was performed using a Wilcoxon test. The function `ordinate` was used to perform Principal Coordinates Analysis (PCoA) on weighted and unweighted Unifrac, Bray-Curtis and Jaccard indices. Multivariate analysis of dissimilarity was performed using Adonis with 10000 permutations.

IV. Indicator species analysis

The indicator species analysis was performed using R package Indicspecies (De Cáceres and Legendre, 2009). The indicator values (IndVal) method (Dufrêne and Legendre, 1997) implemented in the `multipatt` function was applied to the ASV data. The statistical significance of this relationship was tested using 10000 permutations. The p-values were adjusted using "BH" method (Benjamini and Hochberg, 1995).

V. Clone libraries

Ten samples were selected for cloning: two of each mouse population (AH, CB, MC, KH, and WI) representing the highest diversity of *Bacteroides* ASVs.

V.1 Cloning procedure

Initial PCR

To amplify *Bacteroides* an approx. 700 bp fragment 16S rRNA gene from the DNA of cecum content samples, genus-specific primers Bac 32F and Bac 708R (Table 11) were used. PCR reactions contained 3.6 µl H_2O, 5 µl of Multiplex mixture (MP), 0.2 µl of 2 µM Primer and 1 µl DNA template. The amplification program is presented in Table 12.

Table 11. *Bacteroides* genus-specific primers used to sequence nearly full-length 16S rDNA genes (Bernhard and Field, 2000).

Primers	Sequence	Amplicon size (bp)
Bac32F	AACGCTAGCTACAGGCTT	676
Bac708R	CAATCGGAGTTCTTCGTG	

Table 12. PCR program for initial PCR

Temperature	Time	Number of cycles
95°C	15 min	
94°C	30 sec	× 35
58°C	90 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

Cloning and transformation

For the cloning procedure CloneJET PCR Cloning Kit (Thermo Scientific) was used. Blunting reaction and ligation were performed according to the sticky-end cloning protocol provided with the kit. For the transformation, 1 µl of ligation mixture was mixed with 25 µl of NEB 5-alpha *E. coli* competent cells (New England Biolabs). The mixture was first incubated on ice for 30 min, and then *E. coli* cells were transformed by heat shock at 42°C for 30 sec. The transformation mixture was placed on ice for 2 min, mixed with 125 µl of SOC medium and subsequently incubated at 37°C for 1 hour 30 min. Forty microliters were plated on Luria-Bertani (LB) agar medium, containing 100 µg/ml ampicillin, and incubated at 37°C for overnight growth.

Transformant confirmation

To confirm the presence of the cloned fragment, colony screening PCR was applied, using pJET1.2 forward and reverse primers. First, colonies were picked from the LB agar plates with the help of the pipet tip, and resuspended in 10 µl of H_2O. PCR reactions contained 6.9 µl of H_2O, 7.5 µl of MP, 0.3 µl of each primer and 1 µl of the colony suspension. The amplification program is presented in Table 13.

Table 13. PCR program for transformant confirmation

Temperature	Time	Number of cycles
95°C	15 min	
94°C	30 sec	× 35
60°C	90 sec	
72°C	90 sec	
72°C	10 min	
12°C	∞	

V.2 Sanger sequencing

ExoSAP cleaning and cycle sequencing

To remove excess primers and dNTPs, PCR products were subjected to enzymatic cleaning using ExoSAP kit (New England Biolabs). Each reaction contained 1.215 µl of H_2O, 0.06 µl of Exo I, 0.225 µl of SAP and 5 µl of PCR product. The treatment was carried out at 37°C for 20 min followed by an inactivation of the enzymes at 80°C for 20 min.

Cycle sequencing reactions were performed using Bac 32F and Bac 708R primers (Table 11). Each reaction contained 6.25 µl of H_2O, 0.5 µl of 2 µM Primer (forward or reverse), 1.75 µl of sequencing buffer (Thermo Scientific), 0.5 µl of BigDye (Thermo Scientific) and 1 µl of PCR product. The amplification program is presented in Table 14.

Table 14. Cycle sequencing PCR program

Temperature	Time	Number of cycles
96°C	1 min	
96°C	10 sec	× 30
55°C	15 sec	
60°C	4 min	
12°C	∞	

Extension product purification

Enzymatically cleaned PCR products were purified using BigDye® XTerminator™ purification kit (Applied Biosystems). Each reaction contained 45 µl of SAM solution, 10 µl of XTerminator solution and 10 µl of PCR product. The mixture was vortexed for 30 min at maximum speed to capture and immobilize the unwanted components. After vortexing, the reactions were centrifuged at 1000 x g for 2 min.

Sequencing and data processing

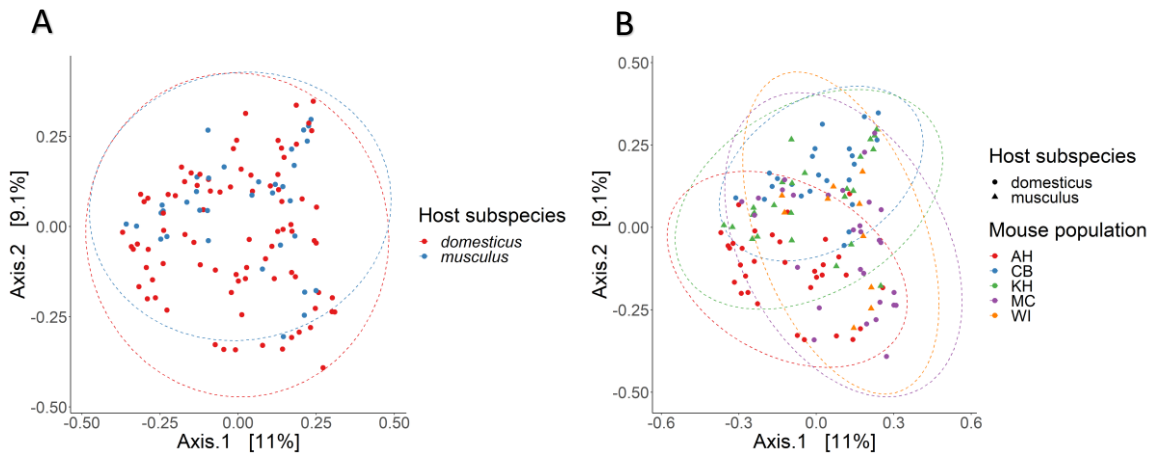
A 700 bp portion of 16S rRNA gene of the confirmed transformants was sequenced using classical Sanger sequencing. The reactions were performed using ABI Dye (v.3.1) sequencing chemistry (Applied Biosystems) and run on an ABI 3730 automated sequencer. Sequenced raw data AB1 files were visualized and edited using Geneious (v.11.0) (Kearse *et al.*, 2012), forward and reversed reads were aligned using progressive pairwise Geneious aligner. The sequences were identified by the online RDP classifier (Cole *et al.*, 2014).

V.3 Taxonomic classification of *Bacteroides* ASVs

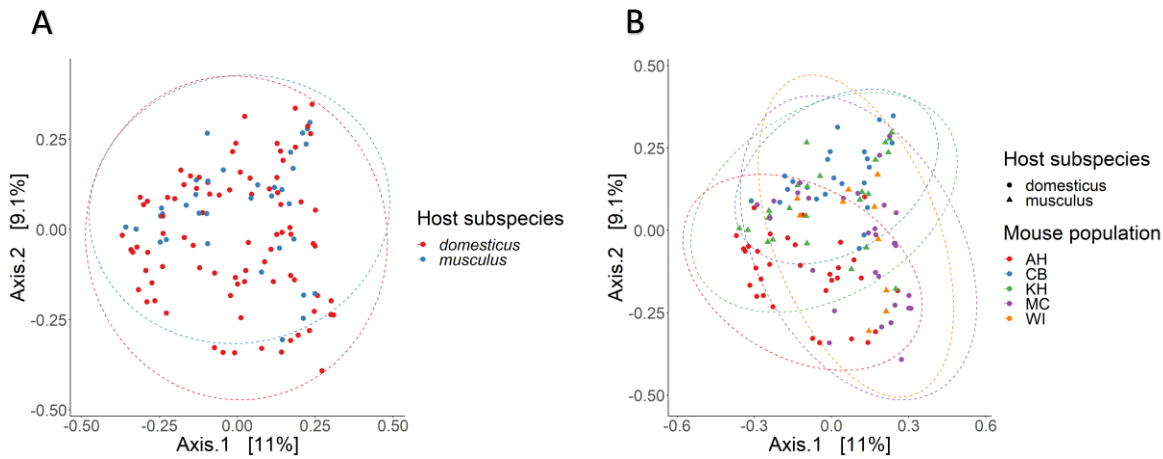
To obtain feasible taxonomic classification of candidate *Bacteroides* ASVs, sequenced ASV fragments (300 bp) were aligned to longer Sanger reads (750 bp) using progressive pairwise Geneious aligner. The best matches of Sanger to ASVs reads were then selected and matched to the table with taxonomic classification of clone sequences from the previous section.

Supplementary material

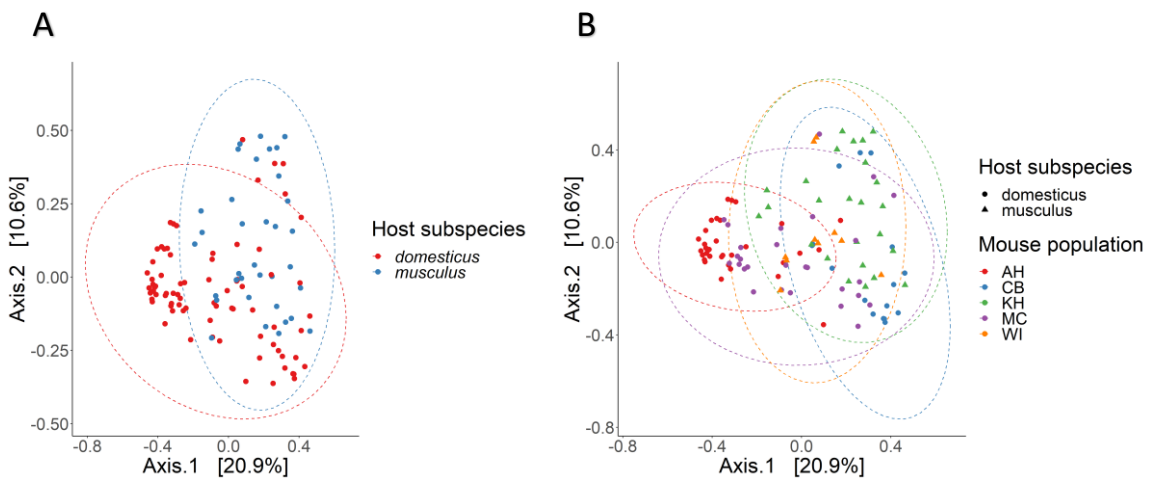
I. Supplementary figures



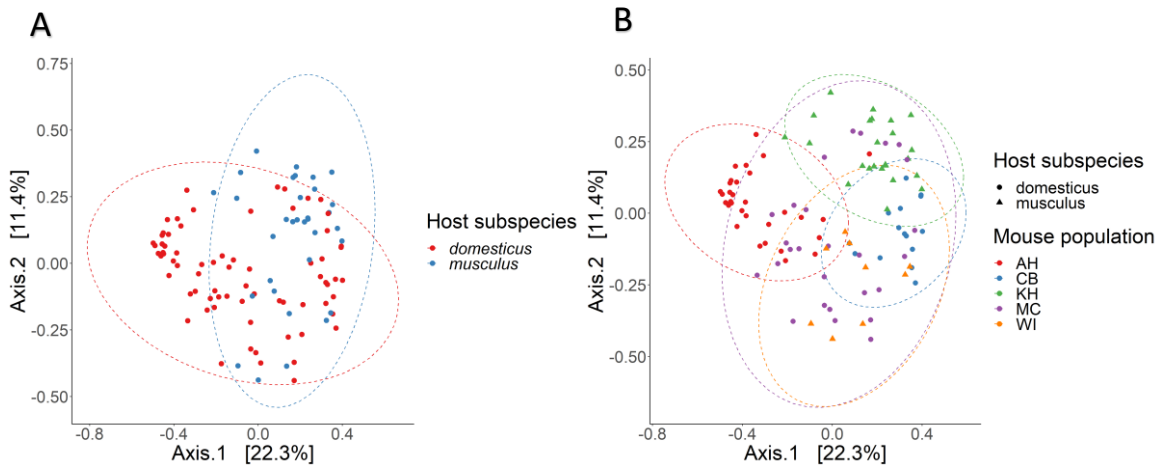
Supplementary figure 1. Principal coordinate analysis (PCoA) of Bray-Curtis index. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0270$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.1425$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.



Supplementary figure 2. Principal coordinate analysis (PCoA) of Jaccard index. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0194$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.0996$, $p = 9.999 \times 10^{-5}$), based on 10000 permutations. Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC –



Supplementary figure 3. Principal coordinate analysis (PCoA) of Bray-Curtis index applied to *Bacteroides* taxa. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0680$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.3149$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.



Supplementary figure 4. Principal coordinate analysis (PCoA) of Jaccard index applied to *Bacteroides* taxa. PCoA between **A -** *musculus* and *domesticus* host subspecies (Adonis, $r^2 = 0.0526$, $p = 9.999 \times 10^{-5}$) and **B -** between mouse populations (Adonis, $r^2 = 0.2412$, $p = 9.999 \times 10^{-5}$). Abbreviations of geographic locations: AH – Ahvaz, Iran; CB – Cologne/Bonn, Germany; KH - Almaty, Kazakhstan; MC – Massif Central, France; WI – Vienna, Austria.

II. Supplementary tables

Supplementary table 1. Summary statistics of beta-diversity among groups of mice. P-values and r^2 were calculated using analysis of dissimilarity (Adonis) with 10000 permutations. Significant p-values (≤0.05) are indicated in bold.

	Bray-Curtis		Jaccard		Unweighted Unifrac		Weighted Unifrac	
	r^2	p-value	r^2	p-value	r^2	p-value	r^2	p-value
Bacterial taxa abundance between *mus* (n=35) and *dom* (n=85)	0.0270	**9.999×10^{-5}**	0.0194	**9.999×10^{-5}**	0.0292	**9.999×10^{-5}**	0.0261	**9.999×10^{-5}**
Bacterial taxa abundance between house mouse populations	0.1425	**9.999×10^{-5}**	0.0996	**9.999×10^{-5}**	0.1446	**9.999×10^{-5}**	0.1284	**9.999×10^{-5}**
Bacteroides abundance between *mus* (n=35) and *dom* (n=85)	0.0680	**9.999×10^{-5}**	0.0526	**9.999×10^{-5}**	0.0541	**9.999×10^{-5}**	0.0703	**9.999×10^{-5}**
Bacteroides abundance between house mouse populations	0.3149	**9.999×10^{-5}**	0.2412	**9.999×10^{-5}**	0.2694	**9.999×10^{-5}**	0.3296	**9.999×10^{-5}**

Supplementary table 2. Summary statistics of major phyla and genera abundances in *mus* (n=35) and *dom* (n=85) mice. Unclass: unclassified; STD: standard deviation.

	Taxon	**Rank**	**Min**	**Max**	**Mean**	**STD**
domesticus	Bacteroidetes	Phylum	0.06	0.62	0.33	0.12
	Firmicutes	Phylum	0.20	0.79	0.50	0.15
	Proteobacteria	Phylum	0	0.49	0.12	0.13
	Alistipes	Genus	0	0.26	0.06	0.05
	Helicobacter	Genus	0	0.49	0.12	0.13
	Paraprevotella	Genus	0	0.43	0.10	0.11
	Eisenbergiella	Genus	0	0.65	0.12	0.12
	Unclass. Bacteroidales	Genus	0	0.44	0.08	0.07
	Unclass. Lachnospiraceae	Genus	0.04	0.53	0.24	0.13
musculus	Bacteroidetes	Phylum	0.07	0.61	0.35	0.13
	Firmicutes	Phylum	0.20	0.77	0.45	0.15
	Proteobacteria	Phylum	0	0.52	0.12	0.12
	Alistipes	Genus	0.01	0.20	0.06	0.05
	Helicobacter	Genus	0	0.51	0.12	0.12
	Paraprevotella	Genus	0	0.37	0.06	0.10
	Eisenbergiella	Genus	0	0.45	0.07	0.09
	Unclass. Bacteroidales	Genus	0.01	0.27	0.11	0.05
	Unclass. Lachnospiraceae	Genus	0.04	0.55	0.22	0.12

Supplementary table 3. Summary statistics of twenty most abundant *Bacteroides* ASVs in *mus* (n=35) and *dom* (n=85) mice. STD: standard deviation.

ASV	***domesticus***				***musculus***			
	Min	Max	Mean	STD	Min	Max	Mean	STD
ASV 35	0	0.0554	0.0036	0.0145	0	0.0840	0.0100	0.0122
ASV 22	0	0.0322	0.0015	0.0086	0	0.1147	0.0057	0.0124
ASV 268	0	0.0827	0.0007	0.0139	0	0.0376	0.0042	0.0041
ASV 163	0	0.0314	0.0018	0.0062	0	0.0234	0.0029	0.0035
ASV 128	0	0.0406	0.0002	0.0085	0	0.0081	0.0023	0.0010
ASV 411	0	0.0157	0.0008	0.0042	0	0.0245	0.0022	0.0032
ASV 261	0	0.0261	0.0019	0.0048	0	0.0229	0.0020	0.0043
ASV 214	0	0.0255	0.0012	0.0059	0	0.0331	0.0019	0.0041
ASV 110	0	0.0270	0.0005	0.0053	0	0.0164	0.0019	0.0019
ASV 691	0	0.0298	0.0002	0.0050	0	0.0150	0.0014	0.0017
ASV 822	0	0.0250	0.0002	0.0042	0	0.0133	0.0011	0.0015
ASV 525	0	0.0258	0.0026	0.0047	0	0.0258	0.0010	0.0054
ASV 19	0	0.0158	0.0033	0.0033	0	0.0705	0.0010	0.0085
ASV 242	0	0.0069	0.0004	0.0019	0	0.0066	0.0009	0.0010
ASV 2917	0	0.0106	0	0.0024	0	0	0.0007	0.0001
ASV 1299	0	0.0028	0.0002	0.0008	0	0.0171	0.0003	0.0019
ASV 311	0	0.0015	0.0003	0.0004	0	0.0033	0.0001	0.0007
ASV 2406	0	0.0020	0.0004	0.0004	0	0.0056	0.0001	0.0010
ASV 2348	0	0.0024	0.0004	0.0004	0	0.0064	0.0001	0.0010
ASV 3527	0	0	0.0002	0	0	0.0029	0	0.0006

Chapter II:

Candidate *Bacteroides* genome isolation and sequencing

Introduction

In the previous chapter it was shown that the influence of geography on gut community structure, and on the *Bacteroides* genus in particular, was higher than the influence of the host subspecies. Despite this observation, the subsequent results revealed interesting *Bacteroides* candidates to be differentially abundant in the *M. m. musculsus (mus)* host subspecies, independent of the geographic location, suggesting these to be promising candidates for further characterization. Moreover, an independent genetic mapping study displays evidence for the existence of a genetic basis for *Bacteroides* ASV 35's association to *mus* mice. These results now lead to the following questions: are there differences in the *Bacteroides* genomes of the isolates derived from the *mus* and *M. m. domesticus (dom)* host subspecies? And which of these differences potentially contribute to bacterial adaptation to different mouse subspecies?

In order to answer these questions, the present study aimed to isolate candidate *Bacteroides* taxa, fully sequence their genomes and perform comparative genomics on the level of protein families to identify potential strain-level variation with respect to the host subspecies. The combination of the culturing and genomics methods yielded genome sequences and the identity of 146 *Bacteroides* isolates. Taxonomic classification based on the full genomes indicates that two potentially new *Bacteroides* species were isolated, along with *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains. Furthermore, a perfect match between the sequence of ASV 35 and both unclassified isolates was detected, suggesting the involvement of potentially new *Bacteroides* species in the intriguing host-microbe association, mentioned above. The differences in protein family content in *B. acidifaciens* and *B. caecimuris* with regard to *mus* or *dom* subspecies were also assessed.

Results

I. Isolation, whole genome sequencing and taxonomy of *Bacteroides* isolates

To further characterize the candidate *Bacteroides* ASVs identified by indicator species analysis, bacteria were isolated from the cecum content of 18 mice (8 *dom* and 10 *mus*). In total, 146 colonies were successfully isolated and confirmed to belong to *Bacteroides* by Sanger sequencing, using genus-specific primers to amplify a 750 bp fragment 16S rRNA gene. To get an approximate idea of the taxonomic annotation of the obtained isolates, the 16S rRNA gene sequences were classified by RDP. Most of the isolates show high similarity to *B. acidifaciens* (76.7%) and the remaining 23.3% showed the closest match to *Bacteroides* sp. (Suppl. Table 1). A closer look at the phylogeny of sequenced 16S rRNA gene fragments revealed a pattern of clustering according to the mouse subspecies of origin (Figure 1). There is a clade of isolates from *dom* mice that cluster close to *B. acidifaciens* A40 type strain and another clade of the isolates mostly from *mus* mice. However, sequences from samples AH 251 and AH 755 (*dom*) cluster separately from the rest of *dom* mouse isolates. MC 158 and MC 527 from *dom* mice seem to be more closely related to *mus* isolates KH 365 and KH 569 than to the samples from the same mouse subspecies.

Subsequently, all 146 isolates were fully sequenced to a level of ≥30X coverage, in addition to the *B. acidifaciens* A40 type strain as a positive control (Table 1). The genomes were then assembled into contigs and annotated. The number and length of the contigs vary, with an average of approximately 500 contigs per genome with an average size of 11kb (Suppl. Table 2). Nearly 4051 protein sequences corresponding to the *Bacteroides* genome data from the NCBI were identified per genome.

Table 1. Summary of isolated *Bacteroides* strains. Mouse populations MC (France), AH (Iran), and CB (Germany) belong to the *dom* subspecies; KH (Kazakhstan) and WI (Austria) belong to the *mus* subspecies.

Sample abbreviation	Number of isolates per sample	Number of isolates per geographic location	Number of isolates per mouse subspecies
MC 083	1	6	56
MC 701	2		
MC 946	3		
AH 598	2	42	
AH 251	20		
AH 763	20		
CB010	5	8	
CB076	3		
KH 353	18	29	90
KH 365	2		
KH 346	8		
KH 569	1		
WI 296	8	61	
WI 395	2		
WI 535	18		
WI 693	9		
WI 852	12		
WI 587	12		

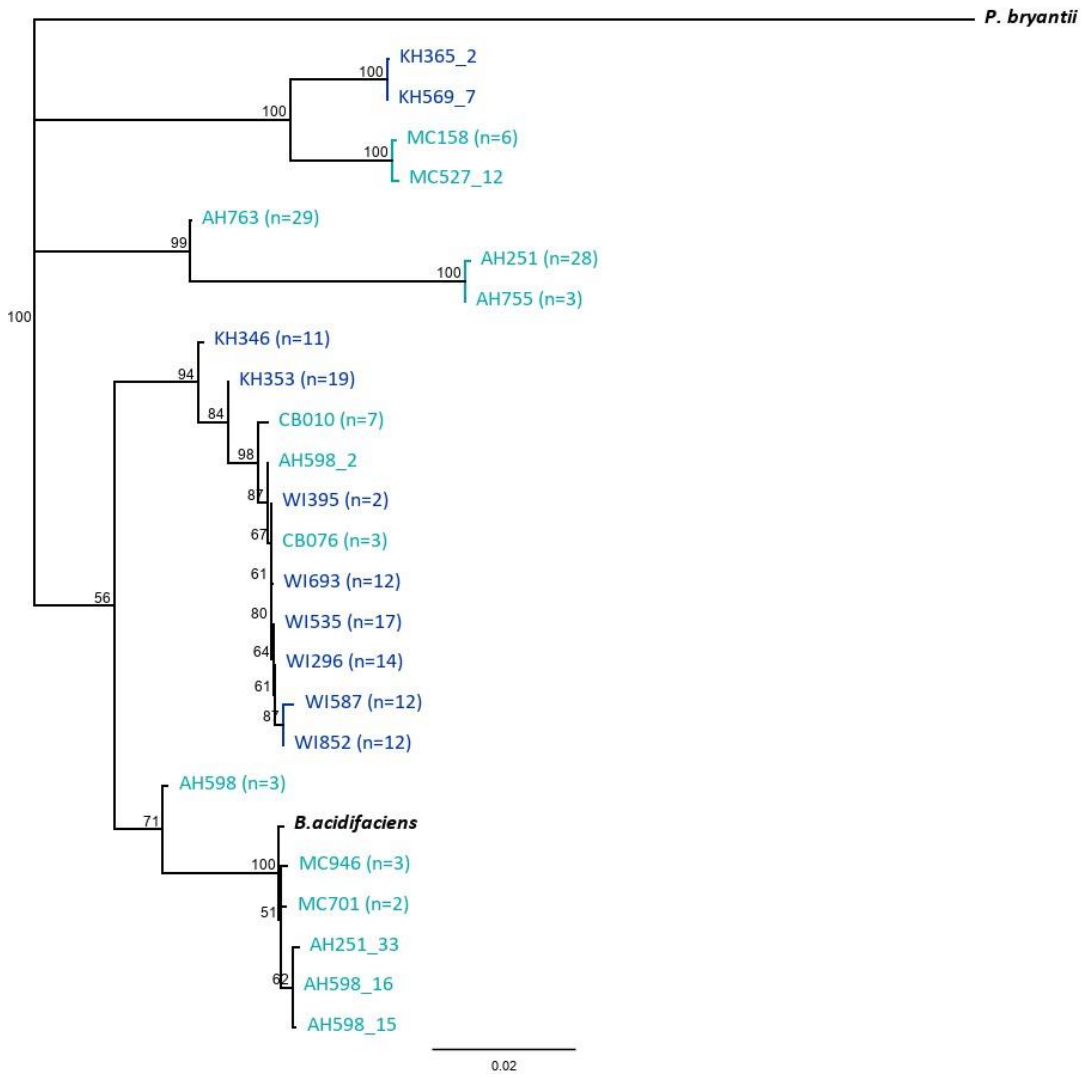


Figure 1. Neighbor-joining tree of the *Bacteroides* isolates. The tree is color coded according to the mouse subspecies of origin: light blue – isolates from domesticus; dark blue – isolates from musculus subspecies. Phylogeny is based on 750 bp alignments of 16S rRNA gene sequences, amplified with *Bacteroides* genus-specific primers Bac32F and Bac708R. *Prevotella bryantii* was used as an outgroup. The isolates with identical 16S rRNA gene sequences were grouped; n – the number of grouped sequences. The bootstrapping was used with 10 000 iterations and a threshold of 50% support. The numbers represent the percentages of consensus support (bootstrap values).

I.1 Taxonomic classification of the isolates using TYGS and GTDB-Tk

To gain deeper insight into the taxonomy of the isolated and sequenced *Bacteroides,* two different tools were used. First, the set of 10 closely related type strains were determined by TYGS (see Methods) via the 16S rRNA gene sequences extracted from uploaded *Bacteroides* genomes, and BLASTed against all available type strains in the TYGS database. The results revealed most of the isolates belonging to *B. caecimuris* (89), following by 38 *B. sartorii* and 17 *B. acidifaciens* strains (Suppl. Table 3). However, calculated digital DNA-DNA hybridization (dDDH) values were all lower than 70%, which is the generally accepted species boundary. Intriguingly, TYGS detected two potentially new species/strains (KH365_2 and KH569_7) which do not belong to any species found in the database.

To validate taxonomic classification based on the nearly full length 16S rRNA gene, the GTDB-Tk toolkit was used (see Methods). In agreement with the outcome of TYGS analysis, GTDB-Tk classified the isolates as *B. caecimuris*, *B. sartorii* and *B. acidifaciens,* with ANI values higher than 95% (Suppl. Table 4). Similar to the TYGS method, no clear taxonomic classification was found for the KH365_2 and KH569_7 strains.

Overall, the results of taxonomic classification reveal *B. caecimuris* and *B. acidifaciens* to be the most abundant isolates in the present dataset. It was possible to isolate 61 *B. caecimuris* strains from WI, 18 from KH, 8 from CB and 2 from AH. *B. acidifaciens* is represented by 9 KH, 6 MC and 2 AH strains. Thus, isolates that belong to each of the bacterial species originate from both *dom* and *mus*. On the other hand, *B. sartorii* was recovered only from AH *dom* mice. The unclassified strains KH365_2 and KH569_7 belong to the same mouse subspecies (*mus*), and originate from two different mouse samples.

I.2 Genomic similarity of *Bacteroides* isolates based on ANI

Average nucleotide identity was also calculated based only on the core genomes of 146 *Bacteroides* isolates together with the reference genomes of *B. acidifaciens* A40 (sequenced with the isolates), *B. sartorii*, *B. caecimuris* and *B. uniformis* (Table 6). *B. uniformis* was included because it was shown to be the closest match to unclassified isolates when BLASTed against NCBI database. At the 95% ANI threshold, the analysis subdivides the sequences into four distinct species-level groupings (Figure 2). Three of them cluster together with *B. acidifaciens*, *B. sartorii* and *B. caecimuris*, respectively, and two unclassified isolates cluster separately from the rest and from *B. uniformis*, confirming the previous results.

In summary, the results reveal that the isolated *Bacteroides* strains belong to *B. sartorii*, *B. caecimuris* and *B. acidifaciens*, the latter two also being the most abundant in the present dataset. Interestingly, two of the strains were not classified by the tools used in this study, suggesting they could belong to a new species. Also, *B. caecimuris* and *B. acidifaciens* isolates were found in both *mus* and *dom* mice, which enables further comparative analysis of these bacterial species between host subspecies.

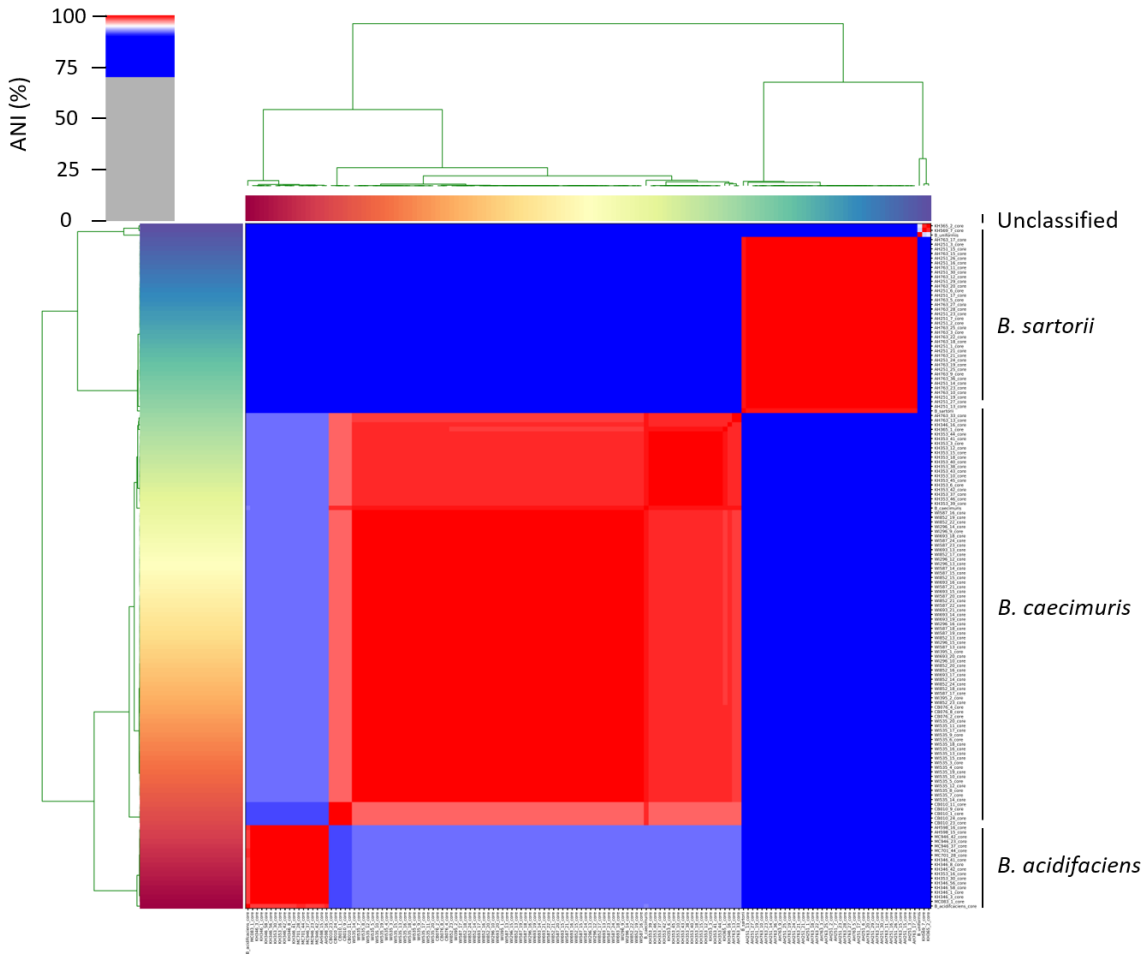


Figure 2. Heatmap of ANIm percentage identity of 146 *Bacteroides* genomes. Isolate IDs are given as row and column labels. Cells in the heatmap corresponding to 95% ANIm sequence identity (and thus the same species) are colored red. Blue cells correspond to ANIm comparisons indicating that the corresponding organisms do not belong to the same species. Color bars above and to the left of the heatmap correspond to source species-level assignments for each isolate in the analysis. Hierarchical clustering of the analysis results in two dimensions is represented by dendrograms, constructed by simple linkage of ANIm percentage identities.

II. Classification of the indicator ASVs based on the isolate genome sequences

In order to determine which of the *Bacteroides* ASVs identified in the Chapter I were included in the isolation and genome sequencing of this chapter, as well as the proportion of isolated ASVs to the total detected by 16S rRNA gene sequence analysis, the following approach was taken. First, *Bacteroides* ASVs were BLASTed against genomes of the isolated strains and the respective nucleotide identities were calculated. For this, only the alignments of full-length ASVs (approx. 300 bp) to the isolate genomes were considered. The results indicate that all 33 *Bacteroides* ASVs identified in Chapter I display a complete alignment to 100 out of 146 sequenced genomes, and show nucleotide identities between 80% and 100%. This suggests that some low-abundant strains captured through cultivation appear to have been undetected through the 16S rRNA gene profiling in Chapter I. Furthermore, however, the ASV is defined as an individual DNA sequence, and the method used to infer ASVs accordingly resolves between sequences that differ by a single nucleotide. Thus, the BLAST results should display 100% nucleotide identity of the ASV to the genome. As such, only 8 ASVs (24% of the total) are identical to a sequenced genome in that specific 300 bp region of the 16S rRNA gene (Suppl. Table 5). Three ASVs (ASV 110, ASV 242 and ASV 311) appear to belong to *B. caecimuris*, three (ASV 2406, ASV 2348 and ASV 5872) to *B. sartorii* and one indicator ASV 268 to *B. acidifaciens.* Interestingly, the indicator ASV 35 shows 100% identity to the unclassified genomes of KH365_2 and KH569_7 (Figure 3, Suppl. Table 5). The nucleotide identities of the other 4 indicator ASVs to *Bacteroides* isolates genomes in that specific 16S rRNA gene region range from 84.6% to 99.7%. The ASV 822 and ASV 691 are 99.7 % identical to the sequenced isolates, and ASV 22 and ASV 411 have the lowest identities to *Bacteroides* genomes, displaying 92.5% and 89.3%, respectively. (Figure 3, Table 2).

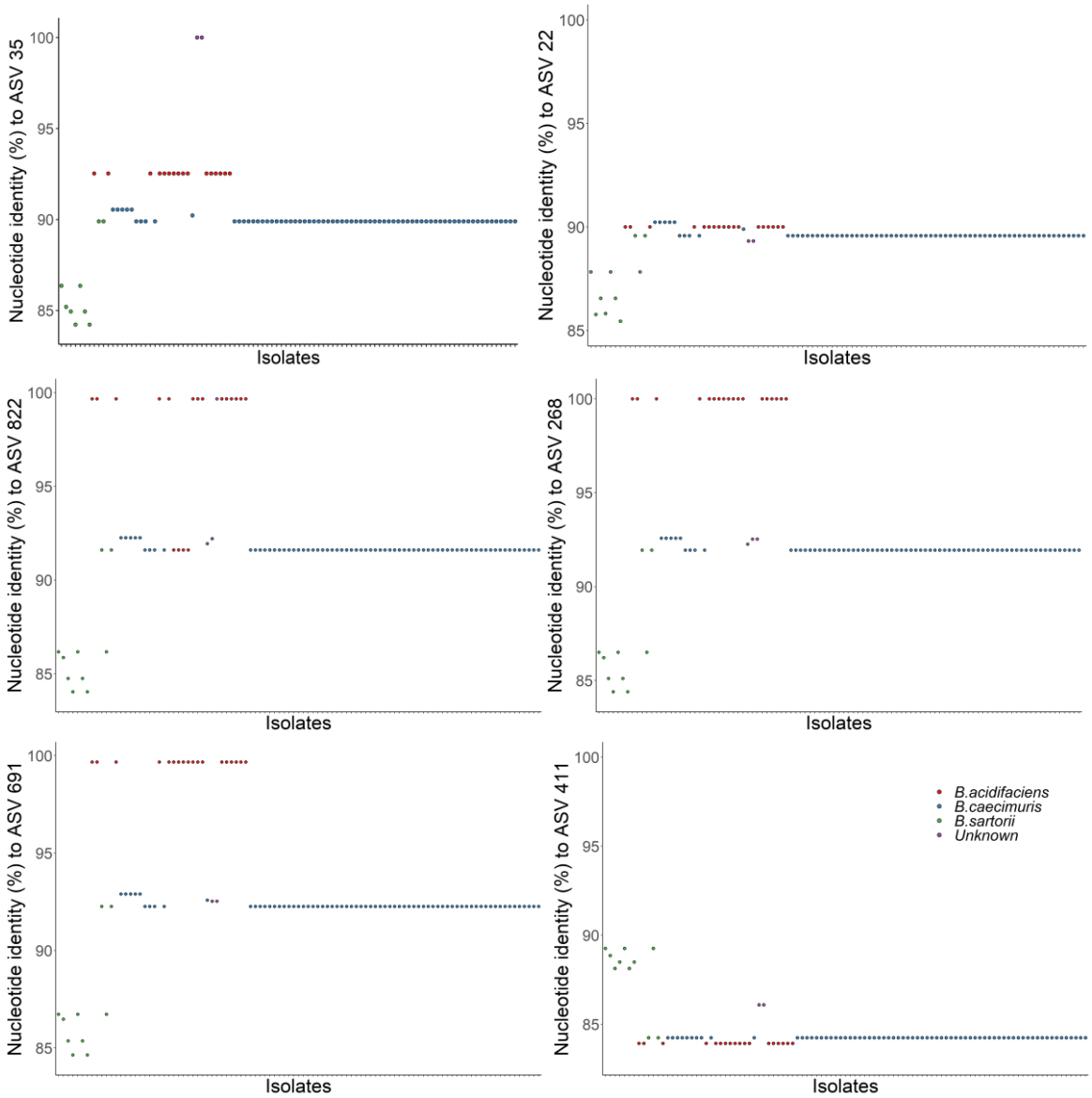


Figure 3. The overview of nucleotide identity between indicator *Bacteroides* ASVs sequences and genomes of the isolates. The ASVs sequences were BLASTed against sequenced *Bacteroides* genomes and only nucleotide identities of minimum 84% are shown.

Table 2. Summary of nucleotide identities between candidate *Bacteroides* ASVs and the sequenced genomes. Nucleotide identities from 99% to 100% are shown in bold. NA: nucleotide identity value is not available due to very short alignment length.

	Nucleotide identity (%)				
	ASV 22	ASV 35	ASV 268	ASV 691	ASV 822
KH365_2	89.3	**100**	92.5	92.5	92.2
KH569_7	89.3	**100**	92.5	92.5	**99.7**
AH598_15	90	92.5	**100**	**99.7**	**99.7**
AH598_16	90	92.5	**100**	**99.7**	**99.7**
KH346_1	90	92.5	**100**	**99.7**	**99.7**
KH346_3	90	91.8	**100**	**99.7**	**99.7**
KH346_8	90	92.5	**100**	**99.7**	**99.7**
KH346_41	90	92.5	**100**	**99.7**	**99.7**
KH346_42	90	92.5	**100**	**99.7**	**99.7**
KH346_56	90	92.5	**100**	**99.7**	**99.7**
KH346_58	90	92.5	**100**	**99.7**	**99.7**
KH353_16	90	92.5	**100**	**99.7**	**99.7**
KH353_30	90	92.5	**100**	**99.7**	**99.7**
MC083_1	90	92.5	**100**	**99.7**	**99.7**
MC701_28	90	92.5	**100**	**99.7**	**99.7**
MC701_44	90	92.5	**100**	**99.7**	**99.7**
MC946_23	90	92.5	**100**	**99.7**	**99.7**
MC946_37	90	92.5	**100**	**99.7**	**99.7**
MC946_42	90	92.5	**100**	**99.7**	**99.7**
AH251_1	87.8	84.6	86.5	86.7	86.2
AH251_26	87.8	86.4	86.5	86.7	86.2
AH763_18	87.8	86.4	86.5	86.7	86.2
KH353_37	92.5	NA	NA	NA	NA
KH353_38	92.5	NA	NA	NA	NA
KH353_39	92.5	NA	NA	NA	NA
KH353_40	92.5	NA	NA	NA	NA
KH353_41	92.5	NA	NA	NA	NA

To gain insight into the proportions of isolated *Bacteroides* ASVs compared to those predicted based on 16S rRNA gene sequencing, 16S profiles of the samples used for cultivation were plotted and compared to the actually isolated ASVs. In total, the 16S data analysis identified 22 *Bacteroides* ASVs among 18 samples used for bacteria isolation (Figure 4).

In some cases, the bacterial isolates derived from a given sample correspond to the most abundant ASVs detected for the respective sample based on 16S rRNA gene sequencing. For example, ASV 35 is the most abundant sequence for the samples KH 365 and KH 569 (Figure 4), and based on the

nucleotide identities values this ASV matches the isolated strains KH365_2 and KH569_7 (Table 2). A similar pattern is observed for the sample WI 535, from which *B. caecimuris* ASV 110 was successfully isolated (Figure 4, Suppl. table 5). However, in most cases the isolated ASVs correspond to low-abundant taxa in the 16S data. This is the case for sample CB 010, which yielded only ASV 311 via cultivation, whereas ASV 35 and ASV 22 are the most abundant for this sample based on its 16S profile. The sample AH 763 likewise yielded the lowest abundant sequences ASV 2406 and ASV 110 as isolates (Figure 4, Suppl. table 5). Moreover, the samples AH 251 and AH 598 yielded ASV 5872 and ASV 268 as isolates, respectively, which are not present in corresponding 16S profiles (Figure 4). Both ASV sequences showed 100% identity to the isolates AH 251_7, AH 598_15 and AH 598_16 (Suppl. table 5).

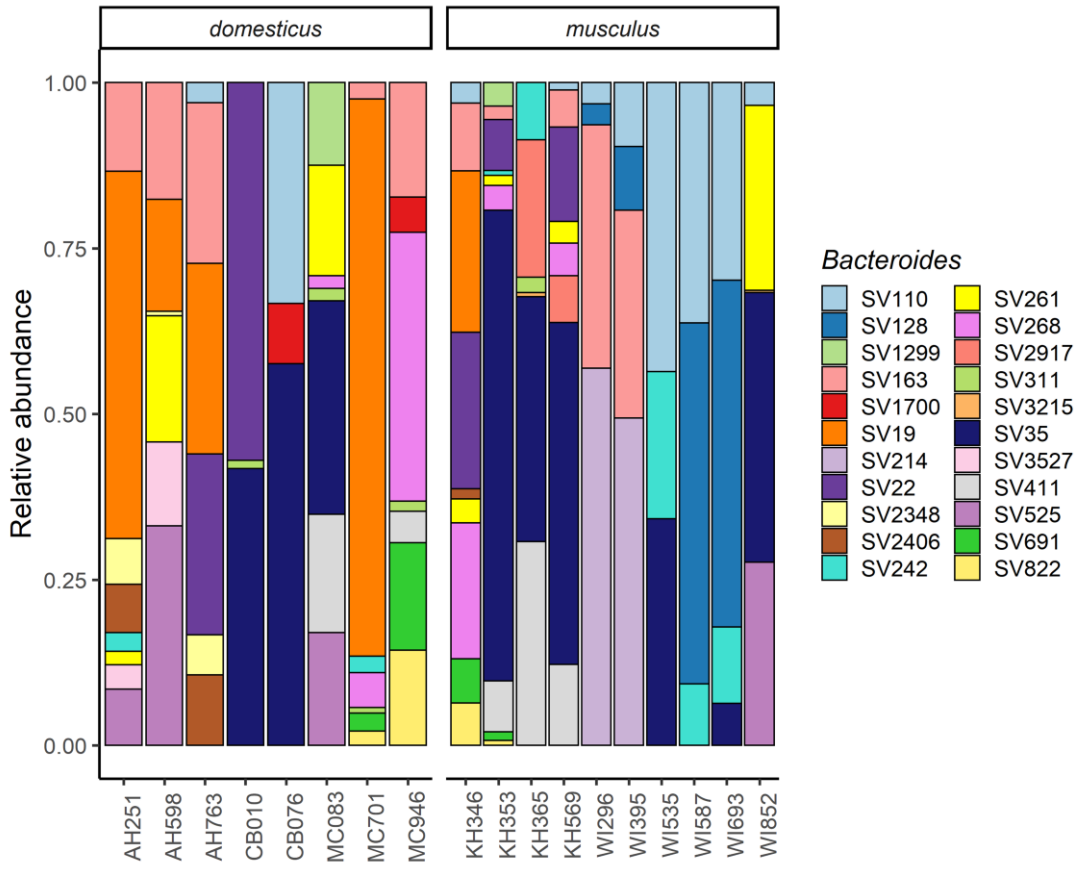


Figure 4. Relative abundances of *Bacteroides* ASVs among mouse samples used for bacteria isolation. The samples grouped according to the host subspecies.

In summary, the results show that with the cultivation conditions used in the study, it was not possible to retrieve all representative *Bacteroides* ASVs, independent of how abundant they are. In fact, most of the isolated ASVs were identified as low abundant in the respective samples based on their 16S profile.

III. Bacteroides pan-genome

III. 1 Protein family distribution

First, the distribution of protein families in isolated *Bacteroides* strains was characterized. Sequenced genomes in the dataset carry on average 4051 proteins that cluster into 9881 protein families. Moreover, 1496 core protein families (15,1%) were identified (protein families found in every sequenced genome), from which 1178 were single-copy (represented only once in every genome) (Figure 5). Additionally, 2287 singleton protein families (present only in single genomes) were found. In total, the accessory protein families comprise 84.9% of total protein content (8385 families).

To further describe the distribution of protein families, a splits network based on the presence – absence patterns of all protein families was computed (Figure 6). The results show that *Bacteroides* isolates classified as the same species cluster strongly together. The analysis of the *B. caecimuris* species highlights that it appears to be more diverse than the other species in the dataset. While *B. acidifaciens* and *B. sartorii* samples are strongly clustered with few splits, *B. caecimuris* samples contain multiple splits, where the division seem to be related to the different geographical locations. In addition, the results suggest that there is a slight divergence according to the host subspecies of origin.

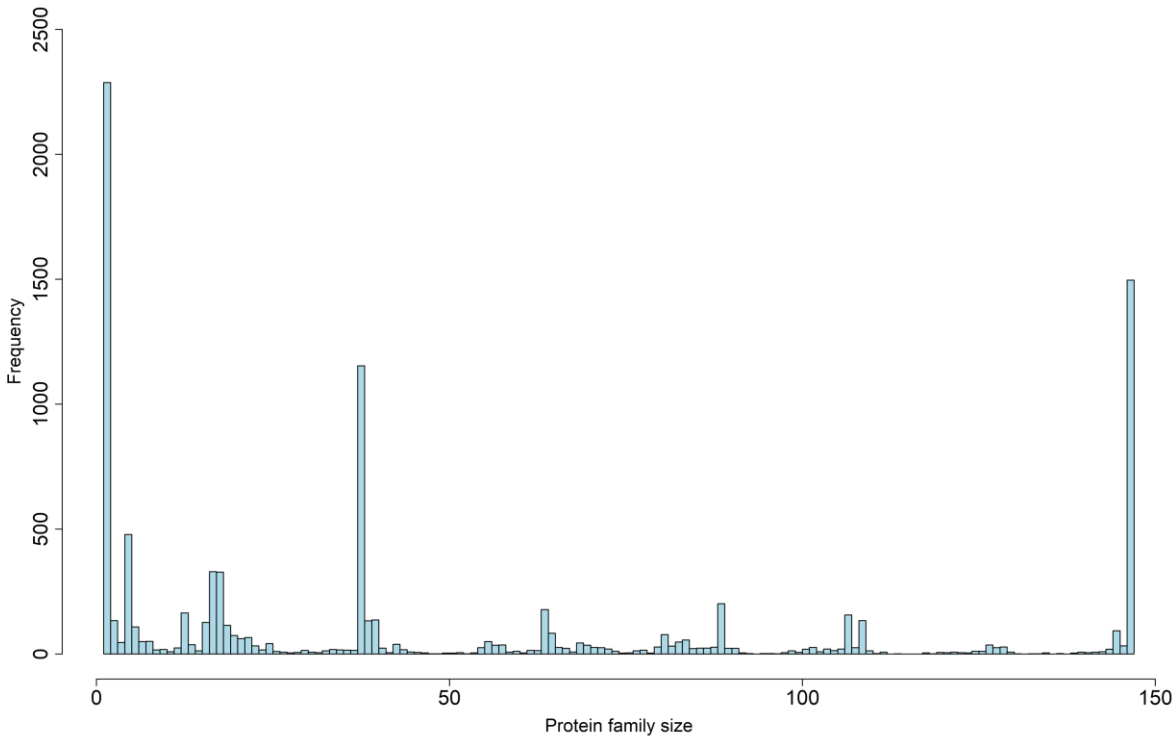


Figure 5. Distribution of the protein family sizes. The histogram is based on the presence-absence pattern of the protein families.

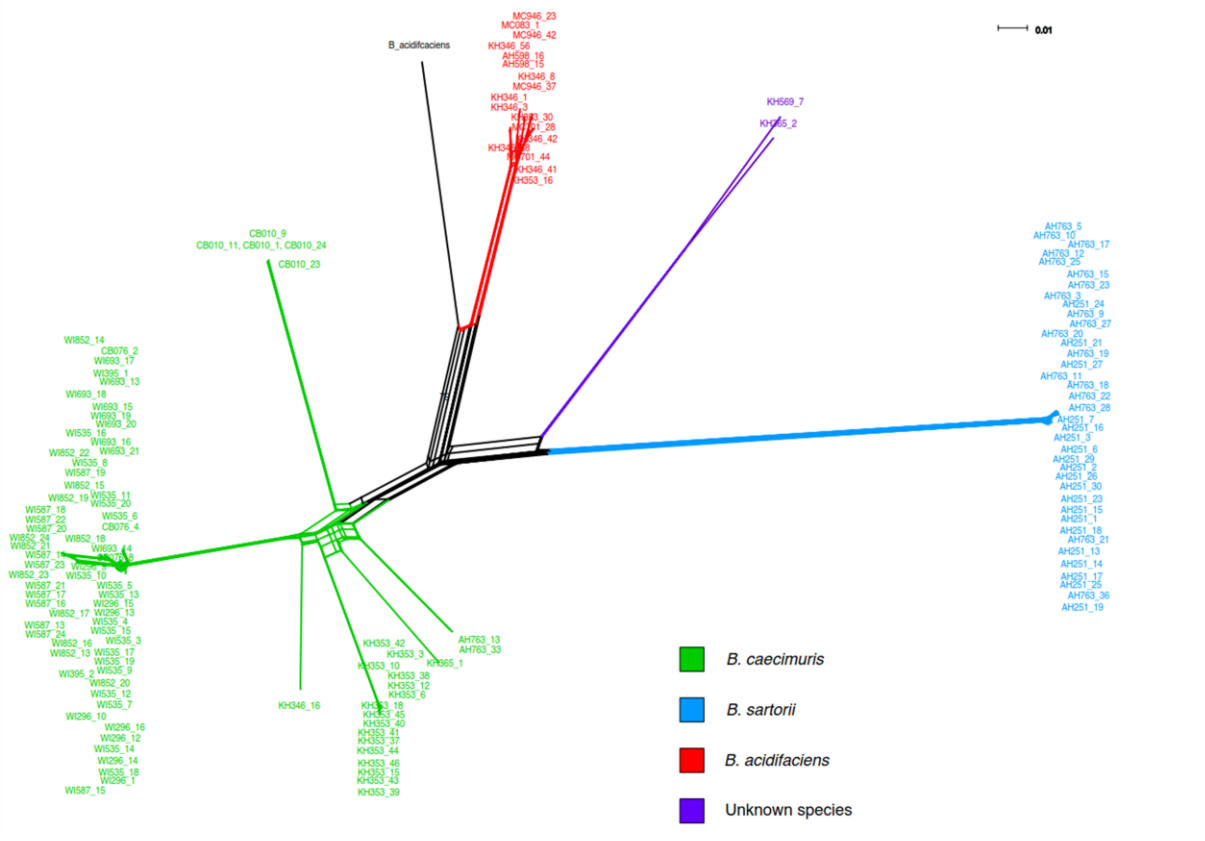


Figure 6. Splits network estimated from the presence-absence patterns of the complete set of protein families. Isolates classified as *B. sartorii* (blue), *B. caecimuris* (green), *B. acidifaciens* (red), unclassified isolates (purple). Reference *B. acidifaciens* A40 strain is represented in black.

III.2 Phylogeny based on single-copy protein families

To further assess the genomic diversity in the dataset, a phylogenetic tree from the complete single-copy protein families was computed (Figure 7). The resulting topology is consistent with splits network based on protein family content, where clustering based on *Bacteroides* species is observed. The strains classified as *B. sartorii* appear to be limited to *domesticus* AH mice (Iran), whereas *B. caecimuris* and *B. acidifaciens* strains were isolated from multiple geographical regions and both mouse subspecies, exhibiting isolates from four (AH, KH, CB and WI) and

three (KH, AH and MC) mouse lines, respectively. Moreover, the overall topology of the phylogenetic tree based on *Bacteroides* full genomes (Figure 8) is similar to the topology of the tree computed from the single-copy protein sequences alignment (Figure 7). However, the unclassified isolates are placed close to the *B. acidifaciens* type strain on the single-copy protein-based tree, and there is no clear separation between *B. acidifaciens* and *B. caecimuris* isolates. In some cases, strains cluster together according to geographic location, but not in the *B. acidifaciens* and *B. caecimuris* clades, which are represented by the isolates from the mouse lines belonging to different subspecies (Figures 7 and 8). Thus, these strains are good candidates to perform the comparative analysis between *domesticus* and *musculus* host subspecies.

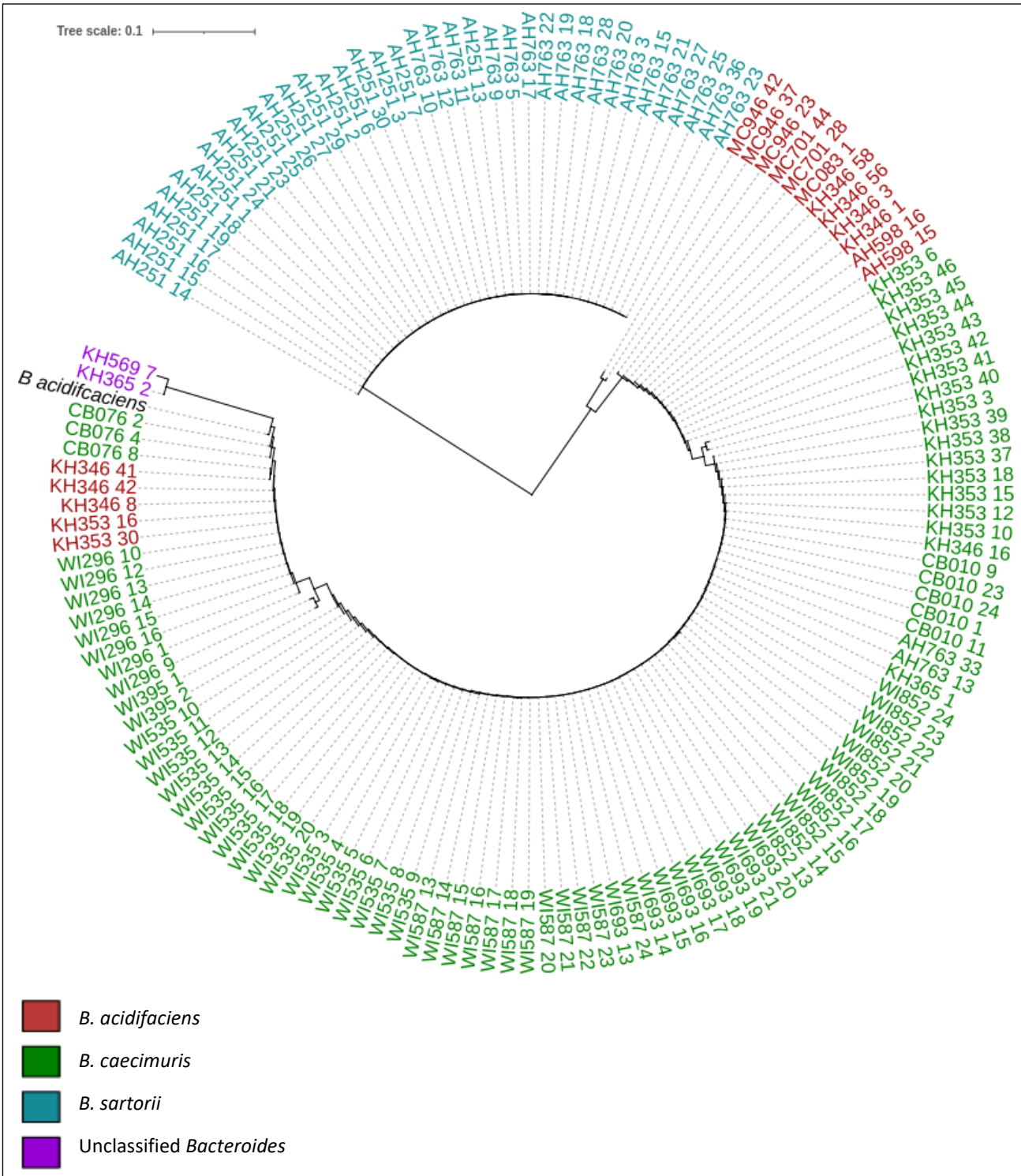


Figure 7. Phylogenetic tree based on the concatenated alignment of the complete single-copy protein families. The sequences were aligned with MAFFT, the phylogeny was computed with IQ-TREE and the tree was rooted using the MAD algorithm.

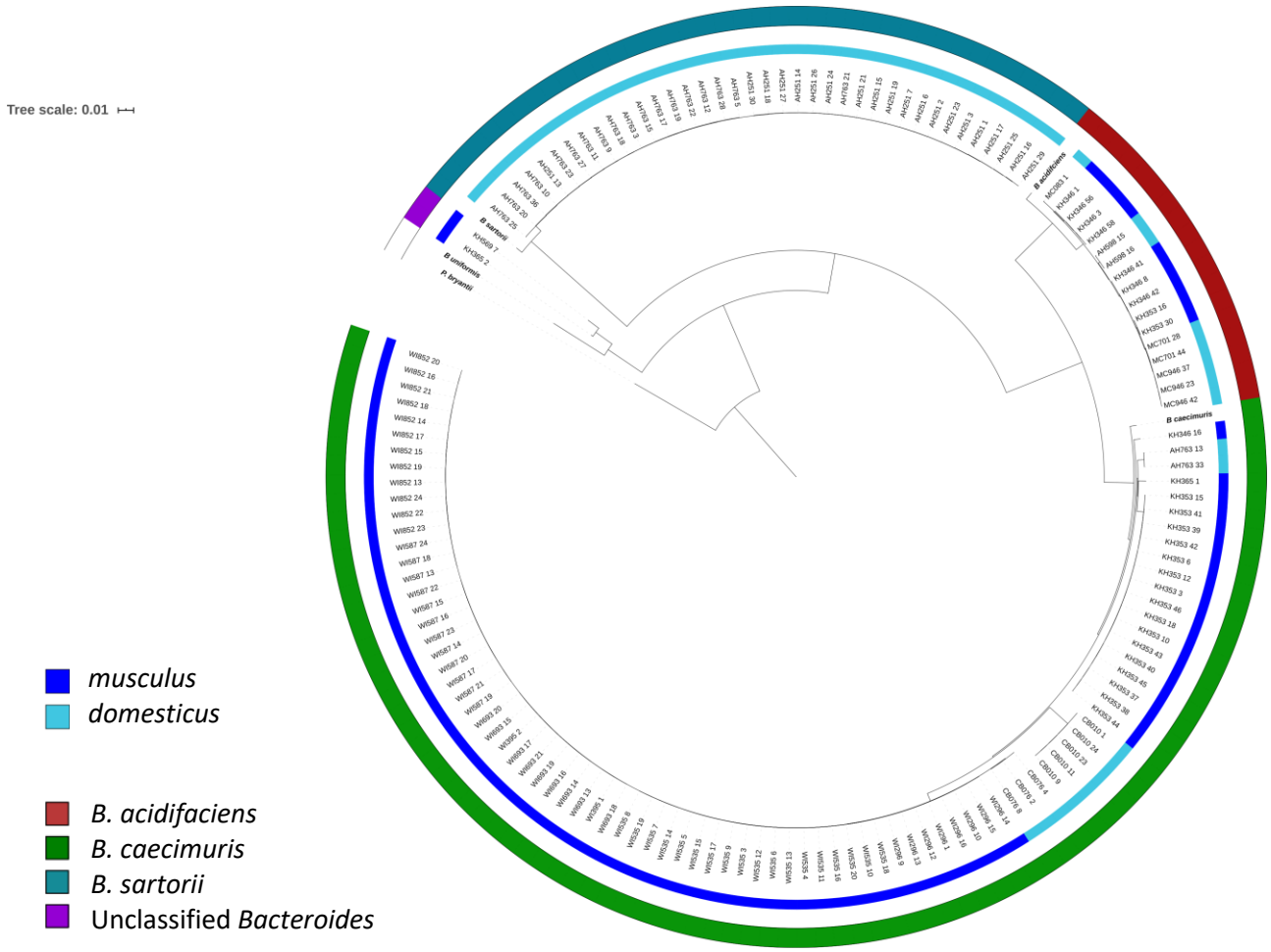


Figure 8. Phylogeny of *Bacteroides* isolates based on the genome sequences. Evolutionary distances were estimated using ANDI, the phylogeny was visualized using iTOL web tool. The first inner ring shows the mouse subspecies of origin, the second ring shows the taxonomic classification of the isolates determined before. *Prevotella bryantii* was used as an outgroup.

III.3 *Bacteroides* protein families across *dom* and *mus* mice

In order to assess the differences in protein family content among host subspecies, families were classified as unique or common to *dom* and *mus,* separately for *B. acidifaciens* and *B. caecimuris* strains. Accordingly, the protein families that were present in at least one *B. acidifaciens* or *B. caecimuris* strain from *dom* or from *mus* mice were extracted, and from these the common and host subspecies-specific protein families were identified. A total of 3748 protein families were retrieved for *B. acidifaciens*, 3447 (92%) of which are common to both mouse subspecies, whereas 243 are specific to *dom* and 58 are specific to *mus* (Figure 9A). The total number of protein families for *B. caecimuris* is higher, reaching 5744 families, but the proportion of protein families shared between *dom* and *mus* is smaller, with 3836 (67%) families common to both host subspecies and 871 and 1037 specific to *dom* and *mus*, respectively (Figure 9A).

Next, the protein families that were present in all the strains of *B. acidifaciens* or *B. caecimuris* from *dom* or from *mus* mice were selected and the host subspecies-specific or common families were determined. For *B. acidifaciens* the total of 3288 protein families were identified. From these 3143 (96 %) are common to both mouse subspecies, 60 are unique to *dom* and 85 to *mus* mice (Figure 9B). Similarly, for *B. caecimuris* most of the protein families (86 %) were identified as common to both subspecies, whereas 171 and 177 are unique to *dom* and *mus*, respectively (Figure 9B).

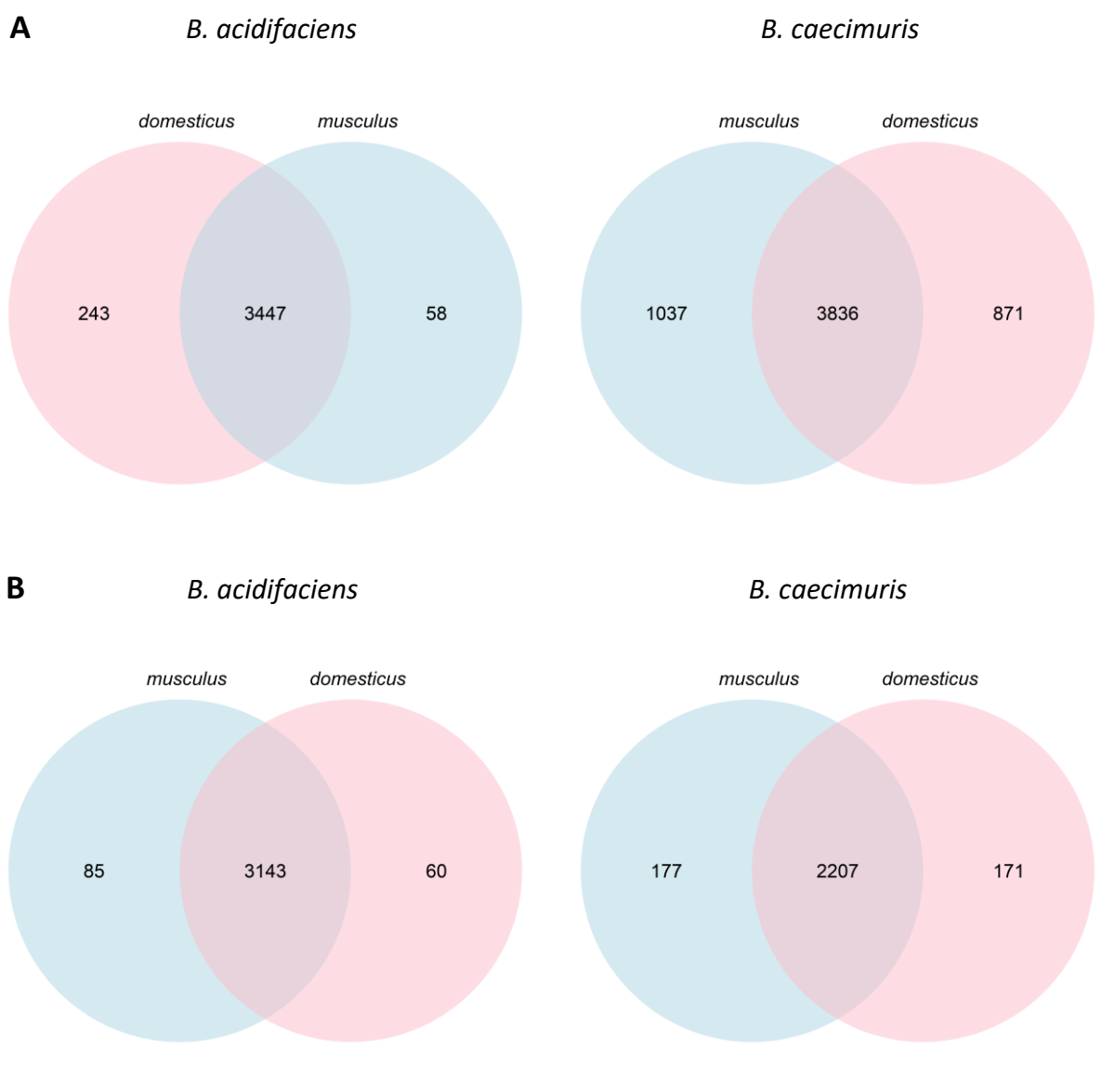


Figure 9. Host subspecies-specific protein families of *B. acidifaciens* and *B. caecimuris* strains. Venn diagrams show shared and mouse subspecies-specific protein families present in at least one in *B. acidifaciens* and *B. caecimuris* strains (A) and present in all *B. acidifaciens* and *B. caecimuris* strains (B).

To compare host subspecies-specific protein families at the functional level, the annotations were extracted for protein families present in all *B. acidifaciens* and *B. caecimuris* strains. Most of the proteins were annotated as hypothetical, ranging from 68% for *B. caecimuris* from *dom* mice to 94% for *B. acidifaciens* from *mus* mice (Table 3).

Table 3. Summary of host-specific protein families identified for candidate *Bacteroides* strains from *dom* and *mus* mice. The numbers correspond to the total identified protein families and the numbers/percentages of protein families annotated as hypothetical protein or translated product with defined function.

Host subspecies	Bacteria	Total	Hypothetical protein		Translated product	
			Number	Percentage	Number	Percentage
domesticus	*B. acidifaciens*	60	49	82 %	11	18 %
	B. caecimuris	171	117	68 %	54	32 %
musculus	*B. acidifaciens*	85	80	94 %	5	6 %
	B. caecimuris	177	130	73 %	47	27%

A closer look a protein family functions reveals many proteins with different IDs to be involved in the same biological processes and have similar functions. This observation was made not only among *B. acidifaciens* and *B. caecimuris* isolated from the same-, but also for strains isolated from different host subspecies. For instance, all *B. acidifaciens* strains of *mus* host origin have protein families 278 and 2979 involved in tyrosine recombinase activity (Suppl. table 6). Similarly, eight *B. caecimuris* protein families with different IDs (five identified in strains from *domesticus* and three from *musculus* mice) were identified as TonB-dependent receptors (Suppl. table 6). Interestingly, *B. acidifaciens* protein family 3042 and *B. caecimuris* protein family 3172, annotated as anaerobic sulfatase-maturating enzymes involved in sulfatase oxidation pathway, were identified exclusively in strains isolated from *domesticus* mice (Suppl. table 6).

Among *B. acidifaciens* proteins unique either to *dom* or *mus* mice, some are involved in cell metabolic processes such as glycolysis (protein family 6337 in *mus*) or lipid metabolism (protein family 3730 in *domesticus*) (Suppl. table 6). Several protein families unique to *dom* mice were identified to be involved in pathogenesis and drug resistance (Fragilysin 6522 and 6525, and efflux pump membrane transporter 503). For *B. caecimuris,* in addition to *dom*- and *mus*-specific protein families involved in cell metabolic processes and antibiotic resistance, several proteins with functions in metal-binding were detected: protein families 1863, 1874 and 1875 (colicin I receptor) from *musculus* and 1856 from *domesticus* hosts (Suppl. table 6).

In summary, analysis of pan protein families among the *Bacteroides* genome pool reveals a very high proportion (84.9%) to be represented by accessory protein families. The protein family content of *B. acidifaciens* and *B. caecimuris* isolated from *mus* and *dom* mice yielded high numbers of shared, and a smaller number of proteins unique to a given host subspecies. Anaerobic sulfatase-maturating enzymes, which contribute to the colonization of the intestinal tract, were found in *B. acidifaciens* and *B. caecimuris* of exclusively *dom* origin.

Discussion

Bacteroides is an important genus inhabiting the mammalian intestine involved in many health-related traits of the host. It is already known that gut microbiota evolves and works in tandem with their hosts (Moeller *et al.*, 2014, 2017). However, the forces that drive host-microbiota coadaptation are poorly understood. This study aimed to perform a broad characterization of *Bacteroides* genus among the *domesticus* and *musculus* house mouse subspecies. After the identification of interesting candidates, 146 isolates belonging to *B. acidifaciens*, *B. caecimuris*, *B. sartorii* and two unclassified species were fully sequenced, and the differences in protein family content were assessed with respect to the host subspecies.

Bacteroides are relatively easy to cultivate, however, some challenges were encountered at the stage of isolation from the mouse cecum content samples. The use of well-known *Bacteroides* Bile Esculin (BBE) (Livingston, Kominos and Yee, 1978) medium was not successful. No colonies were observed on the plates, while the pure *B. ovatus* and *B. thetaiotaomicron* strains available in the lab displayed good growth on BBE. One of the possible explanations for this could be that dilutions used for plating were too high. In later related work conducted in our lab, *Bacteroides* from human fecal samples were isolated on BBE plates using less diluted samples (10^{-2}), while for the KV plates the optimal dilutions of the same samples were of 10^{-5} – 10^{-6}.

Despite our efforts, the within-genus diversity of *Bacteroides* detected through 16S rRNA gene sequencing was not entirely covered with the isolation method used in this study. The isolates comprised *B. caecimuris* and *B. acidifaciens*, the most abundant and present in both mouse subspecies, and *B. sartoii*, represented only by *dom* mice. Each of these three species are common gut commensals. Moreover, *B. acidifaciens* was previously reported to prevent obesity in mice (Yang *et al.*, 2017), and the relative abundance of *B. caecimuris* was increased by intermittent fasting in a mouse model of multiple sclerosis (Cignarella *et al.*, 2018). Caballero *at al*. (2017) showed *B. sartorii*

to play a role in the restoration of colonization resistance to vancomycin-resistant *Enterococcus faecium.* Intriguingly, two strains with no clear taxonomic annotation were also isolated, suggesting potentially new species. Similar genomes were previously assembled from the rat gut (Parks *et al.*, 2017), however the bacterium was not isolated and cultivated.

Another important result concerns the *Bacteroides* ASVs identified in Chapter I. Only eight out of 33 *Bacteroides* ASVs were identified to be present among cultivated strains. However, the length of the ASV sequence is very short and it is possible to have a 100% match at the V1-V2 region of 16S rRNA gene, but still not belong to the same strain/species. Interestingly, indicator ASV 35 displays 100% identity to the unclassified genomes KH365_2 and KH569_7. This ASV was misclassified as *B. acidifaciens* in Chapter I, based on an approximately 700 bp fragment of the 16S rRNA gene. Unfortunately, ASV 35 is represented only by two strains in the dataset, belonging to the same mouse subspecies. Thus, to perform the comparative analysis among *mus* and *dom* mice, more strains matching ASV 35 must be obtained and sequenced from both host subspecies. Moreover, it was not possible to obtain isolated representatives of all *Bacteroides* ASVs, independently on how abundant they are. In fact, most of the isolated ASVs display a low abundance in the respective samples, indicating that the isolation media or growth conditions might stimulate the growth of certain strains and slow down the growth of the others, even if they belong to the same bacterial species. Even though *Bacteroides* strains are very similar physiologically, differences in susceptibility to antibiotics were reported (Wexler, 2007). *B. massiliensis*, for instance, can grow in a larger range of temperatures (25-42°C) compared to other *Bacteroides* species (Fenner *et al.*, 2005). On the other hand, some of the isolated *Bacteroides* ASVs were not detected in the 16S data analysis. This outcome might be explained by insufficient sequencing coverage for low abundant taxa.

The phylogeny inferred from the whole genome data and from single-copy protein families show that *Bacteroides* isolates cluster together according to their taxonomic classification rather than to host subspecies of origin, suggesting that little or no host subspecies-specific preferences exist at

the level of these *Bacteroides* species. The discrepancy observed in the topology of phylogenetic trees based on core protein families (Figure 7) and whole genome data (Figure 8) might be explained by high similarity between the core genomes of *B. acidifaciens* and *B. caecimuris*, which interferes with the phylogenetic resolution. Moreover, an outgroup and several *Bacteroides* type strains were used to infer phylogenetic distances for the genome-based tree.

The study of pan protein families reveals a larger proportion of the *Bacteroides* genome pool to be represented by accessory protein families (84.9%), while the core corresponds to 1496 (15.1%) of total proteins. Similar observations were made by Karlsson *et al*. (2011), who detected 1116 core protein families for a set of 31 *Bacteroides* genomes. In most cases, the accessory genome constitutes s larger part of bacterial pan genome, whose size is dependent on the amount of genomes under analysis as well as the chosen methodology (Tettelin *et al.*, 2008).

The analysis of *B. acidifaciens* and *B. caecimuris* protein family content among *mus* and *dom* mice yielded high numbers of shared proteins in both cases (Figure 9), since the strains are closely related within each of the bacterial species. The number of *mus*- or *dom*-unique protein families are significantly smaller and is made up of entirely accessory proteins. Considerably higher numbers of unique protein families were observed for *B. caecimuris* when compared to *B. acidifaciens*. This result might be influenced by the unbalanced sample sizes, with 89 *B. caecimuris* isolates compared to only 17 *B. acidifaciens* isolates.

A closer look into the annotated functions for protein families reveals a number of interesting candidate functions, such as the *domesticus*-specific anaerobic sulfatase-maturating enzymes identified in *B. acidifaciens* and *B. caecimuris*. This enzyme is involved in sulfatase oxidation pathway, linking the heparin to chondroitin sulfate utilization pathways, and contributing to the colonization of the intestinal tract (Cheng, Hwa and Salyers, 1992; Benjdia et al., 2008). Also, one of *B. caecimuris* *mus*-specific protein families was annotated as colicin I receptor – the outer membrane receptor IA and IB colicin toxins, produced by *E.coli*, also known to participate in iron transport (Lazdunski *et al.*,

1998). According to the NCBI database, the gene *cir*A encoding this receptor, was already identified in several other *Bacteroides* species, possibly acquired by horizontal gene transfer.

In conclusion, this is the first study to systematically investigate the *Bacteroides* genus among two house mouse subspecies *dom* and *mus*. The whole genome sequencing of the isolated strains sheds light on the *Bacteroides* pan genome in terms of protein content and functions, which in some cases could represent specialization to host subspecies. Future experimental studies that are also guided by the identification of potentially adaptively evolving genes in both host and *Bacteroides* genomes may help further identify and confirm signatures of coadaptation within the mammalian metaorganism.

Methods

I. Isolation of candidate *Bacteroides* from cecal content

I.1 Selective medium

Schaedler Anaerobe KV (SKV) Selective Agar with Lysed Horse Blood (Thermo Scientific) was used for the isolation of *Bacteroides* from the cecal samples. The plates were purchased ready to use and stored at 4°C. Prior to inoculation, the SKV medium was reduced by placing the plates overnight under anaerobic conditions at room temperature.

I.2 Isolation procedure and growth conditions

All the steps were performed under an anaerobic atmosphere (gas mixture: 5% H_2, 5%CO_2, 90% N_2) in a vinyl anaerobic chamber (Coy Lab Products). Cecal contents of 18 mice (Table 5) were homogenized by vortexing and serial 10-fold dilutions in anaerobic BHI medium were made. Next, 50 µl of the last three dilutions (10^{-6}, 10^{-5} and 10^{-4}) were plated on SKV plates. The plates were incubated at 37°C for 48 hours. If the growth was insufficient after 48 hours of incubation, the plates were incubated for an additional 24 hours.

Table 5. Cecum content samples used for candidate *Bacteroides* isolation.

Abbreviation	Subspecies	Origin	Number of mice
CB	*domesticus*	Cologne/Bonn, Germany	2
MC		Massif Central, France	3
AH		Ahvaz, Iran	3
KH	*musculus*	Almaty, Kazakhstan	4
WI		Vienna, Austria	6

I.3 Approximate taxonomic classification of the isolates

First, grown colonies were picked from agar plates according to their morphology: *Bacteroides* forms circular, white or beige, shiny, smooth colonies that are approximately 2-3 mm in diameter.

Next, *Bacteroides* genus-specific primers (Chapter I, Methods: Table 11) were used for the amplification of approximately 750 bp portions of 16S rRNA gene by colony PCR (Chapter I, Methods: Table 13). Colonies confirmed to belong to *Bacteroides* were streaked on SKV agar plates. *Bacteroides* taxonomic status was double-checked by Sanger sequencing of the 750 bp 16S rRNA gene fragments and classification using RDP classifier (see Methods in Chapter I), yielding 146 *Bacteroides* isolates.

The phylogeny of the isolates was assessed using Geneious Tree Builder with an HKY genetic distance model and the Neighbor-joining tree building method. The reference sequences of the *B. acidifaciens* and *Prevotella bryantii* 16S rRNA gene were obtained from NCBI (Table 6). The tree was resampled using the bootstrap method with 10,000 iterations and a threshold of 50% support, using *P. bryantii* as an outgroup.

Table 6. Summary of the reference genomes and 16S rRNA genes used for phylogenetic analysis.

Genus	Species	Strain	Type	Identification number
Bacteroides	*acidifaciens*	A40	16S rRNA	NR_028607
Prevotella	*bryantii*	B14	16S rRNA	NR_028866
Bacteroides	*caecimuris*	I48	Genome	NZ_CP015401
Bacteroides	*sartorii*	DSM 21941	Genome	NZ_BAKM01000159
Bacteroides	*uniformis*	ATCC 8492	Genome	GCF_000154205

II. Genomic DNA extraction and whole genome sequencing

Genomic DNA of *Bacteroides* isolates was extracted using the DNeasy UltraClean Microbial Kit from Qiagen. The *B. acidifaciens* A40 type strain purchased from the "Deutsche Sammlung von Mikroorganismen und Zellkulturen" (DSMZ) was used as a positive control. It was processed together with the isolates under the same experimental conditions. Bacterial biomass from isolates grown on SKV agar plates was resuspended directly in 300 µl of PowerBead Solution (Qiagen) and vortexed to mix. All following steps were performed according to the manufacturer's instructions. Extracted DNA was stored at -20°C.

DNA samples were prepared according to Illumina Nextera XT protocol, which uses transposome to simultaneously fragment and tag input DNA, adding the unique adapters in the process. The final DNA library was supplemented with 1% PhiX and sequenced on an Illumina NextSeq 500 system using NextSeq 500/550 High Output Kit v2.5 with 300 cycles.

III. Genome assembly and annotation

From the BCL files containing base calls obtained from sequencing machine, the demultiplexed fastq files were generated. For this, The Illumina bcl2fastq2 Conversion Software v2.20 was used, allowing 1 mismatch in the barcodes. Read quality was checked with the FastQC tool, version 0.11.6 (Andrews, 2010). Next, the adapters were removed using Cutadapt (Martin, 2011). The quality trimming was performed using the option `--nextseq-trim`, removing 10 bases from low quality ends of the reads. The filtering criteria was chosen to apply to both reads (forward and reverse) with the option `--pair-filter=both`. In order to eliminate empty reads, the option `--minimum-length` was applied. All the reads shorter than 75 bp were discarded.

The filtered reads were then assembled to contigs by using the SPAdes genome assembler (Bankevich *et al.*, 2012). The option `--careful` was used to minimize the number of mismatches in the contigs. Subsequently, the contigs were annotated by Prokka (Seemann, 2014) using the standard options.

IV. Taxonomic classification and phylogeny of the isolates

IV.1 TYGS and GTDB

Sequenced isolates were first classified using TYGS - Type (Strain) Genome Server (Meier-Kolthoff and Göker, 2019), an online tool available from the DSMZ. The taxonomic classification is based on the full length 16S rRNA gene sequences extracted from the uploaded genomes and compared to the existing database of type strains of bacterial species. All pairwise comparisons among

the set of genomes are conducted using Genome Blast Distance Phylogeny approach (Henz *et al.*, 2005).

The taxonomic annotation of *Bacteroides* genomes was validated using GTDB-Tk, the Genome Taxonomy Database Toolkit (Chaumeil *et al.*, 2019). Here, the bacterial reference tree was inferred from the multiple sequence alignments of 120 phylogenetically informative marker genes. Next, maximum-likelihood placement of each *Bacteroides* genome in the reference tree was found, based on its average nucleotide identity (ANI) to reference genomes.

IV.2 ANI calculation

Average Nucleotide Identity (ANI) analysis was applied to the core genomes of *Bacteroides* isolates. All core genomes were retrieved from FAA files, using a python script (Van Rossum G and Drake FL, 2009). This analysis was performed by means of the ANIm method, which uses the MUMmer system for sequence alignment (Richter and Rosselló-Móra, 2009) and is implemented in the Python3 module PYANI (0.2.9) (Pritchard *et al.*, 2016). The method consists of the alignment of two genomes, identification of the matching regions and calculation of the average percent nucleotide identity of these matching regions. The percentage threshold for species boundary is 95% ANI.

IV.3 Phylogenetic reconstruction based on ANDI

Evolutionary distances between isolated *Bacteroides* isolates and the reference genomes were estimated using ANDI (Klötzl and Haubold, 2016) with a Jukes-Cantor model and 10000 iterations. The output matrixes from ANDI were then imported to R, and a neighbor joining tree for each iteration was built using the function `nj` from the APE package. Finally, the tree was visualized using iTOL web tool (http://huttenhower.sph.harvard.edu/galaxy/). *Prevotella bryantii* was used as an outgroup.

IV.4 Classification of *Bacteroides* ASVs based on genomic sequences of the isolates

Bacteroides ASV sequences were BLASTed against sequenced genomes. Only the nucleotide identities of the alignments of full-length ASVs (300 bp) to the isolate genome and ≥ 99 % were considered.

V. Pan-genome analysis

V.1 Homologous protein identification and clustering into families

First, all the FAA files containing the protein information of the translated CDS sequences were merged to one FASTA file. Next, using the `makeblastdb` command of blastp v2.5.0+ (Altschul *et al.*, 1990), the protein database was created, containing all the protein information from the concatenated FAA files. To identify homologous proteins, all-against-all local alignments were performed using the `blastp` command with an e-value threshold of 1e-5. Every pair of significant hits was aligned using the global alignment tool, powerneedle from the EMBOSS package (Needleman and Wunsch, 1970; Rice, Longden and Bleasby, 2000). Significant hits sharing at least 30% global amino acid identity were retrieved and clustered into homologous families using MCL v14-137 (Enright, Dongen and Ouzounis, 2002), option -l 2.0. All identified protein families were summarized in the matrix consisting of binary patterns of presences and absences (PAP). The core protein families were defined as being represented in all studied genomes, whereas accessory protein families were absent in at least one of the genomes.

V.2 Splits network and phylogenetic tree inference

To reconstruct the network, a PAP matrix was used in SplitsTree with the uncorrected P distance model (Huson and Bryant, 2006). The single copy protein families (i.e. protein families that are present once in every samples) were retrieved and aligned with MAFFT (Katoh *et al.*, 2002). The resulting alignments were concatenated and the phylogeny was computed with IQ-TREE (Minh *et al.*,

2020) under default options (`-m LG`). We then rooted the tree with the MAD algorithm (Domingues Kümmel Tria, Landan and Dagan, 2017).

V.3 Protein family content among mouse subspecies

First, protein family PAPs were obtained for all pairs of host subspecies-bacteria in R. Using the Python command (Van Rossum G and Drake FL, 2009), all the proteins present in at least one *Bacteroides* strain, or in all the strains from *dom* or *mus,* were extracted. Next, the protein families of each *Bacteroides* strain shared between *dom* and *mus* or those that are host subspecies-specific were identified. Venn diagrams showing shared and unique *Bacteroides* protein families between *dom* and *mus* were plotted using the "VennDiagram" R package (v.3.6.3) (Chen and Boutros, 2011).

Supplementary material

I. Supplementary tables

Supplementary table 1. Summary of *Bacteroides* isolates classification by RDP, based on nearly-full length of 16S rRNA gene. The DNA fragments were sequenced by Sanger.

Isolate	Mouse line	Mouse subspecies	S_ab score	Best match
AH2511	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH2512	Iran/AH	*domesticus*	0.95	*Bacteroides* sp. TP-5; AB499846
AH2513	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH2516	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH2517	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25113	Iran/AH	*domesticus*	0.95	*Bacteroides* sp. TP-5; AB499846
AH25114	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25115	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25116	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25117	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25118	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25119	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25121	Iran/AH	*domesticus*	0.95	*Bacteroides* sp. TP-5; AB499846
AH25123	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH25124	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25125	Iran/AH	*domesticus*	0.95	*Bacteroides* sp. TP-5; AB499846
AH25126	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25127	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25129	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH25130	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH7633	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH7635	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH7639	Iran/AH	*domesticus*	0.97	*B. acidifaciens*; A43; AB021165
AH76310	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76311	Iran/AH	*domesticus*	0.97	*B. acidifaciens*; AB021157

AH76312	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76313	Iran/AH	*domesticus*	1.00	*B. acidifaciens* A43; AB021165
AH76315	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76317	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76318	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76319	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH76320	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76321	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76322	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76323	Iran/AH	*domesticus*	1.00	*B. acidifaciens* A43; AB021165
AH76325	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH76327	Iran/AH	*domesticus*	0.96	*Bacteroides* sp. TP-5; AB499846
AH76328	Iran/AH	*domesticus*	0.97	*Bacteroides* sp. TP-5; AB499846
AH76333	Iran/AH	*domesticus*	1.00	*B. acidifaciens* A43; AB021165
AH76336	Iran/AH	*domesticus*	1.00	*B. acidifaciens* A43; AB021165
AH59815	Iran/AH	*domesticus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
AH59816	Iran/AH	*domesticus*	0.94	*B. acidifaciens*; A1; AB021158
MC70128	France/MC	*domesticus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
MC70144	France/MC	*domesticus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
MC0831	France/MC	*domesticus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
MC94623	France/MC	*domesticus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
MC94637	France/MC	*domesticus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
MC94642	France/MC	*domesticus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
CB0762	France/MC	*domesticus*	1.00	*B. acidifaciens*; A43; AB021165
CB0764	France/MC	*domesticus*	1.00	*B. acidifaciens*; A43; AB021165
CB0768	France/MC	*domesticus*	1.00	*B. acidifaciens*; A43; AB021165
CB0101	France/MC	*domesticus*	0.98	*B. acidifaciens*; A43; AB021165
CB0109	France/MC	*domesticus*	0.98	*B. acidifaciens*; A43; AB021165
CB01011	France/MC	*domesticus*	0.98	*B. acidifaciens*; A43; AB021165
CB01023	France/MC	*domesticus*	0.99	*B. acidifaciens*; A43; AB021165
CB01024	France/MC	*domesticus*	0.98	*B. acidifaciens*; A43; AB021165

KH3651	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH3652	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
KH3533	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH3536	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH35310	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35312	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35315	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35316	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
KH35318	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35330	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH35337	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH35338	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35339	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35340	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35341	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35342	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35343	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35344	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35345	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH35346	Kazakhstan/KH	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
KH3461	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
KH3463	Kazakhstan/KH	*musculus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
KH3468	Kazakhstan/KH	*musculus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
KH34616	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
KH34641	Kazakhstan/KH	*musculus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
KH34642	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
KH34656	Kazakhstan/KH	*musculus*	0.98	*B. acidifaciens*; JCM 10556; AB510696
KH34658	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
KH5697	Kazakhstan/KH	*musculus*	0.99	*B. acidifaciens*; JCM 10556; AB510696
WI2961	Austria/WI	*musculus*	0.97	*B. acidifaciens*; A43; AB021165
WI2969	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165

WI29610	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI29612	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI29613	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI29614	Austria/WI	*musculus*	0.98	*B. acidifaciens*; A43; AB021165
WI29615	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI29616	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI3951	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI3952	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI5353	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI5354	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI5355	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI5356	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI5357	Austria/WI	*musculus*	0.98	*B. acidifaciens*; A43; AB021165
WI5358	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI5359	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI53510	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI53511	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53512	Austria/WI	*musculus*	0.98	*B. acidifaciens*; A43; AB021165
WI53513	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI53514	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53515	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53516	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53517	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53518	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53519	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI53520	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69313	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI69314	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69315	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69316	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI69317	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165

WI69318	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69319	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69320	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI69321	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85213	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85214	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85215	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85216	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85217	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI85218	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85219	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85220	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI85221	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85222	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85223	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI85224	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58713	Austria/WI	*musculus*	0.98	*B. acidifaciens*; A43; AB021165
WI58714	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI58715	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58716	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58717	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58718	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58719	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58720	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58721	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58722	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165
WI58723	Austria/WI	*musculus*	0.99	*B. acidifaciens*; A43; AB021165
WI58724	Austria/WI	*musculus*	1.00	*B. acidifaciens*; A43; AB021165

Supplementary table 2. Summary statistics for 146 sequenced *Bacteroides* isolates genomes. Proteins seqs: number of protein sequences; contigs: number of contigs per genome; contigs max: maximum length of the contig; contigs min – minimum length of contigs.

Genome ID	Protein seqs	Contigs	Contigs max (bp)	Contigs min (bp)
AH251_1	4140	308	395447	78
AH251_2	4373	315	363659	78
AH251_3	4384	320	382455	78
AH251_6	4379	287	363659	78
AH251_7	4371	300	381580	78
AH251_13	4382	279	291759	78
AH251_14	4371	274	336054	78
AH251_15	4377	294	364097	78
AH251_16	4375	282	336050	78
AH251_17	4383	307	363659	78
AH251_18	4380	267	363659	78
AH251_19	4353	280	363659	78
AH251_21	4379	277	363659	78
AH251_23	4424	300	387837	78
AH251_24	4354	316	363659	78
AH251_25	4374	320	363659	78
AH251_26	4354	296	363659	78
AH251_27	4343	325	382454	78
AH251_29	4381	297	390065	78
AH251_30	4345	256	390065	78
AH76_33	4453	285	363659	78
AH76_35	4486	269	363659	78
AH76_39	4478	267	363659	78
AH763_10	4473	297	363659	78
AH763_11	4476	263	363659	78
AH763_12	4485	281	336286	78
AH763_13	3540	417	279561	78
AH763_15	4475	275	336290	78
AH763_17	4478	252	336249	78
AH763_18	4485	289	336290	78
AH763_19	4483	248	363659	78
AH763_20	4480	264	336292	78
AH763_21	4474	284	363655	78
AH763_22	4454	269	336261	78
AH763_23	4471	324	363659	78
AH763_25	4478	318	336294	78
AH763_27	4480	280	336290	78
AH763_28	4455	293	336292	78
AH763_33	3547	370	279561	78
AH763_36	4479	284	454778	78
AH598_15	3966	588	208242	78
AH598_16	3967	449	156019	78
MC701_28	3857	399	208228	78
MC701_44	3856	437	208464	78
MC08_31	4011	462	177943	78
MC946_23	3917	452	208231	78
MC946_37	3911	448	232097	78
MC946_42	3928	428	208461	78
CB07_62	3985	621	264223	78
CB07_64	3986	618	264223	78

CB07_68	3986	719	264223	78
CB010_1	4036	372	322156	78
CB010_9	4039	339	313795	78
CB010_11	2386	361	313795	78
CB010_23	4038	347	322156	78
CB010_24	4039	359	322156	78
KH346_1	3670	462	251702	78
KH346_3	3864	403	205077	78
KH346_8	3885	428	214854	78
KH346_16	3861	683	267227	78
KH346_41	3887	465	211486	78
KH346_42	3889	469	208063	78
KH346_56	3860	443	177952	78
KH346_58	3860	445	289877	78
KH365_1	3731	555	248022	78
KH365_2	3670	403	297506	78
KH353_3	3846	563	226145	78
KH353_6	3763	569	226145	78
KH353_10	3788	581	226676	78
KH353_12	3772	595	226675	78
KH353_15	3787	567	221640	78
KH353_16	3845	496	261029	78
KH353_18	3772	591	213265	78
KH353_30	3784	416	210431	78
KH353_37	3768	567	226147	78
KH353_38	3781	551	226145	78
KH353_39	3766	562	226145	78
KH353_40	3767	587	226676	78
KH353_41	3763	581	226145	78
KH353_42	3763	570	212029	78
KH353_43	3784	573	165286	78
KH353_44	3794	564	226146	78
KH353_45	3767	583	226683	78
KH353_46	3786	538	226675	78
KH569_7	3721	284	231469	78
WI296_1	3949	656	288929	78
WI296_9	3945	755	288765	78
WI296_10	3949	684	288929	78
WI296_12	3946	683	245886	78
WI296_13	3947	733	252991	78
WI296_14	3946	720	253123	78
WI296_15	3928	647	289029	78
WI296_16	3950	727	252993	78
WI395_1	3943	740	252991	78
WI395_2	4006	642	288850	78
WI535_3	3939	706	245508	78
WI535_4	3938	697	288463	78
WI535_5	3945	703	288851	78
WI535_6	3950	632	288851	78
WI535_7	3947	695	252991	78
WI535_8	3944	686	288178	78
WI535_9	3946	666	289495	78
WI535_10	3941	679	288199	78
WI535_11	3944	699	288851	78
WI535_12	3941	719	252991	78
WI535_13	3942	680	289417	78
WI535_14	3940	697	288851	78

WI535_15	3946	639	289495	78
WI535_16	3949	654	288199	78
WI535_17	3944	695	252991	78
WI535_18	3950	665	252991	78
WI535_19	3951	645	289417	78
WI535_20	3945	691	288199	78
WI693_13	3938	691	288198	78
WI693_14	3940	662	252991	78
WI693_15	3934	665	252991	78
WI693_16	3945	699	288198	78
WI693_17	3931	699	289016	78
WI693_18	3938	665	253123	78
WI693_19	3944	670	288928	78
WI693_20	3937	705	288462	78
WI693_21	3942	653	289482	78
WI852_13	4013	697	288589	78
WI852_14	3949	685	288589	78
WI852_15	3935	727	247596	78
WI852_16	4003	716	252989	78
WI852_17	4019	683	288589	78
WI852_18	3946	670	288178	78
WI852_19	4023	708	288589	78
WI852_20	3944	664	288589	78
WI852_21	4159	704	288589	78
WI852_22	4012	697	288589	78
WI852_23	4153	743	288589	78
WI852_24	4158	693	288178	78
WI587_13	4098	669	252989	78
WI587_14	4096	705	288589	78
WI587_15	3952	724	253121	78
WI587_16	4083	661	288589	78
WI587_17	4090	706	253121	78
WI587_18	4096	702	289155	78
WI587_19	3952	686	288851	78
WI587_20	4088	701	289417	78
WI587_21	4087	685	288851	78
WI587_22	4091	657	288199	78
WI587_23	4093	725	253121	78
WI587_24	4089	735	288851	78
B. acidifaciens A40	4140	632	167112	78

Supplementary table 3. Summary of TYGS classification. Mouse subspecies - mouse subspecies of origin; dDDH – digital DNA-DNA Hybridization; Diff. G+C – GC content difference. Only the highest dDDH values are shown.

Genome ID	Classification	Mouse subspecies	dDDH (%)	Diff. G+C (%)
KH346_1	*B. acidifaciens*	*musculus*	59.60	0.45
KH346_3	*B. acidifaciens*	*musculus*	59.60	0.39
KH346_8	*B. acidifaciens*	*musculus*	59.40	0.82
KH346_41	*B. acidifaciens*	*musculus*	59.40	0.84
KH346_42	*B. acidifaciens*	*musculus*	59.40	1.05
KH346_56	*B. acidifaciens*	*musculus*	59.60	0.29
KH346_58	*B. acidifaciens*	*musculus*	59.60	0.79
KH353_16	*B. acidifaciens*	*musculus*	60.00	1.12
KH353_30	*B. acidifaciens*	*musculus*	60.00	1.03
AH598_15	*B. acidifaciens*	*domesticus*	59.70	1.11
AH598_16	*B. acidifaciens*	*domesticus*	60.30	1.05
MC701_28	*B. acidifaciens*	*domesticus*	59.90	0.98
MC701_44	*B. acidifaciens*	*domesticus*	59.90	0.62
MC083_1	*B. acidifaciens*	*domesticus*	59.90	0.82
MC946_23	*B. acidifaciens*	*domesticus*	61.10	1.67
MC946_37	*B. acidifaciens*	*domesticus*	61.00	0.38
MC946_42	*B. acidifaciens*	*domesticus*	61.10	0.79
KH365_2	Unclassified	*musculus*	39.60	1.62
KH569_7	Unclassified	*musculus*	38.10	3.67
AH251_1	*B. sartorii*	*domesticus*	69.00	2.17
AH251_2	*B. sartorii*	*domesticus*	68.80	2.51
AH251_3	*B. sartorii*	*domesticus*	68.90	2.24
AH251_6	*B. sartorii*	*domesticus*	68.80	1.85
AH251_7	*B. sartorii*	*domesticus*	68.80	2.22
AH251_13	*B. sartorii*	*domesticus*	68.90	2.54
AH251_14	*B. sartorii*	*domesticus*	68.80	1.36
AH251_15	*B. sartorii*	*domesticus*	68.80	1.85
AH251_16	*B. sartorii*	*domesticus*	68.80	1.90
AH251_17	*B. sartorii*	*domesticus*	68.80	2.23
AH251_18	*B. sartorii*	*domesticus*	68.90	1.77
AH251_19	*B. sartorii*	*domesticus*	68.90	2.12
AH251_21	*B. sartorii*	*domesticus*	68.80	1.55
AH251_23	*B. sartorii*	*domesticus*	68.80	1.59
AH251_24	*B. sartorii*	*domesticus*	68.80	2.43
AH251_25	*B. sartorii*	*domesticus*	69.00	2.46
AH251_26	*B. sartorii*	*domesticus*	68.80	1.53
AH251_27	*B. sartorii*	*domesticus*	68.80	2.01
AH251_29	*B. sartorii*	*domesticus*	68.90	2.17
AH251_30	*B. sartorii*	*domesticus*	68.90	2.21
AH763_3	*B. sartorii*	*domesticus*	68.60	2.06
AH763_5	*B. sartorii*	*domesticus*	68.50	1.83
AH763_9	*B. sartorii*	*domesticus*	68.50	2.08
AH763_10	*B. sartorii*	*domesticus*	68.60	1.93
AH763_11	*B. sartorii*	*domesticus*	68.50	2.21
AH763_12	*B. sartorii*	*domesticus*	68.60	2.70
AH763_15	*B. sartorii*	*domesticus*	68.50	2.03
AH763_17	*B. sartorii*	*domesticus*	68.60	2.00
AH763_18	*B. sartorii*	*domesticus*	68.50	2.09
AH763_19	*B. sartorii*	*domesticus*	68.60	1.52
AH763_20	*B. sartorii*	*domesticus*	68.50	1.61

AH763_21	*B. sartorii*	*domesticus*	68.50	2.22
AH763_22	*B. sartorii*	*domesticus*	68.60	2.00
AH763_23	*B. sartorii*	*domesticus*	68.50	2.13
AH763_25	*B. sartorii*	*domesticus*	68.60	2.34
AH763_27	*B. sartorii*	*domesticus*	68.60	2.11
AH763_28	*B. sartorii*	*domesticus*	68.7	1.86
AH763_36	*B. sartorii*	*domesticus*	68.50	2.03
WI296_1	*B. caecimuris*	*musculus*	65.90	0.41
WI296_9	*B. caecimuris*	*musculus*	65.80	0.71
WI296_10	*B. caecimuris*	*musculus*	65.80	0.70
WI296_12	*B. caecimuris*	*musculus*	65.80	0.75
WI296_13	*B. caecimuris*	*musculus*	65.90	0.70
WI296_14	*B. caecimuris*	*musculus*	65.90	0.64
WI296_15	*B. caecimuris*	*musculus*	65.80	0.39
WI296_16	*B. caecimuris*	*musculus*	65.80	0.53
WI395_1	*B. caecimuris*	*musculus*	66.10	0.63
WI395_2	*B. caecimuris*	*musculus*	65.30	0.80
WI535_3	*B. caecimuris*	*musculus*	65.80	0.63
WI535_4	*B. caecimuris*	*musculus*	65.80	0.47
WI535_5	*B. caecimuris*	*musculus*	65.90	0.55
WI535_6	*B. caecimuris*	*musculus*	65.80	0.51
WI535_7	*B. caecimuris*	*musculus*	65.90	0.59
WI535_8	*B. caecimuris*	*musculus*	65.80	0.29
WI535_9	*B. caecimuris*	*musculus*	65.80	0.35
WI535_10	*B. caecimuris*	*musculus*	65.80	0.28
WI535_11	*B. caecimuris*	*musculus*	65.80	0.62
WI535_12	*B. caecimuris*	*musculus*	65.90	0.65
WI535_13	*B. caecimuris*	*musculus*	65.90	0.51
WI535_14	*B. caecimuris*	*musculus*	65.90	0.75
WI535_15	*B. caecimuris*	*musculus*	65.80	0.72
WI535_16	*B. caecimuris*	*musculus*	65.80	0.29
WI535_17	*B. caecimuris*	*musculus*	65.80	0.36
WI535_18	*B. caecimuris*	*musculus*	65.90	0.60
WI535_19	*B. caecimuris*	*musculus*	65.80	0.69
WI535_20	*B. caecimuris*	*musculus*	65.80	0.82
WI693_13	*B. caecimuris*	*musculus*	66.10	0.78
WI693_14	*B. caecimuris*	*musculus*	66.00	0.78
WI693_15	*B. caecimuris*	*musculus*	65.90	0.97
WI693_16	*B. caecimuris*	*musculus*	66.10	0.79
WI693_17	*B. caecimuris*	*musculus*	66.00	0.88
WI693_18	*B. caecimuris*	*musculus*	66.00	0.39
WI693_19	*B. caecimuris*	*musculus*	66.00	0.88
WI693_20	*B. caecimuris*	*musculus*	66.10	1.00
WI693_21	*B. caecimuris*	*musculus*	66.00	0.76
WI852_13	*B. caecimuris*	*musculus*	65.40	0.26
WI852_14	*B. caecimuris*	*musculus*	66.10	0.55
WI852_15	*B. caecimuris*	*musculus*	66.10	0.76
WI852_16	*B. caecimuris*	*musculus*	65.50	0.45
WI852_17	*B. caecimuris*	*musculus*	65.30	0.69
WI852_18	*B. caecimuris*	*musculus*	66.10	0.53
WI852_19	*B. caecimuris*	*musculus*	65.30	0.57
WI852_20	*B. caecimuris*	*musculus*	66.10	0.51
WI852_21	*B. caecimuris*	*musculus*	63.80	0.30
WI852_22	*B. caecimuris*	*musculus*	65.40	0.82
WI852_23	*B. caecimuris*	*musculus*	63.80	0.87
WI852_24	*B. caecimuris*	*musculus*	63.80	0.86
WI587_13	*B. caecimuris*	*musculus*	64.40	0.80

WI587_14	*B. caecimuris*	*musculus*	64.40	0.83
WI587_15	*B. caecimuris*	*musculus*	66.00	0.72
WI587_16	*B. caecimuris*	*musculus*	64.40	0.59
WI587_17	*B. caecimuris*	*musculus*	64.40	0.73
WI587_18	*B. caecimuris*	*musculus*	64.40	0.87
WI587_19	*B. caecimuris*	*musculus*	65.90	0.62
WI587_20	*B. caecimuris*	*musculus*	64.40	0.51
WI587_21	*B. caecimuris*	*musculus*	64.30	0.57
WI587_22	*B. caecimuris*	*musculus*	64.40	0.84
WI587_23	*B. caecimuris*	*musculus*	64.40	0.67
WI587_24	*B. caecimuris*	*musculus*	64.40	0.66
CB076_2	*B. caecimuris*	*domesticus*	65.30	0.39
CB076_4	*B. caecimuris*	*domesticus*	65.30	0.63
CB076_8	*B. caecimuris*	*domesticus*	65.30	0.36
CB010_1	*B. caecimuris*	*domesticus*	52.90	1.88
CB010_9	*B. caecimuris*	*domesticus*	52.90	1.83
CB010_11	*B. caecimuris*	*domesticus*	52.90	1.91
CB010_23	*B. caecimuris*	*domesticus*	52.90	1.44
CB010_24	*B. caecimuris*	*domesticus*	52.90	1.55
AH763_13	*B. caecimuris*	*domesticus*	69.60	2.04
AH763_33	*B. caecimuris*	*domesticus*	69.60	2.15
KH346_16	*B. caecimuris*	*musculus*	68.70	1.23
KH365_1	*B. caecimuris*	*musculus*	69.30	0.11
KH353_3	*B. caecimuris*	*musculus*	72.50	0.34
KH353_6	*B. caecimuris*	*musculus*	72.50	0.41
KH353_10	*B. caecimuris*	*musculus*	72.30	0.03
KH353_12	*B. caecimuris*	*musculus*	72.40	0.45
KH353_15	*B. caecimuris*	*musculus*	72.30	0.27
KH353_18	*B. caecimuris*	*musculus*	72.50	0.54
KH353_37	*B. caecimuris*	*musculus*	72.30	0.38
KH353_38	*B. caecimuris*	*musculus*	72.50	0.09
KH353_39	*B. caecimuris*	*musculus*	72.00	0.19
KH353_40	*B. caecimuris*	*musculus*	72.50	0.22
KH353_41	*B. caecimuris*	*musculus*	72.50	0.26
KH353_42	*B. caecimuris*	*musculus*	72.40	0.47
KH353_43	*B. caecimuris*	*musculus*	72.20	0.57
KH353_44	*B. caecimuris*	*musculus*	72.30	0.09
KH353_45	*B. caecimuris*	*musculus*	72.40	0.40
KH353_46	*B. caecimuris*	*musculus*	72.30	0.09

Supplementary table 4. Summary of GTDB-Tk classification. Mouse subspecies - mouse subspecies of origin; ANI – average nucleotide identity; Ref. ANI – the accession number of the closest reference genome as determined by ANI.

Genome ID	Classification	Mouse subspecies	ANI (%)	Ref. ANI
KH346_1	*B. acidifaciens*	*musculus*	98.09	GCA000613385.1
KH346_3	*B. acidifaciens*	*musculus*	98.08	GCA000613385.1
KH346_8	*B. acidifaciens*	*musculus*	98.05	GCA000613385.1
KH346_41	*B. acidifaciens*	*musculus*	98.06	GCA000613385.1
KH346_42	*B. acidifaciens*	*musculus*	98.07	GCA000613385.1
KH346_56	*B. acidifaciens*	*musculus*	98.05	GCA000613385.1
KH346_58	*B. acidifaciens*	*musculus*	98.11	GCA000613385.1
KH353_16	*B. acidifaciens*	*musculus*	98.06	GCA000613385.1
KH353_30	*B. acidifaciens*	*musculus*	98.10	GCA000613385.1
AH598_15	*B. acidifaciens*	*domesticus*	98.19	GCA000613385.1
AH598_16	*B. acidifaciens*	*domesticus*	98.18	GCA000613385.1
MC701_28	*B. acidifaciens*	*domesticus*	98.13	GCA000613385.1
MC701_44	*B. acidifaciens*	*domesticus*	98.14	GCA000613385.1
MC083_1	*B. acidifaciens*	*domesticus*	98.22	GCA000613385.1
MC946_23	*B. acidifaciens*	*domesticus*	98.11	GCA000613385.1
MC946_37	*B. acidifaciens*	*domesticus*	98.02	GCA000613385.1
MC946_42	*B. acidifaciens*	*domesticus*	98.09	GCA000613385.1
KH365_2	Unclassified	*musculus*	98.18	GCA002491635.1
KH569_7	Unclassified	*musculus*	98.23	GCA002491635.1
AH251_1	*B. sartorii*	*domesticus*	98.25	GCA000614185.1
AH251_2	*B. sartorii*	*domesticus*	98.25	GCA000614185.1
AH251_3	*B. sartorii*	*domesticus*	98.29	GCA000614185.1
AH251_6	*B. sartorii*	*domesticus*	98.27	GCA000614185.1
AH251_7	*B. sartorii*	*domesticus*	98.28	GCA000614185.1
AH251_13	*B. sartorii*	*domesticus*	98.27	GCA000614185.1
AH251_14	*B. sartorii*	*domesticus*	98.24	GCA000614185.1
AH251_15	*B. sartorii*	*domesticus*	98.30	GCA000614185.1
AH251_16	*B. sartorii*	*domesticus*	98.29	GCA000614185.1
AH251_17	*B. sartorii*	*domesticus*	98.20	GCA000614185.1
AH251_18	*B. sartorii*	*domesticus*	98.26	GCA000614185.1
AH251_19	*B. sartorii*	*domesticus*	98.26	GCA000614185.1
AH251_21	*B. sartorii*	*domesticus*	98.31	GCA000614185.1
AH251_23	*B. sartorii*	*domesticus*	98.25	GCA000614185.1
AH251_24	*B. sartorii*	*domesticus*	98.28	GCA000614185.1
AH251_25	*B. sartorii*	*domesticus*	98.30	GCA000614185.1
AH251_26	*B. sartorii*	*domesticus*	98.25	GCA000614185.1
AH251_27	*B. sartorii*	*domesticus*	98.28	GCA000614185.1
AH251_29	*B. sartorii*	*domesticus*	98.29	GCA000614185.1
AH251_30	*B. sartorii*	*domesticus*	98.24	GCA000614185.1
AH76_33	*B. sartorii*	*domesticus*	98.24	GCA000614185.1
AH76_35	*B. sartorii*	*domesticus*	98.30	GCA000614185.1
AH76_39	*B. sartorii*	*domesticus*	98.26	GCA000614185.1
AH763_10	*B. sartorii*	*domesticus*	98.28	GCA000614185.1
AH763_11	*B. sartorii*	*domesticus*	98.23	GCA000614185.1
AH763_12	*B. sartorii*	*domesticus*	98.31	GCA000614185.1
AH763_15	*B. sartorii*	*domesticus*	98.30	GCA000614185.1
AH763_17	*B. sartorii*	*domesticus*	98.24	GCA000614185.1
AH763_18	*B. sartorii*	*domesticus*	98.27	GCA000614185.1
AH763_19	*B. sartorii*	*domesticus*	98.26	GCA000614185.1
AH763_20	*B. sartorii*	*domesticus*	98.26	GCA000614185.1
AH763_21	*B. sartorii*	*domesticus*	98.23	GCA000614185.1
AH763_22	*B. sartorii*	*domesticus*	98.25	GCA000614185.1
AH763_23	*B. sartorii*	*domesticus*	98.30	GCA000614185.1
AH763_25	*B. sartorii*	*domesticus*	98.28	GCA000614185.1
AH763_27	*B. sartorii*	*domesticus*	98.29	GCA000614185.1
AH763_28	*B. sartorii*	*domesticus*	98.25	GCA000614185.1

AH763_36	*B. sartorii*	*domesticus*	98.21	GCA000614185.1
WI296_1	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI296_9	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI296_10	*B. caecimuris*	*musculus*	98.11	GCF001688725.2
WI296_12	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI296_13	*B. caecimuris*	*musculus*	98.03	GCF001688725.2
WI296_14	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI296_15	*B. caecimuris*	*musculus*	98.03	GCF001688725.2
WI296_16	*B. caecimuris*	*musculus*	98.01	GCF001688725.2
WI395_1	*B. caecimuris*	*musculus*	98.03	GCF001688725.2
WI395_2	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI535_3	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI535_4	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI535_5	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI535_6	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI535_7	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI535_8	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI535_9	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI535_10	*B. caecimuris*	*musculus*	98.03	GCF001688725.2
WI535_11	*B. caecimuris*	*musculus*	98.13	GCF001688725.2
WI535_12	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI535_13	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI535_14	*B. caecimuris*	*musculus*	98.08	GCF001688725.2
WI535_15	*B. caecimuris*	*musculus*	98.08	GCF001688725.2
WI535_16	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI535_17	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI535_18	*B. caecimuris*	*musculus*	98.11	GCF001688725.2
WI535_19	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI535_20	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI693_13	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI693_14	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI693_15	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI693_16	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI693_17	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI693_18	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI693_19	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI693_20	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI693_21	*B. caecimuris*	*musculus*	97.97	GCF001688725.2
WI852_13	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI852_14	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI852_15	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI852_16	*B. caecimuris*	*musculus*	98.01	GCF001688725.2
WI852_17	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI852_18	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI852_19	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI852_20	*B. caecimuris*	*musculus*	98.03	GCF001688725.2
WI852_21	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI852_22	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI852_23	*B. caecimuris*	*musculus*	98.05	GCF001688725.2
WI852_24	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI587_13	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI587_14	*B. caecimuris*	*musculus*	98.06	GCF001688725.2
WI587_15	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI587_16	*B. caecimuris*	*musculus*	98.02	GCF001688725.2
WI587_17	*B. caecimuris*	*musculus*	98.01	GCF001688725.2
WI587_18	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI587_19	*B. caecimuris*	*musculus*	98.04	GCF001688725.2
WI587_20	*B. caecimuris*	*musculus*	98.09	GCF001688725.2
WI587_21	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI587_22	*B. caecimuris*	*musculus*	98.10	GCF001688725.2
WI587_23	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
WI587_24	*B. caecimuris*	*musculus*	98.07	GCF001688725.2
CB076_2	*B. caecimuris*	*domesticus*	98.01	GCF001688725.2

CB076_4	*B. caecimuris*	*domesticus*	98.04	GCF001688725.2
CB076_8	*B. caecimuris*	*domesticus*	98.04	GCF001688725.2
CB010_1	*B. caecimuris*	*domesticus*	96.67	GCF001688725.2
CB010_9	*B. caecimuris*	*domesticus*	96.65	GCF001688725.2
CB010_11	*B. caecimuris*	*domesticus*	96.65	GCF001688725.2
CB010_23	*B. caecimuris*	*domesticus*	96.69	GCF001688725.2
CB010_24	*B. caecimuris*	*domesticus*	96.69	GCF001688725.2
AH763_13	*B. caecimuris*	*domesticus*	98.19	GCF001688725.2
AH763_33	*B. caecimuris*	*domesticus*	98.20	GCF001688725.2
KH346_16	*B. caecimuris*	*musculus*	98.26	GCF001688725.2
KH365_1	*B. caecimuris*	*musculus*	97.89	GCF001688725.2
KH353_3	*B. caecimuris*	*musculus*	98.22	GCF001688725.2
KH353_6	*B. caecimuris*	*musculus*	98.20	GCF001688725.2
KH353_10	*B. caecimuris*	*musculus*	98.18	GCF001688725.2
KH353_12	*B. caecimuris*	*musculus*	98.21	GCF001688725.2
KH353_15	*B. caecimuris*	*musculus*	98.18	GCF001688725.2
KH353_18	*B. caecimuris*	*musculus*	98.20	GCF001688725.2
KH353_37	*B. caecimuris*	*musculus*	98.24	GCF001688725.2
KH353_38	*B. caecimuris*	*musculus*	98.20	GCF001688725.2
KH353_39	*B. caecimuris*	*musculus*	98.17	GCF001688725.2
KH353_40	*B. caecimuris*	*musculus*	98.21	GCF001688725.2
KH353_41	*B. caecimuris*	*musculus*	98.15	GCF001688725.2
KH353_42	*B. caecimuris*	*musculus*	98.22	GCF001688725.2
KH353_43	*B. caecimuris*	*musculus*	98.21	GCF001688725.2
KH353_44	*B. caecimuris*	*musculus*	98.21	GCF001688725.2
KH353_45	*B. caecimuris*	*musculus*	98.22	GCF001688725.2
KH353_46	*B. caecimuris*	*musculus*	98.22	GCF001688725.2

Supplementary table 5. Summary of the *Bacteroides* ASVs and respective 100% matches among sequenced isolates. The indicator ASVs are shown in bold.

***Bacteroides* ASV**	**Isolate ID**	**Isolate classification**
ASV 268	AH598_15 AH598_16 KH346_1 KH346_3 KH346_8 KH346_41 KH346_42 KH346_56 KH346_58 KH353_16 KH353_30 MC083_1 MC701_28 MC701_44 MC946_23 MC946_37 MC946_42	*B. acidifaciens*
ASV 35	KH365_2 KH569_7	Unclassified
ASV 5872	AH251_7	*B. sartorii*
ASV 2406	AH251_1 AH251_26 AH251_30 AH763_18	
ASV 2348	AH251_13	
ASV 242	AH251_19 KH365_1	*B. caecimuris*
ASV 311	CB010_1 CB010_9 CB010_11 CB010_23 CB010_24	
ASV 110	AH763_13 AH763_33 CB076_2 CB076_4 CB076_8 KH346_16 WI296_1 WI296_9 WI296_10 WI296_12 WI296_13 WI296_14 WI296_15 WI296_16 WI395_1 WI395_2 WI535_3 WI535_4 WI535_5 WI535_6 WI535_7 WI535_8 WI535_9 WI535_10 WI535_11	

WI535_12
WI535_13
WI535_14
WI535_15
WI535_16
WI535_17
WI535_18
WI535_19
WI535_20
WI693_13
WI693_14
WI693_15
WI693_16
WI693_17
WI693_18
WI693_19
WI693_20
WI693_21
WI852_13
WI852_14
WI852_15
WI852_16
WI852_17
WI852_18
WI852_19
WI852_20
WI852_21
WI852_22
WI852_23
WI852_24
WI587_13
WI587_14
WI587_15
WI587_16
WI587_17
WI587_18
WI587_19
WI587_20
WI587_21
WI587_22
WI587_23
WI587_24

Supplementary table 6. Summary of the functions predicted for the protein families identified as host subspecies-specific.

Host subspecies	Bacteria	Protein family ID	Name (UniProt)	Biological process/ function
domesticus	*B. acidifaciens*	503	Multidrug resistance protein MexB	Efflux transmembrane transporter activity
		2534	Ribosomal protein S12 methylthiotransferase accessory factor YcaO	Beta-methylthiolation of ribosomal protein S12
		2797 2922 3014	Fluoroquinolones export permease	Part of the ABC transporter complex involved in fluoroquinolones export. Confers resistance to ciprofloxacin and, to a lesser extent, norfloxacin, moxifloxacin and sparfloxacin
		3042	Anaerobic sulfatase-maturating enzyme	Involved in 'Ser-type' sulfatase maturation under anaerobic conditions. Links the heparin and the chondroitin sulfate utilization pathways which contribute to the colonization of the intestinal tract
		3045	Thiol-disulfide oxidoreductase ResA	Oxidoreductase, involved in cell redox homeostasis
		3046	Lactococcin-G-processing and transport ATP-binding protein LagD	Antibiotic biosynthetic process. Might be involved in export of the bacteriocin lactococcin G
		3730	3-oxoacyl-[acyl-carrier-protein] synthase 3	Fatty acid biosynthesis pathway, lipid metabolism
		6522 6525	Fragilysin	Pathogenesis. Diarrheal toxin that hydrolyzes gelatin, azocoll, actin, tropomyosin, and fibrinogen
	B. caecimuris			
		196	Cellobiose 2-epimerase	Cellobiose epimerase activity
		232 327 497 3150 3154	TonB-dependent receptor SusC	Mediates transport of starch oligosaccharides from the surface of the outer membrane to the periplasm for subsequent degradation
		297	Thiol disulfide oxidoreductase ResA	Oxidoreductase activity
		441	Multidrug resistance protein MdtA	Transmembrane transporter activity, response to antibiotics
		468	Tyrosine recombinase XerD	DNA recombination (binds DNA)
		473	ECF RNA polymerase sigma factor SigE	DNA transcription, response to heat, response to host immunity

503	Multidrug resistance protein MexB	The inner membrane transporter component of the MexAB-OprM efflux system that confers multidrug resistance
1773	D-inositol 3-phosphate glycosyltransferase	Mycothiol biosynthetic process
1781	Dihydroanticapsin 7-dehydrogenase	Involved in the pathway bacilysin biosynthesis, which is part of antibiotic biosynthesis
1783	Polyphosphate kinase	Polyphosphate biosynthetic process, protein autophosphorylation
1791	Putative ribosomal N-acetyltransferase YdaF	tRNA aminoacylation
1792	4-hydroxythreonine-4-phosphate dehydrogenase	Pyridoxine biosynthetic process
1854 1993	Serine/threonine-protein kinase HipA	Calcium ion binding, kinase activity
1856	Thermophilic serine proteinase	Metal ion binding
1865	HTH-type transcriptional activator RhaR	DNA-binding transcription factor activity
1866	Molecular chaperone Hsp31 and glyoxalase 3	Glutamine metabolic process
1877 1948	N-acetylglucosaminyl-diphospho-decaprenol L-rhamnosyltransferase	Extracellular polysaccharide biosynthetic process
1916	Chloramphenicol acetyltransferase 3	Chloramphenicol O-acetyltransferase activity
1940	Tyrosine recombinase XerC	DNA recombination (binds DNA)
1946	Serine/threonine-protein kinase HipA	Calcium ion binding; kinase activity
1991	Glutaminase 1	Glutamate biosynthetic process, negative regulation of growth, response to acidic pH
1992	Glutamate decarboxylase	Glutamate metabolic process
2006	2-C-methyl-D-erythritol_4-phosphate_cytidylyltransferase	Oxidoreductase activity
2021	dTDP-glucose 4,6-dehydratase	Enterobacterial common antigen biosynthetic process
2064 2559 3153	SusD-like protein	Xyloglucan degradation pathway, a part of Glucan metabolism. Involved in symbiotic process benefiting host
2073	Glucitol operon repressor	DNA-binding transcription factor activity
2082	Dipeptidyl-peptidase 5	Peptide catabolic process
2083	Alkyl hydroperoxide reductase subunit C	Cell redox homeostasis
2084	Endo-polygalacturonase	Carbohydrate metabolic process
2376	Thiol-disulfide oxidoreductase ResA	Cell redox homeostasis, cytochrome complex assembly
2456	Plasmid recombination enzyme	DNA recombination
2509	Putative type I restriction enzymeP M protein	DNA binding, endonuclease activity
2510	Type-1 restriction enzyme R protein	DNA restriction-modification system
3047	Mannosylfructose-phosphate synthase	Disaccharide biosynthetic process
3129	Inositol 2-dehydrogenase	Inositol catabolic process, NADPH regeneration

		3134	Glycerol-3-phosphate dehydrogenase 2	Carbohydrate metabolic process
		3135	Glycerophosphodiester phosphodiesterase	Lipid metabolic process
		3136	Glycerol-1-phosphate dehydrogenase [NAD(P)+]	Phospholipid biosynthetic process
		3139	RNA polymerase sigma factor FliA	Regulation of transcription, DNA-templated
		3145	Unsaturated rhamnogalacturonyl hydrolase YteR	Metabolic process
		3157	Alkyl hydroperoxide reductase subunit F	Cell redox homeostasis, response to reactive oxygen species
		3159	Endo-1, 4-beta-xylanase Z	Xylan catabolic process
		3160	Endo-1, 4-beta-xylanase/feruloyl esterase	Xylan catabolic process, carbohydrate metabolic process
		3169	Beta-hexosaminidase	Carbohydrate metabolic process
		3172	Anaerobic sulfatase-maturating enzyme	Involved in 'Ser-type' sulfatase maturation under anaerobic conditions. Links the heparin and the chondroitin sulfate utilization pathways which contribute to the colonization of the intestinal tract
musculus	*B. acidifaciens*	278 2979	Tyrosine recombinase XerD Tyrosine recombinase XerC	DNA recombination (binds DNA)
		3729	tRNA(Ser)-specific nuclease WapA	Toxic component of a toxin-immunity protein module, functions as a cellular contact-dependent growth inhibition (CDI) system
		5287	Transcriptional regulator ClgR	Transcription, transcription regulation (binds DNA)
		6337	Fructose-bisphosphate aldolase class 1	Glycolysis
	B. caecimuris	88 2423	ECF RNA polymerase sigma factor SigW	DNA-templated transcription, initiation
		110 2539 2805	TonB-dependent receptor SusC	Mediates transport of starch oligosaccharides from the surface of the outer membrane to the periplasm
		129	Putative NAD(P)H-dependent FMN-containing oxidoreductase YwqN	Putative NADPH-dependent oxidoreductase
		264 388	Tyrosine recombinase XerC	DNA recombination (binds DNA)
		278	Tyrosine recombinase XerD	DNA recombination (binds DNA)
		390	Very short patch repair protein	Mismatch repair
		479	Putative RNA polymerase sigma factor FecI	Regulation of transcription
		480	Sensor histidine kinase TodS	Carbohydrate transport
		528	Scyllo-inositol 2-dehydrogenase (NADP(+))	Inositol catabolic process
		1785	Acetyl-coenzyme A synthetase	Acetyl-CoA biosynthetic process, histone acetylation
		1862	Putative TrmH family tRNA/rRNA methyltransferase	RNA methylation

1863	Alcohol dehydrogenase	Metal ion binding, oxidoreductase activity
1874	3', 5'-cyclic adenosine monophosphate phosphodiesterase CpdA	Metal ion binding
1875	Colicin I receptor	Iron transport. Outer membrane receptor for colicins IA and IB
1879 2001	Lipopolysaccharide assembly protein B	Lipopolysaccharide metabolic process, regulation of lipid biosynthetic process
1903	Group II intron-encoded protein LtrA	Multifunctional protein. Promotes group II intron splicing and mobility by acting both on RNA and DNA
1906	Serine recombinase PinR	DNA recombination (binds DNA)
1955	Ribosome-associated ATPase	Positive regulation of translation, transmembrane transport
1960	Virginiamycin A acetyltransferase	Provides resistance to virginiamycin-like antibiotics
1990	N-acetylglucosaminyldiphosphoundecaprenol N-acetyl-beta-D-mannosaminyltransferase	Biosynthetic process
2000	Phosphate-binding protein PstS	Cellular response to phosphate starvation, growth of symbiont in host
2015	Multidrug export protein EmrA	Response to antibiotic
2365	HTH-type transcriptional regulator SinR	DNA-binding transcription factor activity
2371	Endo-beta-N-acetylglucosaminidase F1	Carbohydrate metabolic process
2384	Sensor histidine kinase TmoS	Sensor kinase activity, ATP binding
2385	Endo-1,4-beta-xylanase Z	Xylan catabolic process
2390	Inner membrane protein YbaN	Membrane component
2416	ATP-dependent RecD-like DNA helicase	DNA recombination
2417	Beta-glucanase	Carbohydrate metabolic process
2424	Virulence regulon transcriptional activator VirF	Pathogenesis
2428	Outer membrane protein TolC	Bile acid and bile salt transport, response to antibiotics
2430	HTH-type transcriptional regulator YesS	DNA-binding transcription factor activity
2434	Serine/threonine-protein kinase HipA	Calcium ion binding, kinase activity
2435	Putative HTH-type transcriptional regulator YybR	DNA-binding transcription factor activity
2437	Exoenzyme S synthesis regulatory protein ExsA	Phosphorelay signal transduction system
2787	Chaperone protein DnaK	Protein refolding
2796	ECF RNA polymerase sigma factor EcfG	DNA-templated transcription, initiation
2802	Inositol 2-dehydrogenase	Polyol metabolism
2811	Group II intron-encoded protein LtrA	Multifunctional protein that promotes group II intron splicing and mobility by acting both on RNA and DNA
2820	Nicotinate phosphoribosyltransferase	NAD biosynthetic process

		2870	Modification methylase ApII	DNA methylation on cytosine
		2873	Alpha-D-kanosaminyltransferase	Transferase activity

Chapter III:

Antagonistic bacteria-bacteria interactions among *Bacteroides* isolates

Introduction

The members of the *Bacteroides* genus, which belong to the order Bacteroidales, are dominant Gram-negative bacteria in the mammalian gut. *Bacteroides* exclusively inhabit the gastrointestinal tracts of mammals and establish stable, long-term associations with their hosts, suggesting strong adaptation to the gut environment (Ley *et al.*, 2008; Moeller *et al.*, 2014, 2017). Moreover, these bacteria are physically in contact with each other in the gut (Salyers, Shoemaker and Li, 1995; Coyne *et al.*, 2014). It was already shown that the colonization of the human gut with more than one strain belonging to the same Bacteroidales species is common, indicating the promotion of cocolonization (Bjerke *et al.*, 2011; Zitomersky, Coyne and Comstock, 2011). In addition to the characteristics that permit these bacteria to cocolonize, they also evolved mechanisms to antagonize each other (Wexler and Goodman, 2017). Since Bacteroidetes members naturally live in complex communities, the production of antimicrobial compounds that target related members provides a competitive advantage, and plays an important role in the assembly and maintenance of these microbial communities (García-Bayona and Comstock, 2018).

Bacteroidales were shown to engage in two different types of antagonistic interactions: contact-dependent type VI secretion systems (T6SSs) (Russell *et al.*, 2014; Chatzidaki-Livanis *et al.*, 2016) and secreted diffusible antimicrobial toxins (Chatzidaki-Livanis, Coyne and Comstock, 2014b; Coyne *et al.*, 2019). The previous study by Coyne, Roelofs and Comstock (2016) revealed most of human gut *B. fragilis* strains to carry genetic loci encoding T6SSs. Moreover, some of these systems have been shown to antagonize nearly all gut Bacteroidales species tested (Chatzidaki-Livanis *et al.*, 2016). Similar to T6SSs, a family of diffusible peptide toxins called bacteroidetocins produced by some *Bacteroides* species were identified to have broad spectrum activity and inhibit not only across species or genera, but also across families (Coyne *et al.*, 2019). Furthermore, antimicrobial proteins secreted by Bacteroidales (BSAPs) were revealed to contain membrane attack/perforin (MACPF) domains, and contrary to T6SS systems and

bacteroidetocins, these proteins target a subset of closely related strains (Chatzidaki-Livanis, Coyne and Comstock, 2014a; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Shumaker *et al.*, 2019).

In contrast to some members of the Firmicutes phylum, where antimicrobial compounds have been studied for decades, the production of antimicrobial toxins produced by members of Bacteroidetes remain poorly described (Mattick, Hirsch and Berridge, 1947; Gardner, 1950). The first antimicrobials from the members of this phylum were identified in the last six years and mainly concern the strains of several human gut-associated *Bacteroides*: *B. fragilis*, *B. ovatus*, *B. vulagtus*, *B. thetaiotaomicron*, *B. ovatus*, *B. dorei*, *B. cellulosilyticus*, and *B. stercoris* (Chatzidaki-Livanis, Coyne and Comstock, 2014a; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Coyne *et al.*, 2019)

In Chapter II, the set of *Bacteroides* strains isolated from *M. m. domesticus* and *M. m. musculus* (house mouse species complex described in general introduction) were sequenced and classified taxonomically. This yielded a unique set of house mouse gut-associated *Bacteroides* strains belonging to *B. acidifaciens*, *B. caecimuris* and *B. sartorii*. Furthermore, ASV 35, which was identified as differentially abundant in *mus* compared to *dom* mouse subspecies in Chapter I, shows 100% nucleotide identity to an unclassified *Bacteroides* species represented by two isolates. The *Bacteroides* isolates obtained in Chapter II have yet to been shown to produce antimicrobial toxins and engage in antagonistic interactions. Furthermore, it is unclear whether such antagonistic interactions could mediate host subspecies-specific differences in *Bacteroides* composition.

In the present study I aimed to identify potential contact-independent antagonistic interactions between *dom* and *mus* gut-associated *Bacteroides* strains, and to investigate whether the antagonism between the isolates from different host subspecies is more frequent compared to the antagonism between isolates from the same host subspecies. To achieve this, 23 *Bacteroides* strains were selected in order to cover the overall phylogenetic diversity of the overall sample of isolates and tested for growth

inhibitory activity in a pairwise manner. The assay identified *Bacteroides* isolates to engage in antagonistic interactions. Moreover, inhibitory interactions seem to occur mostly between isolates belonging to different *Bacteroides* species (inter-species antagonism), rather than between strains isolated from different host subspecies.

Results

I. Antagonistic bacteria-bacteria interactions among *Bacteroides* isolates

To investigate the ability of *Bacteroides* strains isolated from *M. musculus* cecum content to inhibit the growth of each other in a contact-independent manner, 23 isolates (11 from *dom* and 12 from *mus* mice subspecies) and two type strains available in the lab, *B. acidifaciens* A 40 and *B. uniformis* DSM 6597 (Table 5, Methods section), were tested using a soft agar overlay assay. The selection of the strains for the experiment aimed to cover the overall phylogenetic diversity of the set of isolates from Chapter II.

The data obtained demonstrate great diversity in both the ability to inhibit- and to be inhibited by secreted antimicrobials of heterologous strains. Nine out of 23 isolates secrete molecules that inhibit the growth of the other *Bacteroides* strains, corresponding to 29 (5.7%) out of 506 possible pairwise interactions (Suppl. table 1). These nine antagonists originate from different mouse samples. Two of them, KH 353_16 and KH 365_1, were able to antagonize co-occurrent strains (the strains from the same mouse sample) KH 353_3 and KH 365_2, respectively (Figure 1). Notably, isolates of both pairwise antagonistic interactions mentioned above belong to different *Bacteroides* species. Moreover, each strain was also tested against itself, and no self-inhibition was observed. On the other hand, some of the pairwise antagonistic interactions were observed only once (the assay was performed in triplicate), suggesting the need of further optimization of the assay. The data presented in this chapter is preliminary and requires additional validation.

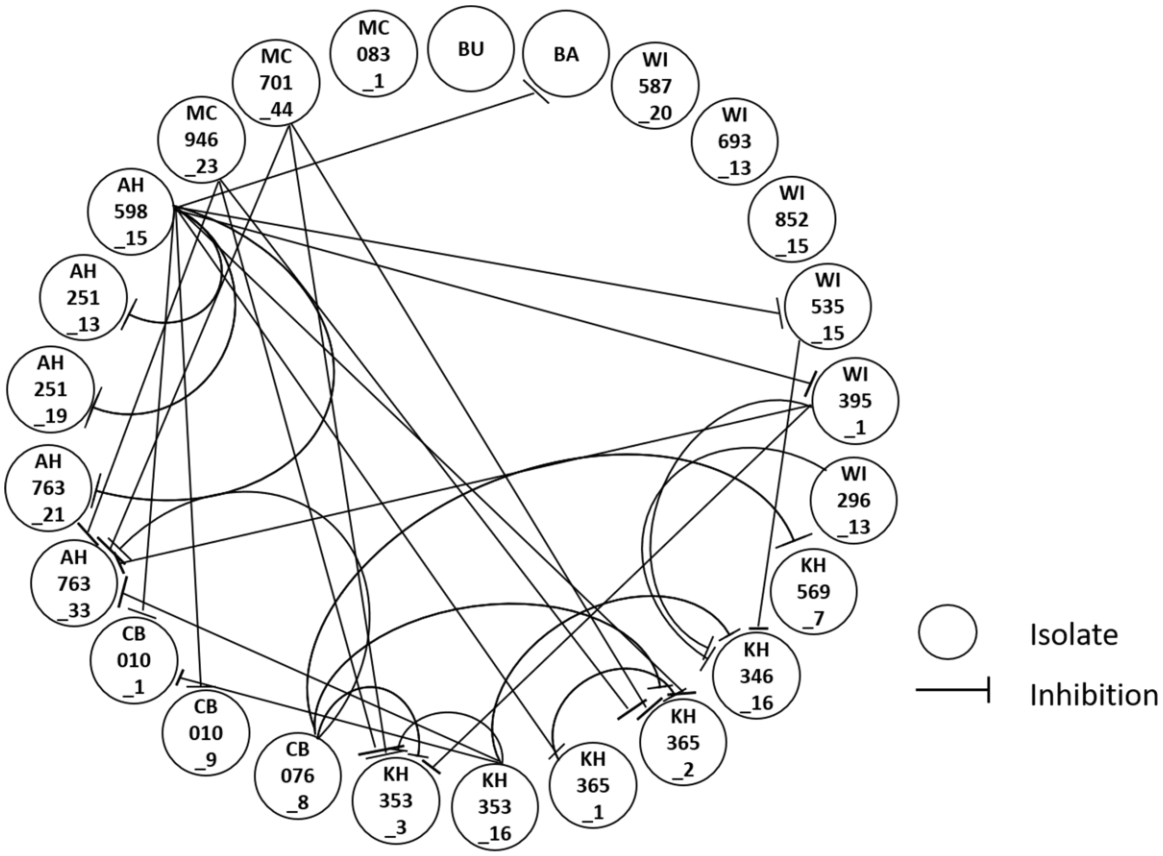


Figure 1. Schematic depiction of antagonistic interactions between *Bacteroides* isolates. Each circle represents an isolate or a type strain. BA: *B. acidifaciens* A40; BU: *B. uniformis* DSM 6597. The lines represent growth inhibition between two strains observed at least once. Soft agar overlay assay was performed in triplicate.

I.1 Phenotypes of *Bacteroides* isolates: antagonistic vs sensitive

The phenotypes of all tested strains were defined as antagonistic (isolates able to inhibit the growth of the others) or sensitive (isolates inhibited by the other strains). Overall, the sensitive phenotype was observed more frequently (13 strains) than the antagonistic (9 strains). Interestingly, the isolate AH 598_15 was able to inhibit 8 other strains, which represents over one third of the total tested strains (Figure 1).

The growth of the isolates KH 365_2, KH 353_3 and AH763_33 was inhibited by 5 other strains, thus these strains seem to be the most sensitive to the secreted antimicrobial molecules. In most cases, there is a clear separation between antagonistic and sensitive strains, meaning that the strain able to inhibit the growth of the others is itself not sensitive to the antagonism by the other isolates. However, three strains (KH 365_1, WI 395_1 and WI 535_15) appear to be antagonistic and sensitive at the same time (Figure 1), suggesting the involvement of different mechanisms of antimicrobial antagonism. For instance, the strain WI 535_15 inhibits the growth of KH 346_16, whereas it is inhibited by the strain AH 598_15. Similarly, isolates WI 395_1 and KH 365_1 inhibit strains KH 353_3 and KH 365_2, respectively, and both are inhibited by the isolate AH 598_15. Furthermore, several *Bacteroides* strains (MC 083_1, WI 693_13, WI 852_15 and WI 587_20), including the type strain *B. uniformis* DSM 6597, did not display any inhibitory activity and were also not inhibited by other heterologous strains (Figure 1).

I.2 Antagonism between *Bacteroides* isolates originated from *dom* and *mus* mice

To investigate whether the antagonism between *Bacteroides* isolates from different mouse subspecies is more frequent compared to the antagonism between isolates from the same mouse subspecies, the number of antagonistic interactions within the same host subspecies and between *mus* and *dom* hosts was assessed. The inhibitory interactions represent 51.7% (15 interactions) within the same mouse subspecies and 48.3% (14 interactions) between isolates from *mus* and *dom* mice (Figure 2, Table 1). A chi-square test of independence showed that there was no significant relationship between the host subspecies and antagonism among the *Bacteroides* isolated from the same or different mouse subspecies (X^2=1.0773, df=1, p-value=0.2993).

Consistent with the total counts of antagonistic and sensitive strains, the isolates display a sensitive phenotype more often than an inhibitory one in both mouse subspecies (3 antagonistic versus 6

sensitive and 5 antagonistic versus 7 sensitive for *dom* and *mus* mice, respectively). Moreover, most of the antagonistic interactions occur between *Bacteroides* isolates from different mouse populations, corresponding to 79.3% (23 out of 29 interactions) (Table 1).

In the two previous chapters, ASV 35 was identified to be differentially abundant in *mus* mice and potentially belongs to the unclassified *Bacteroides* KH 365_2 and KH 569_7 strains. Interestingly, both of the isolates show a sensitive phenotype and were inhibited by the strains isolated from different *dom* hosts, with one exception (KH 365_1 antagonizes KH 365_2) (Figure 2). Another indicator taxon for *mus* mice was *Bacteroides* ASV 268, represented by several strains in the tested set of isolates (AH 598_15, MC 083_1, KH 353_16, MC 701_44 and MC 946_23). In contrast to ASV 35, ASV 268-representative strains all exhibited an antagonistic phenotype, with the exception of MC 083_1. The antagonistic interactions included both the inhibition of strains from the same- and from different host subspecies (Figure 2).

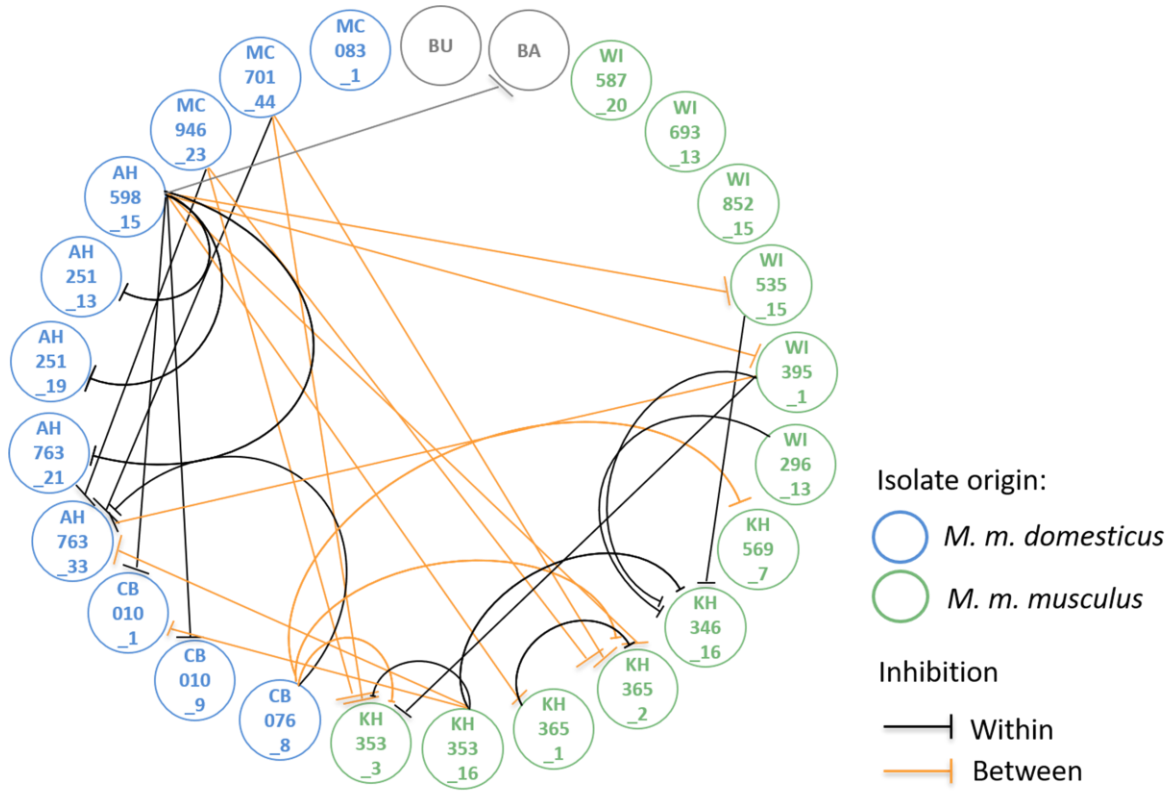


Figure 2. Schematic depiction of antagonistic interactions between *Bacteroides* isolates from *dom* and *mus* mouse subspecies. Each circle represents an isolate or a type strain: blue and green circles correspond to the isolates originated from *dom* and *mus* mice, respectively. BA: *B. acidifaciens* A40; BU: *B. uniformis* DSM 6597. The lines represent growth inhibition between two strains observed at least once. Black lines denote antagonism between isolates originated from the same host subspecies; yellow lines denote antagonism between isolates originated from different host subspecies; grey line donates growth inhibition of *B. acidifaciens* A40 type strain. Soft agar overlay assay was performed in triplicate.

Table 1. Summary of antagonistic bacteria-bacteria interactions between 23 *Bacteroides* isolates. Host subspecies: *dom – M. m. domesticus*; *mus – M. m. musculus*. Mouse populations: AH - Ahvaz, Iran; CB - Cologne/Bonn, Germany; MC - Massif Central, France; KH - Almaty, Kazakhstan; WI - Vienna, Austria.

		Number of interactions					
Host subspecies		*dom*			*mus*		Total
	Within	8			7		15
	Between	11			3		14
Mouse lines		AH	CB	MC	KH	WI	
	Within	3	0	0	3	0	6
	Between	6	4	6	2	5	23
***Bacteroides* species**		*B. acidifaciens**	*B. caecimuris*	*B. sartorii*	*B. uniformis* DSM 6597		
	Within	1	7	0	–		8
	Between	12	0	0	–		21

* includes *B. acidifaciens* A40 type strain

I.3 *Bacteroides* inter- and intra-species antagonism

To investigate inter- and intra-species antagonism among isolates, the numbers of antagonistic interactions between isolates belonging to the same and different *Bacteroides* species was assessed. The results show that inter-species antagonism is more frequent (21 interactions corresponding to 72.4%) compared to intra-species, and involves all three *Bacteroides* species: *B. acidifaciens*, *B. sartorii* and *B. caecimuris* (Figure 3, Table 1).

Bacteroides acidifaciens strains mostly inhibited the growth of the other two species, except for the antagonistic interaction between AH 598_15 and the *B. acidifaciens* A40 type strain (Figure 3). None of the *B. acidifaciens* isolates were antagonized by the other strains. Isolates belonging to *B. caecimuris* displayed both, inter- and intra-species antagonism, while *B. sartorii* isolates exhibited no antagonism against other tested *Bacteroides*.

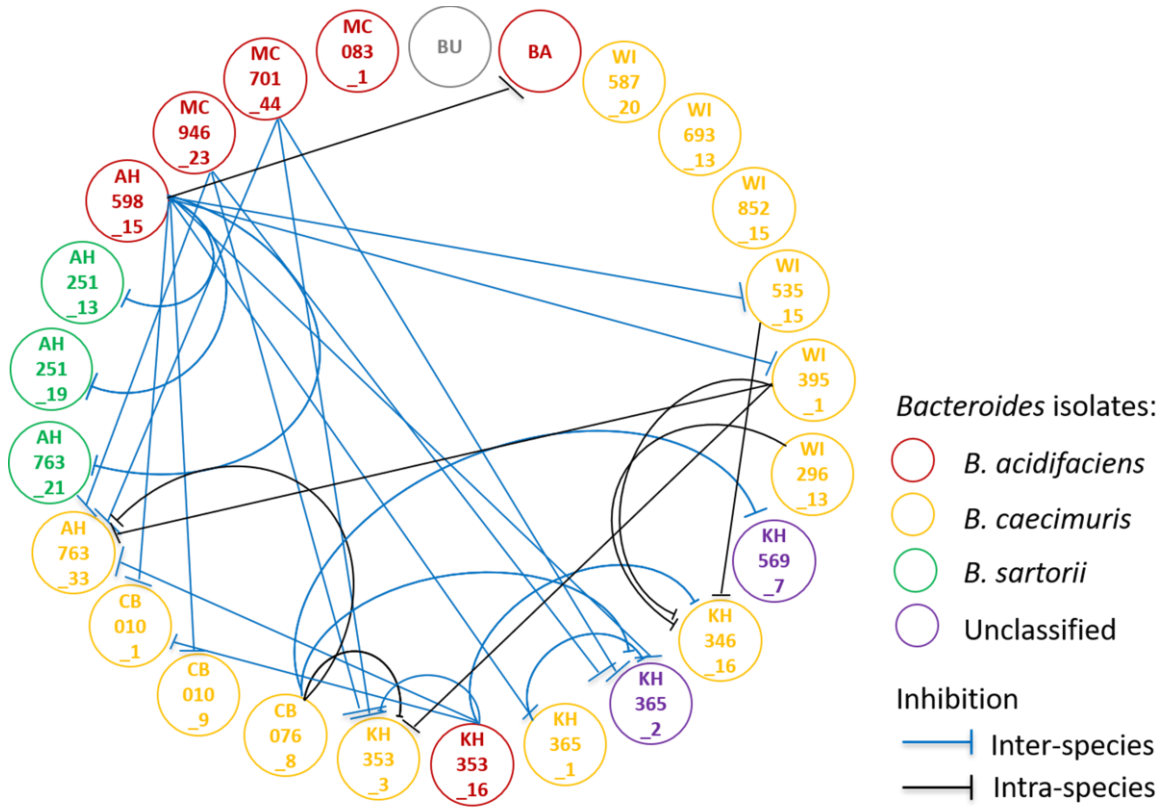


Figure 3. Schematic depiction of intra- and inter-species antagonistic interactions between *Bacteroides* isolates. Each circle represents an isolate or a type strain: red – *B. acidifaciens*; yellow – *B. caecimuris*; green – *B. sartorii* and purple – unclassified isolates. BA: *B. acidifaciens* A40; BU: *B. uniformis* DSM 6597. The lines represent growth inhibition between two strains observed at least once. Blue and black lines denote inter- and intra-species antagonism, respectively. Soft agar overlay assay was performed in triplicate.

II. Antimicrobial activity measurement

To measure the antimicrobial activity of toxin(s) secreted by *Bacteroides* isolates, diameters of the inhibition zones were measured. The size of the zone of inhibition is usually related to the level of antimicrobial activity. A larger inhibition halo usually means that the antimicrobial is more potent. However, this method does not allow to estimate the quantity of the toxin diffused into agar. Also, it is not possible to know if the measured inhibitory activity is caused by one or by a combination of toxins.

The data obtained demonstrate variation in the size of the inhibition zone, ranging from 5.00 mm to 9.83 mm in diameter (Table 2). Moreover, some of the results display great variability in measured diameters, mainly for the activity of AH598_15 on the bacterial lawns of CB010_1, CB010_9 and WI535_15 strains (Table 2, Figure 4). The difference in halo size was observed between the measurements of different experiments. However, the inhibition zone sizes for other bacterial lawns vary less (AH251_13 and A40) (Table 2, Figure 4). The method has some natural variability, and zones of microbial inhibition do not always have clear or regular boundaries.

Table 2. Growth inhibition effect of nine *Bacteroides* strains. The values represent mean ± standard deviation of the inhibition zone diameter of three independent experiments.

		Growth inhibition zone diameter (mm)								
		Spent medium of producer strains								
		AH598_15	CB076_8	KH353_16	MC701_44	MC946_23	WI395_1	KH365_1	WI296_13	WI535_15
Bacterial lawns of sensitive strains	AH251_13	7.44±0.59	–	–	–	–	–	–	–	–
	AH251_19	7.33±0.29	–	–	–	–	–	–	–	–
	AH763_21	9.83±0.29	–	–	–	–	–	–	–	–
	AH763_33	–	6.33±0.58	6.67±0.29	6.67±0.58	7.33±0.58	8.33±0.58	–	–	–
	CB010_1	7.17±1.88	–	5.67±0.58	–	–	–	–	–	–
	CB010_9	6.58±1.06	–	–	–	–	–	–	–	–
	KH346_16	–	–	8.67±0.58	–	–	9.00±0.00	–	7.17±0.29	7.67±0.58
	KH353_3	–	6.67±0.58	7.08±0.35	5.00±0.00	6.33±0.58	5.17±0.29	–	–	–
	KH365_1	5.75±0.35		–	–	–	–	–	–	–
	KH365_2	5.33±0.58	7.50±0.50	–	8.17±0.29	8.00±0.00	–	8.67±0.58	–	–
	KH569_7	–	7.17±0.29	–	–	–	–	–	–	–
	WI395_1	5.83±0.29	–	–	–	–	–	–	–	–
	WI535_15	5.42±1.53	–	–	–	–	–	–	–	–
	A40	7.44±0.38	–	–	–	–	–	–	–	–

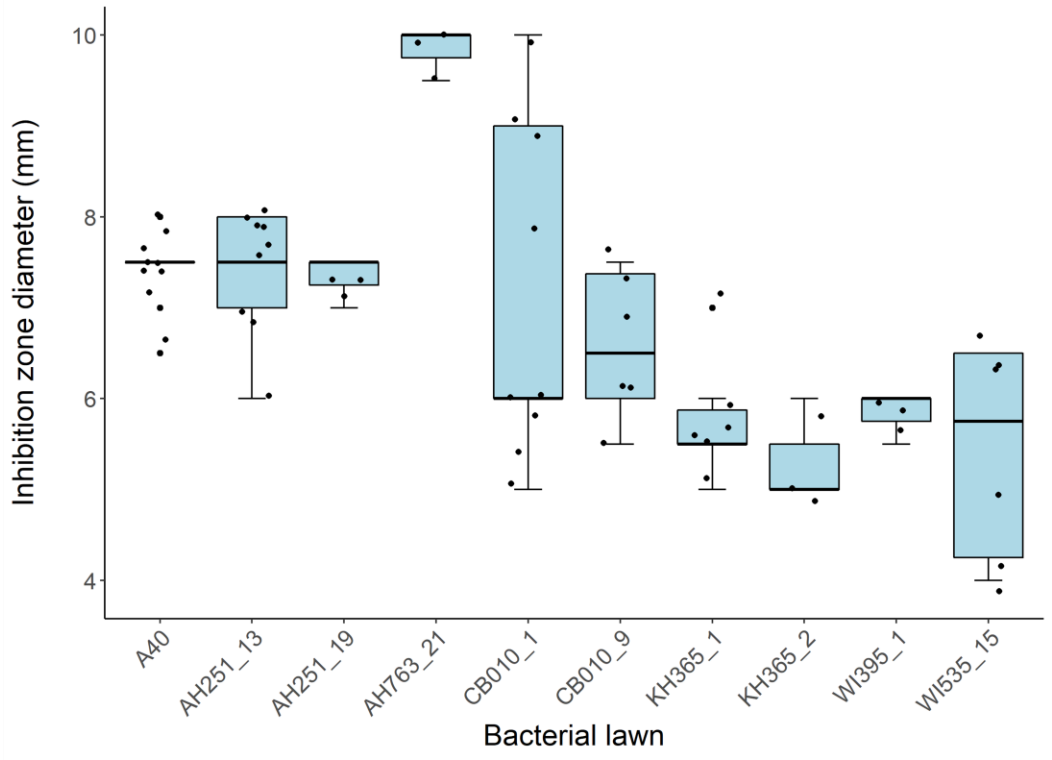


Figure 4. Inhibition zone diameter of *Bacteroides* AH598_15 isolate supernatant on different bacterial lawns. Box-plot is based on soft agar overlay assay and shows mean diameter (mm) of the inhibition zone for each bacterial lawn on sensitive *Bacteroides* isolates (A40: *B. acidifaciens* A40 type strain). The line represents the median, the top of the box is the 75th percentile, the bottom of the box is the 25th percentile and the whiskers represent the maximum and minimum values. Soft agar overlay assay was performed in triplicate.

According to the assumption that the potency of the toxin is related to the size of the inhibition zone, the most potent antimicrobials seem to be produced by *B. acidifaciens* AH 598_15 (which also inhibits the growth of the largest number of strains) and *B. caecimuris* WI395_1. The inhibition halo diameters are 9.83 mm and 9.00 mm for the antagonistic-sensitive pairs AH598_15–AH763_21 and WI395_1–KH346_16, respectively (Table 2, Figure 4, Suppl. figure 1E). Moreover, the inhibition potency of the spent medium from the same isolate seems to be different, and depends on the bacterial lawn of the tested sensitive strain. This was observed for all *Bacteroides* isolates tested in the present study that antagonize more than one strain (Table 2, Figure 4, Suppl. figure 1A-E). Similar observations were made when comparing the inhibition halo diameters induced by different antagonists on the bacterial lawn of

the same sensitive strain. For instance, the inhibition zone sizes vary from 5.38 to 8.67 mm for KH365_2 and from 7.17 to 9.00 mm for KH346_16 (Table 2, Figure 4, Suppl. figure 1).

III. Screen for the described toxins

To investigate whether *Bacteroides*-produced molecules, shown to inhibit the growth of the other sensitive strains, potentially belong to already described toxin classes secreted by *Bacteroides* species, sequences of 5 antimicrobial proteins and one small peptide (Table 3) were BLASTed (pBLAST) against the annotated genomes of 23 *Bacteroides* isolates used in the soft agar overlay assay. Only full-length toxin sequence alignments to the annotated genomes were considered, and the pairwise identity percentages are presented in the Table 4.

Table 3. Summary of *Bacteroides* toxins described previously.

Name	Producing strain	Gene name	Protein length (aa)	Reference
Bacteroidetocin B	*B. thetaiotaomicron* CL15T12C11	EH213_01844	57	Coyne *et al*., 2019
BSAP-1	*B. fragilis* 638R	BF638R_1646	372	Chatzidaki-Livanis, Coyne and Comstock, 2014
BSAP-2	*B. uniformis* CL03T00C23	HMPREF1072_01167	508	Roelofs *et al*., 2016
BSAP-3	*B. vulgatus* CL09T03C04	HMPREF1058_01765	491	McEneany *et al*., 2018
BSAP-4	*B. fragilis* 638R	BF638R_2714	506	Shumaker *et al*., 2019

Table 4. Percent identity of the toxin amino acid sequences produced by different *Bacteroides* species to the annotated genomes of the isolates used in the present study. The identities calculated based on full-length alignments are presented; NA: nucleotide identity value is not available due to very short alignment length. BSAP: *Bacteroides* secreted antimicrobial protein; Bact B: bacteroidetocin B. Antagonist *Bacteroides* isolates are depicted in bold.

Isolate ID	Identity (%)				
	Toxins				
	Bact B	BSAP-1	BSAP-2	BSAP-3	BSAP-4
MC 083_1	100	NA	NA	NA	NA
MC 701_44	100	NA	NA	NA	NA
MC 946_23	100	NA	NA	NA	NA
AH 598_15	100	NA	NA	NA	NA
AH 251_13	NA	NA	NA	NA	NA
AH 251_19	NA	NA	NA	NA	NA
AH 763_21	NA	NA	NA	NA	NA
AH 763_33	NA	NA	NA	NA	NA
CB 010_1	NA	NA	NA	30.88	NA
CB 010_9	NA	NA	NA	30.88	NA
CB 076_8	100	NA	NA	30.88	NA
KH 353_3	NA	NA	NA	NA	NA
KH 353_16	100	NA	NA	NA	NA
KH 365_1	NA	NA	NA	NA	NA
KH 365_2	NA	NA	NA	NA	NA
KH 346_16	NA	NA	27.50	30.83	NA
KH 569_7	NA	NA	NA	NA	NA
WI 296_13	100	NA	NA	30.88	NA
WI 395_1	100	NA	NA	30.88	NA
WI 535_15	100	NA	NA	30.88	NA
WI 693_13	100	NA	NA	30.88	NA
WI 852_15	100	NA	NA	30.88	NA
WI 587_20	100	NA	NA	30.88	NA

A broad spectrum bacteroidetocin B antimicrobial peptide produced by *B. thetaiotaomicron* was recently identified to kill not only bacteria from the same species and genus, but also across families (Coyne *et al.*, 2019). BLAST results showed bacteroidetocin B to be present in half of the tested strains (Table 4). Notably, all of the *Bacteroides* isolates identified to inhibit the growth of the others were also shown to possess bacteroidetocin B, except for KH365_1. Furthermore, the spent medium of the producer isolate KH 353_16 was incubated at high temperature and frozen prior the soft agar overlay assay to test

the tolerance of the putative toxins to heat and cold. Figure 5 represents the results of this experiment, where the spent medium of KH 353_16 was tested on the bacterial lawn of KH353_3 isolate. The results showed no significant effect of heat or cold on the presence of the inhibitory activity of the producer strain, nor on the potency of the secreted antimicrobial molecule (Wilcoxon test, p-value = 0.17 and 0.20 for boiled and frozen compared to fresh supernatant, respectively). Because small peptides are insensitive to heat or cold, these results suggest that the toxin activity might be mediated by a small peptide. Interestingly, bacteroidetocin B was also detected in isolates MC083_1, WI 693_13, WI 852_15 and WI 587_20. However, no antagonistic activity was observed for these strains (Table 4, Figure 1).

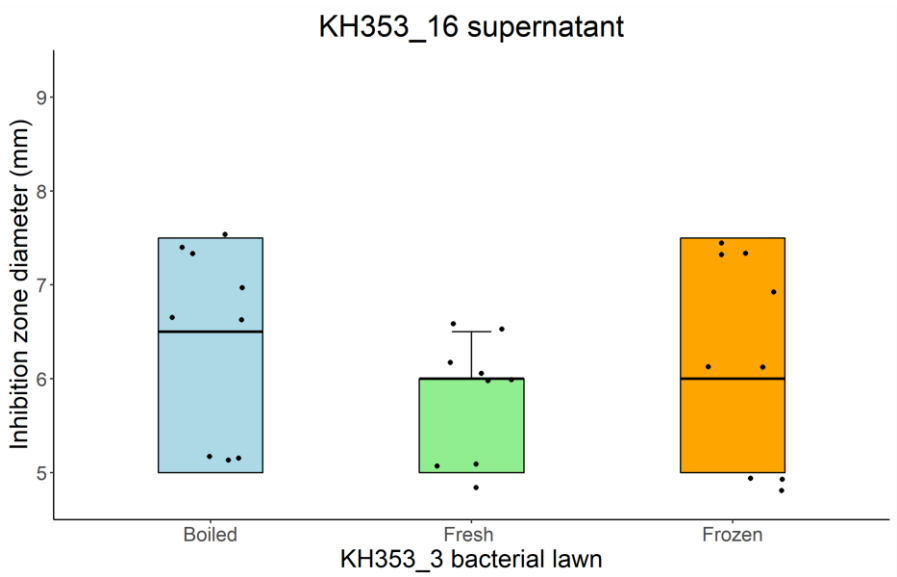


Figure 5. Inhibition zone diameter of *Bacteroides* KH353_16 isolate supernatant on KH353_3 bacterial lawn. Box-plot is based on soft agar overlay assay and shows mean diameter (mm) of the inhibition zone. Boiled: supernatant incubated at 99°C; frozen – supernatant kept at -20°C; fresh – supernatant not subjected to any treatment. The line represents the median, the top of the box is the 75th percentile, the bottom of the box is the 25th percentile and the whiskers represent the maximum and minimum values. The calculation of pairwise comparisons was performed using Pairwise Wilcoxon. Soft agar overlay assay was performed in triplicate.

Unlike bacteroidetocin B, *Bacteroides* secreted antimicrobial proteins (BSAPs) are proteins known to be produced by *B. fragilis*, *B. uniformis* and *B. vulgatus* strains and are able to kill sensitive members of the same species (Chatzidaki-Livanis, Coyne and Comstock, 2014; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Shumaker *et al.*, 2019). A screen for BSAPs resulted in lower identities to the annotated genomes of tested strains. BSAP-3 was detected in most of the isolates, compared to the others screened BSAPs. It is 30.9% identical to ten *Bacteroides* strains, four of which were shown to be antagonists (Table 4). BSAP-2 was only detected in KH 346_16 with an identity of 27.5%, while BSAP-1 and BSAP-4 molecules were not detected in any of the tested isolates (Table 4).

In summary, some of *Bacteroides* strains isolated from *dom* and *mus* were found to engage in contact-independent antagonistic interactions. Most of the inhibitory interactions seem to occur between isolates belonging to different *Bacteroides* species than between strains isolated from different host subspecies. The potency of the toxin(s) produced by the isolates varies and possibly depends on the sensitive strain. Based on the inhibition zone size, the most potent antimicrobial activity is produced by *B. caecimuris* WI395_1 and *B. acidifaciens* AH 598_15 strain, which also inhibits the growth of the largest number of isolates in the present dataset. Moreover, the previously described bacteroidetocin B was detected in most of the identified antagonistic *Bacteroides* strains of this study.

Discussion

The present chapter first aimed to identify contact-independent antagonistic interactions among *Bacteroides* strains isolated from the house mouse species complex. Second, inhibitory interactions were analyzed in order to gain insight into whether bacteria-bacteria antagonism is more frequent between strains isolated from the same- or from different mouse subspecies. This study provides a screen for antagonism among different gut-associated *Bacteroides* strains and the quantification of the respective antimicrobial activity of the producer strains. This is the first screen for antagonistic interactions among mouse-associated *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains isolated from two closely related host subspecies.

Antagonistic bacteria-bacteria interactions between the present set of *Bacteroides* strains isolated from the mouse gut were largely expected, since this was shown before for the isolates from other *Bacteroides* species. Several studies revealed human *B. uniformis*, *B. vulgatus*, *B. fragilis* and *B. thetaiotaomicron* isolates to produce and secret BSAPs, which target sensitive strains of the same and different species (Chatzidaki-Livanis, Coyne and Comstock, 2014a; Roelofs *et al.*, 2016; McEneany *et al.*, 2018; Coyne *et al.*, 2019; Shumaker *et al.*, 2019). Interestingly, the present study identified three isolates that showed both a sensitive- and an antagonistic phenotype. This suggests that different mechanisms and toxic molecules might mediate antagonism, since the producer strains are usually immune to self-produced toxins, and the presence of the immunity determinant seems to be widespread among gut Bacteroidales (Cotter, Hill and Ross, 2005; Ross *et al.*, 2019).

The host species-specific differences in *Bacteroides* taxa identified in Chapter I were proposed to be reliable, since they were consistently detected across geographic locations. This leads to the following question: are *Bacteroides* indicators for *mus* mouse subspecies, strongly co-adapted to their hosts,

different enough from the strains co-adapted to *dom* mouse subspecies such that they show more antagonistic interactions against each other compared to the antagonism within the same host subspecies? The present study did not detect a difference in the number of antagonistic interactions within the same or between different host subspecies. Notably, most of the observed antagonistic events occurred between isolates belonging to different *Bacteroides* species. The inhibition of the isolates KH 365_2 and KH 569_7 (representing indicator ASV 35 for *mus* mice) by strains from different *dom* hosts seems to be explained by their belonging to different *Bacteroides* species rather by their host origin. Unfortunately, with respect to *B. acidifaciens* (4 *dom* and 1 *mus* isolates) and *B. caecimuris* (4 *dom* and 9 *mus* isolates), the selection of the isolates did not allow the systematic comparison of the number of antagonistic interactions within and between the same mouse subspecies. To validate these results of antagonism with respect to host subspecies, the sample selection should be more equal and more isolates need to be screened for bacteria-bacteria inhibition.

The soft agar overlay assay (see Methods section) used in this study is a fast, inexpensive and simple technique to screen for bacteria-bacteria antagonism mediated by secreted antimicrobial molecules. It allows a large number of isolates to be screened simultaneously. However, this method is limited to identify contact-independent antagonism, excluding potential positive interactions between tested strains. Zones of inhibition observed in this experiment indicate the growth inhibition of the sensitive strain provoked by the spent medium of the antagonist strain. Thus, it is not possible to conclude if the antagonism observed in the present study between *Bacteroides* isolates resulted in actual killing of the sensitive strain. The measurement of the inhibition zone diameters enables the quantification and estimation of the potency of the secreted antimicrobial compound. On the other hand, due to the natural variability of the method, zones of microbial inhibition do not always have clear or regular boundaries, leading to imprecise quantification. This might also explain the variability in the inhibition zone diameters, measured to quantify the potency of the toxins in the present study.

Following the identification and quantification of the inhibitory interactions, a screen for previously characterized *Bacteroides*-produced toxins was performed. The results indicate that BSAPs might not be the proteins responsible for the antagonism between *Bacteroides* isolates, due to the low amino acid identity between BSAPs and annotated isolates genomes, and due to the fact that antimicrobial activity of spent media remains after heat treatment and freezing. Considering that all described BSAPs are large proteins, their activity would be abolished as a result of denaturation by heat or interruption of noncovalent interactions by frozen water (Table 3). Moreover, the screen for described toxins revealed bacteroidetocin B to be present in essentially all antagonist *Bacteroides* isolates. In contrast to BSAPs, this toxin is a small peptide, and thus more likely to be resistant to high and low temperatures and keep its activity after boiling or freezing the spent medium. However, Coyne *et al*. (2019) observed self-intoxication of the producer strains by bacteroidetocin B, which is in disagreement with the observations of the present study. Here, no self-inhibition was observed for all 23 tested *Bacteroides* strains. These results together suggest that the toxin responsible for the antagonistic interaction between mouse gut-associated *Bacteroides* might be a novel molecule yet to be described.

In conclusion, this is the first study investigating antagonistic interactions among mouse-associated *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains isolated from two closely related host subspecies. The soft agar overlay assay identified antagonism between different *Bacteroides* strains, and the toxin responsible for the inhibitory interactions seems to be thus far uncharacterized. Nonetheless, additional studies on a larger set of the isolates are needed to validate these findings and also screen for positive interactions. The investigation of the antagonism between gut-associated *Bacteroides* and pathogens would shed light on contributions to colonization resistance. Finally, the screen for toxin biosynthetic gene clusters among the sequenced *Bacteroides* genomes and the characterization of the toxic compound and its mechanism of action are of particular interest.

Methods

I. Antimicrobial activity screening among *Bacteroides* isolates

I.1 Strain selection, culture media and growth conditions

To cover the overall phylogenetic diversity, 23 *Bacteroides* isolates representing one from each clade of the phylogenetic tree (Chapter II, Figure 8) were selected for the antimicrobial activity screening assay together with two type strains available in the lab (Table 5). All strains were grown anaerobically at 37°C for 48-72 hours (Chapter II, Methods section) as single cultures, in Chopped Meat medium (CM) from DSMZ with the following modification: the lean beef was replaced by 20 g/L meat extract (BD Difco).

Table 5. *Bacteroides* isolates selected for antimicrobials screening assay.

Isolate ID	Taxonomic classification	Mouse subspecies	Origin
MC 083_1	*B. acidifaciens*	*domesticus*	Massif Central, France
MC 701_44	*B. acidifaciens*	*domesticus*	Massif Central, France
MC 946_23	*B. acidifaciens*	*domesticus*	Massif Central, France
AH 598_15	*B. acidifaciens*	*domesticus*	Ahvaz, Iran
AH 251_13	*B. sartorii*	*domesticus*	Ahvaz, Iran
AH 251_19	*B. sartorii*	*domesticus*	Ahvaz, Iran
AH 763_21	*B. sartorii*	*domesticus*	Ahvaz, Iran
AH 763_33	*B. caecimuris*	*domesticus*	Ahvaz, Iran
CB 010_1	*B. caecimuris*	*domesticus*	Cologne/Bonn, Germany
CB 010_9	*B. caecimuris*	*domesticus*	Cologne/Bonn, Germany
CB 076_8	*B. caecimuris*	*domesticus*	Cologne/Bonn, Germany
KH 353_3	*B. caecimuris*	*musculus*	Almaty, Kazakhstan
KH 353_16	*B. acidifaciens*	*musculus*	Almaty, Kazakhstan
KH 365_1	*B. caecimuris*	*musculus*	Almaty, Kazakhstan
KH 365_2	Unclassified	*musculus*	Almaty, Kazakhstan
KH 346_16	*B. caecimuris*	*musculus*	Almaty, Kazakhstan
KH 569_7	Unclassified	*musculus*	Almaty, Kazakhstan
WI 296_13	*B. caecimuris*	*musculus*	Vienna, Austria
WI 395_1	*B. caecimuris*	*musculus*	Vienna, Austria
WI 535_15	*B. caecimuris*	*musculus*	Vienna, Austria
WI 693_13	*B. caecimuris*	*musculus*	Vienna, Austria
WI 852_15	*B. caecimuris*	*musculus*	Vienna, Austria
WI 587_20	*B. caecimuris*	*musculus*	Vienna, Austria

B. acidifaciens A40	Type strain
B. uniformis DSM 6597	Type strain

I.2 Soft agar overlay assay

The ability of one *Bacteroides* isolate to inhibit the growth of another in a contact-independent manner by the secretion of inhibitory molecules was assayed by the soft agar overlay technique (Hockett and Baltrus, 2017). The OD_{600} of all bacterial cultures grown in 5 ml of CM medium was measured (the control tube containing 5 ml of pure medium was used as a blank), the spent medium was centrifuged at 6500 x g for 10 min to separate the biomass, and then filter sterilized (0.2 µm syringe filter, VWR™). In order to normalize for the amount of the antimicrobial molecule secreted to the medium, an OD_{600} = 1.0 (one of the most frequently measured) was chosen as a reference. Hence, 600 µl of the cultures with an OD_{600} = 1.0 were spun down to collect supernatant. For the OD_{600} above or below 1.0, the culture volume to collect was calculated accordingly (for instance, 1200 µl of spent media were collected for cultures with an OD_{600} = 0.5 and 300 µl for cultures with an OD_{600} = 2).

The antimicrobial activity of cell-free spent media was assessed on lawns of each tested strain in a pairwise manner (all-against-all, including self as a control). The soft agar 0.5% (w/v) was melted and maintained at 40°C before the start of the experiment. *Bacteroides* cultures were carefully mixed with 15 ml of soft agar in order to get a final OD_{600} of 0.04, and poured onto square Greiner CM agar plates 12×12cm (Merck). The plates were allowed to solidify. Next, drops of cell-free supernatant were spotted onto the lawn of bacteria-soft agar and allowed to dry completely. A 5 µl drop of the pure medium used as a control was spotted on each plate. The plates were incubated at 37°C for 48 hours. After the incubation period, microbial growth inhibition halos were observed. The experiments were performed in triplicate.

I.3 Bacterial growth inhibition measurements

Bacterial growth inhibition was determined as an average diameter (in mm) of the inhibition zones around the spent media drops, measured in 3 different points of each inhibition halo. The standard deviation was also calculated.

I.4 Spent media treatment for the heat- and cold-susceptibility experiment

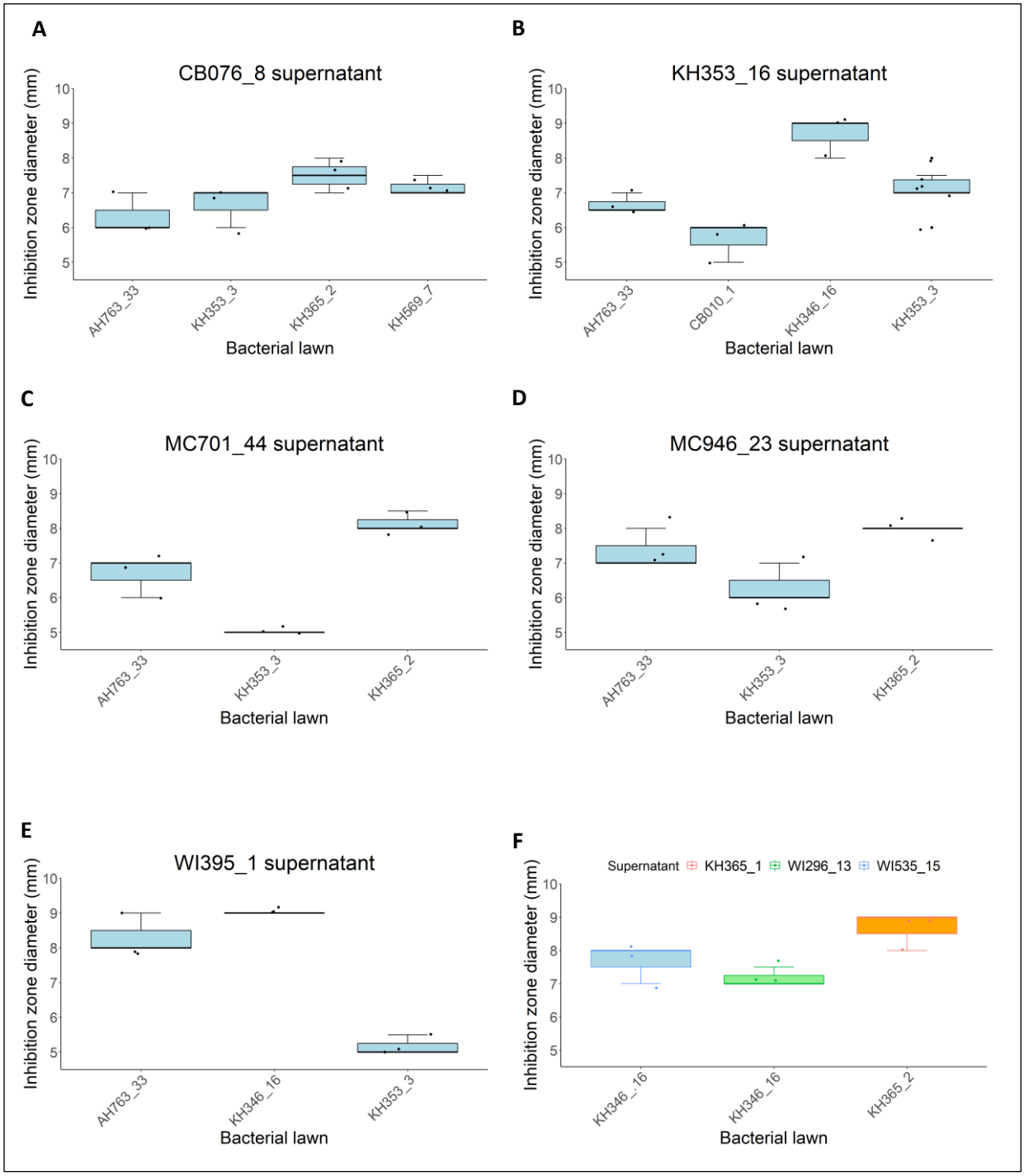
The spent media of the producer strains was incubated at 99°C for 15 min, and at -20°C prior the experiment. Subsequently, the soft agar overlay assay using pre-treated and respective freshly collected media was performed as described above. The presence of the inhibition activity was recorded and the diameters of the inhibition zones were measured.

II. Screen for the described toxins

Amino acid sequences of six previously described toxins produced by *Bacteroides* strains (Results, Table 3) were obtained from the NCBI database. Next, toxin sequences were BLASTed against annotated genomes of 23 *Bacteroides* isolates (Table 5) using the `blastp` command and an e-value threshold of 1e-5. Only nucleotide identities of toxin-to-isolate full-length alignments were considered for further analysis.

Supplementary material

I. Supplementary figures



Supplementary figure 1. Inhibition zone diameter of antagonist *Bacteroides* isolates' supernatant on different bacterial lawns. Box-plot is based on soft agar overlay assay and shows the mean diameter (mm) of the inhibition zone for each bacterial lawn on sensitive *Bacteroides* isolates. **A-B**: supernatants of CB076_8 and KH353_16 isolates inhibit the growth of four different strains each; **C-E**: supernatants of MC701_44, MC946_23 and WI395_1 isolate inhibit the growth of three different sensitive strains each; **F**: supernatants of KH365_1, WI296_13 and WI535_15 isolates inhibit the growth of one sensitive strain each. The line represents the median, the top of the box is the 75th percentile, the bottom of the box is the 25th percentile and the whiskers represent the maximum and minimum values. Soft agar overlay assay was performed in triplicate.

II. Supplementary tables

Supplementary table 1. Analysis of growth inhibition among 23 *Bacteroides* isolates. Growth inhibition (sensitive) is indicated by "+", no growth inhibition is indicated by "-". BA: *B. acidifaciens* A40; BU: *B. uniformis* DSM 6597.

		***Bacteroides* producer strains**																								
		MC 083_1	MC 701_44	MC 946_23	AH 598_15	AH 251_13	AH 251_19	AH 763_21	AH 763_33	CB 010_1	CB 010_9	CB 076_8	KH 353_3	KH 353_16	KH 365_1	KH 365_2	KH 346_16	KH 569_7	WI 296_13	WI 395_1	WI 535_15	WI 693_13	WI 852_15	WI 587_20	BA	BU
***Bacteroides* strains grown in soft agar**	MC 083_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	MC 701_44	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	MC 946_23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	AH 598_15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	AH 251_13	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	AH 251_19	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	AH 763_21	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	AH 763_33	–	+	+	–	–	–	–	–	–	–	+	–	+	–	–	–	–	–	+	–	–	–	–	–	–
	CB 010_1	–	–	–	+	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–
	CB 010_9	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	CB 076_8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

KH 353_3	–	+	+	–	–	–	–	–	–	–	+	–	+	–	–	–	–	–	+	–	–	–	–	–	–
KH 353_16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
KH 365_1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
KH 365_2	–	+	+	+	–	–	–	–	–	–	+	–	–	+	–	–	–	–	–	–	–	–	–	–	–
KH 346_16	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	+	+	+	–	–	–	–	–
KH 569_7	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 296_13	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 395_1	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 535_15	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 693_13	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 852_15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
WI 587_20	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
BA	–	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
BU	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

General conclusion

The present work represents the first study systematically investigating gut-associated *Bacteroides* among two house mouse subspecies *M. m. domesticus* and *M. m. musculus*. Using the house mouse species complex as a model to study the signatures of differentiation in *Bacteroides* taxa according to host subspecies, I analyzed bacterial 16S rRNA gene profiles of gut bacterial communities from both mouse subspecies and identified candidate *Bacteroides* ASVs for characterization at the strain level. By performing whole genome sequencing of the isolated strains, I was able to describe the *Bacteroides* pan genome and compare protein content and function according to the host subspecies. Finally, I obtained novel and intriguing data regarding antagonistic interactions between different mouse gut-associated *Bacteroides* isolates.

In Chapter I, I aimed to identify potential signatures of differentiation in microbial taxon abundance according to host subspecies, focusing on *Bacteroides* as a candidate genus to identify potential coevolutionary processes. To achieve this, I used multiple wild-derived outbred mouse colonies originating from five locations across the geographic range of the subspecies of *M. m. domesticus* (Germany, France, and Iran) and *M. m. musculus* (Austria and Kazakhstan). The analysis of 16S rRNA gene profiles revealed *Bacteroides* genus-level diversity to appear similar in both mouse subspecies. Despite all the mice being maintained under the same housing conditions, the geographic location of the founding mice influences the variability of the respective gut microbiota communities. Host subspecies, in turn, seems to play a comparatively minor role in gut community structure. Furthermore, 16S rRNA gene survey and indicator species analysis applied to the *Bacteroides* genus yielded several strong *Bacteroides* indicators for *M. m. musculus*, suggesting that these host-*Bacteroides* associations are consistent across different geographic locations and represent promising candidates for further characterization. The

abundance of one of the strong indicators identified for *M. m. musculus, Bacteroides* ASV 35, might have changed since the common ancestor of studied mouse subspecies.

In Chapter II I combined culturing and genomics methods to further characterize candidate *Bacteroides* at the strain-level, and identify the differences in bacterial genomes that might contribute to the adaptation to different mouse subspecies. These approaches yielded fully sequenced genomes of 146 *Bacteroides* isolates, classified as *B. acidifaciens*, *B. caecimuris* and *B. sartorii*, along with two potentially new *Bacteroides* species. Furthermore, a perfect match between a candidate indicator *Bacteroides* ASV, which strongly associates to *M. m. musculus*, and both unclassified isolates was detected, suggesting the involvement of this potentially new *Bacteroides* species in the intriguing host-microbe association. Additionally, the whole genome sequencing of the isolated strains sheds light on the *Bacteroides* pan genome in terms of protein content and functions, which in some cases could represent specialization to host subspecies.

My findings in Chapter III provide novel insights into the antagonistic interactions between mouse gut-associated *B. acidifaciens*, *B. caecimuris* and *B. sartorii* strains isolated from two closely related host subspecies. I aimed to investigate whether the antagonism between the isolates from different host subspecies is more frequent compared to the antagonism between isolates from the same host subspecies. To achieve this, I used the soft agar assay applied to a subset of 23 *Bacteroides* isolates, which enabled me to test for inhibitory activity between strains in a pairwise manner. The identified antagonistic interactions seem to occur mostly between isolates belonging to different *Bacteroides* species and to different mouse populations, rather than between strains isolated from different host subspecies. Moreover, a screen for previously characterized *Bacteroides*-produced toxins suggests that the toxin or toxins responsible for the inhibitory interactions observed in this study appear to be thus far uncharacterized.

For future research I would study adaptively evolving genes in both host and *Bacteroides* genomes to further identify and confirm signatures of coadaptation between the host subspecies and bacterial strains in the context of the metaorganism. Moreover, to validate the findings of Chapter 3, I would perform additional studies on a larger set of the isolates and also screen for positive interactions among *Bacteroides* isolates. The investigation of the antagonism between gut-associated *Bacteroides* and pathogens would shed light on contributions to colonization resistance. Finally, I would purify and characterize toxins involved in the antagonistic interactions by performing experiments and by screening for toxin biosynthetic gene clusters among the sequenced *Bacteroides* genomes.

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr. John F. Baines, for constant support, guidance and understanding throughout all my PhD stages. Thanks to Prof. Baines, I had a chance to dive into the fascinating field of gut microbiome and evolution, which was completely new to me. Also, I would like to thank my thesis committee members, Prof. Ruth Schmitz-Streit and Prof. Tal Dagan, who gave a valuable constructive criticism on my thesis project, and were always available for the discussion of ideas and project designs. A big thanks to Prof. Daniel Unterweger for the support and lively discussions concerning *Bacteroides*.

A special thanks goes to my postdoc colleagues Dr. Marie Vallier and Dr. Meriem Belheouane for introducing me into the world of bioinformatics and for being always there to discuss experimental designs and ideas. Moreover, I would like to thank my colleagues Cecilia Chung and Shauni Doms for sharing with me long days of mouse dissections. I thank Dr. Sven Künzel for all his help and efficiency, as well as Katja Cloppenborg-Schmidt, Jan Schubert, Silke Carstensen and Olga Eitel for the excellent technical support. I also would like to thank the rest of Baines' Group for the support, kindness and friendship.

I acknowledge the Collaborative Research Centre "Origin and Function of Metaorganisms" (CRC1182) for giving me an opportunity to be a part of it, for organizing exciting retreats, seminars and supporting early career scientists.

I would like to thank the mouse facility manager of the Max Planck Institute for Evolutionary Biology Christine Pfeifle, as well as all the mouse caretakers for being always efficient to help and advice in mouse work. I thank all the IMPRS and CRC 1182 students with whom I had not only lively scientific

discussions, but also a lot of fun together. Special thanks to our small Portuguese community, Ana Teles, Filipa Moutinho and Carolina Peralta, you made me feel like home. Obrigada!

I specially thank my parents for giving me a good start in life, for being always on my side, supporting and advising. Thank you for making me the person I am today. I thank my lovely sister for always patiently listening to all my biology conversations. And finally, I thank my boyfriend for helping me discover Germany and German language, for being my love, my best friend and my greatest supporter in whatever I do.

Curriculum Vitae

Name: Hanna Fokt

Date of Birth: 29.08.1985

Nationality: Portuguese

Current residence: Plön, Germany

Education

October 2005 – July 2008: Bachelor (B.Sc.) in Applied Biology, University of Minho, Biology Department, Braga, Portugal. Number of semesters: 6.

November 2008 - May 2011: Master's degree (M.Sc.) in Biotechnology and Bio-Entrepreneurship in Medicinal and Aromatic Plants, University of Minho, Biology Department, Braga, Portugal. Number of semesters: 4. Master's Thesis "Modulatory activity of *Ginkgo biloba* extract on *Saccharomyces cerevisiae* cell cycle and DNA damage repair ability in cells under replicative stress".

May 2016 – current: PhD student at the Max-Planck-Institute for Evolutionary Biology, Plön and at the Collaborative Research Centre "Origin and Function of Metaorganisms" (CRC1182), Kiel.

Academic work experience

May 2011 - October 2013: Research fellow at Department of Biological Engineering, University of Minho, Braga, Portugal. The project Micro2Micro - microbial products for biofilm control, focused on the study of mixed species biofilms to gain a deeper knowledge on the type of interactions

established between the bacteria with the purpose to foreseen novel molecules with antimicrobial properties and biofilm control properties.

Declaration

Hereby I declare that,

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

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

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

Authors' contributions

Chapter I: John Baines designed the study. Hanna Fokt collected mouse cecum content samples, performed nucleic acid extractions, library preparation for 16S rRNA gene sequencing (with technical support from Katja Clöppenborg-Schmidt) and cloning. Sven Künzel performed Miseq Illumina sequencing, Hanna Fokt performed all analysis and wrote the chapter, with editing from John Baines.

Chapter II: Hanna Fokt performed bacterial strains isolation. The optimization of the isolation technique was elaborated with valuable input from Daniela Prasse and Ruth Schmitz-Streit. Sven Künzel performed

Illumina NextSeq sequencing, Hanna Fokt performed genomic data analysis and taxonomic classification of the isolates (with the contribution of Malte Rühlemann), Maxime Godfroid performed comparative protein analysis with the contribution of Rahul Unni. Hanna Fokt wrote the chapter, with editing from John Baines.

Chapter III: Hanna Fokt, Daniel Unterweger and John Baines designed the study. Hanna Fokt performed all laboratory work, data analysis and wrote the chapter, with editing from Daniel Unterweger and John Baines.

Plön, April 2021

Hanna Fokt	Prof. Dr. John F. Baines

Bibliography

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology* 215(3): 403–410.

Bäckhed, F. *et al.* (2005) 'Host-bacterial mutualism in the human intestine', *Science* 307(5717): 1915–1920.

Bankevich, A. *et al.* (2012) 'SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing', *Journal of Computational Biology* 19(5): 455–477.

Bencivenga-Barry, N. A. *et al.* (2020) 'Genetic manipulation of wild human gut bacteroides', *Journal of Bacteriology* 202(3).

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple', *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1): 289-300.

Benjdia, A. *et al.* (2008) 'Anaerobic sulfatase-maturating enzymes, first dual substrate radical S-adenosylmethionine enzymes', *Journal of Biological Chemistry* 283(26): 17815–17826.

Bernhard, A. E. and Field, K. G. (2000) 'Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16s ribosomal dna genetic markers from fecal anaerobes', *Applied and Environmental Microbiology* 66(4):1587-94.

Bjerke, G. A. *et al.* (2011) 'Mother-to-child transmission of and multiple-strain colonization by *Bacteroides fragilis* in a cohort of mothers and their children', *Applied and Environmental Microbiology* 77(23): 8318–8324.

Bonder, M. J. *et al.* (2016) 'The effect of host genetics on the gut microbiome', *Nature Genetics* 48(11): 1407–1412.

Bubier, J. A. *et al.* (2020) 'A microbe associated with sleep revealed by a novel systems genetic analysis of the microbiome in collaborative cross mice', *Genetics* 214(3): 719-733.

De Cáceres, M. and Legendre, P. (2009) 'Associations between species and groups of sites: Indices and statistical inference', *Ecology* 90(12): 3566–3574.

Callahan, B. J. *et al.* (2016) 'DADA2: High-resolution sample inference from Illumina amplicon data', *Nature Methods* 13(7): 581–583.

Chatzidaki-Livanis, M. *et al.* (2016) '*Bacteroides fragilis* type VI secretion systems use novel effector and immunity proteins to antagonize human gut Bacteroidales species', *Proceedings of the National Academy of Sciences of the United States of America* 113(13): 3627–3632.

Chatzidaki-Livanis, M., Coyne, M. J. and Comstock, L. E. (2014a) 'An antimicrobial protein of the gut symbiont *Bacteroides fragilis* with a MACPF domain of host immune proteins', *Molecular Microbiology* 94(6): 1361–1374.

Chaumeil, P.-A. *et al.* (2019) 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics* 36(6): 1925–1927.

Chen, H. and Boutros, P. C. (2011) 'VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R', *BMC Bioinformatics* 12: 35.

Cheng, Q., Hwa, V. and Salyers, A. A. (1992) 'A locus that contributes to colonization of the intestinal tract by *Bacteroides thetaiotaomicron* contains a single regulatory gene (chuR) that links two polysaccharide utilization pathways', *Journal of Bacteriology* 174(22): 7185–7193.

Cignarella, F. *et al.* (2018) 'Intermittent fasting confers protection in cns autoimmunity by altering the gut microbiota', *Cell Metabolism* 27(6): 1222-1235.

Cole, J. R. *et al.* (2014) 'Ribosomal Database Project: Data and tools for high throughput rRNA analysis', *Nucleic Acids Research* 42(D1): D633-42.

Comstock, L. E. (2009) 'Importance of glycans to the host-*Bacteroides* mutualism in the mammalian intestine', *Cell Host and Microbe* 5(6): 522–526.

Cotter, P. D., Hill, C. and Ross, R. P. (2005) 'Bacteriocins: developing innate immunity for food', *Nature Reviews Microbiology* 3(10): 777–788.

Coyne, M. J. *et al.* (2014) 'Evidence of extensive DNA transfer between Bacteroidales species within the human gut', *mBio* 5(3): e01305-14.

Coyne, M. J. *et al.* (2019) 'A family of anti-Bacteroidales peptide toxins wide-spread in the human gut microbiota', *Nature Communications* 10(1): 3460.

Cucchi, T., Vigne, J.-D. and Auffray, J.-C. (2005) '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(3): 429–445.

David, L. A. *et al.* (2014) 'Diet rapidly and reproducibly alters the human gut microbiome', *Nature* 505(7484): 559–563.

Domingues Kümmel Tria, F., Landan, G. and Dagan, T. (2017) 'Phylogenetic rooting using minimal ancestor deviation', *Nature Ecology & Evolution* 1: 0193.

Donaldson, G. P. *et al.* (2020) 'Spatially distinct physiology of *Bacteroides fragilis* within the proximal colon of gnotobiotic mice', *Nature Microbiology* 5(5): 746–756.

Donaldson, G. P., Lee, S. M. and Mazmanian, S. K. (2015) 'Gut biogeography of the bacterial microbiota', *Nature Reviews Microbiology* 14: 20–32.

Dufrêne, M. and Legendre, P. (1997) 'Species assemblages and indicator species: the need for a flexible asymmetrical approach', *Ecological Monographs* 67(3): 345–366.

Enright, A. J., Dongen, S. Van and Ouzounis, C. A. (2002) 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Research* 30(7):1575-84.

Fenner, L. *et al.* (2005) '*Bacteroides massiliensis* sp. nov., isolated from blood culture of a newborn'. *International Journal of Systematic and Evolutionary Microbiology* 55(3): 1335-1337.

García-Bayona, L. and Comstock, L. E. (2018) 'Bacterial antagonism in host-associated microbial communities' *Science* 361: eaat2456.

Gardner, J. F. (1950) 'Some antibiotics formed by bacterium coli', *British journal of experimental pathology* 31(1): 102–111.

Gensollen, T. *et al.* (2016) 'How colonization by microbiota in early life shapes the immune system', *Science* 352(6285): 539–544.

Guénet, J.-L. and Bonhomme, F. (2003) 'Wild mice: an ever-increasing contribution to a popular mammalian model', *Trends in Genetics* 19(1): 24–31.

Henz, S. R. *et al.* (2005) 'Whole-genome prokaryotic phylogeny', *Bioinformatics* 21(10): 2329–2335.

Hiergeist, A. *et al.* (2016) 'Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability', *International Journal of Medical Microbiology* 306(5): 334-342.

Hockett, K. L. and Baltrus, D. A. (2017) 'Use of the soft-agar overlay technique to screen for bacterially produced inhibitory compounds', *Journal of Visualized Experiments* 119: e55064.

Huson, D. H. and Bryant, D. (2006) 'Application of phylogenetic networks in evolutionary studies', *Molecular Biology and Evolution* 23(2): 254–267.

Katoh, K. *et al.* (2002) 'MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research* 30(14): 3059–3066.

Kearse, M. *et al.* (2012) 'Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics* 28(12): 1647–1649.

Klötzl, F. and Haubold, B. (2016) 'Support values for genome phylogenies', *Life* 6(1): 11.

Kurilshikov, A. *et al.* (2017) 'Host Genetics and gut microbiome: challenges and perspectives', *Trends in Immunology*. 38(9): 633–647.

Lazdunski, C. J. *et al.* (1998) 'Colicin import into *Escherichia coli* cells', *Journal of Bacteriology* 180(19): 4993-5002.

Ley, R. E. *et al.* (2008) 'Evolution of mammals and their gut microbes', *Science* 320(5883): 1647-51.

Linnenbrink, M. *et al.* (2013) 'The role of biogeography in shaping diversity of the intestinal microbiota in house mice', *Molecular Ecology* 22(7): 1904–1916.

Livingston, S. J., Kominos, S. D. and Yee, R. B. (1978) 'New medium for selection and presumptive identification of the *Bacteroides fragilis* group', *Journal of Clinical Microbiology* 7(5): 448–453.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal* 17(1): 10.

Mattick, A. T. R., Hirsch, A. and Berridge, N. J. (1947) 'Further observations on an inhibitory substance (nisin) from lactic streptococci', *The Lancet* 250(6462): 5–8.

Mazmanian, S. K. *et al.* (2005) 'An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system', *Cell* 122(1): 107–118.

McEneany, V. L. *et al.* (2018) 'Acquisition of MACPF domain-encoding genes is the main contributor to LPS glycan diversity in gut *Bacteroides* species', *The ISME Journal* 12(12): 2919–2928.

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

Meier-Kolthoff, J. P. and Göker, M. (2019) 'TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy', *Nature Communications* 10(1): 1–10.

Minh, B. Q. *et al.* (2020) 'IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution* 37(5): 1530–1534.

Moeller, A. H. *et al.* (2014) 'Rapid changes in the gut microbiome during human evolution', *Proceedings of the National Academy of Sciences of the United States of America* 111(46): 16431–16435.

Moeller, A. H. *et al.* (2017) 'Dispersal limitation promotes the diversification of the mammalian gut microbiota', *Proceedings of the National Academy of Sciences of the United States of America* 114(52): 13768–13773.

Moeller, A. H. *et al.* (2018) 'Transmission modes of the mammalian gut microbiota', *Science* 362(6413): 453–457.

Moya, A. and Ferrer, M. (2016) 'Functional redundancy-induced stability of gut microbiota subjected to disturbance', *Trends in Microbiology* 24(5): 402–413.

Nakata, T. *et al.* (2017) 'Inhibitory effects of soybean oligosaccharides and water-soluble soybean fibre on formation of putrefactive compounds from soy protein by gut microbiota', *International Journal of Biological Macromolecules* 97: 173–180.

Needleman, S. B. and Wunsch, C. D. (1970) 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology* 48(3): 443–453.

Neme, R. and Tautz, D. (2016) 'Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence', *eLife* 5: e09977.

Ochoa-Repáraz, J. *et al.* (2010) 'A polysaccharide from the human commensal *Bacteroides fragilis* protects against CNS demyelinating disease', *Mucosal Immunology* 3(5): 487–495.

Parks, D. H. *et al.* (2017) 'Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life', *Nature Microbiology* 2: 1533–1542.

Pérez-Cobas, A. E. *et al.* (2013) 'Gut microbiota disturbance during antibiotic therapy: a multi-omic approach', *Gut* 62(11): 1591–1601.

Pickard, J. M. *et al.* (2017) 'Gut microbiota: role in pathogen colonization, immune responses, and inflammatory disease', *Immunological Reviews* 279(1): 70–89.

Pritchard, L. *et al.* (2016) 'Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens', *Analytical Methods* 8: 12–24.

Rehman, A. *et al.* (2016) 'Geographical patterns of the standing and active human gut microbiome in health and IBD', *Gut* 65(2): 238–248.

Reyes, A. *et al.* (2010) 'Viruses in the faecal microbiota of monozygotic twins and their mothers', *Nature* 466(7304): 334–338.

Rice, P., Longden, L. and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics* 16(6): 276–277.

Richter, M. and Rosselló-Móra, R. (2009) 'Shifting the genomic gold standard for the prokaryotic species definition', *Proceedings of the National Academy of Sciences of the United States of America* 106(45): 19126–19131.

Roediger, W. E. W. (1980) 'Role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in

man', *Gut* 21(9): 793–798.

Roelofs, K. G. *et al.* (2016) 'Bacteroidales secreted antimicrobial proteins target surface molecules necessary for gut colonization and mediate competition *in vivo*', *mBio* 7(4): e01055-16.

Ross, B. D. *et al.* (2019) 'Human gut bacteria contain acquired interbacterial defence systems', *Nature* 575(7781): 224–228.

Van Rossum, G. and Drake Jr, F.L. (2009) '*Python 3 Reference Manual*', Scotts Valley, CA: CreateSpace.

Round, J. L. and Mazmanian, S. K. (2009) 'The gut microbiota shapes intestinal immune responses during health and disease', *Nature Reviews Immunology* 9: 313–323.

Rühlemann, M. C. *et al.* (2021) 'Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome', *Nature Genetics* 53: 147–155.

Russell, A. B. *et al.* (2014) 'A type VI secretion-related pathway in bacteroidetes mediates interbacterial antagonism', *Cell Host and Microbe* 16(2): 227–236.

Salyers, A. A., Shoemaker, N. B. and Li, L.-Y. (1995) 'In the driver's seat: The *Bacteroides* conjugative transposons and the elements they mobilize', *Journal of Bacteriology* 177(20): 5727-31.

Seemann, T. (2014) 'Genome analysis Prokka: rapid prokaryotic genome annotation', *Bioinformatics* 30(14): 2068–2069.

Shumaker, A. M. *et al.* (2019) 'Identification of a fifth antibacterial toxin produced by a single *Bacteroides fragilis* strain', *Journal of Bacteriology* 201: e00577-18.

Tettelin, H. *et al.* (2008) 'Comparative genomics: the bacterial pan-genome', *Current Opinion in Microbiology* 11(5): 472–477.

Turpin, W. *et al.* (2016) 'Association of host genome with intestinal microbial composition in a large healthy cohort', *Nature Genetics* 48: 1413–1417.

Wang, J. *et al.* (2014) 'Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice', *Proceedings of the National Academy of Sciences of the United States of America* 111(26): E2703–E2710.

Wang, J. *et al.* (2015) 'Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome', *Nature Communications* 6(1): 6440.

Wang, J. *et al.* (2019) 'Core gut bacteria analysis of healthy mice', *Frontiers in Microbiology* 10: 887.

Wexler, A. G. and Goodman, A. L. (2017) 'An insider's perspective: *Bacteroides* as a window into the microbiome', *Nature Microbiology* 2: 17026.

Wexler, H. M. (2007) '*Bacteroides*: thegood, the bad, and the nitty-gritty', *Clinical Microbiology Reviews* 20(4): 593–621.

Wickham, H. (2016) 'ggplot2: elegant graphics for data analysis', Springer-Verlag New York.

Yang, J.-Y. *et al.* (2017) 'Gut commensal *Bacteroides acidifaciens* prevents obesity and improves insulin sensitivity in mice', *Mucosal Immunology* 10(1): 104-116.

Yatsunenko, T., Rey, Federico E., *et al.* (2012) 'Human gut microbiome viewed across age and geography',

Nature 486: 222–227.

Zitomersky, N. L., Coyne, M. J. and Comstock, L. E. (2011) 'Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order Bacteroidales in the human gut', *Infection and Immunity* 79(5): 2012–2020.